Issues in Quantitative Linguistics 3

edited by

Reinhard Köhler
Gabriel Altmann

Dedicated to Karl-Heinz Best
on the occasion of his 70th birthday

2013
RAM-Verlag

Studies in quantitative linguistics

Editors
Fengxiang Fan (fanfengxiang@yahoo.com)
Emmerich Kelih (emmerich.kelih@uni-graz.at)
Reinhard Köhler (koehler@uni-trier.de)
Ján Mačutek (jmacutek@yahoo.com)
Eric S. Wheeler (wheeler@ericwheeler.ca)

1. U. Strauss, F. Fan, G. Altmann, *Problems in quantitative linguistics 1*. 2008, VIII + 134 pp.
2. V. Altmann, G. Altmann, *Anleitung zu quantitativen Textanalysen. Methoden und Anwendungen.* 2008, IV+193 pp.
3. I.-I. Popescu, J. Mačutek, G. Altmann, *Aspects of word frequencies*. 2009, IV +198 pp.
4. R. Köhler, G. Altmann, *Problems in quantitative linguistics 2.* 2009, VII + 142 pp.
5. R. Köhler (ed.), *Issues in Quantitative Linguistics.* 2009, VI + 205 pp.
6. A. Tuzzi, I.-I. Popescu, G.Altmann, *Quantitative aspects of Italian texts*. 2010, IV+161 pp.
7. F. Fan, Y. Deng, *Quantitative linguistic computing with Perl.* 2010, VIII + 205 pp.
8. I.-I. Popescu et al., *Vectors and codes of text*. 2010, III + 162 pp.
9. F. Fan, *Data processing and management for quantitative linguistics with Foxpro***.** 2010, V + 233 pp.
10. I.-I. Popescu, R. Čech, G. Altmann, *The lambda-structure of texts*. 2011, II + 181 pp
11. E. Kelih et al. (eds.), *Issues in Quantitative Linguistics Vol. 2*. 2011, IV + 188 pp.
12. R. Čech, G. Altmann, *Problems in quantitative linguistics 3*. 2011, VI +168 pp.
13. R. Köhler, G. Altmann, *Issues in Quantitative Linguistics Vol 3.* 2013. II + 403 pp.

ISBN: 978-3-942303-12-5

RAM-Verlag
Stüttinghauser Ringstr. 44
D-58515 Lüdenscheid
RAM-Verlag@t-online.de
http://ram-verlag.de

Contents

Karl-Heinz Best

was born on January 28, 1943 in Koblenz, Germany. He studied general linguistics and German linguistics at the University of Bonn (1964 - 1966), then in Bochum (1966 -1971) where he extended his study to Scandinavian languages. In Bochum he wrote his PhD thesis on analogy. From 1974 he taught at the University of Göttingen.

There are many important aspects of Best's scientific life. The first is the application of quantitative methods to writing, morphology, lexicon, syntax, language learning, language change, borrowing, language comprehension and history of quantitative linguistics. His main research field is German and Northern Germanic languages. His immortal merit is the testing of candidates for language laws on an enormous amount of data. The data were either collected by him from modern texts or found by him on places where nobody had ever sought them: in the history of German linguistics. In this way he has shown that German linguists were always interested in quantitative evaluation of linguistic data. The 26 historiographic articles in his bibliography were published mostly in the journal Glottometrics, which enjoys his membership in the editorial board. Bests's biographies bring to light the fact that linguistics was not that qualitative as presented in books about its history. There were ideas and works which were simply ignored because of the weight of other paradigms such as the neogrammarians and structuralism. The same was the fate of G.K. Zipf, whose philosophy passed almost unnoticed along the army of structural and generative linguists, who were sure to have discovered the "truth". Today, the traces of Zipf's work can be found in every scientific discipline, his hypotheses form the basis of frequentism and synergetic linguistics. Karl-Heinz Best detected the son of the famous physician S.V. Čebanov, who introduced the Poisson process into linguistics and wrote a historical note about his father together with him.

The second important activity of K.H. Best is the propagation of quantitative linguistics on the Internet. There are hundreds of articles written by him about different issues both in qualitative and quantitative linguistics. Not everybody is ready or takes the time to write popular articles in Wikipedia or Wiktionary but K.H. Best does it with relentless enthusiasm. In linguistic circles one says that after having read something about linguistics on the Internet, one obtained the "best" information.

The third aspect is Best's editorial activity. This is not restricted to his membership in the editorial boards of *Glottotheory*, *Glottometrics* and *Göttinger Beiträge zur Sprachwissenschaft,* where he cannot impel the colleagues directly but he edits also omnibus volumes where he can play the role of a slave driver. His volumes are always ready on time and concern both synchronic and diachronic linguistics, everything nicely wrapped in a quantitative dress.

K.-H. Best cooperates at the conferences on *Knowledge transfer* in Göttingen and Halle.

Since we all begin to be active somewhere at the university and know how difficult it is to persuade colleagues and students that exact linguistics striving for theories is perhaps more advanced than concept coining, classifications and rule description, we cannot sufficiently emphasize the enormous success with which K.-H. Best won both students and colleagues for quantitative linguistics. Under normal circumstances one needs a Smith & Wesson, calibre 8 mm, to persuade a linguist that there are many ways to truth (not only generative linguistics) but mostly not even a revolver helps. But K.-H. Best succeeded without bodily violence and both his students and his colleagues published a number of articles (see References) which would fill several thick volumes. He called his collection *Göttinger Projekt Quantitative Linguistik* which can be found also on the Internet. Göttingen is one of the few German universities accepting quantitative analyses as bachelor, master or doctoral theses. His own work comprises more than 140 articles and more then 100 book reviews in the journal *Germanistik*.

Bibliography of Karl-Heinz Best

Books

Best, Karl-Heinz. (2001[1], 2003[2], 2006[3]). *Quantitative Linguistik. Eine Annäherung.* Göttingen: Peust & Gutschmidt.

Best, Karl-Heinz. (2008). *LinK: Linguistik in Kürze mit einem Ausblick auf die Quantitative Linguistik.* 5., durchgesehene Ausgabe. Luedenscheid: RAM-Verlag.

Popescu, Ioan-Iovitz; Emmerich Kelih; Ján Mačutek; Radek Čech; Best, Karl-Heinz; Gabriel Altmann 2010. *Vectors and Codes of Text*. Lüdenscheid: RAM-Verlag.

Editions

Best, Karl-Heinz, & Kohlhase, Joerg (Hrsg.) (1983). *Exakte Sprachwandelforschung*. Goettingen: edition herodot.

Best, Karl-Heinz (Hrsg.). (1997). *Glottometrika 16.* Trier: Wissenschaftlicher Verlag Trier.

Best, Karl-Heinz (Hrsg.). (2001). *Häufigkeitsverteilungen in Texten*. Göttingen: Peust & Gutschmidt.

Articles

Best, Karl-Heinz. 2001. Kommentierte Bibliographie zum Göttinger Projekt. In Best, Karl-Heinz (Hrsg.), *Häufigkeitsverteilungen in Texten*: *284-310*. Göttingen: Peust & Gutschmidt.

Best, Karl-Heinz. 2005. Quantitative Linguistik: Ein Plädoyer. In: Altmann, Gabriel, Levickij, Viktor, & Perebyinis, Valentina (Hrsg.), *Problemy kvantytatyvnoi linhvistyky/ Problems of Quantitative Linguistics: zbirnyk naukovych prac*: *76-88*. Cernivci: Ruta.

Best, Karl-Heinz. 2009. Kann man Transfer messen? In: Stenschke, Oliver, & Wichter, Sigurd (Hrsg.). *Wissenstransfer und Diskurs: 13-24*. Frankfurt: Peter Lang.

Best, Karl-Heinz. 2009. Sprachliche Kürzungen. *Glottotheory 2/1, 12-17.*

Best, Karl-Heinz. 2001. Silbenlängen in Meldungen der Tagespresse. In: Best, Karl-Heinz (Hrsg.), *Häufigkeitsverteilungen in Texten*: *15-32*. Göttingen: Peust & Gutschmidt.

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

Best, Karl-Heinz. 2001. Zur Länge von Morphen in deutschen Texten. In: Best, Karl-Heinz (Hrsg.), *Häufigkeitsverteilungen in Texten*: *1-14*. Göttingen: Peust & Gutschmidt.

Best, Karl-Heinz. 2000. Morphlängen in Fabeln von Pestalozzi. *Göttinger Beiträge zur Sprachwissenschaft 3, 19-30*.

Best, Karl-Heinz. 2005. Morphlänge. In: Koehler, Reinhard, Altmann, Gabriel, & Piotrowski, Raimund (Hrsg.) *Quantitative Linguistik - Quantitative Linguistics. Ein internationales Handbuch* (S. 255-260). Berlin/New York: de Gruyter.

Best, Karl-Heinz. 2006. Wie viele Morphe enthalten Wörter in deutschen Pressetexten? *Glottometrics 13, 47-58.*

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

Best, Karl-Heinz. 2000. Unser Wortschatz. Sprachstatistische Untersuchungen. In: R. Hoberg, & K.M. Eichhoff-Cyrus (Hrsg.), *Die deutsche Sprache zur Jahrtausendwende*: *35-52*. Mannheim/ Leipzig/Wien/Zürich: Dudenverlag.

Best, Karl-Heinz. 2007. Frequenz und Länge von Woertern. In: Kaliuscenko, Volodymir, Koehler, Reinhard, & Levickij, Viktor (eds.), *Problemy typolohicnoi ta kvantytatyvnoi leksikologii - Problems of Typological and Quantitative Lexicology: 83-90.* Cernivci: Ruta.

Best, Karl-Heinz. 1994. Word class frequencies in contemporary German short prose texts. *Journal of Quantitative Linguistics 1, 144-147*.

Best, Karl-Heinz. 1997. Zur Wortartenhäufigkeit in Texten deutscher Kurzprosa der Gegenwart. In: Best, Karl-Heinz (Hrsg.), *Glottometrika 16, 276-285.* Trier: Wiss. Verlag Trier.

Best, Karl-Heinz. 1998. Zur Interaktion der Wortarten in Texten. *Papiere zur Linguistik 58, 83-95*.

Best, Karl-Heinz. 2000. Verteilungen der Wortarten in Anzeigen. *Göttinger Beiträge zur Sprachwissenschaft 4, 37-51*.

Best, Karl-Heinz. 2001. Zur Gesetzmässigkeit der Wortartenverteilungen in deutschen Pressetexten. *Glottometrics 1, 1-26*.

Overbeck, Anja, Best, Karl-Heinz. 2008. Wortartenverteilungen in italienischen Opernlibretti. *Göttinger Beiträge zur Sprachwissenschaft 16, 39-46*. (Appeared 2010)

Popescu, Ioan-Iovitz, Best, Karl-Heinz; Altmann, Gabriel. 2007. On the dynamics of word classes in text. *Glottometrics 14, 58-71*.

Zhu, Jinyang, & Best, Karl-Heinz. 1992. Zum Wort im modernen Chinesisch. *Oriens Extremus 35, 45-60*.

Altmann, Gabriel, Best, Karl-Heinz, & Kind, Bernd. 1987. Eine Verallgemeinerung des Gesetzes der semantischen Diversifikation. In: Fickermann, Ingeborg (Hrsg.), *Glottometrika 8, 130-139*. Bochum: Brockmeyer.

Best, Karl-Heinz. 1990. Die semantische Diversifikation eines Wortbildungsmusters im Fruehneuhochdeutschen. In: Hřebíček, Luděk (Hrsg.), *Glottometrika 11, 107-110*. Bochum: Brockmeyer.

Best, Karl-Heinz. 1991. Von: Zur Diversifikation einer Partikel des Deutschen. In: Rothe, Ursula (Hrsg.), *Diversification Processes in Language: Grammar*: *94-104*. Hagen: Margit Rottmann Medienverlag.

Best, Karl-Heinz. 2005. Diversifikation der Fremd- und Lehnwoerter im Türkischen. *Archiv Orientalní 73, 291-298*.

Best, Karl-Heinz. 2005. Ein Modell fuer das etymologische Spektrum des Wortschatzes. *Naukovyj Visnyk Cernivec'koho Universytetu: Herman'ska filolohija. Vypusk 266, 11-21*.

Best, Karl-Heinz. 2007. Diversifikation bei Eigennamen. In: Peter Grzybek & Reinhard Koehler (eds.), *Exact Methods in the Study of Language and Text. Dedicated to Gabriel Altmann on the Occasion of his 75th Birthday: 21-31*. Berlin/ New York: Mouton de Gruyter.

Best, Karl-Heinz. 2007. Kürzungstendenzen im Deutschen aus der Sicht der Quantitativen Linguistik. In: Baer, Jochen, A., Roelcke, Thorsten, & Steinhauer, Anja (Hrsg.): *Sprachliche Kuerze. Konzeptuelle, strukturelle und pragmatische Aspekte*: *45-62*. Berlin/ New York: de Gruyter.

Best, Karl-Heinz. 2008. Rangordnungen deutscher Eigennamen. In: Altmann, Gabriel, Zadorozhna, Iryna, & Matskulyak, Yuliya (eds.), *Problemy zagal'noho, hermans'koho ta slov'janskoho movoznavstva do 70-riccja professora V.V. Levic'koho/ Problems of General, Germanic and Slavic Languages. Papers for 70-th Anniversary of Professor V. Levickij*: *454-460*. Chernivtsi: Books - XXI.

Best, Karl-Heinz. 2008. Das Fremdwortspektrum im Tuerkischen. *Glottometrics 17, 8-11*.

Best, Karl-Heinz. 2008. Zur Diversifikation deutscher Hexameter. *Naukovyj Visnyk Cernivec'koho Universytetu: Herman'ska filolohija. Vypusk 431, 172-180.*

Best, Karl-Heinz. 2008. Zur Diversifikation lateinischer und griechischer Hexameter. *Glottometrics 17, 43-50.*

Best, Karl-Heinz. 2008. Verteilungen von Fugenelementen im Deutschen. *Göttinger Beiträge zur Sprachwissenschaft* 16, 7-16. (Erschienen 2010)

Best, Karl-Heinz. 2009. Diversifikation des Phonems /r/ im Deutschen. *Glottometrics 18, 26-31*.

Best, Karl-Heinz. 2009. Zum etymologischen Spektrum des Hundeshagener Kochums. *Göttinger Beiträge zur Sprachwissenschaft 19, 25-29*. (Appeared 2011)

Best, Karl-Heinz. 2010. Zum Fremdwortspektrum im Japanischen. *Glottotheory 3(1), 5-8*.

Best, Karl-Heinz. 2011. Diversification of a single sign of the Danube script. *Glottometrics 22, 1-4.*

Boschtan, Aljona, & Best, Karl-Heinz. 2010. Diversification of simple attributes in German. *Glottotheory 3(2), 5-9.* (Appeared 2011)

Altmann, Gabriel, Beöthy, Erszébet, & Best, Karl-Heinz. 1982. Die Bedeutungskomplexität der Wörter und das Menzerathsche Gesetz. *Zeitschrift für Phonetik, Sprachwissenschaft und Kommunikationsforschung 35, 537-543*.

Asleh, Laila, Best, Karl-Heinz. 2005. Zur Überprüfung des Menzerath-Altmann-Gesetzes am Beispiel deutscher (und italienischer) Wörter. *Göttinger Beiträge zur Sprachwissenschaft 10(11), 9-19.*

Best, Karl-Heinz, & Altmann, Gabriel. 1996. Project Report. *Journal of Quantitative Linguistics 3, 85-88.*

Best, Karl-Heinz. 1997. "Zum Stand der Untersuchungen zu Wort- und Satzlängen." *Tischvorlage fuer QUALICO 3, Helsinki 1997*.

Best, Karl-Heinz. 1997. Warum nur: Wortlänge? Nicht nur ein Vorwort. In: Best, Karl-Heinz (Hrsg.), *Glottometrika* 16. *The Distribution of Word and Sentence Length: 5-12*. Trier: WVT.

Best, Karl-Heinz. 1998. Results and perspectives of the Goettingen project on quantitative Linguistics. *Journal of Quantitative Linguistics* 5, 155-162.

Best, Karl-Heinz. 1999. Quantitative Linguistik: Entwicklung, Stand und Perspektive. *Göttinger Beiträge zur Sprachwissenschaft* 2, 7-23.

Best, Karl-Heinz. 2000. Unser Wortschatz. Sprachstatistische Untersuchungen. In: R. Hoberg, & K. M. Eichhoff-Cyrus (Hrsg.), *Die deutsche Sprache zur Jahrtausendwende: 35-52*. Mannheim/Leipzig/Wien/Zürich: Dudenverlag.

Best, Karl-Heinz. 2000. Verteilungen sprachlicher Einheiten in Texten und im Sprachsystem. *Tischvorlage für QUALICO IV, Prag, 24.-26.8.2000.*

Best, Karl-Heinz. 2001a. Zur Einführung. In: Best, Karl-Heinz (Hrsg.), *Häufigkeitsverteilungen in Texten*: *5-20*. Göttingen: Peust & Gutschmidt.

Best, Karl-Heinz. 2001b. Kommentierte Bibliographie zum Göttinger Projekt. In: Best, Karl-Heinz (Hrsg.), *Häufigkeitsverteilungen in Texten*: *284-310*. Göttingen: Peust & Gutschmidt.

Best, Karl-Heinz. 2001d. Probability Distributions of Language Entities. *Journal of Quantitatitve Linguistics 8, 1-11*.

Best, Karl-Heinz. 2005. Wortlängen. In: Koehler, Reinhard, Altmann, Gabriel, & Piotrowski, Raimund (Hrsg.), *Quantitative Linguistik - Quantitative Linguistics. Ein internationales Handbuch*: *260-273*. Berlin/New York: de Gruyter.

Best, Karl-Heinz, & Cebanov, Sergej Viktorovic. 2001. *Biographische Notiz: Sergej Grigor'evic Cebanov (1897-1966).* In: Best, Karl-Heinz (Hrsg.), *281-283*. Göttingen: Peust & Gutschmidt.

Best, Karl-Heinz. 1996. Word Length in Old Icelandic Songs and Prose Texts. *Journal of Quantitative Linguistics 3, 97-105*.

Best, Karl-Heinz, & Zhu, Jinyang. 2001. Wortlängenverteilungen in chinesischen Texten und Woerterbuechern. In: Best, Karl-Heinz (Hrsg.), *Häufigkeitsverteilungen in Texten: 101-114*. Goettingen: Peust & Gutschmidt.

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. Wortlängen in chinesischen Briefen. In: Hongjun Cai (Hg.), *Neue Forschungen chinesischer Germanisten in Deutschland: 121-129*. Frankfurt: Peter Lang.

Zhu, Jinyang, & Best, Karl-Heinz. 1997. Zur Modellierung der Wortlängen im Chinesischen. In: Best, Karl-Heinz (Hrsg.), *Glottometrika 16. The Distribution of Word and Sentence Length*: *185-194*. Trier: Wiss. Vlg. Trier. (revised in Best (Hrsg.), Glottometrika 16)

Zhu, Jinyang, & Best, Karl-Heinz. 1997. Zur Modellierung der Wortlängen im Chinesischen. In: Best, Karl-Heinz (Hrsg.), *Glottometrika 16, 185-194.* Trier: WVT.

Zhu, Jinyang, & Best, Karl-Heinz. 1998. Wortlängenhäufigkeiten in chinesischen Kurzgeschichten. *Asian and African Studies 7, 45-51*.

Best, Karl-Heinz. 2009. Wortlängen im Dänischen. *Göttinger Beiträge zur Sprachwissenschaft 19, 7-23*. (Appeared 2011)

Altmann, Gabriel, & Best, Karl-Heinz. 1996. Zur Länge der Woerter in deutschen Texten. In: Schmidt, Peter (Hrsg.), *Glottometrika 15, 166-180*. Trier: WVT.

Best, Karl-Heinz. 1996. Zur Bedeutung von Wortlängen, am Beispiel althochdeutscher Texte. *Papiere zur Linguistik 55, 141-152*.

Best, Karl-Heinz. 1997. Wortlängen in mittelhochdeutschen Texten. In: Best, Karl-Heinz (Hrsg.), *Glottometrika 16. The Distribution of Word and Sentence Length: 40-54*. Trier: WVT.

Best, Karl-Heinz. 1997. Zur Wortlängenhäufigkeit in deutschsprachigen Pressetexten. In: Best, Karl-Heinz (Hrsg.), *Glottometrika 16. The Distribution of Word and Sentence Length: 1-15*. Trier: Wissenschaftlicher Verlag Trier.

Best, Karl-Heinz. 2000. Wie viele Morphe enthalten deutsche Woerter? Am Beispiel einiger Fabeln Pestalozzis. In: S. Ondrejovič & M. Považaj (Hrsg.), Lexicographica '99. *Sborník na počesť Kláry Buzássyovej*: *258-270*. Bratislava: Veda.

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

Best, Karl-Heinz. 2006. Wortlängen im Deutschen. *Göttinger Beiträge zur Sprachwissenschaft 13, 23-49*.

Best, Karl-Heinz. 2007. Quantitative Untersuchungen zum deutschen Wörterbuch. *Glottometrics 14, 32-45.*

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

Best, Karl-Heinz, & Zhu, Jinyang. 1994. Zur Häufigkeit von Wortlängen in Texten deutscher Kurzprosa (mit einem Ausblick auf das Chinesische). In: Klenk, Ursula (Hrsg.), *Computatio Linguae II: 19-30*. Stuttgart: Steiner.

Best, Karl-Heinz. 2009. Wortlängen im Englischen. *Glottometrics 19, 1-10*.

Bartens, Hans-Hermann, & Best, Karl-Heinz. 1997. Wortlängen in erzamordwinischen Texten. *Linguistica Uralica 23, 5-13.*

Bartens, Hans-Hermann, & Best, Karl-Heinz. 1996. Wortlängen in estnischen Texten. *Ural-altaische Jahrbücher N.F. 14, 112-128.*

Best, Karl-Heinz, & Kaspar, Ingolf. 1998. Wortlängen in färoeischen Briefen. *Naukovyj Visnyk Cernivec'koho Universytetu: Hermans'ka filolohija. Vypusk 41, 3-14.*

Best, Karl-Heinz, & Kaspar, Ingolf. 2001. Wortlängen im Färoeischen. In: Best, Karl-Heinz (Hrsg.), *Häufigkeitsverteilungen in Texten*: *92-100*. Göttingen: Peust & Gutschmidt.

Vettermann, Aniko, & Best, Karl-Heinz. 1997. Wortlängen im Finnischen. *Suomalais-ugrilaisen seuran aikakauskirja/Journal de la Societé Finno-Ougrienne 87, 249-262*.

Best, Karl-Heinz. 2011. Word length distribution in French. *Glottometrics 22, 57-61*.

Best, Karl-Heinz, & Brynjolfsson, Einar. 1997. Wortlängen in isländischen Briefen und Pressetexten. *Skandinavistik 27, 24-40*.

Best, Karl-Heinz, & Medrano, Paulina. 1997. Wortlängen in Ketschua-Texten. In: Best, Karl-Heinz (Hrsg.), *Glottometrika 16. The Distribution of Word and Sentence Length*: *204-212.* Trier: Wissenschaftlicher Verlag Trier.

Best, Karl-Heinz, & Song, Hea-Yean. 1996. Wortlängen im Koreanischen. *Asian and African Studies 5, 39-49.*

Bartens, Hans-Hermann, & Best, Karl-Heinz. 1997. Word Length Distribution in Sami Texts. In: Altmann, Gabriel, Mikk, Jaan, Saukkonen, Pauli, & Wimmer, Gejza (eds.), *Festschrift in Honour of Juh. Tuldava. Journal of Quantitative Linguistics 4, 45-52*.

Best, Karl-Heinz. 2006. Quantitative Untersuchungen zum Niederdeutschen und Niederländischen. *Göttinger Beiträge zur Sprachwissenschaft 13, 51-71*.

Best, Karl-Heinz. 2011. Wortlängen im Norwegischen. In: Emmerich Kelih, Viktor Levickij, Yuliya Matskulyak (eds). *Issues in Quantitative Linguistics 2. Dedicated to Reinhard Koehler on the occasion of his 60th birthday: 129-135*. Luedenscheid: RAM-Verlag.

Best, Karl-Heinz. 2008. Word length in Persian. *Glottometrics 16, 27-30.*

Altmann, Gabriel, Best, Karl-Heinz, & Wimmer, Gejza. 1997. Wortlänge in romanischen Sprachen. In: Gather, Andreas, & Werner, Heinz (Hrsg.), *Semiotische Prozesse und natürliche Sprache. Festschrift für Udo L. Figge zum 60. Geburtstag*: *1-13*. Stuttgart: Steiner.

Best, Karl-Heinz, & Zinenko, Svetlana. 1998. Wortlängenverteilungen in Briefen A. T. Twardowskis. *Göttinger Beiträge zur Sprachwissenschaft 1, 7-19.*

Best, Karl-Heinz, & Zinenko, Svetlana. 2001. Wortlängen in Gedichten A.T. Twardowskis. In: Uhlirova, Ludmila, Wimmer, Gejza, Gabriel Altmann & Reinhard Koehler (eds.), *Text as a Linguistic Paradigm: Levels, Constituents, Constructs. Festschrift in Honour of Luděk Hřebiček*: *21-28.* Trier: WVT.

Best, Karl-Heinz. 1996. Zur Wortlängenhäufigkeit in schwedischen Pressetexten. In: Schmidt, Peter (Hrsg.), *Glottometrika 15, 147-157*. Trier: WVT.

Best, Karl-Heinz. 2008. Wortlängen im Schwedischen. *Naukovyj Visnyk Cernivec'koho Universytetu: Herman'ska filolohija. Vypusk 370-371, 155-162*.

Best, Karl-Heinz. 2007. Wortlängen im Tschechischen. Ein Nachtrag. *Göttinger Beiträge zur Sprachwissenschaft 14, 35-39.*

Bartens, Hans-Hermann, & Best, Karl-Heinz. 1997. Wortlängen im Tscheremissischen (Mari). *Finnisch-Ugrische Mitteilungen 20, 1-20.*

Best, Karl-Heinz, & Oezmen, Elif. 1996. Wortlängenhäufigkeiten in tuerkischen Texten und ihre linguistischen Implikationen. *Archiv Orientalni 64, 19-30*.

Best, Karl-Heinz, & Zinenko, Svetlana. 1999. Wortkomplexität im Ukrainischen und ihre linguistische Bedeutung. *Zeitschrift fuer Slavische Philologie 58, 107-123.*

Best, Karl-Heinz, & Zinenko, Svetlana. 1999. Wortlängen in Gedichten des ukrainischen Autors Ivan Franko. In: Jozef Genzor, & Slavomír Ondrejovič (eds.), *Pange Lingua. Zbornik na Pocest' Viktora Krupu*: *201-213*. Bratislava: Veda, Vydavatel'stvo SAV.

Best, Karl-Heinz. 2005. Wortlängen im Ungarischen (und anderswo). In: *Lihkkun lehkos! Beiträge zur Finnougristik aus Anlass des sechszigsten Geburtstages von Hans-Hermann Bartens: 43-56*. Hrsg. v. Cornelius Hasselblatt, Eino Koponen u. Anna Widmer. Wiebaden: Harrassowitz.

Best, Karl-Heinz. 2004. Wortschatzwachstum. In: *Wissenstransfer und gesellschaftliche Kommunikation. Festschrift für Sigurd Wichter zum 60. Geburtstag*: *333-342*. Hrsg. v. Albert Busch & Oliver Stenschke. Frankfurt: P. Lang.

Best, Karl-Heinz. 2004. Zum Wortschatzwachstum und -umfang in Texten. *Naukovyj Visnyk Cernivec'koho Universytetu: Hermans'ka filolohija. Vypusk 206-207, 31-43*. Cernivci.

Best, Karl-Heinz. 2006. Zum Computerwortschatz im Deutschen. *Naukovyj Visnyk Cernivec'koho Universytetu: Herman'ska filolohija. Vypusk 289, 10-24.*

Best, Karl-Heinz. 2008. Wortschatzwachstum und -umfang in Texten und Textkorpora. In: Sibyla Misličová (ed.), *Jazyk a jazykoveda v pohybe. Zborník štúdií na počest' Slavomíra Ondrejoviča pri príležitosti jeho životného jubilea*: *422-437*. Bratislava: VEDA, vydavatel'stvo SAV.

Best, Karl-Heinz. 2001. Zur Verteilung rhythmischer Einheiten in deutscher Prosa. In: Best, Karl-Heinz (Hrsg.), *Häufigkeitsverteilungen in Texten*: *162-166.* Göttingen: Peust & Gutschmidt.

Best, Karl-Heinz. 2001. Probability Distributions of Language Entities. *Journal of Quantitative Linguistics 8, 1-11*.

Best, Karl-Heinz. 2002. The distribution of rhythmic units in German short prose. *Glottomometrics* 3 (= To Honor G.K. Zipf), *135-142*.

Best, Karl-Heinz. 2005. Längen rhythmischer Einheiten. In: Koehler, Reinhard, Altmann, Gabriel, & Piotrowski, Raimund (Hrsg.), *Quantitative Linguistik*

- *Quantitative Linguistics. Ein internationales Handbuch: 208-214*. Berlin/New York: de Gruyter.

Best, Karl-Heinz. 2006. Rhythmische Einheiten im Altgriechischen. *Göttinger Beiträge zur Sprachwissenschaft 13, 73-76.*

Best, Karl-Heinz. 2007. Quantitative Untersuchungen zum Rhythmus. *Göttinger Beiträge zur Sprachwissenschaft 15, 7-14*.

Best, Karl-Heinz. 2009. Rhythmische Einheiten in Huelsen, Natur-Betrachtungen. (1800). In: Emmerich Kelih, Viktor Levickij, Gabriel Altmann (Hrsg.), *Metody analizu tekstu/ Methods of Text Analysis: 53-62*. Cernivci : Cerniveckyj nacional'nyj universitet.

Best, Karl-Heinz. 2008. Zur Diversifikation lateinischer und griechischer Hexameter. *Glottometrics 17, 43-50.*

Best, Karl-Heinz. 2008. Zur Diversifikation deutscher Hexameter. *Naukovyj Visnyk Cernivec'koho Universytetu: Herman'ska filolohija. Vypusk 431, 172-180.*

Best, Karl-Heinz. 2001. Wie viele Wörter enthalten Sätze im Deutschen? Ein Beitrag zu den Sherman-Altmann-Gesetzen. In: Best, Karl-Heinz (Hrsg.), *Häufigkeitsverteilungen in Texten*: *167-201*. Göttingen: Peust & Gutschmidt.

Best, Karl-Heinz. 2002. Satzlängen im Deutschen: Verteilungen, Mittelwerte, Sprachwandel. *Göttinger Beiträge zur Sprachwissenschaft 7, 7-31*.

Best, Karl-Heinz. 2005. Satzlänge. In: Köhler, Reinhard, Altmann, Gabriel, & Piotrowski, Raimund (Hrsg.), *Quantitative Linguistik - Quantitative Linguistics. Ein internationales Handbuch*: *298-394*. Berlin/New York: de Gruyter.

Best, Karl-Heinz. 2006. Verteilung von Phrasen- und Subsatzlängen in deutscher Fachsprache. *Naukovyj Visnyk Cernivec'koho Universytetu: Herman'ska filolohija. Vypusk 319-320, 113-120*.

Best, Karl-Heinz. 2006. Quantitative Untersuchungen zum Niederdeutschen und Niederländischen. *Göttinger Beiträge zur Sprachwissenschaft 13, 51-71*.

Busch, Andrea. 2002. Zur Entwicklung der Satzlänge in deutscher Fachsprache. *Staatsexamensarbeit*, Göttingen.

Best, Karl-Heinz. 2004. Sind Wort- und Satzlänge brauchbare Kriterien zur Bestimmung der Lesbarkeit von Texten? Erscheint im *Tagungsband des Transfer-Kolloquiums* Göttingen 2003.

Best, Karl-Heinz. 2004. Wortschatzwachstum. In: *Wissenstransfer und gesellschaftliche Kommunikation. Festschrift für Sigurd Wichter zum 60. Geburtstag: 333-342*. Hrsg. v. Albert Busch & Oliver Stenschke. Frankfurt/M.: Peter Lang.

Best, Karl-Heinz. 2004. Zum Wortschatzwachstum und -umfang in Texten. *Naukovyj Visnyk Cernivec'koho Universytetu: Hermans'ka filolohija. Vypusk 231, 119-127*. Cernivci.

Best, Karl-Heinz. 2005. Buchstabenhäufigkeiten im Deutschen und Englischen. *Naukovyj Visnyk Cernivec'koho Universytetu: Hermans'ka filolohija. Vypusk 206-207, 31-43*. Cernivci.

Best, Karl-Heinz. 2005. Sprachliche Einheiten in Textblöcken. *Glottometrics 9, 1-12.*

Best, Karl-Heinz. 2005. Laut- und Phonemhäufigkeiten im Deutschen. *Göttinger Beiträge zur Sprachwissenschaft 10/11, 21-32.*

Best, Karl-Heinz. 2004. Das Fremdwort aus der Sicht der Quantitativen Linguistik. In: *Theorie, Steuerung und Medien des Wissenstransfers*: 89-99. Hrsg. v. Sigurd Wichter & Oliver Stenschke in Zusammenarbeit m. Manuel Tants. Frankfurt u.a.: Peter Lang.

Best, Karl-Heinz. 2004. Wortschatzwachstum. In: *Wissenstransfer und gesellschaftliche Kommunikation. Festschrift für Sigurd Wichter zum 60. Geburtstag: 333-342*. Hrsg. v. Albert Busch und Oliver Stenschke. Frankfurt: P. Lang.

Best, Karl-Heinz. 2004. Zum Wortschatzwachstum und -umfang in Texten. *Naukovyj Visnyk Cernivec'koho Universytetu: Hermans'ka filolohija. Vypusk 206-207, 31-43*. Cernivci.

Best, Karl-Heinz. 2006. Sind Wort- und Satzlänge brauchbare Kriterien zur Bestimmung der Lesbarkeit von Texten? In: Wichter, Sigurd, & Busch, Albert (Hrsg.), *Wissenstransfer - Erfolgskontrolle und Rueckmeldungen aus der Praxis: 21-31*. Frankfurt/M. u.a.: Lang.

Best, Karl-Heinz. 2009. Kann man Transfer messen? In: Stenschke, Oliver, & Wichter, Sigurd (Hrsg.). *Wissenstransfer und Diskurs*: *13-24*. Frankfurt u.a.: Peter Lang.

Best, Karl-Heinz. 2004/05. Laut- und Phonemhäufigkeiten im Deutschen. *Göttinger Beiträge zur Sprachwissenschaft 10/11, 21-32.*

Best, Karl-Heinz. 2005. Buchstabenhäufigkeiten im Deutschen und Englischen. *Naukovyj Visnyk Cernivec'koho Universytetu: Herman'ska filolohija. Vypusk 231, 119-127*.

Best, Karl-Heinz. 2005. Zur Häufigkeit von Buchstaben, Leerzeichen und anderen Schriftzeichen in deutschen Texten. *Glottometrics 11, 9-31*.

Best, Karl-Heinz. 2006. Quantitative Untersuchungen zum Niederdeutschen und Niederländischen. *Göttinger Beiträge zur Sprachwissenschaft 13, 51-71*.

Best, Karl-Heinz. 2008. Gesetzmässigkeiten der Lautdauer. *Glottotheory* 1, 1-9.

Best, Karl-Heinz. 2009. Diversifikation des Phonems /r/ im Deutschen. *Glottometrics 18, 26-31*.

Best, Karl-Heinz. 2011. Zur Gesetzmässigkeit der Vokalquantität im Deutschen. *Naukovyj Visnyk Cernivec'koho Universytetu: Hermans'ka filolohija. Vypusk 532, 3-13*.

Best, Karl-Heinz. 2011. Diversification of a single sign of the Danube script. *Glottometrics 22, 1-4.*

Best, Karl-Heinz, & Altmann, Gabriel. 2005. Some properties of graphic systems. *Glottometrics 9, 1-12.*

Best, Karl-Heinz, & Altmann, Gabriel. 2008. Script ornamentality. In: Altmann, Gabriel, & Fan, Fengxiang (eds.), *Analyses of Script. Properties of Characters and Writing Systems: 91-104*. Berlin/ New York: Mouton de Gruyter.

Best, Karl-Heinz, & Zhu, Jinyang. 2010. Ein Modell für die Zunahme chinesischer Schriftzeichen. *Glottometrics 20, 29-33.*

Historiography

Best, Karl-Heinz, & Cebanov (= Chebanov), Sergej Viktorovic. 2001. Biographische Notiz: Sergej Grigor'evic Cebanov (1897-1966). In: Best, Karl-Heinz (Hrsg.), *Häufigkeitsverteilungen in Texten*: *281-283*. Göttingen: Peust & Gutschmidt.

Best, Karl-Heinz. 2005. Karl Marbe (1869-1953). *Glottometrics 9, 74-76*.

Best, Karl-Heinz. 2005. Georg von der Gabelentz (1840-1893). *Glottometrics 9, 77-79.*

Best, Karl-Heinz. 2005. Georg Philipp Harsdoerffer (1607-1658). *Glottometrics 9, 86-88.*

Best, Karl-Heinz. 2005. Gottfried Wilhelm Leibniz (1646-1716). *Glottometrics 9, 79-82.*

Best, Karl-Heinz, & Kotrasch, Brita. 2005. Albert Thumb (1865-1915). *Glottometrics 9, 82-84*.

Best, Karl-Heinz. 2006. Jean Paul (1763-1825). *Glottometrics 12, 75-77.*

Best, Karl-Heinz. 2006. Ernst Wilhelm Foerstemann (1822-1906). *Glottometrics 12, 77-86.*

Best, Karl-Heinz. 2006. Karl Knauer (1906-1966). *Glottometrics 12, 86-94*.

Best, Karl-Heinz. 2006. August Friedrich Pott (1802-1887). *Glottometrics 12, 94-96.*

Best, Karl-Heinz. 2006. August Schleicher (1821-1868). *Glottometrics 13, 73-75*.

Best, Karl-Heinz. 2006. Hans Arens (1911-2003). *Glottometrics 13, 79-84*.

Best, Karl-Heinz. 2006. Siegfried Behn (1884-1970). *Glottometrics 13, 85-88*.

Best, Karl-Heinz. 2006. Adolf Lucas Bacmeister (1827-1873). *Glottometrics 13, 73-75.*

Best, Karl-Heinz. 2007. Lorenzo Bianchi (1889-1960). *Glottometrics 14, 72-74.*

Best, Karl-Heinz. 2007. Manfred Faust (1936-1997). *Glottometrics 14, 74-78.*

Best, Karl-Heinz. 2007. Erwin Kunath (1899-1984). *Glottometrics 14, 78-80.*

Best, Karl-Heinz. 2007. Otto Behaghel (1854-1936). *Glottometrics 14, 80-86*.

Best, Karl-Heinz. 2007. Paul Menzerath (1883-1954). *Glottometrics 14, 86-98.*

Best, Karl-Heinz. 2008. Helmut Meier (1897-1973). *Glottometrics 16, 122-124*.

Best, Karl-Heinz. 2008. Adolf Busemann (1887-1967). *Glottometrics 16, 124-127.*

Best, Karl-Heinz. 2008. Kaj Brynolf Lindgren (1922-2007). *Glottometrics 16, 127-131*.

Best, Karl-Heinz. 2008. Moritz Wilhelm Drobisch (1802-1896). *Glottometrics 17, 109-114.*

Best, Karl-Heinz. 2009. William Palin Elderton (1877-1962). *Glottometrics 19, 99-101*.

Best, Karl-Heinz. 2009. Herdan, Gustav. In: Stammerjohann, Harro (Hrsg.): *Lexicon Grammaticorum. A Bio-Bibliographical Companion to the History of Linguistics. Volume II: A - K.*: *641f.* Tübingen: Niemeyer 2009.

Best, Karl-Heinz. 2010. Laut- und Buchstabenzählungen im frühen 19. Jahrhundert. *Glottometrics 20, 110-114*.

Works of pupils stimulated by K.-H. Best (Göttingen Projekt)

Ahlers, Astrid. 1994. *Untersuchungen zur Wortlängenhäufigkeit an verschiedenen Textsorten des Niederdeutschen.* Staatsexamensarbeit, Göttingen.

Ahlers, Astrid. 2001. The Distribution of Word Length in Different Types of Low German Texts. In: Best, Karl-Heinz (Hrsg.), *Häufigkeitsverteilungen in Texten: 43-58.* Göttingen: Peust & Gutschmidt.

Ahn, Suenje von, & Potthast, Karl-Uwe. 1993. *Wortlängenhäufigkeiten in spanischsprachigen Pressetexten.* Seminararbeit, Göttingen.

Ammermann, Stefan, & Bengtson, Malin. 1997. Zur Wortlängenhäufigkeit im Schwedischen: Gunnar Ekeloefs Briefe. In: Best, Karl-Heinz (Hrsg.), *Glottometrika 16. The Distribution of Word and Sentence Length: 88-97.* Trier: WVT.

Ammermann, Stefan. 1996. *Zur Wortlänge in deutschen Briefen seit fnhd. Zeit*. Staatsexamensarbeit, Göttingen.

Ammermann, Stefan. 1997. Untersuchungen zur Wortlängenhäufigkeit in Briefen Kurt Tucholskys. In: Best, Karl-Heinz (Hrsg.), *Glottometrika 16. The Distribution of Word and Sentence Length: 63-70.* Trier: WVT.

Ammermann, Stefan. 2001. Zur Wortlängenverteilung in deutschen Biefen ueber einen Zeitraum von 500 Jahren. In: Best, Karl-Heinz (Hrsg.), *Häufigkeitsverteilungen in Texten: 59-91*. Göttingen: Peust & Gutschmidt.

Arlt, Iris. 2006. Zur Wortlängenverteilung in SMS-Texten. *Göttinger Beiträge zur Sprachwissenschaft 13, 9-21.*

Balschun, Claudia. 1997. Wortlängenhäufigkeiten in althebräischen Texten. In: Best, Karl-Heinz (Hrsg.), *Glottometrika 16. The Distribution of Word and Sentence Length: 174-179.* Trier: WVT.

Barbaro, S. 2000. Word Length Distribution in Italian Letters of Pier Paolo Pasolini. *Journal of Quantitative Linguistics 7, 115-120.*

Bartels, Olaf, & Strehlow, Michael. 1997. Zur Häufigkeit von Wortlängen in deutschen Briefen im 19. Jahrhundert und in der ersten Hälfte des 20. Jahrhunderts (Bismarck, Brecht, Kafka, Th. Mann, Tucholsky). In: Best, Karl-Heinz (Hrsg.), *Glottometrika 16. The Distribution of Word and Sentence Length: 71-76*. Trier: WVT.

Bartens, Hans-Hermann, & Zoebelin, Thomas. 1997. Wortlängenhäufigkeiten im Ungarischen. In: Best, Karl-Heinz (Hrsg.), *Glottometrika 16. The Distribution of Word and Sentence Length: 195-203.* Trier: WVT.

Becker, Carmen. 1996. Word Lengths in Letters of the Chilenic Author Gabriela Mistral. *Journal of Quantitative Linguistics 3, 128-131.*

Behrmann, Gabi. 1997. Die Wortlängenhäufigkeiten von deutschsprachigen naturwissenschaftlichen Publikationen. In: Best, Karl-Heinz (Hrsg.), *Glottometrika 16. The Distribution of Word and Sentence Length: 77-87*. Trier: WVT.

Brandt, Ingrid Christine Reindorff. 1994. *Quantitative Untersuchung dänischer Brieftexte.* Seminararbeit, Göttingen.

Brueers, Nina, & Heeren, Anne. 2004. Pluralallomorphe in Briefen Heinrich von Kleists. *Glottometrics 7, 85-90.*

Cassier, Falk-Uwe. 1998. *Silbenlängen in Meldungen der deutschen Tagespresse*. Staatsexamensarbeit, Göttingen.

Cassier, Falk-Uwe. 2001. Silbenlängen in Meldungen der deutschen Tagespresse. In: Best, Karl-Heinz (Hrsg.), *Häufigkeitsverteilungen in Texten: 33-42*. Göttingen: Peust & Gutschmidt.

Christiansen, Birte. 1997. Wortlängenverteilung in deutschen Barockgedichten. In: Best, Karl-Heinz (Hrsg.), *Glottometrika 16. The Distribution of Word and Sentence Length: 16-39.* Trier: WVT.

Culp, Christine. 1995. *Untersuchung zur Häufigkeit von Wortlängen in ausgewählten Briefen Majakovskijs.* Seminararbeit, Göttingen.

Dieckmann, Sandra, & Judt, Birga. 1996. Untersuchung zur Wortlängenverteilung in franzoesischen Pressetexten und Erzählungen. In: Schmidt, Peter (Hrsg.), *Glottometrika 15: 158-165*. Trier: Wissenschaftlicher Verlag Trier.

Diel, Michael. 1993. *Wortlängen in dänischen Pressetexten*. Seminararbeit, Göttingen.

Diel, Michael. 1993. *Wortlängen in Gedichten von Benny Andersen*. Seminararbeit, Göttingen.

Diel, Michael. 1993. *Wortlängen in Kommentaren Rudolf Augsteins*. Seminararbeit, Göttingen.

Diel, Michael. 1994. *Wortlängen in deutschen und dänischen Texten*. Staatsexamensarbeit, Göttingen.

Dittrich, Heike. 1996. Word Length Frequency in the Letters of G.E. Lessing. *Journal of Quantitative Linguistics 3, 260-264.*

Dittrich, Heike. 1996. *Wortlängenhäufigkeiten in deutschen Briefen des 18. Jhds.* Staatsexamensarbeit, Göttingen.

Drechsler, Jochen. 2001. Häufigkeitsverteilungen von Wortlängen in gälischen Texten. In: Best, Karl-Heinz (Hrsg.), *Häufigkeitsverteilungen in Texten: 115-123.* Göttingen: Peust & Gutschmidt.

Egbers, Jannetje, Groen, Claudia, Rauhaus, Esther, & Podehl, Ralf. 1997. Zur Wortlängenhäufigkeit in griechischen Koine-Texten. In: Best, Karl-Heinz (Hrsg.), *Glottometrika 16. The Distribution of Word and Sentence Length: 108-120.* Trier: WVT.

Feldt, Karin. 1998. *Wortlängen im klassischen Griechisch (ionischer Dialekt).* Seminararbeit, Göttingen.

Feldt, Sabine, Janssen, Marianne, & Kuleisa, Silke. 1997. Untersuchung zur Gesetzmässigkeit von Wortlängenhäufigkeiten in französischen Briefen und Pressetexten. In: Best, Karl-Heinz (Hrsg.), *Glottometrika 16. The Distribution of Word and Sentence Length: 145-151.* Trier: WVT.

Fotiadou-Dubois, Maria. 1997. *Zur Wortlängenverteilung in griechischen Volksmärchen.* Seminararbeit, Göttingen.

Fuchs, Rinje. 1991. Semantische Diversifikation der deutschen Präposition auf. In: Rothe, Ursula (Hrsg.), *Diversification Processes in Language: Grammar: 105-115.* Hagen: Margit Rottmann Medienverlag.

Gerlach, Rainer. 1982. Zur Ueberpruefung des Menzerath'schen Gesetzes in der Morphologie. In: Lehfeldt, Werner, & Strauss, U. (Hrsg.), *Glottometrika 4: 95-113.* Bochum: Brockmeyer.

Girzig, Patricia. 1997. Untersuchung zur Häufigkeit von Wortlängen in russischen Texten. In: Best, Karl-Heinz (Hrsg.), *Glottometrika 16. The Distribution of Word and Sentence Length: 152-162*. Trier: WVT.

Hartig, Andrea. 1996. *Wortlängen in Briefen des griechischen Dichters Nikos Kazantzakis.* Seminararbeit, Göttingen.

Hasse, Alice, & Weinbrenner, Michaela. 1997. Zur Häufigkeit von Wortlängen in englischen Texten. In: Best, Karl-Heinz (Hrsg.), *Glottometrika 16. The Distribution of Word and Sentence Length: 98-107.* Trier: WVT.

Hein, Michaela. 1997. Wortlängen in Briefen des spanischen Dichters Federico Garcia Lorca. In: Best, Karl-Heinz (Hrsg.), *Glottometrika 16. The Distribution of Word and Sentence Length: 138-144.* Trier: Wiss. Vlg. Trier.

Heinicke, Nora. 2008. Wortlängenverteilungen in franzoesischen Briefen eines Autors. *Glottometrics 16, 38-45*.

Heups, Gabriela. 1983. Untersuchungen zum Verhältnis von Satzlänge zu Clauselänge am Beispiel deutscher Texte verschiedener Textklassen. In: Koehler, R., & Boy, J. (Hrsg.), *Glottometrika 5: 113-133*. Bochum: Brockmeyer.

Hollberg, Cecilie. 1997. Wortlängenhäufigkeiten in italienischen Pressetexten. In: Best, Karl-Heinz (Hrsg.), *Glottometrika 16. The Distribution of Word and Sentence Length: 127-137.* Trier: Wiss. Verlag Trier.

Jahn, Thomas, & Uckel, Annika. 2008. Verteilung von Wortlängen in englischen Spam-E-Mails. *Glottometrics 17, 1-17.*

Janssen, Enno, & Suhren, Svenja. 2000. Wortlängenhaufigkeiten in ostfriesisch-niederdeutschen Gedichten von Hans-Hermann Briese. *Göttinger Beiträge zur Sprachwissenschaft 4, 53-62.*

Jenner, Karin. 1997. *Zur Wortkomplexität deutscher und norwegischer Texte*. Staatsexamensarbeit, Göttingen.

Jenner, Karin. 1997. *Zur Wortkomplexität deutscher und norwegischer Texte.* Staatsexamensarbeit, Göttingen.

Jing, Zhuo. 2001. Satzlängenhäufigkeiten in chinesischen Texten. In: Best, Karl-Heinz (Hrsg.), *Häufigkeitsverteilungen in Texten: 202-210*. Göttingen: Peust & Gutschmidt.

Judt, Birga. 1995. *Wortartenhäufigkeiten im Deutschen und Französischen*. Staatsexamensarbeit, Göttingen.

Kahl, Susan. 2002. Wortlängen in wogulischen Texten. *Göttinger Beiträge zur Sprachwissenschaft 7, 51-63.*

Kähler, Dierk. 1994. *Analyse zur Häufigkeit von Wortlängen in irischen Prosatexten.* Seminararbeit, Göttingen.

Kassel, Anja, & Livesey, Eleanor. 2001. Untersuchungen zur Satzlängenhäufigkeit im Englischen: Am Beispiel von Texten aus Presse und Publizistik, Literatur (Belletristik). In: *Glottometrics 1, 27-50.*

Kassel, Anja. 2002. *Zur Verteilung rhythmischer Einheiten in deutschen und englischen Texten.* Staatsexamensarbeit, Göttingen.

Kaydanova, Liliya. 2005. Zur Wortlängenhäufigkeit in usbekischen Texten. *Göttinger Beiträge zur Sprachwissenschaft 10/11, 57-66.*

Kern, Patricia. 1995. *Wortlängenhäufigkeiten in portugiesischen Pressetexten*. Seminararbeit, Göttingen.

Kiefer, Alexander. 2001. Wortlängenverteilung im Pfälzischen. In: Best, Karl-Heinz (Hrsg.), *Häufigkeitsverteilungen in Texten: 124-131*. Göttingen: Peust & Gutschmidt.

Kloepfel, Henning. 2010. *Quantitative Untersuchungen zu Komposita im Deutschen. Zur Länge von Komposita in deutschen Pressetexten.* Staatsexamensarbeit, Göttingen.

Knaus, Marina. 2008. Zur Verteilung rhythmischer Einheiten in russischer Prosa. *Glottometrics 16, 57-62.*

Knopp, Angela. 1998. *Wortlängen in franzoesischen Briefen deutscher und französischer Verfasser*. Staatsexamensarbeit, Göttingen.

Knueppel, Anke. 1997. *Untersuchungen zum Zipf-Mandelbrot-Gesetz im Deutschen.* Staatsexamensarbeit, Göttingen.

Knueppel, Anke. 2001. *Untersuchungen zum Zipf-Mandelbrot-Gesetz im Deutschen.* Staatsexamensarbeit, Göttingen. Abbr. in: Best, Karl-Heinz (Hrsg.), *Häufigkeitsverteilungen in Texten: 248-280*. Göttingen: Peust & Gutschmidt.

Kohfeldt, Anja. 1999. *Satzlängenverteilung im Französischen*. Seminararbeit, Göttingen.

Kücüker, Yüksel. 2000. *Satzlängen in türkischen Erzählungen*. Seminararbeit, Göttingen.

Kuhr, Saskia, & Mueller, Barbara. 1997. Zur Wortlängenhäufigkeit in Luthers Briefen. In: Best, Karl-Heinz (Hrsg.), *Glottometrika 16. The Distribution of Word and Sentence Length: 55-62*. Trier: WVT.

Kuhr, Saskia. 2001. Wortlängen in Liedern und Fabeln Luthers. In: Best, Karl-Heinz (Hrsg.), *Häufigkeitsverteilungen in Texten: 132-141*. Göttingen: Peust & Gutschmidt.

Laass, Francoise. 1996. Zur Verteilung der Wortlänge in deutschen Lesebuchtexten. In: Schmidt, Peter (Hrsg.), *Glottometrika 15: 181-194.* Trier: WVT.

Lips, Michael. 2003. Zur Wortlängenhäufigkeit in jiddischen Texten. *Göttinger Beiträge zur Sprachwissenschaft 9, 20-25.*

Livesey, Eleanor. 2001. *Satzlängen im Deutschen und Englischen*. Staatsexamensarbeit, Göttingen.

Marx, Michael. 2001. Zu den Wortlängen in polnischen Briefen. In: *Glottometrics 1, 52-62.*

Meier, Inga. 1997. *Wortlängenhäufigkeiten in Briefen des kubanischen Schriftstellers Jose Julian Marti*. Seminararbeit, Göttingen.

Mueller, Folke. 2001. Wortlängen in finnischen E-Mails und Briefen. *Göttinger Beiträge zur Sprachwissenschaft 8, 71-85.*

Nedtwig, Katja. 1996. *Wortkomplexität in deutschen Briefen des 19. und 20. Jhds.* Staatsexamensarbeit, Göttingen.

Niehaus, Brigitta. 1994. *Untersuchung zur Satzlängenhäufigkeit im Deutschen*. Staatsexamensarbeit, Göttingen.

Niehaus, Brigitta. 1997. Untersuchung zur Satzlängenhäufigkeit im Deutschen. In: Best, Karl-Heinz (Hrsg.), *Glottometrika 16. The Distribution of Word and Sentence Length: 213-275.* Trier: WVT.

Niehaus, Brigitta. 1997. *Wortlängen in Lesebuchtexten*. Seminararbeit, Göttingen.

Niehaus, Brigitta. 2001. Die Satzlängenverteilung in literarischen Prosatexten der Gegenwart. In: Uhlířová, Ludmila; Wimmer, Gejza; Gabriel Altmann & Reinhard Köhler (eds.), *Text as a Linguistic Paradigm: Levels, Constituents, Constructs. Festschrift in Honour of Luděk Hřebíček: 196-214*. Trier: WVT.

Nitsch, Olaf. 1997. *Wortkomplexität in Texten der Computerfachpresse*. Staatsexamensarbeit, Göttingen.

Nolte, Mahnaz. 2003. *Untersuchung zur Wortlängenverteilung in persischen Lehrbuchtexten.* Seminararbeit, Göttingen.

Ommen, Enno. 2003. *Quantitative Untersuchungen zur Syntax des Deutschen.* Staatsexamensarbeit, Göttingen.

Poppe, Stefanie. 2007. Die Verteilung von Kompositalängen in deutschen journalistischen Texten. *Göttinger Beiträge zur Sprachwissenschaft 15, 2007, 79-85.*

Prill, Natascha. 1995. *Wortlängen in Luthers Tischreden*. Seminararbeit, Göttingen.

Rheinländer, Nicole. 2000. *Sprachstatistische Untersuchungen zur Syntax des Niederdeutschen und Niederländischen.* Staatsexamensarbeit, Göttingen.

Rheinländer, Nicole. 2001. Die Wortlängenhäufigkeit im Niederländischen. In: Best, Karl-Heinz (Hrsg.), *Häufigkeitsverteilungen in Texten: 142-152*. Göttingen: Peust & Gutschmidt.

Riedemann, Gesa. 1997. Wortlängenhäufigkeiten in japanischen Pressetexten. In: Best, Karl-Heinz (Hrsg.), *Glottometrika 16. The Distribution of Word and Sentence Length: 180-184*. Trier: WVT.

Riedemann, Hagen. 1994. *Wortlängen in der Sprache der deutschen und englischen Tages- und Wochenpresse*. Staatsexamensarbeit, Göttingen.

Riedemann, Hagen. 1996. Word Length Distribution in English Press Texts. *Journal of Quantitative Linguistics 3, 265-271*.

Roettger, Winfred, & Schweers, Anja. 1997. Wortlängenhäufigkeiten in Plinius-Briefen. In: Best, Karl-Heinz (Hrsg.), *Glottometrika 16. The Distribution of Word and Sentence Length: 121-126.* Trier: Wiss. Verlag Trier.

Roettger, Winfred. 1996. The Distribution of Word Length in Ciceronian Letters. *Journal of Quantitative Linguistics 3, 68-72*.

Roettger, Winfred. 1996. *Wortlängenhäufigkeiten lateinischer Texte deutschsprachiger Autoren.* Staatsexamensarbeit, Göttingen.

Schneemann, Okke F. 2001. *Sprachstatistische Untersuchungen zu Wort- und Silbenlängen in deutschen Musikzeitschriften.* Staatsexamensarbeit, Göttingen.

Schroeder, Ulla. 1996. *Zur Wortlängenhäufigkeit im Finnischen*. Seminararbeit, Göttingen.

Schweers, Anja, & Zhu, Jinyang. 1991. Wortartenklassifizierung im Lateinischen, Deutschen und Chinesischen. In: Rothe, Ursula (Hrsg.), *Diversification processes in language: grammar: 157-165*. Hagen: Margit Rottmann Medienverlag.

Stark, Alexandra. 2001. Die Verteilung von Wortlängen in schweizerdeutschen Texten. In: Best, Karl-Heinz (Hrsg.), *Häufigkeitsverteilungen in Texten: 153-161*. Göttingen: Peust & Gutschmidt.

Stitz, Katrin. 1994. *Untersuchungen zu den Wortlängen in deutschen und russischen Briefen des 19. Jahrhunderts.* Staatsexamensarbeit, Göttingen.

Strehlow, Michael. 1997. *Satzlängen in pädagogischen Fachartikeln des 19. Jahrhunderts.* Staatsexamensarbeit, Göttingen.

Strobel, Heike. 1996. *Wortlängen in Briefen und Erzählungen von Böll und Hemingway.* Staatsexamensarbeit, Göttingen.

Suhren, Svenja. 2000. *Wortlängen in ostfriesischen Pressetexten (Kolumnen).* Seminararbeit, Göttingen.

Suhren, Svenja. 2002. *Untersuchung zum Gesetz von Zwirner, Zwirner und Frumkina am Beispiel des niederdeutschen De Lütte Prinz.* Staatsexamensarbeit, Göttingen.

Wachlin, Katharina. 2002. *Die Verteilung von Wortlängen in gesprochener Sprache.* Seminararbeit, Göttingen.

Wittek, Martin.1995. Zur Entwicklung der Satzkomplexität im gegenwärtigen Deutschen. Staatsexamensarbeit, Göttingen. Gekürzt in: Best, Karl-Heinz (Hrsg.) 2001, *Häufigkeitsverteilungen in Texten: 219-247*. Göttingen: Peust & Gutschmidt.

Wossidlo, Arnd. 1997. *Wortlängen in Briefen des argentinischen Autors Julio Cortazar*. Seminararbeit, Göttingen

Yu, Xiaoli. 2001. Zur Komplexität chinesischer Schriftzeichen. *Göttinger Beiträge zur Sprachwissenschaft 5, 121-129.*

Yu, Xiaoli. 2002. *Quantitative Aspekte in pädagogischen Fachartikeln des 19. Jahrhunderts.* Magisterarbeit, Göttingen.

Zauner, Thomas. 2003. *Die Entwicklung der Wortlänge in der Sprachgeschichte des Englischen.* Magisterarbeit, Göttingen.

Zuse, Maria. 1996. The Distribution of Word Length in English Letters of Sir Philip Sidney. *Journal of Quantitative Linguistics 3, 272-276*.

Zuse, Maria. 1998. *Silbenlängen in deutschen und englischen Pressetexten der Gegenwart.* Staatsexamensarbeit, Göttingen.

Aspects of word length

Gabriel Altmann, Lüdenscheid

The study of word length is as old as quantitative linguistics itself. It is a property situated at the surface of language, it is placed in the centre of Köhler's (1986, 2005) control cycle, i.e. it has links to many other properties, it is important for language typology and, even more interesting, it is easily accessible even if the linguist must make sometimes conventional decisions. Word length is measured in terms of the number of syllables because the syllable is the immediate (phonetic/phonological) constituent of the word. Measuring length using phonemes/graphemes/letters as units would skip an intermediate level and the Menzerathian link would show much variance. Using morph(eme)s as units of measurement would not reflect length but rather the morphological complexity of the word. Measuring in terms of morae is more or less a conventional compromise between phonetics and script (e.g. in Japanese).

Based on the long development of word length studies, word length could be seen from different points of view and for each of them some theoretical attempts were made. Under theoretical attempts we mean the subsumption of word length behaviour under some law-like hypotheses. In practice, it means the creation of models and in turn their testing on the given data. As time passes, the models become generalized and linked with other ones, boundary conditions are discovered and at last a theory may arise in which Köhler's requirements play the role of forces, and difference and differential equations show the interaction of these forces and the mechanisms which are in action in text generation and language development.

In the sequel we show ten aspects of word length investigation. Needless to say, the same can be done with every other linguistic property.

1. The spectrum of word lengths. The specific word length distribution associated with a text is obtained by counting all the words in the text with the given lengths. A large variety of theoretical probability distributions is available as potential models of such an empirical distribution. It has been observed that there is a certain regularity of distribution kinds which correspond to languages and eras. A great number of models have been tested by K.-H. Best (1996, 1996a, 1996b, 1999, 2001; Best, Brynjólfsson 1997; Best, Kaspar 1998; Best, Özmen 1996; Best, Song 1996; Best, Zhu 1994; Zhu, Best 1992, 1998; Bartens, Best 1997) who investigated about 50 languages and processed about 3000 texts. The "best" empirical model can be chosen mechanically using a software (e.g. the *Altmann Fitter*) but the theoretical reflection should be focused on the general model (cf. Wimmer, Altmann 2005) in which there is place for interpreted parameters and which follows the historical development of language: In early stages of a language, words may have been more or less monosyllabic, then a proportion of monosyllables was, in the process of word-formation, prolonged by

affixes; later on, affixes became parts of the words and new affixes or inflections arose, etc. Thus, if there are disyllabic words, their proportion is related to the proportion of monosyllables, i.e. $P_2 = aP_1$. Unfortunately, the proportionality is not constant, hence one writes rather $P_2 = g(2)P_1$ and in general $P_x = g(x)P_{x-1}$. The solution to this difference equation depends on the choice of the proportionality $g(x)$. The simplest $g(x)$-functions are e.g. $g(x) = 1 + a_0 + a_1/x + a_2/x^2 + \ldots$, or $g(x) = (1 + a_0 + a_1/x)^b$, etc. which are easy to solve and interpretable. One of the parameters may play the role of the ceteris paribus condition or a language constant. In this way, one obtains several discrete probability mass functions, which can be further generalized (e.g. by Feller-generalization), or in which one of the parameters itself follows a special distribution (compounding). Many of these models can be found in the articles by K.-H. Best and his pupils. It is recommended to be careful with the direct use of Fuck's model in form of mixed distributions because the number of parameters is frequently that great that the degrees of freedom do not suffice for a statistical test. Nevertheless, mixing signalizes stratification, and this can represent historical strata or borrowed words, or word-classes, etc. Even here, the decision must be made by the researcher. For example, in some Slavic languages there are words of zero syllabic length: if we consider these words as proclitics, nothing changes; but if we want zero-syllables to be taken explicitly into account, we obtain a modified distribution in which the zero-class has the probability 1-α, and all the others have αP_x. If the mode of the empirical distribution is not at $x = 1$, the given language tends to syntheticity. Thus one of the parameters can express the degree of synthetism.

2. Rank distribution. Here the "variable" is the rank and the distribution is simply a reorganization of the spectrum as discussed above. The first rank represents the class of the longest words, etc. It is in any case a decreasing function. Even if the spectrum is bell-shaped, the rank distribution is decreasing. The transformation has been shown by Boroda and Zörnig (1992). However, here also the stratificational approach (cf. Popescu, Altmann, Köhler 2010) is possible because different rank classes may coincide with classes of foreign words, with classes of derived or compound words, of parts-of-speech, etc.

Simple empirical investigations can show that using the criterion of Ord (1972) and plotting the points $<I, S>$ in a Cartesian coordinate system, we obtain different areas or straight lines for analytic languages and for synthetic languages. Here I = variance/mean, and S = third central moment/variance. For other purposes, it has been shown that the ranking fitted by the Zipf-function is a reflection of the traces of the language type: in strongly synthetic languages, Zipf's (power) function is situated below the hapax legomena; in strongly analytic languages it is situated above the hapax legomena; and in balanced languages (those not approaching the extremes), it crosses the hapax legomena (cf. Popescu, Mačutek, Altmann 2009: 104 f.). It should be tested whether this holds also for rank distributions of word lengths.

3. Position in sentence. According to a hypothesis by L. Uhlířová (1997, cf. also Fenk, Fenk-Oczlon 2006; Kelih 2012a; Fan, Grzybek, Altmann 2010), mean word length increases from the beginning to the end of a sentence. This is caused by the information structure of sentences (cf. publications by the Prague school), the "theme – rheme" structure. A sentence begins in general with the theme, the known part of its meaning, which can be kept short and continues with the rheme, the new information, which will be much longer, contain more and also less frequent words. The trend may be linear but here new hypotheses must be set up. The regularity may differ with texts, text-sorts, styles and languages, and this phenomenon could be applied also in typology.

For testing one should treat all sentences of the same length as one group and compute the mean word length in individual positions. One should not put together sentences of different lengths, however, this is another possible option. The minimum number of sentences in a length-group should be 5. Smaller groups should be rather omitted even if they may corroborate the hypothesis because data in such a small class are statistically unreliable. Different results may be obtained if one takes the means of all first words, then of all second words, etc. The interplay of clauses is evident: if the same tendency holds also for clauses, then the trend must be oscillating because clause ends may be different in each sentence. Besides, the position of the clause in the sentence may be relevant, too.

4. Position in text. Since a text is a kind of deployment of information, the same trend should appear also from the "vertical" point of view, that is, as the text grows, the mean word length in the sentences should grow, too. In order to test this hypothesis, in each sentence the mean word length must be computed separately and the regression between position of sentence in text and mean word length in sentence must be scrutinized. If the trend exists, any function must converge to a fixed value because word length cannot increase without limit, and its lower limit is 1

Another possibility of measuring the change of word length with deploying text is the partitioning of text in equally long parts, e.g. each 10 sentences long, or "natural" parts like paragraphs and compute the mean word length in individual parts. This way is more reliable and may help to reveal hidden tendencies. A special view is the analysis of a stage play in which one can study either the change of length in individual acts or in the speech of individual persons and associate it with the development of the plot.

5. Time series. If we simply measure word length from the beginning to the end of texts – ignoring sentence boundaries – we obtain a time series whose properties can be evaluated by means of well-known techniques (cf., e.g., Pawłowski 1997, 2001). Here one should not forget that whatever result one obtains, it must be interpreted linguistically. Still better, one should state a hypothesis before one begins to count. A nice result in autocorrelation of lags is merely nice,

but if it is not based on a linguistic hypothesis, it has no further use and cannot be systematized.

Consider for example the course of entropy of word lengths measured for each sentence of the text separately. Is it only a horizontal straight line? But even if no trend can be observed, there is the possibility that the course is oscillating or that there are breaks in the trend. This aspect can be of importance especially in studying dialogues or conversations, e.g. in stage plays, since here there are several speaking persons and all of them may have different styles. If we consider the distribution of word length in each sentence separately, we obtain a multiple time series; or we can compare pairs of subsequent distributions using non-parametric methods and learn some background mechanisms of text construction.

6. Word length motifs. Whatever property we consider, units can be joined into motifs, representing sequences consisting of units with a not-decreasing degree of the given property. If we find a sequence of word lengths, say, 1,2,2, 4,2,3,1,5,2 then we can set up the motifs 1-2-2-4, 2-3, 1-5, 2. These motifs have again their own lengths and can be treated as quite normal linguistic entities. In music they were introduced with a different background by M.G. Boroda (1982), in linguistics they were used for the first time by R. Köhler (2006, 2008a,b) and R. Köhler and S. Naumann (2008). The combinatorial background has been shown by J. Mačutek (2009), who applied it to motifs in Slovak poetry.

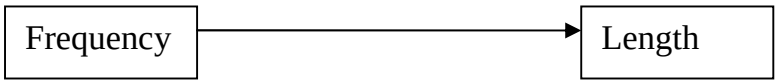
It can be conjectured that there is some relation between the form of the motifs and their placing in sentence and in text. An obvious example of such a hypothesis is the assumption that word length and word frequency motifs should reflect, to some extent, the positions of syntactic structures; for these structures correspond to parts-of-speech patterns, and words of different parts-of-speech differ in their typical lengths and frequencies. Since word length motifs are analogues of word length, one can strive both for capturing their spectrum and their rank-frequency distribution. The first tests performed by Mačutek (2009) show that the spectra abide by the hyper-Pascal distribution while the rank-frequencies can be captured by the usual Zipf-Mandelbrot distribution.

7. Distances between equal lengths. If we replace the words in the text by their syllabic lengths, just as in Aspect 5, we obtain a sequence of numbers. If the lengths are placed by the author randomly, then the distances between equal numbers abide by the distribution derived by P. Zörnig (1987). If they do not, there is a special mechanism which must be found and substantiated linguistically. Some of the causes may be anchored in the grammar of the given language, but other ones may originate from style or text sort.

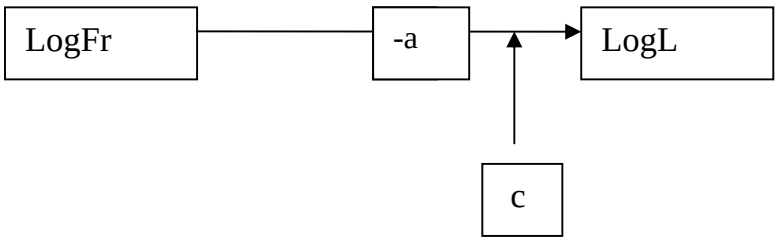
The same computation can be performed with any other property and it can be expected that some of them will have very peculiar forms. Since Zörnig's derivation this branch of research did not make pertinent progress (cf. Zörnig, in this volume).

8. Word length and phoneme inventory. Since speech needs sufficient redundancy, words should differ at least in one phoneme, but the greater the difference, the more redundancy is produced. In order to secure redundancy, languages either create longer words in dependence on the size of their phoneme inventories, or use tones or accents or vowel lengths for differentiation. In a thorough study, Nettle (1996, 1998; cf. also Čech, Altmann 2011: 77 f.) has shown that the greater is the inventory of phonemes, the shorter are the words of the language on the average. This study is not yet finished because each inventory size must be represented by more than one language. However, this aspect is in any case part of Köhler's control cycle and joins two different levels. Further investigations are necessary. Since the phonemic inventories of languages are seldom crisp sets - much depends on interpretation - this relationship may help to make decisions concerning the inventory in agreement with a law.

9. Köhler's control cycle. Length is one of the many properties of the word. In language there are no isolated properties, each of them is associated with at least another one. This association can usually be captured by a simple function. The links can be displayed by systems theoretical figures as currently used in synergetic linguistics. If, e.g., length is linked with frequency, whereby frequency is the driving force, we may symbolize it as



However, such a simple presentation does not lead to a reasonable formula, hence Köhler takes into account both the Weber-Fechner law, the requirements of the speaker (in this case the production effort) and a language constant which can express the ceteris paribus condition, and according to his prescription we would obtain



which can be solved quite simply

$$\mathrm{Log}L = -a(\mathrm{Log}F) + c,$$

from which

$$L = bF^{-a},$$

where $b = exp(c)$. Now, length can be influenced also by polysemy, morphological type of language, phoneme inventory, word class, age, number of grammatical categories, abstractness, specificity, number of synonyms, emotionality, valency, lexicon size, etc., thus we obtain a quite complex control cycle which will never be complete and whose parameters may differ with respect to individual languages. The links between properties express the kind of influence and yield a plastic picture of the mechanism of self-regulation (cf. Köhler 1986, 1990a,b,c; Zörnig, Köhler, Brinkmöller 1990).

The same results can be obtained if we conjecture that the relative rate of change of property *y*, i.e. dy/y is proportional to a function of the rate of change of *x*, i.e.

$$dy/y \approx g(x)dx$$

as has already be shown above. If we set $g(x) = -a/x$, we obtain again

$$y = bx^{-a}$$

as above. The power function - representing in some cases the Zipf-function - is frequently used in many sciences. A very complex control cycle encompassing word length is presented in Köhler (2005). Its counterpart presented in form of differential and difference equations can be found in Wimmer, Altmann (2005).

10. Laws. If one starts from the control cycle or from the approach with differential equations and tests the individual links on data from many languages adding boundary conditions if necessary, one obtains a theory in which some of the links (well corroborated empirically!) represent language laws. Of course, these are the most valuable statements in any science. If to each of the links all requirements of speaker and hearer (cf. Köhler 2005) are added in form of forces, we obtain a theory yielding functional explanations which cannot be obtained in a different way.

If the measurement of word length should be relevant for a language theory, then one must be able to show that

(a) it is not arbitrary but abides by laws,

(b) it is no isolated property that could be formed independently of other properties, on the contrary, it is linked with at least one other property and

(c) it underlies evolution and diversification.

Point (a) simply means that one postulates the existence of mechanisms controlling the formation of length. Statements about these mechanisms are called hypotheses; if they are theoretically derived and empirically well corroborated for *all languages* fulfilling certain conditions, we call them laws. Since “every-

thing abides by laws" (Bunge 1977: 17), it is not difficult to accept this assumption, on the contrary, it stimulates research even if finding these laws is not quite easy. Needless to say, one should be very cautious with forming analogies to natural laws because the forces are different. In linguistics laws, special kinds of statements are involved. In the history of linguistics (and science in general) the concept of law changed.

Point (b) is much more complicated because here we must search for all connections of length to other properties whose number is potentially infinite. They are of diverse character.

First, there are **hierarchic** relations expressing the links of word length to that of constructs (e.g. clauses, sentences, hrebs) and to the components of which words consist (e.g. syllables, morphemes). These mechanisms are known as Menzerath's law and Arens' law (cf. Menzerath 1954; Altmann 1980; Geršić, Altmann 1980; Köhler 1982, 1984; Heups 1983; Rothe 1983; Sambor 1984; Teupenhayn, Altmann 1984; Altmann, Schwibbe 1989; Hřebíček 1990, 1997; Boroda, Altmann 1991; Nemcová 1994; Prün 1994; Krott 1996; Best 2001; Weber 1998; Andres, Benešová 2011).

Second, there are **collateral** relations capturing the influence of length on other properties of the word or vice versa, e.g. on meaning/polysemy, meaning generality, frequency, polytextuality, compositionality, synonymy, some grammatical properties such as agreement, valency, full valency, etc. They form complex control cycles which are the subject of synergetic linguistics (cf. Köhler 1986, 1990a,b, 1999; 2005; Köhler, Altmann 1986; Hammerl 1991).

Third, there are **sequential** relations arising from the fact that in speech/ writing the units follow one after another and create different length configurations or distance patterns (cf. Zörnig 1987; Altmann 1988; Hřebíček 2000; Pawlowski 1997, 2001; Andersen 2005; Köhler 2006, 2007).

Fourth, the relations of word length with segmental or suprasegmental **inventories** such as the phoneme inventory, the number of different suprasegmentals, different kinds of accent, etc. (cf., e.g. Kelih 2012b).

Point (c) has two aspects associated with variability, viz. (i) the **historical** aspect comprehending the development of lengths in one language or in the human language in general. Besides, in all the circumstances mentioned in (a) and (b) time plays the role of the independent variable (cf. Wittek 2001; Ammermann 2001). Thus the age of words may be a factor influencing word length but not vice versa. (ii) Besides the historical aspect one must consider also the synchronic variation, i.e. the **diversification** because lengths can develop differently e.g. in poetry and in scientific texts and this can lead to the necessity of setting up different models in the synchronic aspect. Here the kind of text, style, idiolect, dialect, sociolect, etc. may play some role (Altmann 1985a,b, 1991, 1992, 1996a, 2005; Altmann, Best, Kind 1987; Beöthy, Altmann 1984a,b; Best 1994, 1997a; Köhler 1991; Raether, Rothe 1991; Rothe 1991; Ziegler 1998, 2000; Fan, Altmann 2008; Fan, Popescu, Altmann 2008; Popescu, Kelih, Best, Altmann 2008; Popescu, Altmann 2008; Sanada, Altmann 2009; Tuzzi, Popescu, Altmann

2009). The study of diversification is also a means for classification and discrimination of texts.

Inferring from ontogenesis to phylogenesis of length it may be conjectured that in all domains of language entities consisting of one unit were primary e.g. monosyllabic words, monolexemic sentences, etc. The prolongation of units arose by self-organization caused by many different motives. Some of them were the small inventory of phonemes/sounds causing increasing homonymy leading to loss of redundancy, specification of meaning, increasing grammatical complexity, the impact of culture, aesthetic reasons and many others. It will never be possible to elicit the reasons in individual cases particularly because there are opposite tendencies, too, e.g. the loss of word length in the development of language (e.g. Polynesian languages as compared with the Austronesian ones). In any case, a self-organisational jump is always accompanied by self-regulatory changes in other domains which are the object of synergetic linguistics.

Realizing a research program of this extent concerning lengths is a task for generations of linguists. It is difficult not only because of the number of languages and texts but especially because of the collateral relationships of length. Their examination alone requires a construction of at least a partial language theory comprising not only the study of length. It does not play any role whether one begins with research into length or into some other property because in any case one must decipher as much as possible of the complete mechanism of language. Earlier or later the laws of length must be embedded in the theory of language. Nevertheless, it is very advantageous to begin with word length because it is easy to access, easy to operationalize and relatively easy to measure. However, the number of boundary conditions is that overwhelming that in a better developed theory we must reconcile with an extensive battery of models.

The situation can be graphically approximated as shown in Figure 1.1.

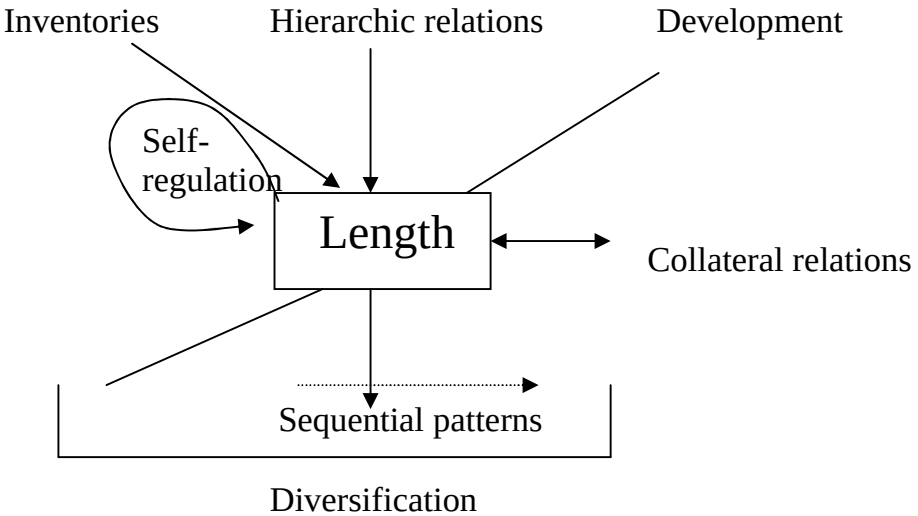


Figure 1.1. The scheme of mechanisms exerting influence on length

None of these aspects is "more important" than the other ones, everything happens simultaneously and is interconnected with everything else, directly or indirectly. It can be, however, expected that as science progresses, still further aspects may appear, e.g. other psycholinguistic or sociolinguistic ones. One should not expect to obtain "smooth solutions", e.g. a unique formula for a certain aspect. This requirement can not hold even for one single language. It is caused by (Zipf´s force of) diversification which steadily disturbs the equilibria and drives the development. The most difficult task will not be the setting up formulas but the interpretation of parameters which in many cases represent only boundary conditions or hardly measurable quantities (e.g. the requirements of communication participants).

The study of word length is a possible beginning. Counting is relatively simple: the boundaries of syllables need not be known, it is sufficient to determine their number in the word. Of course, there are many problems originating from the differences of linguistic schools or opinions of individual researchers, or even from the fuzzy state of a given phenomenon.

Nevertheless, the first two aspects are relatively simple. K.-H. Best and his students came to the conclusion that some distribution models can be considered as laws belonging to the same background model. They and other researchers obtained the binomial d., modified binomial d., extended positive binomial d., negative binomial d., modified negative binomial, Cohen-negative binomial d., hyper-Poisson d., Singh-Poisson d., Hirata-Poisson d., geometric d., Merkyte-binomial d., Poisson d., modified Poisson d., lognormal d., Fucks d., Dacey-Poisson d., Poisson-uniform d., Conway-Maxwell-Poisson d., hyper-Pascal d., Cohen-Poisson d., Pólya d., mixed Poisson d., Consul-Jain-Poisson d., Palm-Poisson d., Pandey-Poisson d., and innovation does not seem to cease.

Here we can list the languages in which word length was measured so far: Arabic, Belorussian, Bulgarian, Chinese, Czech, Danish, Dutch, Early English, English, Estonian, Faroese, Finnish, French, Gaelic, German, Middle German, Swiss, Early New High German, German dialects, Gothic, Greek, Hungarian, Italian, Japanese, Jiddish, Kechua, Korean, Kymric, Latin, Low German, Low Sorbian, Maori, Mordvinian, New High German, New Icelandic, Norwegian, Old Church Slavic, Old High German, Old Greek, Classical Hebrew, Old Icelandic, Persian, Polish, Portuguese, Russian, Saami, Serbocroatian, Slovak, Slovenian, Spanish, Swedish, Turkish, Ukrainian, Uzbek, Vogul (Mansi).

For surveys of word length research see Schmidt (1996), Best (1997, 2001), Grzybek (2006) and the attached references.

Now, the aim of this research is to find two kinds of laws concerning word length: (1) Those in which word length is not associated with other properties, and (2) those in which it is a function of another property or in which another property is the dependent variable, i.e. the links between word length and other properties. Of course, one can perform this research using ad hoc distributions or functions yielding good results, but the aim is to find either a unique background or to substantiate why the given model is adequate for the given data. This is not

a work in synchronic linguistics but a research encompassing both the present state and the stepwise evolution of every language. Nevertheless, one must begin with individual cases and show the forces and requirements influencing the given dependence or distribution. Here we have to do not only with different languages (with different grammars and history) but also with all kinds of variation: individual styles, age and education of the writer, text sorts, dialects, etc. The more we find, the more our horizon disappears, i.e. we find an expanding research space.

If one arrives at the persuasion that a language phenomenon abides by a law, then our task consists in the foundation of this persuasion or hypothesis. A deductive or theoretical foundation consist in deriving the hypothesis from axioms - which can be represented by some preliminary assumptions -, from existing theories or from established laws in form of consequences. An inductive or empirical foundation consists in the agreement of the hypothesis with as many data as possible. A third, very effective way of substantiation is the possibility of deriving consequences from the given hypothesis. Whatever way we choose, we do not arrive at some final truth but only at the level of acceptability indicating the agreement of the hypothesis/law with the criteria whose fulfilment is considered necessary at the given state of the science. The steps on this way are as follows:

On the first step the hypothesis must be explicitly formulated, all possible factors should be declared as known or unknown quantities, and the hypothesis should be translated into a *mathematical form.* In principle it is always possible but the number of alternatives in a new scientific domain is always very great. If there is already a background theory, it is advantageous to show that the hypothesis is one of its consequences. Thus the hypothesis and the theory corroborate one another. We say that the hypothesis has been systematized.

The second step is collecting relevant *data* which are crucial for the hypothesis. In many cases one collects “data” before one formulates the hypothesis, a procedure leading frequently to disappointment. In linguistics there are different data sources, e.g. dictionaries, texts, historical, etymological, frequency and synonymy dictionaries, dialectal, sociolinguistic, psycholinguistic sources, etc. The hypothesis must take into account the known boundary conditions. The way of collecting data is postponed to Chapter 2.

The third step is the *testing* the proposed hypothesis on relevant data. One compares the prediction of the hypothesis with the data and after evaluating the divergence one accepts or rejects it. The iterative procedure can be schematically presented as given in Figure 1.2 (cf. Kliemann, Müller 1972: 463). But a rejection does not mean automatically that the model itself is “wrong”; the error may be hidden in false data, unreliable data, insufficient sampling, a not satisfactory testing, etc. Hence a rejection is always a motivation for checking the whole problem from its beginning.

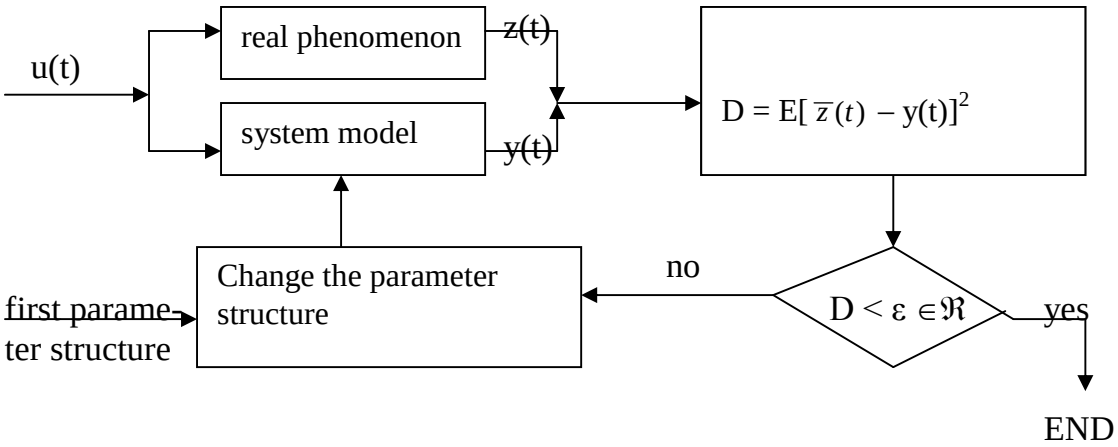


Figure 1.2. Testing and developing a hypothesis (Kliemann, Müller 1972)

Here the real phenomenon will be inserted into the model and a function of the squared deviations from the model will be computed. If it is smaller than a chosen ε, the model will be accepted, otherwise one must change something. Using probability distributions as models, one applies mostly the chi-square test whose problems are well known in statistics; using a function/curve as a model one computes the determination coefficient.

References

Altmann, G. (1980). Prolegomena to Menzerath´s law. *Glottometrika 2, 1-10.*

Altmann, G. (1985a). Semantische Diversifikation. *Folia Linguistica 19, 177-200.*

Altmann, G. (1985b). Die Entstehung diatopischer Varianten. Ein stochastisches Modell. *Zeitschrift für Sprachwissenschaft 4, 139-155.*

Altmann, G. (1991). Modelling diversification phenomena in language. In: Rothe (1991), *33-46.*

Altmann, G. (1992). Two models for word association data. *Glottometrika 13, 105-120.*

Altmann, G. (1996a). Diversification processes of the word. *Glottometrika 15, 102-111.*

Altmann, G. (1988). *Wiederholungen in Texten*. Bochum: Brockmeyer.

Altmann, G. (2005). Diversification processes. In: Köhler, R., Altmann, G., Piotrowski, R.G. (eds.), *Quantitative Linguistics. An International Handbook 648-659*. Berlin: de Gruyter.

Altmann, G., Best, K.-H., Kind, B. (1987). Eine Verallgemeinerung des Gesetzes der semantischen Diversifikation. *Glottometrika 8, 130-139.*

Altmann, G., Schwibbe, M. (1989). *Das Menzerathsche Gesetz in informationsverarbeitenden Systemen*. Hildesheim: Olms.

Ammermann, S. (2001). Zur Wortlängenverteilung in deutschen Briefen über einen Zeitraum von 500 Jahren. In: Best, K.-H. (ed.), *Häufigkeitsverteilungen in Texten:*: *59-91*. Göttingen: Peust & Gutschmidt.

Andersen, S. (2005). Word length balance in texts: Proportion constancy and word-chain-lengths in Proust's longest sentence. *Glottometrics 11, 2005, 31-49.*

Andres, J., Benešová, M. (2011). Fractal analysis of Poe's Raven. *Glottometrics 21, 2011, 73-98*

Bartens, H.-H., Best, K.-H. (1997). Wortlängen im Tscheremissischen (Mari). *Finnisch-Ugrische Mitteilungen 20, 1-20.*

Best, K.-H. (1996). Zur Bedeutung von Wortlängen, am Beispiel althochdeutscher Texte. *Papiere zur Linguistik 55, 141-152.*

Best, K.-H. (1996a). Word length in Old Icelandic songs and prose texts. *Journal of Quantitative Linguistics 3, 97-105.*

Best, K.-H. (1996b). Zur Wortlängenhäufigkeit in schwedischen Pressetexten. In: Schmidt, P. (ed.), *Glottometrika 15, 147-157. Trier: WVT.*

Best, K.H. (ed.) (1997). *Glottometrika 16. The distribution of word and sentence length.* Trier: WVT.

Best, K.-H. (1999). Wortlängen. In: Köhler, R., Altmann, G., Piotrowski, R.G. (eds.), *Quantitative Linguistics. An International Handbook: 260-27*3. Berlin: de Gruyter..

Best, K.-H. (ed.) (2001). *Häufigkeitsverteilungen in Texten.* Göttingen: Peust & Gutschmidt Verlag.

Best, K.-H., Brynjólfsson, E. (1997). Wortlängen in isländischen Briefen und Pressetexten. *Skandinavistik 27, 24-40.*

Best, K.-H., Kaspar, I. (1998). Wortlängen in färöischen Briefen. *Naukovyj visnyk Černiveckogo Universitetu. Vypusk 41. Hermans'ka filolohija 3-14.*

Best, K.-H., Özmen, E. (1996). Wortlängenhäufigkeiten in türkischen Texten und ihre linguistischen Implikationen. *Archiv orientální 64, 19-30.*

Best, K.-H., Song, H.Y. (1996). Wortlängen im Koreanischen. *Asian and African Studies 5, 39-49.*

Best, K.-H., Zhu, J. (1994). Zur Häufigkeit von Wortlängen in Texten deutscher Kurzprosa. In: Klenk, U. (ed.), *Computation Linguae II: 19-30.*. Stuttgart: Steiner.

Best, K.-H., Zinenko, S. (1998). Wortlängenverteilung in Briefen A.T. Twardowskis. *Göttinger Beiträge zur Sprachwissenschaft 1, 7-19.*

Boroda, M.G. (1982). Häufigkeitsstrukturen musikalischer Texte. In: Orlov, Ju.K., Boroda, M.G., Nadarejšvili, I.Š. (eds.), *Sprache, Text, Kunst: Quantitative Analysen: 231-262*. Bochum: Brockmeyer.

Boroda, M.G., Altmann, G. (1991). Menzerath´s law in musical texts. In: Boroda, M.G. (ed.), *Musikometrika 3, 1-13.* Bochum: Brockmeyer.

Bunge, M. (1977). *The furniture of the world*. Dordrecht-Boston: Reidel.

Čech, R., Altmann, G. (2011). *Problems in Quantitative Linguistics 3.* Lüdenscheid: RAM.

Fan, F., Altmann, G. (2008). On meaning diversification in English. *Glottometrics 17, 66-78.*

Fan, F., Grzybek, P., Altmann, G. (2010). Word length in sentence. *Glottometrics 20, 70-109.*

Fan, F., Popescu, I.-I., Altmann, G. (2008). Arc length and meaning diversification in English. *Glottometrics 17, 79-86.*

Fenk, A., Fenk-Oczlon, G. (2006). Within-sentence distribution and retention of content words and function words. In: Grzybek, P. (ed.), *Contributions to the science of text and language. Word length studies and related issues: 157-170*. Dordrecht: Springer.

Geršić, S., Altmann, G. (1980). Laut – Silbe – Wort und das Menzerathsche Gesetz. *Frankfurter Phonetische Beiträge 3, 115-123.*

Grzybek, P. (2006). History and methodology of word length studies. In: Grzybek, P. (ed.), *Contributions to the science of text and language. Word length studies and related issues: 15-90*. Dordrecht: Springer.

Hammerl, R. (1991). *Untersuchungen zur Struktur der Lexik. Aufbau eines lexikalischen Basismodells.* Trier: Wissenschaftlicher Verlag.

Heups, G. (1983). Untersuchungen zum Verhältnis von Satzlänge zu Clauselänge am Beispiel deutscher Texte verschiedener Textklassen. In: Köhler, R., & Boy, J. (eds.), *Glottometrika 5: 113-133*. Bochum: Brockmeyer.

Hřebíček, L. (1990a). The Menzerath-Altmann´s law on the semantic level. In: Köhler, R., Boy, J. (eds.), *Glottometrika 3, 113-133.* Bochum: Brockmeyer.

Hřebíček, L. (1990b). The constants of Menzerath-Altmann law. *Glottometrika 12, 61-71*.

Hřebíček, L. (2000). *Variation in sequences*. Prague: Oriental Institute.

Kelih, E. (2012a). On the dependency of word length on text length. Empirical results from Russian and Bulgarian parallel texts. In: Naumann,S. Grzybek, P., Vulanović, R., Altmann, G. (Eds.), *Synergetic Linguistics. Text and Language as Dynamic Systems: 67-80.* Wien: Praesens.

Kelih, E. (2012b). *Die Silbe in slawischen Sprachen. Von der Optimalitätstheorie zu einer funktionalen Interpretation.* München/Berlin: Sagner (= Specimina philologiae Slavicae, 168).

Kliemann, W., Müller, N. (1976). *Logik und Mathematik für Sozialwissenschaftler 2.* München: Fink.

Köhler, R. (1982). Das Menzerathsche Gesetz auf Satzebene. In Lehfeldt, W., Staruss, U. (eds.), *Glottometrika 4: 103-113.* Bochum: Brockmeyer.

Köhler, R. (1984). Zur Interpretation des Menzerathschen Gesetzes. *Glottometrika 6, 177-183.*

Köhler, R. (1986). *Zur linguistischen Synergetik. Struktur und Dynamik der Lexik.* Bochum: Brockmeyer.

Köhler, R. (1990). Elemente der synergetischen Linguistik. In: Hammerl, R. (ed.), *Glottometrika 12, 179-188*. Bochum: Brockmeyer.

Köhler, R. (1990b). Linguistische Analyseebenen, Hierarchisierung und Erklärung im Modell der sprachlichen Selbstregulation. *Glottometrika 11, 1-18.*

Köhler, R. (1990c). Synergetik und sprachliche Dynamik. In: Koch, Walter A. (ed.), *Natürlichkeit der Sprache und Kultur: 96-112*. Bochum: Brockmeyer.

Köhler, R. (1991). Diversification of coding methods in grammar. In: Rothe (1991), *47-55.*

Köhler, R. (1999). Der Zusammenhang zwischen Lexemlänge und Polysemie im Maori. In: Genzor, J., Ondrejovič, S. (eds.), *Pange lingua: 27-33*. Bartislava: Veda.

Köhler, R. (2005). Synergetic linguistics. In: Köhler, R., Altmann, G., Piotrowski, R.G. (eds.), *Quantitative Linguistics. An International Handbook: 760-774*. Berlin: de Gruyter.

Köhler, R. (2006). The frequency distribution of the lengths of length sequences. In: Genzor, J., Bucková, M. (eds.), *Favete Linguis. Studies in Honor of Viktor Krupa: 145-152*. Bratislava: Slovak Academic Press.

Köhler, R. (2008a). Word length in text. A study in the syntagmatic dimension. In: Mislovičová, S. (ed.), *Jazyk a jazykoveda v pohybe*: 416-421. Bratislava: VEDA, Vydavatel'stvo SAV

Köhler, R. (2008b) Sequences of linguistic quantities. Report on a new unit of investigation. *Glottotheory 1(1), 115-119.*

Köhler, R., **Altmann, G.** (1986). Synergetische Aspekte der Linguistik. *Zeitschrift für Sprachwissenschaft 5, 253-265*.

Köhler, R., **Naumann, S.** (2008). Quantitative text analysis using L-, F- and T-segments. In: Preisach, Ch., Burkhardt, H., Schmidt-Thieme, L., Decker, R. (eds.), *Data Analysis, Machine Learning and Applications: 637-645*. Berlin: Springer.

Krott, A. (1996). Some remarks of the relation between word length and morpheme length. *Journal of Quantitative Linguistics 381), 29-37.*

Menzerath, P. (1954). *Die Architektonik des deutschen Wortschatzes*. Bonn: Dümmler.

Nemcová E. (1994). On two realizations of Menzerath´s law. *J. of Quantitative Linguistics 1, 107-112.*

Nettle, D. (1996). Segmental inventory size, word length, and communicative efficiency. *Linguistics 33, 359-367.*

Nettle, D. (1998). Coevolution of phonology and the lexicon in twelve languages of West Africa. *Journal of Quantitative Linguistics 5(3), 240-245.*

Ord, J.K. (1972). *Families of frequency distributions*. London: Griffin.

Pawlowski, A. (1997). Time series analysis in linguistics: application of the ARIMA method to cases of spoken English. *Journal of Quantitative Linguistics 4(1-3), 203-221.*

Pawlowski, A. (2001) *Metody kwantytatywne w sekwencyjnej analizie tekstu*. Warszawa: Universytet Warszawski, Katedra Lingwistyki Formalnej.

Popescu, I.-I., Altmann, G. (2008). On the regularity of diversification in English. *Glottometrics 17, 94-108.*

Popescu, I.-I., Altmann, G., Köhler, R. (2010). Zipf's law - another view. *Quality and Quantity 44(4), 713-731.*

Popescu, I.-I., Kelih, E., Best, K.-H., Altmann, G. (2008). Diversification of the case. *Glottometrics 18, 32-39.*

Popescu, I.-I., Mačutek, J., Altmann, G. (2009). *Aspects of word frequencies.* Lüdenscheid: RAM.

Prün, C. (1994). Validity of Menzerath-Altmann´s law: graphic representation of language, information processing systems and synergetic linguistics. *Journal of Quantitative Linguistics 1(2), 148-155.*

Rather, A., Rothe, U. (1991). Semantische Diversifikation der deutschen Komposita: `Substantiv plus Substantiv'. In: Rothe, U. (ed), 85-91.

Rothe, U. (ed.) (1991). *Diversification processes in language: grammar*. Hagen: Rottmann.

Rothe, U. (1993). Wortlänge und Bedeutungsmenge: Eine Untersuchung zum Menzerathschen Gesetz an drei romanischen Sprachen. In: Köhler, R., Boy, J. (eds.), *Glottometrika 5, 101-112.* Bochum: Brockmeyer.

Sambor, J. (1984). Menzerath´s law and the polysemy of words. In: Köhler,R., Boy, J. (eds.), *Glottometrika 6, 94-114*. Bochum: Brockmeyer.

Sanada, H., Altmann, **G.** (2009). Diversification of postpositions in Japanese. *Glottometrics 19, 70-79.*

Schmidt, P. (ed.) (1992). *Glottometrika 15. Issues in general linguistic theory and the theory of word length.* Trier: WVT.

Teupenhayn, R., Altmann, G. (1983). Clause length and Menzerath´s law. *Glottometrika 6, 127-138.*

Tuzzi, A., Popescu, I.-I., Altmann, G. (2009). Parts-of-speech diversification in Italian texts. *Glottometrics 19, 42-48*

Uhlířová, L. (1997). O vztahu mezi délkou slova a jeho polohou ve větě. *Slovo a slovesnost 57, 174-184.*

Weber, S. (1998). *Das Menzerathsche Gesetz in gesprochener Sprache*. Trier: Magisterarbeit.

Wimmer, G., Altmann, G. (2005). Unified derivation of some linguistic laws. In: Köhler, R., Altmann, G., Piotrowski, R.G. (eds.), *Quantitative Linguistics. An International Handbook: 791-807*. Berlin: de Gruyter.

Wittek, M. (2001). Zur Entwicklung der Satzlänge im gegenwärtigen Deutschen. In: Best, K.-H. (Hrsg.) 2001, *Häufigkeitsverteilungen in Texten: 219-247.* Göttingen: Peust & Gutschmidt.

Zhu, J., Best, K.-H. (1992). Zum Wort im modernen Chinesisch. *Oriens Extremus 35, 45-60.*

Zhu, J., Best, K.-H. (1998). Wortlängenhäufigkeiten in chinesischen Kurzgeschichten. *Asian and African Studies 7, 45-51.*

Ziegler, A. (1998). Word length in Portuguese texts. *Journal of Quantitative Linguistics 5, 115-120.*

Ziegler, A. (2000). Word length in Romance languages. A complemental contribution. *Journal of Quantitative Linguistics 7, 65-68.*

Zörnig, P. (1987). A theory of distance between like elements in a sequence. *Glottometrika 8, 1-22.*

Zörnig, P. (2013). Distances between words of equal length in a text. *In this volume.*

Zörnig, P., **Boroda, M.** (1992). The Zipf-Mandelbrot law and the interdependencies between frequency structure and frequency distribution in coherent texts. *Glottometrika 13, 205-218.*

Zörnig, P., Köhler, R., Brinkmöller, R. (1990). Differential equation models for the oscillation of the word length as a function of frequency. *Glottometrika 12, 25-40.*

Word Length in Chinese

Lu Wang, Trier

Abstract. This paper presents a study on word length distribution and length-polysemy relationships in Chinese on data from a static corpus, derived from the *Modern Chinese Dictionary (5th Edition)*, and a dynamic corpus, extracted from the *People's Daily* news corpus. The results show that, in modern Chinese, the word length distribution abides by the Positive Cohen-negative binomial model; and word length is related with polysemy.

Keywords: Word length, polysemy, Chinese

1. Introduction

Word length studies in many languages, such as German (Altmann & Best, 1996; Best, 1997), Ketschua (Best & Medrano, 1997), Sami (Bartens & Best, 1997), Icelandic (Best, 1996) and Chinese (Zhu & Best,1997) show that the distribution of this variable is associated with certain distributions. These studies were performed on data from texts (this kind of material will be called here *dynamic*). The present paper studies word length also on dictionary material (in this paper: *static*). Furthermore, a model of the relationship between word length and polysemy is tested.

2. Data

The two corpora adopted are: a static corpus and a dynamic corpus.

Static corpus: This corpus is derived from the *Modern Chinese Dictionary (5th Edition)*, the most authoritative dictionary. Words used merely in classical Chinese are not taken into account. Therefore, instead of the entire dictionary, only 61,969 words are sampled from them to compose this static corpus. Word length is measured in terms of the number of syllables, which equates to the number of Chinese characters.

Dynamic corpus: We compiled the dynamic corpus from *People's Daily* Corpus January 1998, which is a Chinese one-million words news corpus with word segmentation and part-of-speech tagging. Our data was extracted following four rules:

1) All the words containing numbers or alphabetic characters are ignored; only words consisting of Chinese characters are taken into account.

2) Word length is measured in terms of the number of syllables.
3) The extent of polysemy for each word is determined by *Modern Chinese Dictionary (5th Edition).* Words not included in the dictionary are omitted in the length-polysemy study, but still counted for the word length distribution.
4) As mentioned above, the polysemy of many words is unknown, so we select 500 suitable texts from the original 3147 one.

Finally, this dynamic corpus consists of 500 texts, 291278 tokens, 26630 types and 15565 types tagged polysemy.

3. Word length distribution

Table 1 and Figure 1 show the word length distribution in the dictionary. We can see that, in modern Chinese vocabulary, two-syllable words occupy the biggest proportion, nearly two thirds. This fact seems to be an exception to the economic rules. Because it is not the shortest words which take the largest proportion. However, if we attempt to use one-syllable words entirely, the language would have to create more syllables and characters. Then, inevitably, the inventory of characters would be enlarged which would violate the economic rules. Therefore, to save that effort, the most economical method is to extend those words with the minimum length — one more syllable. As a result of the dominating proportion of two-syllable words, the mean word length is 2.18. Words over 4 syllables are less than 1%. This again reveals a universal fact which abides by economic rules: short words are of more preference.

Table 1
Word length distribution in the static corpus.

Word length	**Lexemes**	**Proportion**	**Word length**	**Lexemes**	**Proportion**
1	8240	13.297%	7	58	0.094%
2	41465	66.912%	8	53	0.086%
3	5964	9.624%	9	6	0.01%
4	5758	9.292%	10	3	0.005%
5	279	0.45%	12	1	0.002%
6	142	0.229%			
Mean word length = 2.18					

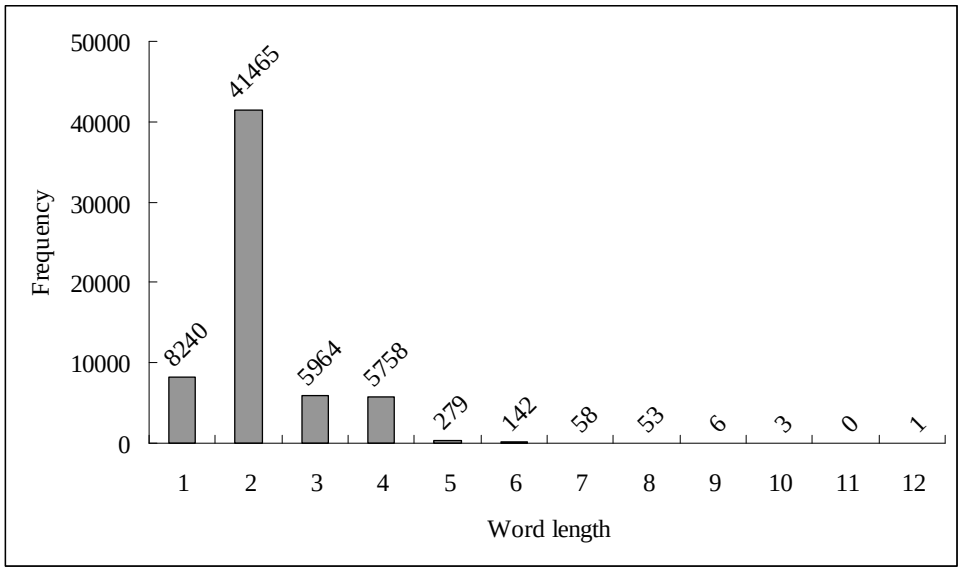


Figure 1. Word length distribution in the static corpus

Table 2 and Figure 2 illustrate the word length distribution from a dynamic point of view. Compared with the situation of static data, the amount of one-syllable tokens is huge. Consequently, the mean token length is lowered to 1.7, shorter than the mean type length. This may be due to segmentation criteria adopted in processing the news corpus. In principle, language as a self-regulating system has its properties interrelated (Köhler, 1986). Here we have data from different sources — static dictionary and dynamic texts. What makes them different is: frequency. As Grotjahn and Altmann (1991) described: "In a running text where all tokens are counted it is frequency which dominates the other factors: in the distribution of word length the frequencies of short words will be high compared to that of long ones." Conversely, frequency plays no role in counting types and lexemes in dictionary. Hence the mean type length is 2.34, close to that of static data. Moreover, they overlap each other.

Table 2
Word length distribution in the dynamic corpus

Word length	**Tokens**	**Types**	**Word length**	**Tokens**	**Types**
1	115269	2170	6	89	63
2	155239	16799	7	43	26
3	13720	4581	8	9	9
4	6219	2727	10	1	1
5	688	253	12	1	1
Mean token length = 1.7			Mean type length = 2.34		

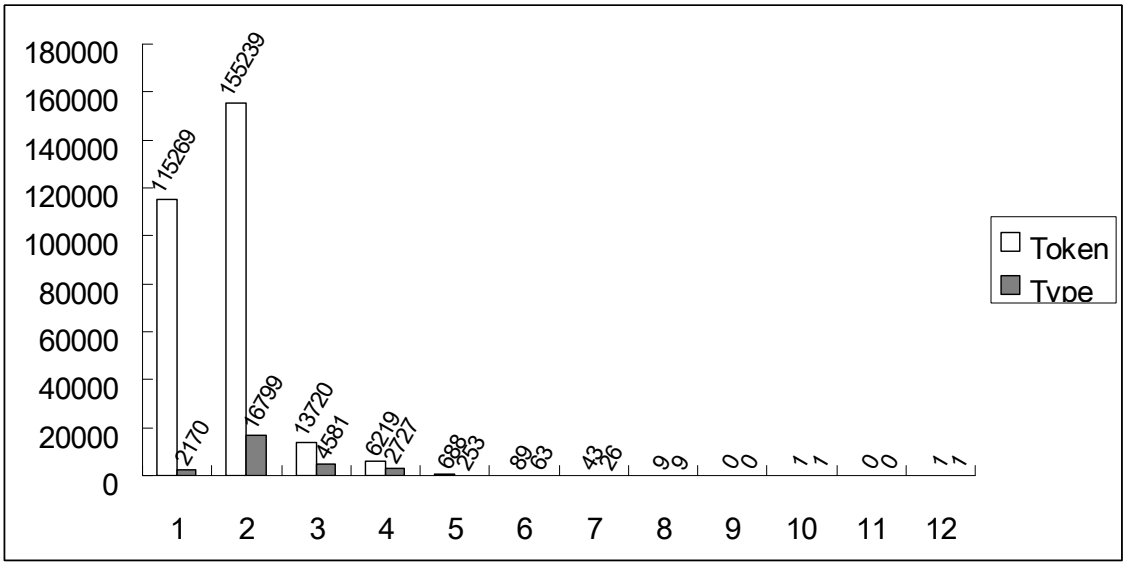


Figure 2. Word length distribution in the dynamic corpus.

The top 10 frequent words are listed in Table 3 with frequency and word classes. Obviously, all are one-syllable words, and 7 among them can play the role of function words, which reveals that high frequent words are short and more often functional.

Table 3
The top 10 frequent words in the dynamic corpus

Word	Frequency	Word classes
的	18678	aux.
在	3773	v. prep. adv.
是	3701	adj. v. n. pron.
了	3624	aux.
和	3470	adj. v. n. prep.
一	2502	num. aux. n.
不	1778	adv.
有	1699	v
上	1343	n. v.
为	1334	prep.

We used the data to test three models (the Extended logarithmic, Right truncated modified Zipf-Alekseev and Positive Cohen-negative binomial), which present the best results and fit all three kinds of data.

Table 4 and Figure 3 show the fitting results of the static (dictionary) data. The Positive Cohen-negative binomial distribution performs better in the tail part, while the other two models overestimated the number of lexemes. All the R^2 values are good, but the problem is: $C > 0.05$ makes them unacceptable. As is well known, the Chi-square and also the coefficient C display this kind of problem with very large data sets. Nevertheless, previous experience with the selected distributions and optical inspection encourage maintaining the models as plausible hypotheses.

Table 4
Fitting results of the static word length distribution

Word length	**Lexemes**	**Extended logarithmic**	**Right truncated modified Zipf-Alekseev**	**Positive Cohen-negative binomial**
1	8240	8240	8240	7929.95
2	41465	38966.55	38894.92	38579.35
3	5964	9622.55	9607.96	10565.85
4	5758	3168.31	3072.5	3255.42
5	279	1173.59	1165.97	1069.89
6	142	463.7	500.07	366.27
7	58	190.85	235.22	128.97
8	53	80.79	118.94	46.36
9	6	34.91	63.76	16.93
10	3	15.33	35.88	6.26
11	0	6.81	21.02	2.34
12	1	5.61	12.75	1.42
		$\theta = 0.4939$ $\alpha = 0.8670$	$a = 2.1186$ $b = 0.7423$ $n = 12.0000$ $\alpha = 0.1330$	$k = 0$ $p = 0.5892$ $\alpha = 0.9578$
		$X^2 = 4719.5460$ $P(X^2) = 0.0000$ $DF = 9$ $C = 0.0762$ $R^2 = 0.9823$	$X^2 = 5114.6111$ $P(X^2) = 0.0000$ $DF = 7$ $C = 0.0825$ $R^2 = 0.9818$	$X^2 = 4929.2814$ $P(X^2) = 0.0000$ $DF = 8$ $C = 0.0795$ $R^2 = 0.9763$

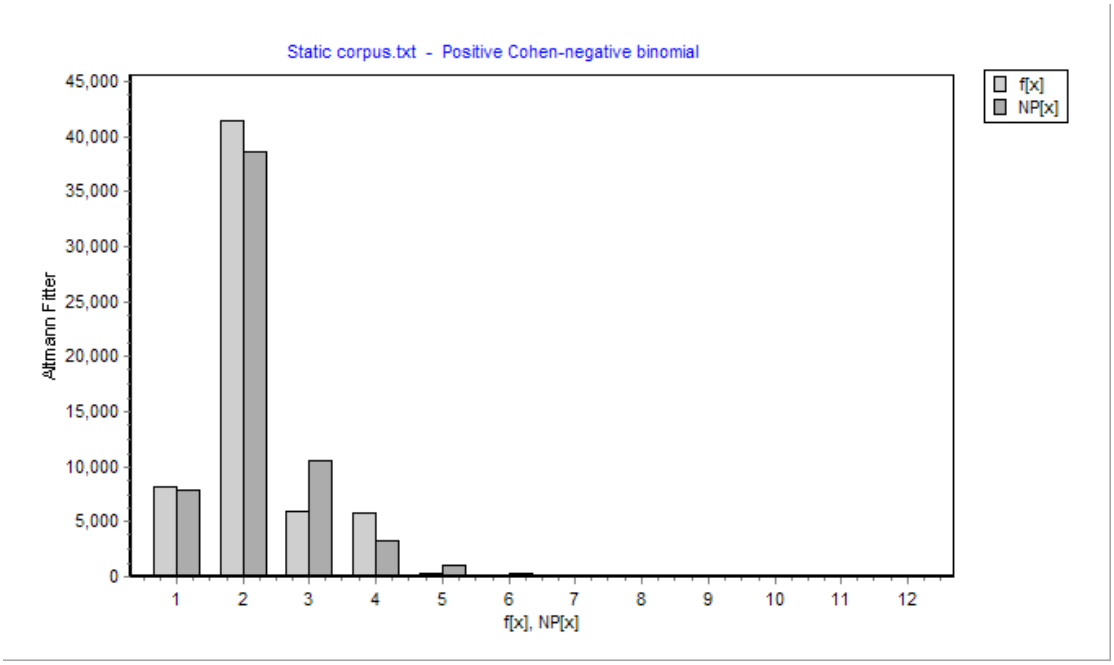


Figure 3. Fitting the Positive Cohen-negative binomial distribution to the static data

The situation of type is demonstrated in the following Table 5 and Figure 4. The discrepancy coefficient values are relatively better, all acceptable. This may be due to a smaller size compared with the dictionary data. Furthermore, the fact that the Positive Cohen-negative binomial distribution fits better also can be seen from C values.

Table 5
Fitting results of the type length distribution

Word length	**Types**	**Extended logarithmic**	**Right truncated modified Zipf-Alekseev**	**Positive Cohen-negative binomial**
1	2170	2170.00	2170.00	2132.41
2	16799	16853.81	16619.12	16325.22
3	4581	4631.62	4856.81	5505.65
4	2727	1697.09	1699.75	1808.60
5	253	699.57	680.37	584.63
6	63	307.60	301.54	186.94
7	26	140.89	144.72	59.31
8	9	66.37	74.06	18.70
9	0	31.92	39.95	5.87
10	1	15.59	22.52	1.84
11	0	7.71	13.19	0.57
12	1	7.82	7.98	0.26

		θ = 0.5496 α = 0.9185	a = 1.4428 b = 0.8880 n = 12.0000 α = 0.0815	k = 1.3473 p = 0.6977 α = 0.9537
		X^2 = 1307.83 P(X^2) = 0.0000 DF = 9 C = 0.0491 R^2 = 0.9948	X^2 = 1329.94 P(X^2) = 0.0000 DF = 7 C = 0.0499 R^2 = 0.9945	X^2 = 936.12 P(X^2) = 0.0000 DF = 6 C = 0.0352 R^2 = 0.9920

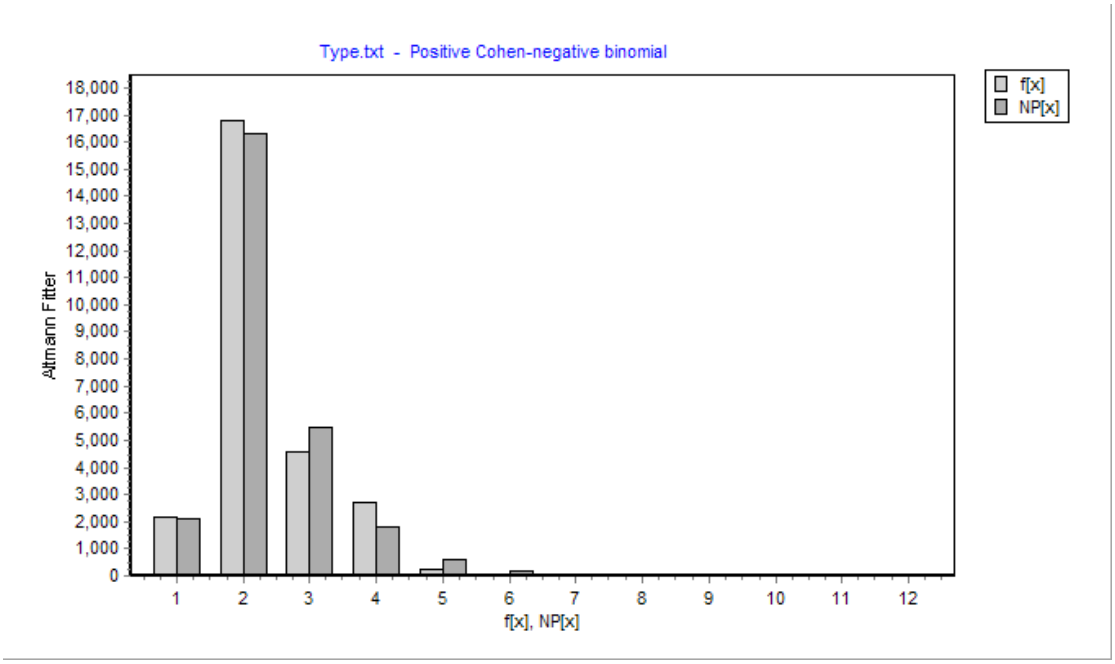


Figure 4. Fitting the Positive Cohen-negative binomial distribution to the type length data

When examining the token data, the Dacey-negative binomial model presents a superior fitting. Although its C value is unacceptable in the test on the other data, here it is smaller than 0.01, which means a "very good fit".

A general remark on modelling is here in order. Usually, one derives different models for discrete and continuous data respectively. Here one mostly assumes that these properties are deeply rooted in the reality and what we perceive are reflections of reality. Unfortunately this is not true. Whether we use discrete or continuous models is not relevant. Thinking in these categories is merely our way of concept formation. We device concepts which allow us the best orientation in the reality, yield good descriptions, classifications and if we have much luck, good concept formation allows us to derive and corroborate laws. But both discrete and continuous models are merely approximations to what we call "truth".

Table 6
Fitting results of the token length distribution

Word length	**Tokens**	**Dacey-negative binomial**	**Right truncated modified Zipf-Alekseev**	**Extended logarithmic**	**Positive Cohen-negative binomial**
1	115269	114781.23	115269	115269	114243.38
2	155239	154674.19	152464.85	150669.89	150431.06
3	13720	15378.59	18132.19	20596.67	22141.67
4	6219	4236.51	3727.99	3754.1	3666.37
5	688	1399.59	1049.46	769.78	647.57
6	89	500.72	362.84	168.37	119.14
7	43	187.42	145.11	38.36	22.55
8	9	72.23	64.71	8.99	4.36
9	0	28.41	31.41	2.15	0.85
10	1	11.35	16.31	0.52	0.17
11	0	4.59	8.96	0.13	0.03
12	1	3.18	5.15	0.04	0.84
		k = 0.1985 p = 0.5575 α= 0.5575	a = 4.6125 b = 0.3565 n = 12 α= 0.3957	θ = 0.2734 α = 0.6043	k = 0 p = 0.7792 α= 0.9162
		X^2 = 2021.62 P(X^2) = 0.00 DF = 8 C = 0.0069 R^2 = 0.9997	X^2 = 3297.67 P(X^2) = 0.00 DF = 7 C = 0.0113 R^2 = 0.9989	X^2 = 4099.83 P(X^2) = 0.00 DF = 6 C = 0.0141 R^2 = 0.9976	X^2= 5175.65 P(X^2) = 0.00 DF = 5 C = 0.0178 R^2= 0.9967

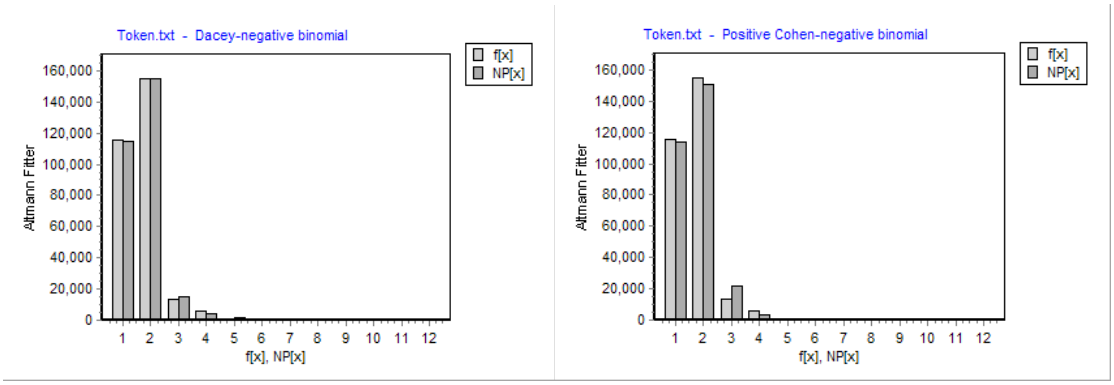


Figure 5. Fitting the Dacey-negative binomial and Positive Cohen-negative binomial distributions to the token length data

There is no problem in transforming a discrete model in a continuous one or vice versa (cf. Mačutek, Altmann 2007). The same holds for the chi-square test and the determination coefficient. Working with chi-square is inadequate if the sample is very large - as is well known - but no statistician tells us what means "large" in relation to the sample size. Thus, the use of the determination coefficient is mostly more effective and more reliable.

4. Word length and polysemy

The quantitative relation between word length and polysemy has been studied in many languages: German, Hungarian and Slovak by Altmann, Beőthy and Best (1982); German, Swedish and Indonesian by Fickermann, Markner-Jäger, Rothe (1984); Maori by Köhler (1999). Therefore, we set up the following hypothesis to test whether it holds also in Chinese:

Hypothesis: *In Chinese, word length (L) is a function of polysemy (P) following a power law and vice versa:*

$$P = aL^b + 1 \tag{1}$$

$$L = cP^d + 1 \tag{2}$$

Here we add 1 to the power functions because both word length and polysemy can not be smaller than 1.

4.1. Fitting $P = aL^b + 1$

Word length and the corresponding mean polysemy in the dictionary are shown in Table 7, from which we can see that the longer a word is, the smaller is its polysemy. When one or more morphemes are added to the word, which make it grow longer, the meaning will become more precise, i.e. its polysemy decreases. In this dictionary, each word with 7 syllables or more has only one meaning. As shown in Table 8, mean polysemy of one-syllable words in dynamic corpus is 4.90917, nearly twice larger than that in the dictionary. One of the important influencing factors is the criterion used in word segmentation. Often, Chinese linguists tend to form as small segments as possible when the continuously written text has to be split into words. This tendency causes a larger proportion of one-syllable words in the data.

Table 7
Word length and mean polysemy in static corpus

Word length	**Mean polysemy**	**Power function (1)**
1	2.50995	2.51008
2	1.25679	1.26024
3	1.11553	1.09304
4	1.03960	1.04485
5	1.01071	1.02546
6	1.00709	1.01603
7	1	1.01084
8	1	1.00773
9	1	1.00573
10	1	1.00439
12	1	1.00276
		a = 1.51008452 b = - 2.53671849 R^2 = 0.9995

Table 8
Word length and mean polysemy in dynamic corpus

Word length	**Mean polysemy**	**Power function (1)**
1	4.90917	4.90793
2	1.40572	1.43549
3	1.18854	1.12065
4	1.03514	1.04853
5	1.05882	1.02395
6	1	1.01344
7	1	1.00825
		a = 3.90793253 b = -3.16570229 R^2 = 0.9994

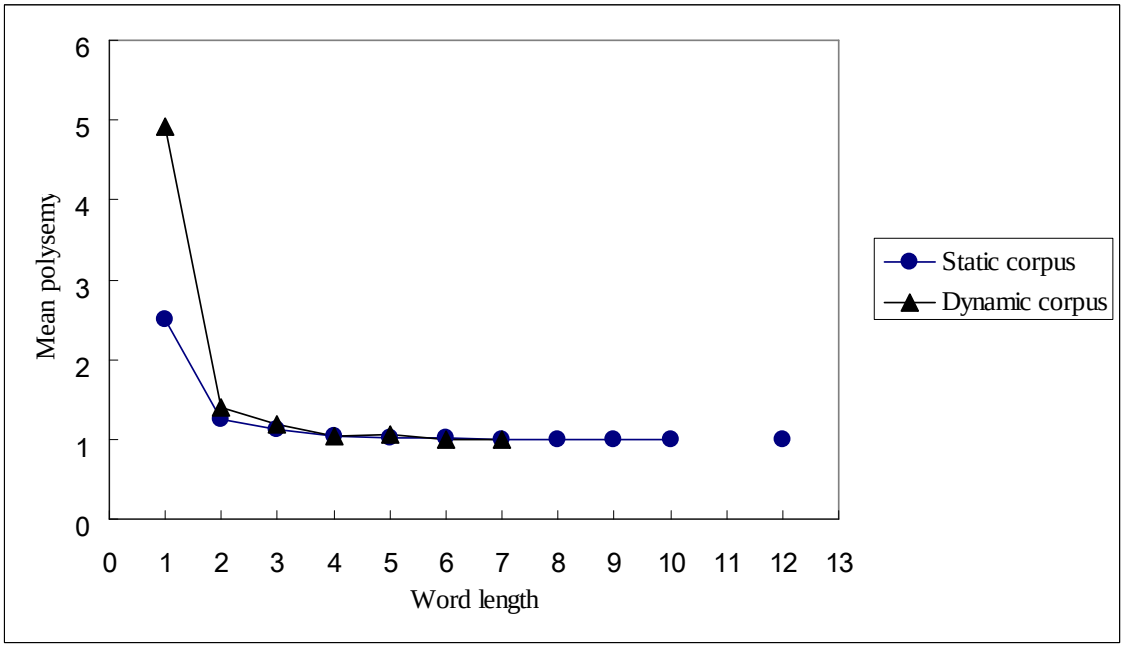


Figure 6. Polysemy as a function of word length in static and dynamic corpora.

The results of fitting the power function (1) to the data in the two corpora are shown in Figure 7 and Figure 8. The coefficients of determination are 0.9995 and 0.9994.

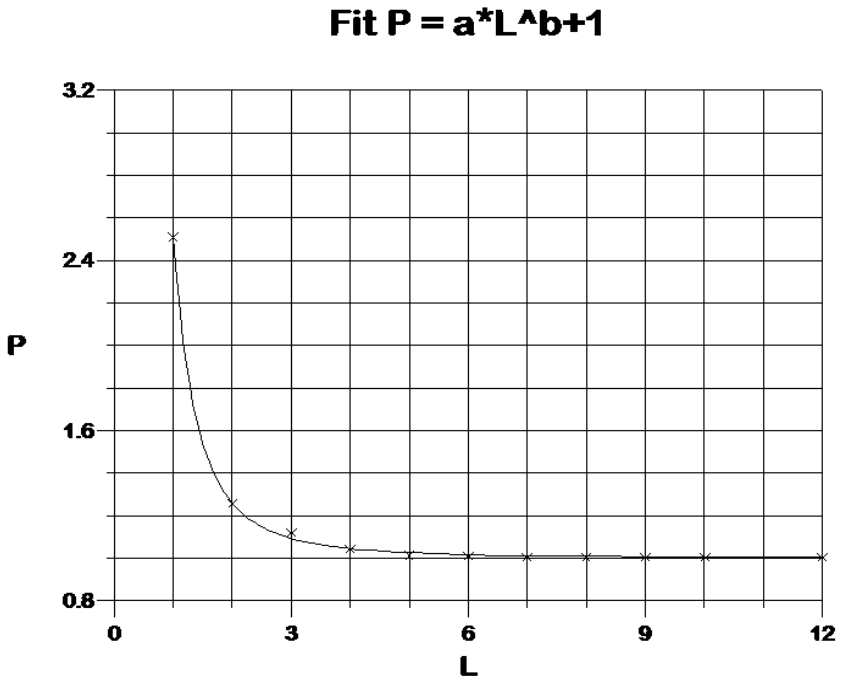


Figure 7. Fitting the power function (1), $P = aL^b + 1$, to static data

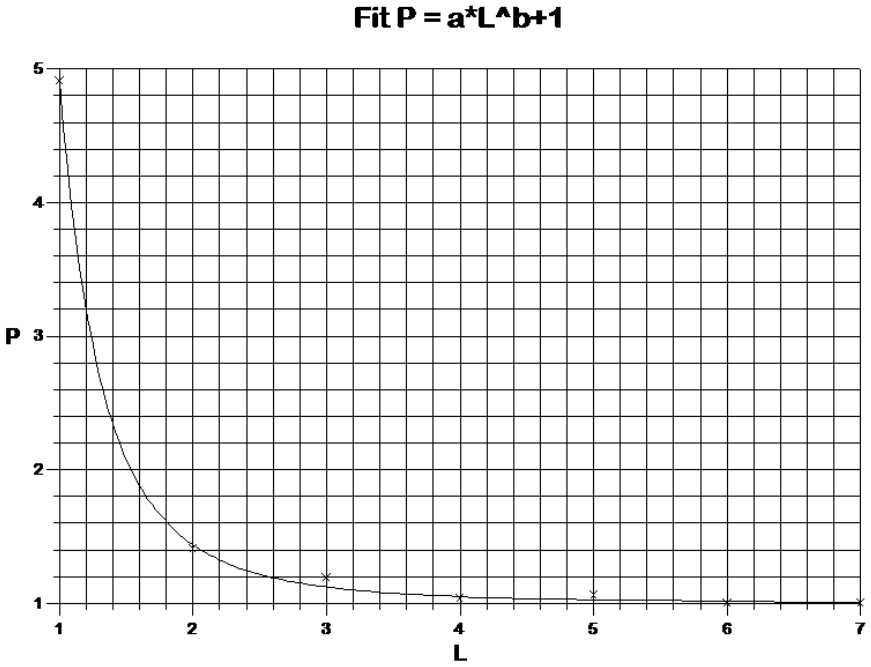


Figure 8. Fitting the power function (1), $P = aL^b + 1$, to the dynamic data

4.2 Fitting $L = cP^d + 1$

Table 9 and Table 10 show the polysemy and mean word length in static and dynamic corpus separately. Figure 9 illustrates the comparison between static and dynamic corpus, which are close to each other.

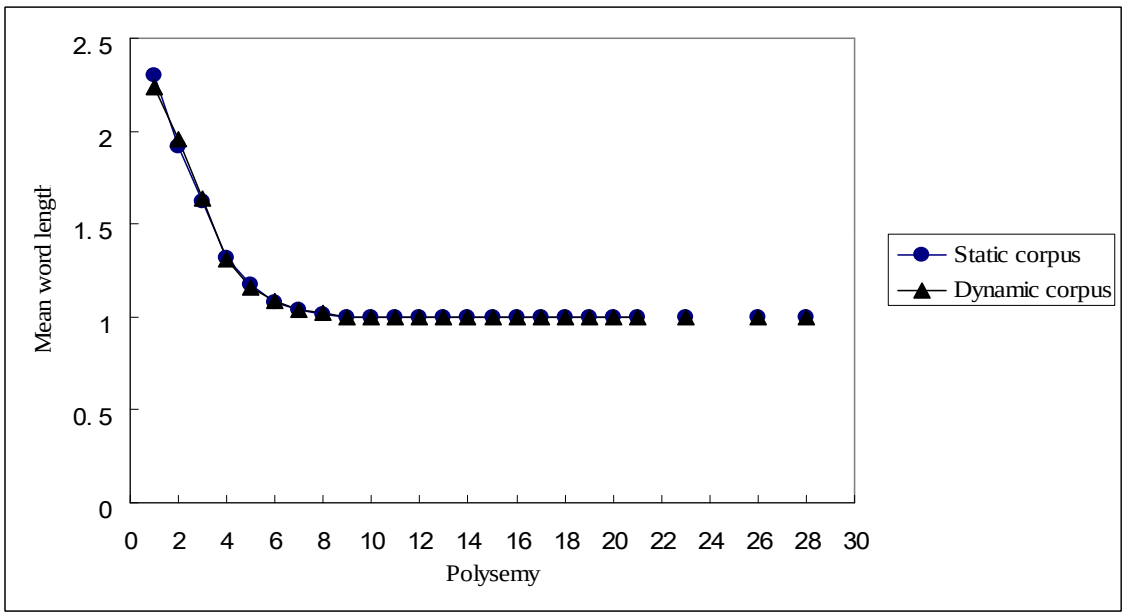


Figure 9. Word length as a function of polysemy in static and dynamic corpora.

Table 9
Polysemy and mean word length in the dictionary

Polysemy	**Mean word length**	**Power function**	**Polysemy**	**Mean word length**	**Power function (2)**
1	2.30375	2.42691	13	1	1.06765
2	1.91338	1.62599	14	1	1.06195
3	1.62368	1.38659	15	1	1.05707
4	1.32128	1.27462	16	1	1.05285
5	1.1758	1.21064	17	1	1.04918
6	1.07519	1.1696	18	1	1.04595
7	1.0400	1.1412	19	1	1.04309
8	1.0177	1.12048	20	1	1.04054
9	1	1.10474	21	1	1.03826
10	1	1.09241	23	1	1.03433
11	1	1.08251	26	1	1.02968
12	1	1.0744	28	1	1.02718
			c = 1.4269; d = -1.1886; R^2 = 0.9041		

Table 10
Polysemy and mean word length in the news corpus

Polysemy	**Mean word length**	**Power function**	**Polysemy**	**Mean word length**	**Power function (2)**
1	2.23969	2.38182	13	1	1.07025
2	1.95455	1.61775	14	1	1.06445
3	1.64041	1.38573	15	1	1.05949
4	1.30769	1.27617	16	1	1.05519
5	1.1571	1.21311	17	1	1.05144
6	1.08411	1.17244	18	1	1.04814
7	1.03676	1.14417	19	1	1.04521
8	1.01961	1.12346	20	1	1.04259
9	1	1.10768	21	1	1.04025
10	1	1.09527	23	1	1.03621
11	1	1.08529	26	1	1.0314
12	1	1.07709	28	1	1.02881
			c = 1.3818; d = -1.1614; R^2 = 0.8813		

The result of fitting the power function (2) to static data is demonstrated in Figure 10 and to the dynamic data in Figure 11. The values of R^2 are 0.9041 and 0.8813, which are good enough. Yet, that of the power function (1) are 0.9995 and 0.9994, which are slightly more significant fits. This is the same finding as for Maori (Köhler, 1999).

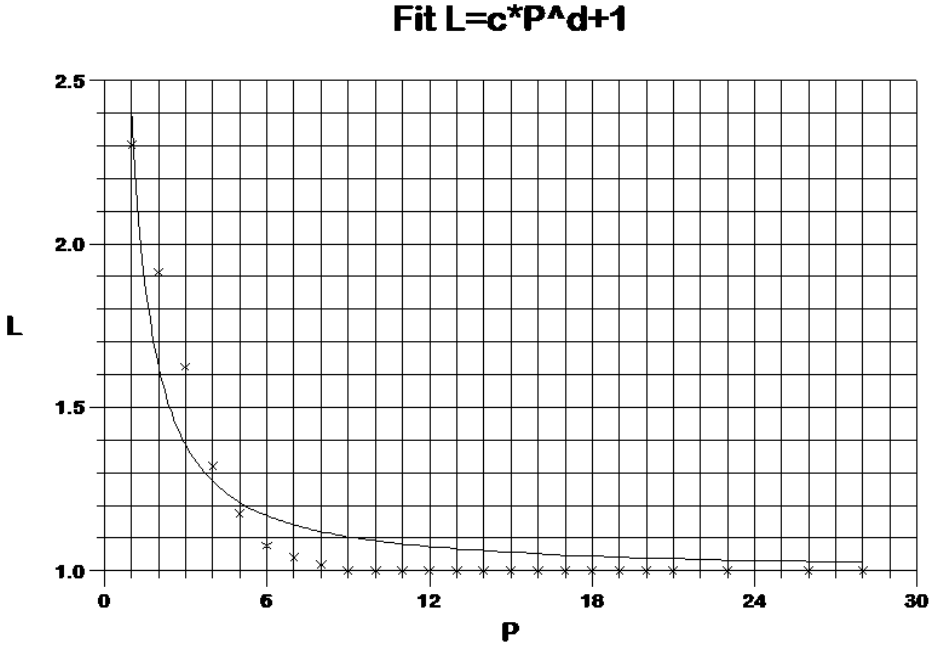


Figure 10. Fitting power function (2) $L = cP^d + 1$ to the static data

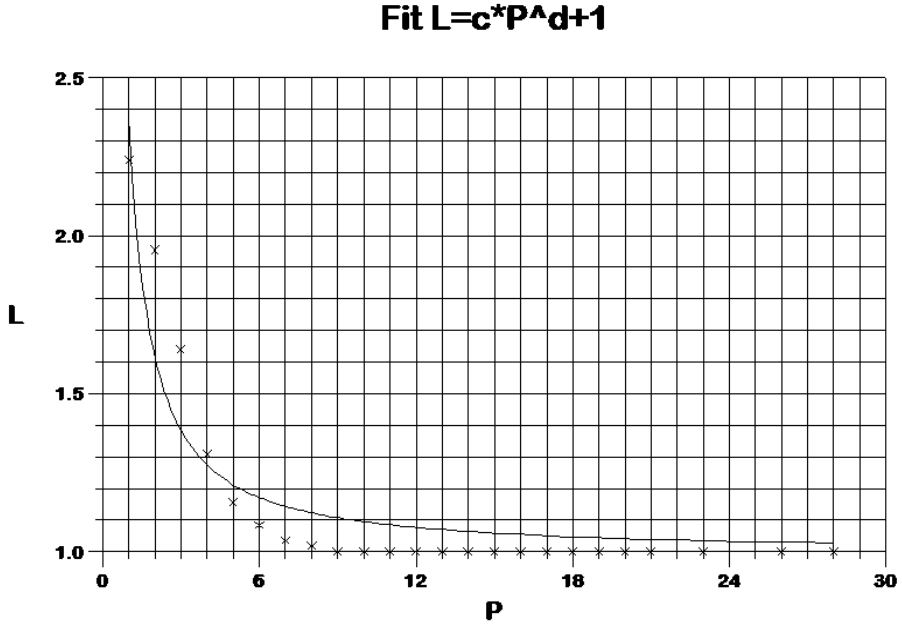


Figure 11. Fitting the power function (2) $L = cP^d + 1$ to dynamic data

5. Conclusion

In this paper, we investigated the word length distribution in Chinese. Both static and dynamic data are distributed lawfully. Among the models, the Positive Cohen-negative binomial distribution can best describe the three kinds of data. The hypothesis of the relationship between word length and polysemy is successfully tested on both corpora.

ACKNOWLEDGEMENTS
This work is supported by the China Scholarship Council and the National Social Science Foundation of China (Grant No. 11&ZD188).

References

Altmann, G., Beőthy, E., Best, K.-H. (1982). Die Bedeutungskomplexität der Wörter und das Menzerathsche Gesetz. *Zeitschrift für Phonetik, Sprachwissenschaft und Kommunikationsforschung 35, 537-543*.

Altmann, G., Best, K-H. (1996). Zur Länge der Wörter in deutschen Texten. *Glottometrika 15, 166-180.*

Bartens, H-H., Best, K-H. (1997). Word Length Distribution in Sami Texts. *Journal of Quantitative Linguistics 4, 45-52.*

Best, K-H. (1996). Word Length in Old Icelandic Songs and Prose Texts. *Journal of Quantitative Linguistics 3, 97-105*.

Best, K-H. (1997). Wörtlängen in mittlehochdeutschen Texten. *Glottometrika 16,40-54*.

Best, K-H. (1997). Wörtlängen in Ketschua-Texten. *Glottometrika 16, 204-212.*

Fickermann, I., Markner-Jäger, B, Rothe, U. (1984). Wortlänge und Bedeutungskomplexität. *Glottometrika 6, 115-126.*

Grotjahn, R., Altmann, G. (1991). Modelling the distribution of word length: some methodological problems. In: Köhler, R., Rieger, B. (eds.), *Contributions to Quantitative Linguistics 141-153*. Dordrecht: Kluwer.

Köhler, R. (1986). *Zur linguistischen Synergetik: Struktur und Dynamik der Lexik.* Bochum: Brokmeyer.

Köhler, R. (1999). Der Zusammenhang zwischen Lexemlänge und Polysemie im Maori. In: Ondrejovič, S., Genzor, J. (eds.), *Pange lingua. Zborník na počest´ Viktora Krupu*: *27-33.* Bratislava: Veda.

Mačutek, J., Altmann, G. (2007). Discrete and continuous modelling in quantitative linguistics. *Journal of Quantitative Linguistics 14(1), 81-94.*

Zhu, J., **Best, K-H.** (1997). Zur Modellierung der Wörtlängen im Chinesischen. *Glottometrika 16, 185-194*.

Bemerkungen zur Wissenschaftsgeschichte der Linguistik: Frequenz und quantitative Methoden bei Baudouin de Courtenay und Hugo Schuchardt

Emmerich Kelih, Vienna

0. Einleitung

Der Jubilar, Karl-Heinz Best hat sich in seinem umfangreichen Œuvre mit einer Vielzahl von unterschiedlichen Forschungsbereichen der Quantitativen Linguistik (wie z.B. der Anwendung exakter Methoden in der Sprachwandelforschung, Entlehnungen, Modellierung von Häufigkeitsverteilungen unterschiedlicher sprachlicher Einheiten, insbesondere der Wortlänge, der Rhythmusforschung, Probleme der sprachlichen Diversifikation) auseinandergesetzt. Neben dieser Vielzahl von Forschungsbereichen der Quantitativen Linguistik hat Karl-Heinz Best auch eine ganze Serie von Artikeln (insbesondere in der Zeitschrift *Glottometrics*) zur Wissenschaftsgeschichte der Quantitativen Linguistik verfasst.

In diesen Arbeiten, die in der Regel als Porträts einzelner Wissenschaftler auf dem Gebiet der Quantitativen Linguistik konzipiert sind, wird eine ganze Reihe von hervorragenden Pionieren, insbesondere aus dem Bereich der deutschsprachigen Linguistik, im Detail vorgestellt[1]. In besonderer Weise ist u.a. die ausführliche Vorstellung von Paul Menzerath (vgl. Best 2007a) zu nennen, der u.a. durch seine empirischen Beobachtungen der Wort- und Silbenstruktur die Basis für eines der wichtigsten Sprachgesetze der Quantitativen Linguistik, das Menzerath'sche Gesetz, geschaffen hat. Vorgestellt werden aber auch Linguisten, die möglicherweise etwas weniger bekannt sind, dafür aber für jeweils ausgewählte Teilbereiche der Quantitativen Linguistik wichtige Überlegungen und Ansätze für weiterführende statistische Untersuchung von Phänomenen geliefert haben. Zu erwähnen ist u.a. in diesem Zusammenhang z.B. Otto Behaghel (Best 2007b), der u.a. mit seinem „Gesetz der wachsenden Glieder" grundsätzliche Prinzipien der syntaktischen Struktur von Sprachen aufgedeckt hat, die mit Ergänzungen erst im gegenwärtigen Kontext in einen synergetischen Kontext eingebettet (vgl. dazu Köhler 2012) wurden.

In besonderer Weise ist auf Karl Marbe (vgl. dazu Best 2005a) und den Indogermanisten und Neogräzisten Albert Thumb (vgl. dazu Best/Kotrasch 2005)

[1] Nicht unerwähnt bleiben sollte die unermüdliche Arbeit von Karl-Heinz Best beim Verfassen von vielen Beiträgen zur Quantitativen Linguistik in der Online-Enzyklopädie Wikipedia. Auch wenn im akademischen Bereich des öfteren Vorbehalte gegenüber derartige Enzyklopädieprojekte angemeldet werden, sollte man nicht die Augen vor der Realität verschließen, dass u.a. die Wikipedia zu einer der ersten Anlaufstellen für Informationssuche und -recherche geworden ist. Daher ist an dieser Stelle dieses Engagement in besonderer Weise hervorzustreichen.

zu verweisen, die – dies wird durch die Arbeiten von Karl-Heinz Best deutlich – in enger Zusammenarbeit an Problemen der Analogieforschung und des sprachlichen Rhythmus gearbeitet haben. Neben der Vorstellung der inhaltlichen und methodologischen Schwerpunkte von ausgewählten Sprachwissenschaftlern zeigt die wissenschaftsgeschichtliche Auseinandersetzung mit Arbeiten aus dem 19. und dem Anfang des 20. Jahrhunderts insbesondere einen erstaunlich hohen Vernetzungsgrad zwischen einzelnen Wissenschaftlern (wie z.B. die Tatsache, dass Paul Menzerath Schüler von Albert Thumb gewesen ist vgl. dazu Best/Kotrasch 2005: 82) und einen bemerkenswerten interdisziplinären Zugang zu den Problemen der quantitativen Sprach- und Textanalyse.

Dieser Punkt, der Vernetzungsgrad und die methodologische Offenheit der Sprachwissenschaft gegenüber quantitativen und statistischen Verfahren am Übergang vom 19. ins 20. Jahrhundert soll im Folgenden anhand zweier bekannter Sprachwissenschaftler demonstriert werden, die bislang im Kontext der Quantitativen Linguistik noch nicht die notwendige Aufmerksamkeit erlangt haben. Es handelt sich hierbei einerseits um Baudouin de Courtenay und andererseits um Hugo Schuchardt, beides Wissenschaftler[2], die die Linguistik am Übergang vom 19. ins 20. Jahrhundert maßgeblich mitgeprägt haben und dabei – wie nun noch zu zeigen sein wird – in jeweils unterschiedlicher Art und Weise Querbezüge zur Quantitativen Linguistik aufweisen. Zuvor sind allerdings einige einleitende Bemerkungen zu wissenschaftshistorischen Zugängen zur Sprachwissenschaft angebracht.

1. Wissenschaftsgeschichte: Einleitende Reflektionen

Als die zentrale Aufgabe der Wissenschaftsgeschichte wird die rationale Rekonstruktion der Erkenntnisgewinnung und Erkenntnisgenerierung gesehen. Es geht also um die Darstellung des historischen Verlaufs einer Wissenschaft als die Abfolge von konkurrierenden Theorien und Forschungsansätzen. In dem Verständnis der analytischen Wissenschaftstheorie sollte sich daher die Wissenschaftsgeschichte insbesondere auf die Rekonstruktion von zu einem gegebenen Zeitpunkt generierten Forschungshypothesen, die im Ideal den Anforderungen der Falsifikation und intersubjektiven Überprüfbarkeit genügen sollen (vgl. dazu Lakatos 1982, Stegmüller 1979), fokussieren.

Es ist hier nicht der Ort die komplexen Schwierigkeiten und Probleme der Übertragung von durch die analytische Wissenschaftstheorie geprägten Ansätze auf die Geschichte der Quantitativen Linguistik zu diskutieren. Es genügt der Hinweis, dass hinsichtlich der Methoden die zur Diskussion stehen (unterschied-

[2] Der im Jahr 2008 erschienene Briefwechsel zwischen Baudouin de Courtenay und Hugo Schuchardt (Eismann/Hurch 2008) ist ein zentraler Ausgangspunkt, um die Gemeinsamkeiten und Unterschiede dieser beiden Wissenschaftlern in Hinsicht auf die Bedeutung statistischer Methoden bzw. der Frequenz als einer zentralen Eigenschaft sprachlicher Systeme herauszuarbeiten.

liche Bereiche der Statistik und Mathematik) ohne Zweifel eine Nähe zu naturwissenschaftlichen Ansätzen gegeben ist, während der Untersuchungsbereich, die Analyse von sprachlichen Systemen und Texten an sich, historisch gesehen in erster Linie vor allem im geisteswissenschaftlichen und philologischem Bereich stattgefunden hat. Oder in anderen Worten ausgedrückt, eine Wissenschaftsgeschichte der Quantitativen Linguistik muss und sollte nicht nur auf einen „harten Kern" von falsifizierbaren Hypothesen reduziert werden, sondern es ist auch durchaus legitim – und dies wird in den wissenschaftshistorischen Arbeiten von Karl-Heinz Best in lebendiger Weise vorgeführt – den breiteren wissenschafts- und kulturgeschichtlichen Hintergrund von zu einem bestimmten Zeitpunkt gewonnenen Erkenntnissen zu beleuchten und zu reflektieren. Dies wäre durchaus im Sinne der konstruktiven Theorie der Wissenschaftsgeschichte von Mittelstraß (1974) zu verstehen.

In vielen Fällen kann sich durch das in den Vordergrundrücken der wissenschaftlichen Tätigkeit einer Person, der von ihr geäußerten Vermutungen, Überlegungen und Andeutungen aus dem Bereich der Sprach- und Textanalyse eine anregende und nachhaltige Wirkung für einen aktuell geführten Diskurs entfalten. Durch eine entsprechende Bezugnahme auf ältere Ansätze und Erkenntnisse wird in vielen Fällen überhaupt erst eine Transformation einer Vermutung bzw. Spekulationen in eine empirisch überprüfbare Hypothese in die Wege geleitet, die wiederum als kleiner Baustein für eine Theorie bzw. ein Forschungsparadigma dienen kann. Unter diesen Voraussetzungen und der Akzeptanz einer deduktiven Herangehensweise an linguistische Probleme eröffnet sich insgesamt für die Geschichte der Quantitativen Linguistik ein umfangreiches Feld, welches trotz einer Vielzahl von Arbeiten zwar weiterhin als ein Desiderat in Erscheinung tritt, aber zumindest Schritt für Schritt bestellt werden kann.

2. Baudouin de Courtenay

Zu beginnen ist mit Jan Baudouin de Courtenay (1845-1929), der als bekanntester Vertreter der Kazaner Schule gilt, und der aufgrund seiner z.T. wegweisenden Untersuchungen in vielen linguistischen Bereichen (Schriftlinguistik, Phonologie, Morphologie, Dialektologie, Lexikographie, Sprachtypologie, Soziolinguistik, kindersprachlicher Erwerb u.v.m.) in vielerlei Hinsicht als Pionier angesehen werden muss und als solcher schon Gegenstand vielfacher Abhandlungen geworden ist, so dass bereits von einer „Boduenistik" gesprochen wird (vgl. dazu u.a. Mugdan 1984, Adamska-Sałaciak 2005, Budziak 1997, Stankiewicz 1972).

Der Beitrag von Baudouin de Courtenay zur Quantitativen Linguistik ist in Kelih (2008: 49-54) ausführlich dargelegt worden. An dieser Stelle sind daher ausschließlich die wichtigsten Eckpunkte in Erinnerung zu rufen. In erster Linie ist es – eine im heutigen Kontext oft vermisste – methodologische und theoretische Offenheit und Breite gegenüber der Anwendung von unterschiedlichen Methoden und Forschungsansätzen. Auch wenn gegen Ende des 19. Jahrhundert

die Sprachwissenschaft hauptsächlich nur mehr als Geisteswissenschaft angesehen wird, hat Baudouin de Courtenay mehrfach für den Einsatz „naturwissenschaftlicher“ (insbesondere statistischer und quantitativer) Methoden plädiert. Die Offenheit gegenüber der Anwendung von unterschiedlichen Methoden ist aber vor allem dadurch begründet, dass sich Baudouin de Courtenay des hybriden onto- und phylogenetischen Status der Sprache bewusst war.

Als konkretes linguistisches Forschungsfeld, welches für eine quantitative Erforschung geeignet angesehen wird, ist die Sprachtypologie im weitesten Sinne. In diesem Punkt unterscheidet sich diese Forderung von Baudouin de Courtenay nach einer quantitativen Sprachtypologie im Grunde genommen wenig von älteren Ansätzen aus dem Bereich der vergleichend-historischen Sprachwissenschaft. Darauf haben u.a. Best (2006) bzw. Grzybek/Kelih (2005) hingewiesen. Trotz der bereits in dieser Hinsicht relativ alten Forderung nach der Anwendung quantitativer Methoden in der Sprachtypologie ist, ausgehend von dem jetzigen Zustand, festzustellen, dass zumindest in einigen Teilbereichen der allgemeinen Sprachtypologie quantitative Verfahren, insbesondere diejenigen, die sich für eine Klassifizierung eignen, herangezogen werden (vgl. u.a. Cysouw 2005). Insgesamt spielen diese Methoden aber weiterhin eine periphere Rolle.

Neben sprachtypologischen Untersuchungen, die u.a. auf der phonologischen Ebene mit der Hilfe von statistischen Methoden durchgeführt werden könnten, verweist Baudouin de Courtenay auf die Möglichkeit von quantitativen schriftlinguistischen Untersuchungen. Die Schriftlinguistik ist eine lange Zeit hindurch ein bewusst peripherisierter Bereich der (insbesondere strukturalistischen) Sprachwissenschaft. Einerseits wurde die Schrift als ein sekundäres System angesehen, und, falls sie doch untersucht wurde, kam es hierbei in der Regel zu einer mechanistischen Übertragung des aus der Phonologie bekannten Methoden- und Begriffsapparates. Baudouin de Courtenay war in vielerlei Hinsicht weitblickender und forderte mit Nachdruck gerade die Untersuchung der Relationen Laut – Phonem (durchaus bereits im Sinne eines Bündels von distinktiven Merkmalen) – Schriftzeichen, welches seiner Ansicht nach idealerweise mit der Hilfe von quantitativen Methoden durchgeführt werden könnte. Eine Forderung, die zumindest in einem kleinen Teilbereich erst in den letzten Jahren in der Form einer quantitativen Schriftlinguistik in Ansätzen geleistet worden ist. Vgl. dazu den Sammelband von Altmann/Fengxiang (2008). Eine schriftlinguistische Analyse des Deutschen, insbesondere zur Graphem-Phonem-Relation im Deutschen, findet sich in dem Beitrag von Best/Altmann (2005).

Aus der Sicht der Quantitativen Linguistik, einer heute vollständig etablierten linguistischen Disziplin, mag der Beitrag von Baudouin de Courtenay zu dieser Forschungsrichtung als bescheiden bezeichnet werden. Dennoch ist der von ihm propagierte Methodenpluralismus und seine Offenheit gegenüber unterschiedlichen theoretischen Ansätzen von unschätzbarem Wert für die Sprachwissenschaft.

Um damit zum zweiten Themenblock des vorliegenden Ansatzes zu kommen: In der allgemeinen Sprachwissenschaft am Ende des 19. Jahrhundert spielte

die Frage von Lautgesetzen, insbesondere aber die von den Junggrammatikern diskutierte Ausnahmslosigkeit, eine zentrale Rolle. In diesem Punkt lässt sich nunmehr der Bogen zwischen Baudouin de Courtenay, der in dieser Frage eine durchaus vermittelnde und konstruktive Position einnahm (vgl. dazu Stankiewicz 1972: 13f) und Hugo Schuchardt spannen, der für seine ablehnende Haltung gegenüber der postulierten Ausnahmslosigkeit bekannt ist.

3. Hugo Schuchardt

Hugo Schuchardt (1842-1927) dürfte als Romanist und Kreolist in der Geschichte der Sprachwissenschaft einen allgemein anerkannten Platz einnehmen. Das Schaffen und die grundsätzlichen linguistischen Positionen von Hugo Schuchardt sind in vielerlei Hinsicht bereits Gegenstand umfangreicher wissenschaftshistorischer Arbeiten[3] (vgl. u.a. Vennemann 1972, den Sammelband Lichem/Simon 1980 mit einer Vielzahl von unterschiedlichen Perspektiven auf das Schaffen von Hugo Schuchardt). Darüber hinaus liegt mit Eismann/Hurch (2008: 1-18) eine umfangreiche Studie zu Gemeinsamkeiten und Unterschieden in den linguistischen Positionen von Hugo Schuchardt und Baudouin de Courtenay vor, die sich aus dem Briefwechsel zwischen diesen beiden Gelehrten extrahieren lassen. Thematisch gesehen ist der Briefwechsel einigen Detailfragen der junggrammatischen Ansätze, grundsätzlichen Problemen des Sprachwandels und vor allem unterschiedlichen Problemen der Sprachmischung gewidmet.

Vor dem Hintergrund einer bereits durchgeführten Analyse der wichtigsten Eckpunkte[4] des Schaffens von Hugo Schuchardt kann im gegebenen Zusammenhang auf Überlegungen eingegangen werden, die in einem direkten Zusammenhang mit der Anwendung von quantitativen Methoden bzw. mit der Häufigkeit als linguistischer Eigenschaft stehen. Die wichtigsten Eckpunkte sind durch Hugo Schuchardt in der berühmten und vielfach zitierten Diskussion um die Ausnahmslosigkeit von Lautgesetzen formuliert worden, in der er immer wieder auf die Bedeutung der Häufigkeit für die sprachwissenschaftliche Analyse hinweist.

In dieser, seiner wohl berühmtesten Schrift („Über die Lautgesetze. Gegen die Junggrammatiker“), in der Schuchardt (1885) gegen die von den Junggrammatikern aufgestellte Behauptung der ausnahmslosen Wirkung von Lautgesetzen argumentiert und polemisiert, wird an mehreren Stellen immer wieder auf die

[3] Eine umfangreiche Bibliographie zu Hugo Schuchardt und seine Originalwerke sind auf der Seite http://schuchardt.uni-graz.at zu finden.

[4] Hugo Schuchardt hatte bei August Schleicher studiert (vgl. dazu den biographischen Abriss in Hurch (2009)). Schleicher war tatsächlich naturwissenschaftlichen Methoden zwar nicht abgeneigt, aber er dürfte wohl keinen direkten Einfluss auf die späteren Ansichten von Schuchardt hinsichtlich der Bedeutung der Frequenz haben. Insgesamt wäre dies ein guter Anlassfall um zu diskutieren, ob es in einigen wissenschaftshistorischen Arbeiten, die sich mit biographischen Aspekten von Wissenschaftlern beschäftigen, nicht zu einer Überhöhung eines Schüler – Lehrer Verhältnisses im akademischen Bereich kommt.

Frequenz sprachlicher Einheiten Bezug genommen. In diesem Zusammenhang wird u.a. auf folgendes verwiesen:

> Die größere oder geringere Häufigkeit im Gebrauche der einzelnen Wörter welche ja bei der Analogiebildung eine so hervorragende Rolle spielt, ist auch für ihre lautliche Umgestaltung von hoher Wichtigkeit, nicht innerhalb kleiner, wohl aber innerhalb bedeutender Differenzen. Sehr selten gebrauchte Wörter bleiben zurück, sehr häufig gebrauchte eilen voran; von beiden Seiten also bilden sich Ausnahmen von den Lautgesetzen (Schuchardt 1885: 24-25).

Der Sprach- bzw. Lautwandel ist insgesamt ein, ohne Zweifel, komplexer Prozess, der durch eine Vielzahl von Faktoren, wie Sprachkontakt, Entlehnung, das Vordringen von dialektalen Varianten, soziolinguistische Faktoren und vieles mehr in Gang gesetzt wird. In diesem Zusammenhang wird, im Übrigen im 19. Jahrhundert doch weit verbreitete Annahme[5], in die Diskussion eingebracht, wonach auch die Gebrauchshäufigkeit den Laut- bzw. Sprachwandel bedingt und begünstigen kann.

Leitende Idee ist, dass erst durch die hohe Verwendungshäufigkeit, durch den hohen Benutzungsgrad die Anzahl von Varianten einer Lautform bzw. sprachlicher Einheiten ansteigt und somit – so die doch recht einfache Vorstellung – Tür und Tor für eine Veränderung einer sprachlichen Einheit geöffnet werden. In anderen Worten, in dieser Konzeption wird die Gebrauchshäufigkeit als erklärende Instanz bzw. auslösender Moment für in Sprachen zu beobachtende Veränderungen herangezogen. Die Gebrauchshäufigkeit ist aber vor diesem Hintergrund eine empirische Größe, die – je nach Art der vorgenommenen Metrisierung – Aussagen über das Ausmaß von Variabilität, Varianz und Diversifikation einer sprachlichen Form zu einem gegebenen Standpunkt zeigt.

Ein weiterer Aspekt, den Schuchardt (1885: 25) ebenfalls als Argument gegen die Ausnahmslosigkeit von Lautgesetzen in die Diskussion einbringt, ist das bekannte Phänomen, dass „[…] in allen Sprachen gerade die allergewöhnlichsten Wörter, von denen man doch am Ersten Gehorsam gegen die Lautgesetze erwarten sollte, am Meisten Neigung zeigen sich von ihnen zu emancipieren […]“. Damit ist in erster Linie das suppletivistische Verhalten von Wortformen gemeint, welches durch eine hohe Verwendungshäufigkeit bzw. durch

[5] Problematisch erscheint an der Polemik um den Begriff „Lautgesetze“ und „Ausnahmslosigkeit“ vor allem zu sein, dass man hierbei nicht zwischen deterministischen und stochastischen Gesetzmäßigkeiten unterscheidet. Paul (1886: 6) bestreitet nicht, dass es einen Zusammenhang zwischen dem Lautwandel und der Häufigkeit gibt: „Eine recht unglückliche, übrigens nicht neue Idee ist es, dass ein Lautwandel sich in häufig gebrauchten Wörtern leichter vollziehen soll als in selteneren“. Darüber hinaus ist aber die Diskussion überschattet von der unterschiedlichen Bewertung der Analogien. Hermann Paul war sich der Bedeutung der Frequenz durchaus bewusst (vgl. dazu Paul 1909: 207ff.) und darüber hinaus hat dieser maßgebliche Impulse für die Sprachökonomie als kognitive und psycholinguistische Kategorie geliefert.

das Alter bedingt sein kann. Es gibt begründeten Grund zur Annahme, Schuchardt (1885) verweist selbst auf diesen Namen, dass die Idee von einem systematischen Zusammenhang zwischen dem Suppletivismus und der Häufigkeit u.a. auf Nikolaj Kruszewski (vgl. dazu Kelih 2009) zurückgeht, der sowohl in engem Kontakt mit Baudouin de Courtenay als auch mit Hugo Schuchardt stand (vgl. dazu die mehrmalige Erwähnung von Kruszweski im Briefwechsel dieser beiden Gelehrten in Eismann/Hurch 2008).

Wichtig wäre allerdings hervorzuheben, dass die Häufigkeit nicht als Ursache für die Bildung suppletivistischer Formen zu verstehen ist, sondern vielmehr bestimmte kognitive Faktoren, die auf eine hohe Differenzierung von naheliegenden und oft gebrauchten Bereichen (und damit extralinguistischen Faktoren) hinauslaufen (vgl. dazu u.a. auch Bittner 1990 und einer ausführlichen Diskussion der Suppletion aus Sicht der Natürlichkeitstheorie). Weitaus wichtiger als eine derartige kognitive Erklärung für ein abweichendes Verhalten von bestimmten sprachlichen Formen ist für Schuchardt die Irregularität von Formen in sprachlichen Systemen an sich. Diese Irregularität wird als globales Argument für die fehlende Systematizität von Sprachwandelprozessen angesehen. Die prinzipielle dynamische und nicht deterministische Natur von sprachlichen Systemen ist allerdings auch Hugo Schuchardt verborgen geblieben.

Neben der Bedeutung der Häufigkeit für den Sprachwandel und für die Bildung von suppletivistischen Formen erwähnt Hugo Schuchardt – auch dies kann wohl als Gemeinplatz der Sprachwissenschaft des 19. Jahrhunderts[6] bezeichnet werden –, dass häufig gebrauchte Wörter insgesamt einem Kürzungsprozess unterliegen. Schuchardt (1885: 25) vermeint dieses Phänomen in allen Sprachen vor allem bei „Titeln und Begrüßungen“ auszumachen; als Beispiel führt er z.B. die im Deutschen häufig vorkommende Kürzung in Form eines g’Morgen statt einem Guten Morgen an. Oder, um es in den Worten von Schuchardt (1885: 25) zu sagen, ist in diesem Fall „[…] freilich das Adjektiv fast ganz um seine Bedeutung gekommen, aber doch nur in Folge der unendlichen Wiederholung“. In diesem Aspekt scheint Schuchardt wiederum auf kognitive Faktoren zu rekurrieren, wie bereits im Falle des postulierten Zusammenhanges zwischen der Vorkommenshäufigkeit und dem Auftreten von Varianz.

Bemerkenswert ist in dieser Hinsicht offenbar nicht nur das Erkennen der engen Verflechtung von (morphologischen) Kürzungsprozessen und der Ge-

[6] Es fehlt bislang eine systematische Auseinandersetzung mit der Geschichte der Quantitativen Linguistik im 19. Jahrhundert. Die Sprachökonomie in allen ihren Facetten könnte hierbei ein wichtiger Anknüpfungspunkt sein, insbesondere die Frage einer kognitiven Motivierung bzw. deren Wechselbeziehungen zur Häufigkeit und zur Länge von sprachlichen Einheiten. In diesem Punkt bietet sich u.a. eine vergleichende Untersuchungen der Ansätze von Hermann Paul, Jan Baudouin de Courtenay, Georg von der Gabelentz und Hugo Schuchardt an. Vgl. dazu auch Roelcke (2002) mit dem Konzept der kommunikativen Effizienz, welche sich mit der Sprachökonomie in Verbindung bringen lässt und somit gemeinsam einen Ausgangspunkt für eine retrospektive Bewertung der Sprachwissenschaft des 19. Jahrhunderts liefern könnte.

brauchshäufigkeit, sondern vor allem die weiterführende Erklärung, wonach Sprecher offenbar eine Abneigung gegenüber „beständig wiederkehrende“ Phänomene und eine allzu hohe Monotonie in der Sprache zeigen und somit den Wunsch bzw. das Bedürfnis nach Varianz und Variabilität verspüren. Als Beispiele für dieses Phänomen führt Schuchardt (1885: 27) phonetische Beispiele (oftmaliges Aussprechen ein und desselben Lautes) und oft zu beobachtende Vereinfachungstendenzen beim Schreiben von hochfrequenten Buchstaben und ähnlichem an. In diesem Sinne war sich Schuchardt durchaus der Bedeutung der Sprachökonomie bzw. der Tendenz zur Bequemlichkeit bewusst – beides Phänomene, die erst Zipf (1935, 1949) bzw. bereits zuvor Jespersen (1922) in nachhaltiger Weise für die Linguistik konzeptualisiert haben.

Insgesamt ist sich Schuchardt (1885: 28) aber der vielschichtigen Rolle der Gebrauchshäufigkeit von sprachlichen Formen durchaus bewusst. Die Gebrauchshäufigkeit zeigt seiner Ansicht nach nicht in allen Fällen einen unidirektionalen Einfluss, in dem Sinne, dass hohe Gebrauchshäufigkeit zwangsläufig zu einer hohen Varianz und Diversifikation führt, sondern es ist seiner Meinung nach auch die entgegengesetzte Tendenz zu beobachten: In Fällen eines sogenannten „verpflanzten“ Lautwandels sei zu beobachten, dass gerade „[…] in den gewöhnlichsten Wörtern die alte Aussprache am Längsten […]“ erhalten bleibt. Auch sei es durchaus möglich, dass gerade selten gebrauchte Wörter eben eine leicht „alterthümliche Gestalt“ aufweisen können (ebda.). In diesem Sinne erscheint die Gebrauchshäufigkeit bei Hugo Schuchardt als ein polyvalentes Konzept, welches in sprachlichen Systemen eine unterschiedliche Rolle spielen kann:

1. Sprachliche Systeme zeichnen sich durch eine unterschiedliche Gebrauchshäufigkeit von sprachlichen Formen aus. Diese ist bedingt durch den Drang und die Notwendigkeit zur Wiederholung, wobei gleichzeitig beim Sprecher eine Abneigung gegenüber der Verwendung von gleichen sprachlichen Zeichen und Formen zu beobachten ist, die zu einer Variabilität und Varianz in sprachlichen Systemen führt.
2. Auftretende sprachliche Varianz, bedingt durch hohe Verwendungshäufigkeit, wird überhaupt erst als Voraussetzung für Sprachwandelprozesse angesehen. Vor diesem Hintergrund kann Sprachwandel nicht als deterministischer Prozess verstanden werden.
3. Die Gebrauchshäufigkeit ist ein entscheidender Faktor für das Verhalten von sprachlichen Einheiten, insbesondere in Hinblick auf die Ausbildung irregulärer phonetischer und morphologischer Formen und in Hinblick auf die Archaizität von Wortformen. In diesem Sinne wird die Gebrauchshäufigkeit als Motor sowohl für Innovation (Bildung neuer, in der Regel irregulärer Formen) als auch für die Bewahrung stabiler Strukturen (z.B. Erhalt von archaischen Wortformen) angesehen.
4. Die Gebrauchshäufigkeit wird in einem engen Zusammenhang zu morphologischen Kürzungsprozessen gebracht, die wiederum mit Bedürfnissen seitens des Sprechers in Verbindung stehen. Hierbei werden vor allem kognitive Faktoren (das bekannte Streben nach Bequemlichkeit, Öko-

nomie usw.) in die Diskussion gebracht bzw. in gelungener Form aus dem damaligen Wissensbestand der Sprachwissenschaft extrahiert und kondensiert.

Die Schrift von Schuchardt „Über die Lautgesetze“ wird in der linguistischen Literatur vor allem als zentrale Absage an das junggrammatische Paradigma interpretiert. Darüber hinaus werden bei Hugo Schuchardt vor allem das „soziale“ und „psychologische“ Moment der Sprache, die Sprachmischung und der Sprachkontakt als entscheidende Positionen hervorgehoben (vgl. dazu in etwa Wandruszka 1980, Hurch 2009). Darüber hinaus sollte – wie sich nun herausstellt – nicht übersehen werden, dass Hugo Schuchardt ebenso der Verdienst zugeschrieben werden muss, in gekonnter Weise die Bedeutung der Vorkommenshäufigkeit von sprachlichen Einheiten, sowohl in Statik als auch Dynamik erkannt und synthetisiert[7] zu haben.

Inwiefern und von welcher Aktualität die grundsätzlichen Überlegungen und Ansätze von Hugo Schuchardt sind, zeigt sich an der gegenwärtigen konzeptuellen Wiederentdeckung der Häufigkeit von sprachlichen Einheiten im Nahbereich des amerikanischen Frequentismus. Vgl. dazu vor allem die Programmschrift in Bybee/Hopper 2001, die sich zu weiten Teilen, wenn auch von den Autoren vermutlich nicht so intendiert, als ein wissenschaftsgeschichtlicher Abriss zur Konzeption der Häufigkeit, durchaus im Sinne des bereits bei Hugo Schuchardt Gesagten, lesen lässt. Insgesamt aber ist mit der von Hugo Schuchardt angedeuteten Relevanz der Häufigkeit ein weiterer kleiner Baustein zur Geschichte der Quantitativen Linguistik gefunden. Gleichzeitig eröffnet sich damit aber auch die Perspektive für eine wissenschaftshistorisch fundierte Neuinterpretation der Häufigkeit als linguistischer Kategorie.

4. Abschließender Vergleich

Das bislang Gesagte lässt bereits einen vorläufigen Vergleich zwischen den Positionen von Hugo Schuchardt und Baudouin de Courtenay hinsichtlich der Verwendung von quantitativen Methoden und der Bedeutung von Frequenz zu. Vorweg lässt sich festhalten, dass weder Hugo Schuchardt noch Baudouin de Courtenay als direkte Vorläufer der Quantitativen Linguistik angesehen werden können. Vielmehr lässt sich auf der Basis des bislang bekannten festhalten, dass Baudouin de Courtenay in erster Linie als Theoretiker und Methodologe in Erscheinung tritt, der mit aller Vorsicht und mit Weitblick entsprechend geeignete Forschungsfelder sondiert. Demgegenüber ist Hugo Schuchardt vor allem an der linguistischen Bedeutung der Frequenz interessiert, die für ihn in erster Linie im Bereich des Sprachwandels von Interesse ist und als Ausdruck eines sprachökonomischen Verhaltens (des Sprechers) interpretiert wird. Vor diesem Hintergrund sind die Ansätze von Hugo Schuchardt und Baudouin de Courtenay als

[7] Ähnliches in Bezug auf die Bedeutung der Häufigkeit von sprachlichen Einheiten im Schaffen von Hugo Schuchardt haben u.a. auch bereits Viereck (1980: 281) und Vennemann (1972: 161ff.) postuliert.

repräsentative Beispiele für die weitgehende Integration der Frequenz und quantitativer Methoden in die Sprachwissenschaft am Übergang vom 19. zum 20. Jahrhundert zu sehen.

Literatur

Adamska-Sałaciak, Arleta (2005). *Language change in the works of Kruszewski, Baudouin de Courtenay and Rozwadowski*. Poznań: Motivex.

Altmann, Gabriel; Fengxiang, Fan (Hg.) (2008). *Analyses of Script. Properties of Characters and Writing Systems.* Berlin, New York: Mouton de Gruyter (Quantitative Linguistics, 63).

Best, Karl-Heinz (2005a). Karl Marbe (1869-1953). *Glottometrics 9, 74–76.*

Best, Karl-Heinz (2006). August Schleicher (1821-1868). *Glottometrics 13, 73–75.*

Best, Karl-Heinz (2007a). Paul Menzerath (1883-1954). *Glottometrics 14, 86–98.*

Best, Karl-Heinz (2007b). Otto Behaghel (1854-1936). *Glottometrics 14, 80–86.*

Best, Karl-Heinz; Kotrasch, Brita (2005). Albert Thumb (1865-1915). *Glottometrics 9, 82–84.*

Best, Karl-Heinz; Altmann, Gabriel (2005): Some properties of graphemic systems. *Glottometrics 9, 29–39.*

Bittner, Andreas (1990). Eine unendliche Geschichte? Nochmal zum Verhältnis von Suppletion und Natürlichkeit. In: Norbert Boretzky, Werner Enninger und Thomas Stolz (Hg.), *Spielarten der Natürlichkeit - Spielarten der Ökonomie. 2 Halband: 227-247.* Bochum: Brockmeyer (Bochum-Essener Beiträge zur Sprachwandelforschung, 8).

Budziak, Renata (1997). *Jan Baudouin de Courtenay als Soziolinguist und Sprachsoziologe.* Bamberg: Univ-Diss.

Bybee, Joan; Hopper, Paul (2001). Introduction to frequency and the emergence of linguistic structure. In: Joan Bybee und Paul Hopper (Hg.), *Frequency and the emergence of linguistic structure: 1-24.* Amsterdam/ Philadelphia: Benjamins (Typological studies in language, 45).

Cysouw, Michael (2005). Quantitative methods in typology. In: Reinhard Köhler, Gabriel Altmann und Rajmund G. Piotrowski (Hg.), *Quantitative Linguistik. Quantitative Linguistics. Ein internationales Handbuch. An International Handbook: 554-578.* Berlin, New York: de Gruyter (Handbücher zur Sprach- und Kommunikationswissenschaft, 27).

Eismann, Wolfgang; Hurch, Bernhard (Hg.) (2008). *Jan Baudouin de Courtenay - Hugo Schuchardt. Korrespondenz.* Heidelberg: Winter (Slavica, 5).

Grzybek, Peter; Kelih, Emmerich (2005). Zur Vorgeschichte quantitativer Ansätze in der russischen Sprach- und Literaturwissenschaft. In: Reinhard Köhler, Gabriel Altmann und Rajmund G. Piotrowski (Hg.), *Quantitative Linguistik. Quantitative Linguistics. Ein internationales Handbuch. An In-*

ternational Handbook: 23-64. Berlin, New York: de Gruyter (Handbücher zur Sprach- und Kommunikationswissenschaft, 27).

Hurch, Bernhard (2009). Von der Peripherie ins Zentrum: Hugo Schuchardt und die Neuerungen der Sprachwissenschaft. In: Karl Acham (Hg.): *Kunst und Wissenschaft aus Graz. Band 2.1: Kunst und Geisteswissenschaft aus Graz: 493-510.* Wien: Böhlau.

Jespersen, Otto (1922). *Language. Its nature, development and origin*. London: Allen and Unwin.

Kelih, Emmerich (2008). *Geschichte quantitativer Verfahren in der russischen Sprach- und Literaturwissenschaft.* Hamburg: Kovač (Studien zur Slavistik, 19).

Kelih, Emmerich (2009). Quantitative Hypothesen bei Mikołaj Kruszewski. *Glottometrics 18, 77–81*.

Köhler, Reinhard (2012). *Quantitative Syntax Analysis.* Berlin, New York: de Gruyter (Quantitative Linguistics, 65).

Lakatos, Imre (1982). *Die Methodologie der wissenschaftlichen Forschungsprogramme.* Braunschweig/Wiesbaden: Vieweg.

Lichem, Klaus; Simon, Hans Joachim (Hg.) (1980). *Hugo Schuchardt. Schuchardt-Symposium 1977 in Graz; Vorträge und Aufsätze.* Österreichische Akademie der Wissenschaften; Schuchardt-Symposium. Wien: Verlag d. Österr. Akad. d. Wiss (Veröffentlichungen der Kommission für Linguistik und Kommunikationsforschung, 10).

Mittelstraß, Jürgen (1974). *Die Möglichkeit von Wissenschaft*. Frankfurt am Main: Suhrkamp.

Mugdan, Joachim (1984). *Jan Baudouin de Courtenay (1845-1929). Leben und Werk.* München: Fink.

Paul, Hermann (1909). *Prinzipien der Sprachgeschichte.* 4. Auflage. Halle/ Saale: Niemeyer.

Paul, Hermann (1886): 'Schuchardt, Hugo, Ueber die Lautgesetze. Gegen die Junggrammatiker'. *Literaturblatt für germanische und romanische Philologie, 7(1), 1–6.*

Roelcke, Thorsten (2002). *Kommunikative Effizienz: eine Modellskizze*. Heidelberg: Winter.

Schuchardt, Hugo (1885). *Über die Lautgesetze. Gegen die Junggrammatiker*. Berlin: Oppenheim.

Stankiewicz, Edward (1972). *A Baudouin de Courtenay Anthology. The Beginnings of Structural Linguistics.* Bloomington, London: Indiana University Press.

Stegmüller, Wolfgang (1979). *Rationale Rekonstruktion von Wissenschaft und ihrem Wandel.* (Mit einer autobiographischen Einleitung). Stuttgart: Reclam.

Viereck, Wolfgang (1980). Sprachwandel. In: Klaus Lichem und Hans Joachim Simon (Hg.), *Hugo Schuchardt. Schuchardt-Symposium 1977 in Graz; Vorträge und Aufsätze: 275-291.* Wien: Verlag d. Österr. Akad. d. Wiss (Veröf-

fentlichungen der Kommission für Linguistik und Kommunikationsforschung, 10).

Vennemann, Theo (1972). Hugo Schuchardt´s theory of phonological change. In: Theo Vennemann und Terence H. Wilbur (Hg.): *Schuchardt, the neogrammarians, and the transformational theory of phonological change. 4 essays by Hugo Schuchardt*: *117-179*. Frankfurt am Main: Athenäum (Linguistische Forschungen, 26).

Wandruszka, Mario (1980). Hugo Schuchardt und die "Linguistik 2000". In: Klaus Lichem und Hans Joachim Simon (Hg.): *Hugo Schuchardt. Schuchardt-Symposium 1977 in Graz; Vorträge und Aufsätze*: *293-314*. Wien: Verl. d. Österr. Akad. d. Wiss (Veröffentlichungen der Kommission für Linguistik und Kommunikationsforschung, 10).

Zipf, George K. (1935). *The Psycho-Biology of Language. An Introduction to Dynamic Philology*. Cambridge: M.I.T. Press.

Zipf, George K. (1949). *Human Behavior and the Principle of Least Effort. An Introduction to Human Ecology*. Cambridge, MA: Addison-Wesley.

Homogeneity and heterogeneity within language(s) and text(s): Theory and practice of word length modeling

Peter Grzybek, Graz

"*Homogeneity in language is merely an idealization, heterogeneity is the 'normal' state and it results from processes all of which are stochastic.*"
Altmann (1987: 231)

0. Introduction

This contribution[1] attempts to shed some light on the consequences the observation of homogeneity and/or heterogeneity of language, or linguistic data, has for the theoretical modeling process in a quantitative linguistics framework. Starting with an introductory discussion of these two key terms (1), the major points of this discussion will then be amplified with regard to concrete linguistic data. To this end, reference will be made to the theory of word length in a synergetic context (2). Given that word length frequencies are regularly organized, it will then be shown that there are systematic variations not only across languages, but also within languages, specific text types, and even within individual texts (3). On the basis of these findings, it will finally be discussed, how such systematic heterogeneities can be taken into account in theoretical modeling, asking for homogeneity (4).

1. Homogeneity and Heterogeneity

The concept and assumption of language being principally characterized by 'homogeneity' has a long history in the development and understanding of linguistics.[2] It may be said to have started with Herder's definition of nation as a community tied together by a common, uniform and, therefore, homogeneous language, and it is also at the basis of Karl Wilhelm von Humboldt's *Über die*

[1] I am sincerely grateful to Gabriel Altmann, Reinhard Köhler, Benedikt Szemerenyi, Bernhard Wälchli, and Rupert Waldenfels, as readers of an earlier version of this text, for their helpful comments.

[2] Unfortunately, the contribution on „Homogenität und Heterogenität der Sprache: Die Entwicklung der Diskussion im 20. Jahrhundert", which had been announced for the third volume of the representative synoptic reference work *History of the Language Sciences*, ed. by Auroux et al. (2006), still appears there with its title, but is missing from the publication and has never been published.

Verschiedenheit des menschlichen Sprachbaus[3] (1836), where he maintains that nation and language completely coincide, and that, despite all individual heterogeneity, only one language prevails throughout a whole nation, eventually diversified into some dialects to a certain degree. It is particularly characteristic for the Saussurian tradition[4], with its dichotomy of speech (*parole*) and language (*langue*), focusing on the alleged homogeneity of *langue* and displacing any kind of heterogeneity to the realm of *parole*.

The assumption of homogeneity is very convenient both for practical purposes (e.g., school grammars) and theoretical objectives (e.g., grammatical models) in linguistics; it is also advantageous for quantitative analyses of language, homogeneous data being an important pre-condition for many statistical tests. This circumstance is well known as the *ceteris paribus* principle, fundamental to descriptive purposes, theoretical modeling, predictive purposes of scientific inquiry, and the formulation of scientific laws. In the framework of scientific experiments, the *ceteris paribus* assumption is realized by controlling all independent variables other than the one(s) under study, so that the effect of the independent variable(s) under observation on the dependent variable can be isolated. In other words, all other relevant factors are kept constant, and all remaining features – which are regarded to possibly affect the data – are considered to be *external* factors, conceived of as being constant for the sample, at least over the period of observation of the sample.

In reality, however, homogeneous data are but a rare case and difficult to obtain: if at all, they can be drawn only from a single population, concentrating on one or more, but usually not all features of the population. As a consequence, contemporary approaches in the field of linguistics (as in other disciplines, too) have increasingly abandoned previously dominating homogeneous concepts and conceptualizations of language, which had largely excluded variation from the description of linguistic systems for methodological reasons. It is particularly the branch of variational linguistics, basically originating in socio-linguistic approaches, which focuses on the usage and function of particular varieties of language, i.e. not only on sociolects, but also dialects, regiolects, registers, etc. Such variations may result from a whole variety of factors, which are not likely to be reduced to spatial differences; they include group-specific linguistic behavior, situational factors (such as formal vs. informal contexts), stages of language

[3] Humboldt's book was first translated into English under the title of *The Heterogeneity of Language and its Influence on the Intellectual Development of Mankind;* a more recent translation is entitled *On Language. On the Diversity of Human Language Construction and Its Influence on the Mental Development of the Human Species* (Cambridge University Press, 1988, 2nd rev. edition 1999).

[4] Cf. Saussure (1916/59): "Taken as a whole, speech is many-sided and heterogeneous" (ibd., 9); "Whereas speech is heterogeneous, language, as defined, is homogeneous" (ibd., 15), "Taken as a whole, speech cannot be studied, for it is not homogeneous" (ibd., 19).

acquisition, speakers‘ age, language contact, and many others. Quite obviously, the specific object, or ‘justification’, of variational linguistics seems to be primarily motivated, or legitimated, by *extra-linguistic* factors.

As such, variational linguistics might seem to be, at first sight, located at the opposite end of the linguistic spectrum as is linguistic typology, if one sees the latter’s major objective to be the comparative study of languages according to their intrinsic (structural) features. At closer sight, however, linguistic typology, aiming at the description of properties (common to various languages), must take into account the structural diversity of the world’s languages, and thus is concerned with variety, too; in its orientation toward the study of universals, linguistic typology cannot but study (possibly) existing differences between languages and features, on the basis of which languages may then be grouped into classes[5], or attributed to types.[6]

Seen from this point, linguistic typology thus is equally concerned with variation as is variational linguistics – more concretely: with variation across, or between, languages, which is assumed to be not random, but subject to specific regularities, or limitations, linguistic typology in this sense being concerned with the question of how (or to what degree) such limitations allow for meaningful sub-divisions into various linguistic groups and sub-categories.

---

[5] It may be reasonable here to refer to the general distinction between typology and classification. Classification, being based on a set of criteria which may concern each element of the classified set or not, captures all elements of a given set and unambigously attributes each of them to exactly one class. Such classes, as established by classificational typology, have no theoretical implication, since no classification has ever been established by theoretical laws, and they all have been by way of inductive procedures only. In typology, as compared to this – albeit as well being the result of a grouping process – the given elements are attributed to groups (or types) in such a way that the elements within a given type are maximally similar to each other with regard to the relevant features (internal homogeneity), the types at the same time being maximally different (external heterogeneity). Therefore, a typology always aims at some specific question, and it may fulfil, among others, some heuristic function in the process of theory formation, but a typology obtained by way of inductive processes only, can never lead to a theory (cf. Altmann 2008).

[6] If by ‘type’ we understand a group (collection, set, class) of objects sharing specific characteristics (attributes, features, properties), it is obvious that the individual objects (the variants) belonging to one type (the invariant) all must share one or more properties reflected in the type, but that each of the variants may, of course, have additional properties not reflected in the type. In other words, the variants share some properties, but they do not share others; as a consequence, they stand in some kind of homomorphic relation to each other. The (invariant) type, however, contains (reflects) only those features which all variants share. With regard to these features, the type may be considered to be an (abstract) model (i.e., a conceptual construct) of the (concrete) variants, standing in an isomorphic relation to them, in this respect.

From the point of view of quantitative linguistics – which aims at the formulation of (stochastic) laws in the field of linguistics and which, in the closely related and indispensable process of hypothesis formulation, must inevitably refer to theoretical models – both variational linguistics and linguistic typology are equally concerned with variants and invariants, i.e. with classes or types, and variations thereof. As a consequence, they are inevitably concerned with questions of homogeneity and heterogeneity, be that with regard to elements between objects, or within a given object under study.

It is just this circumstance, which in this respect places variational linguistics and linguistic typology into one and the same boat: typology is not possible without variants being attributed to a type, and the assumption of variation makes sense only along the assumption of a common invariant, i.e., a type. In both cases, the question of homogeneity and heterogeneity inevitably comes into play. As such, they can be simply accepted, i.e. they can be taken as given; as soon as the aim is theoretical modeling, however, they necessarily have to be adequately taken into account. It is important to note, in this context, that as soon as variants and variations are studied, this can be done only on the basis of (at least assuming) the existence of some higher-order invariant, i.e. a type, to which they belong: in other words, heterogeneity or homogeneity always refer to some super-ordinate system, not its individual elements.[7]

Homogeneity thus refers to the "sameness" of a set of elements with regard to the property, or properties, of one or more (features of) elements of some super-ordinate system, not to the individual elements (or their features) themselves. In other words, within a given system, elements once being attributed to it, may be regarded to be homogeneous with regard to the system they constitute (eventually along with other, heterogeneous elements); across systems, selected elements may be homogeneous (eventually, again, along with other, heterogeneous elements) with regard to some (abstract, theoretical) super-ordinate system. As a consequence, variants and variations are by definition heterogeneous and can be attributed to a type only on some higher level; different types, in turn, are again heterogeneous, and may eventually conceived of as belonging to some super-type.

We may thus conclude that in any kind of linguistic analysis, we are inevitably concerned with the question of homogeneity and heterogeneity of the linguistic material. The material itself, as our linguistic object, will always be characterized by internal heterogeneity; for the purpose of, and in the process of model building, however, it can and, in fact, must be reduced to (some) relevant aspects, which then allows us to ask the question of homogeneity.

The possible focus on either homogeneity or heterogeneity has been interpreted by Altmann, as early as in the mid 1980s, in terms of levels of analysis;

[7] With regard to linguistics in general, and linguistic typology specifically, this has been very clearly stated by Skalička as early as in 1966, in his seminal essay "Ein typologisches Konstrukt".

it may also, however, *cum grano salis*, be interpreted with regard to the history of linguistics. According to Altmann (1987), homogeneity in language is but an idealization, heterogeneity being the "normal" state, resulting from stochastic processes:

1. The assumption of homogeneity, considering language as a homogeneous whole; it leads to the examination of rules, to determinism, and classification (not going beyond monothetic classes); it uses only nominal (in extreme cases dichotomic) scales; at a more progressive stage of this level one uses for descriptions such methods of qualitative mathematics as algebra, two-valued logic, set theory, the theory of automata, etc.
2. The recognition of heterogeneity, both in synchrony and diachrony, considering language as a diversified whole, leads to the need to somehow order the variation(s) observed.

Recognizing (and accepting) heterogeneity opens the doors in two directions: (a) backwards to homogeneity, using reduction procedures (i.e. norming, boundaries, types, classes, dichotomies, categories, etc.), or (b) forwards to the next level, focusing research on latent mechanisms bringing about the heterogeneity. In the first case[8] (a), we are concerned with some elaborated kind of "homogeneous"

[8] Although absurd at first sight, this tendency seems to be characteristic for corpus linguistics, too, at least for its later developments. Subsequent to its initial emphasis on inductive and empirical methods, concentrating on performance rather than competence, corpus linguistics became increasingly impressed by the notion of 'representativeness', accompanied by the illusion of the 'the-more-the-better principle', which would make it possible to (re-)construe of the (descriptive, or statistical, rather than prescriptive) "norm" of a given language. The naïve assumption was, at least at that time, that a corpus, if only "large enough", is representative of a given language as a whole. There is, however, from a theoretical point of view, a major logical flaw in this argumentation, due to the inappropriateness of the law of large numbers in the field of linguistics: the basic dictum of this law, saying that the relative frequency of a random event approximates its probability by the repetition of events, is restricted to the repetition of equivalent events, only – and no individual text can ever be equivalent to some other text, unless it is reduced to specific aspects focused. But in this case we are already concerned with a model of the text, which may be said to be homogeneous to the given language, or rather, to a given model of that language. We are thus again facing the problem of homogeneity and heterogeneity; it turns out that the problem is equally relevant for any study of sub-systems, or sub-models. Therefore, it also concerns so-called "domain-specific" corpora, which do not claim (any more) to represent language as a whole, but specific (thematic) domains of it. But neither language nor any of its specific (sub-)domains can be seen as the simple sum of all texts (to be) produced; therefore, no (random, balanced, domain-specific, etc.) corpus can reasonably be claimed to be representative for something beyond the material observed. Conclusions to be drawn beyond the object observed necessarily ask for a model: in this case, and only in this case, scientific hypotheses may be formulated; else, we are concerned with no more

linguistics, which attempts to grasp the heterogeneity, the 'chaos', by means of sampling small segments of language or by means of homogenization; in the second case (b), more theoretical branches of linguistics, particulary processual and systems theoretical (synergetic) linguistics, try to investigate the laws of this "chaos" (including the boundary conditions being at work), thus leading to the construction of theories and attempting to yield scientific explanations (in a strict understanding of this term). It is here, with language being understood as a process of self-regulation (or a result of this process), that quantitative linguistics comes into play, with its ambition to formulate the laws controlling this process, including the boundary (or antecedent) conditions, which are responsible for a large part of the variation involved (cf. Altmann 1985, 1987).

It is a major concern of the present study to demonstrate this in detail. The practical implications of the problems theoretically outlined above, and the relevance and need to pay due attention to the homogeneity and heterogeneity factor, shall be illustrated in the following empirical analyses. It shall be seen that heterogeneity is far from being relevant for what usually is being conceived of as a variety of language: not only the system of language as a whole, and not only any of its (sub)-systems, but each individual text is, in fact, principally characterized by internal heterogeneity, what represents a crucial methodological problem for any quantitative analysis of language and text.

By way of an example, the following analyses will concentrate on word length. This is by no means to be understood that word length is, or should be, considered to be a crucial (or even the only) factor for linguistic classification and/or typology. In a way, word length can be considered to be an arbitrarily chosen factor here, which could be replaced (or complemented) by many others. Yet, it is not a completely arbitrarily chosen example since the word, and its length, have been in the center of linguistic attention for a long time.

2. Word length: The word in a synergetic framework

Although the study of word length has a more than 150-year long history[9] it was only in the mid-1990s that a theory of word length came to be developed. Such a development was possible, of course, due to the fact that at that time, many "local" studies were available which had not only shown that the frequency with which words of a given length occur in texts, or languages, is not arbitrary, but

---

and no less than the observation and description of delimited objects. And about these objects, empirical (but not theoretical) hypotheses use to be formed; they are helpful for scientific progress, but not sufficient. It is just here, where we find the difference between empirical and theoretical sciences, between statistics of language and quantitative linguistics.

[9] For a survey of this history cf. Grzybek (2005).

rule-based, and that word length is no isolated category in a theory of language, but related to other linguistic units and levels.[10]

What was not clear, if there is a universal model with which word length frequencies can generally be theoretically described (and if so, which model), or if language-specific models are needed (and if so, how they are interrelated): Elderton (1949), for example, analyzing passages from various writers, discussed the geometric distribution with regard to word length in English; as compared to this, Čebanov (1947) propagated the (1-shifted) Poisson distribution, referring to the analyses from 127 Indo-European languages; and Fucks, in the mid-1950s, would speak of a "general law of word-formation" (1955a: 88, 1957: 34), or, more exactly, as the "mathematical law of the process of word-formation from syllables for all those languages, which form their words from syllables" (Fucks 1955b: 209).

An important step in the discussion of possibly adequate distribution models for word length frequencies was Grotjahn's (1982) contribution, who argued in favor of the negative binomial distribution which, under certain circumstances, converges against both the geometric and the Poisson distribution; the importance of Grotjahn's contribution has to be seen in the suggestion that, instead of looking for one general model, one should rather try to concentrate on a variety of distributions which are able to represent a valid "law of word formation from syllables" (ibd., 73).

This idea was later taken up by Grotjahn/ Altmann (1993) and elaborated by Wimmer et al. (1994) and Wimmer/Altmann (1996). The assumption brought forth in these papers was that the frequency of a given class (P_x) is determined by its preceding class (P_{x-1}), thus resulting in the proportionality relation $P_x \sim P_{x-1}$. Further assuming that this relation is characterized by a specific proportionality function $f(x)$, one obtains $P_x = f(x)P_{x-1}$. Depending on which concrete function is chosen, different frequency distribution models are being obtained. In the above-mentioned papers, the function $f(x) = ax^{-b}$ – i.e., the Menzerathian function, well-known to have an important function in linguistic self-regulation – was assumed

[10] After all, it is not by chance, that word length has played a crucial role in Greenberg's (1960) approach to language typology of that time. Notwithstanding the intensive discussions, modifications and improvements of his quantitative approach, going on still today – for discussions and further developments of this issue see: Krupa (1965), Krupa/Altmann (1966), Altmann/Lehfeldt (1973), Kempgen/Lehfeldt (2004), Kelih (2011) – his approach shows the importance which has been attached to the word and its characteristics: according to his definition, an index of synthesis (I_S), i.e. an index for the degree of syntheticity, is defined as the ratio of the number of words (f_W) and the number of morphs (f_M) in a given text: $I_S = f_M / f_W$. Following Krupa (1965), or Krupa and Altmann (1966), it is reasonable to change numerator and denominator – otherwise I_S would theoretically tend to infinity – and to interpret the result as an index of analyticity $I_A = f_W / f_M$, the index of syntheticity consequently equaling $I_S = 1 – I_A$. As can be seen, this index is specifically related to word length, originally being based on the average length of a word, measured in the number of morphemes per word.

to be the basic function, leading to the so-called Conway-Maxwell-Poisson distribution. There is no need to go into details here; what is more important is the fact that this approach provided a good starting point for a flexible system of distributions.

Thus, the function $f(x) = ax^{-b}$ was not the only one taken into consideration; rather a whole system of modifications, extensions, and generalizations was described, resulting in a number of different distribution models.[11]

Later[12], this approach was integrated into Wimmer and Altmann's (2005) even more general "Unified Derivation of Some Linguistic Laws". It would lead too far here to discuss this approach in detail; in short, for a discrete variable *X*, this general approach leads to recurrence formula (1) from which, among many others, the above-mentioned distributions can easily be derived:

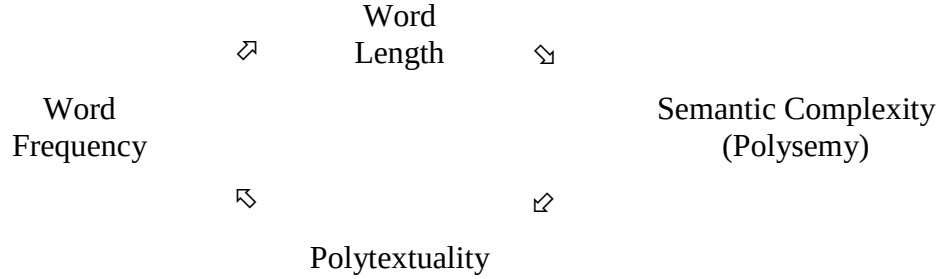
$$(1) \qquad P_x = \left(1 + a_0 + \frac{a_1}{x} + \frac{a_2}{x^2} + \ldots\right) P_{x-1}$$

[11] Thus, to give but a few examples: With $b = 1$ in the basic form mentioned above, one obtains $f(x) = a/x$, leading to the Poisson distribution, with $b = 0$ ($0 < a < 1$), the geometric distribution; from $f(x) = (a+bx)/cx$, the negative binomial distribution results, etc.

[12] At the same time, it had increasingly become clear – not only from structuralist approaches to language, but, first and foremost, from synergetic linguistics – that the word is no isolated entity within a language system. Elaborating on Zipf's works from the 1940s, in which a systematic relation between the frequency and the length of words had already been shown to exist, a number of further relations had been reliably proven to exist, concerning, among others, semantic complexity (polysemy), contextual connectivity (polytextuality), etc.:

1. The more frequent words are, the shorter they tend to be.
2. The shorter words are, the more meanings they tend to have.
3. The more meanings words have, the more likely are they to occur in different (con)texts.
4. In the more different texts a word occurs, the more frequently it tends to be used.

With these selected relations, we are thus facing a circle of interrelations – which, in fact, are much more complex and include many more factors –, being integrated into a synergetic concept (cf. Köhler 1986).



Over the last decades, much empirical evidence has been gathered corroborating hypotheses deduced from this approach. It is not the place here to go into further details; what is important, however, is that this approach can be said to provide the basis of deductive reasoning in quantitative linguistics. As a consequence, it is a matter of boundary conditions, how many and which parameters are needed, and which distribution model results from this. By way of an illustration, Wimmer et al. (1994: 100), referring to individual languages, authorship, and genre, as the three most important factors, have conceived of the situation as a cube (cf. Figure 1).

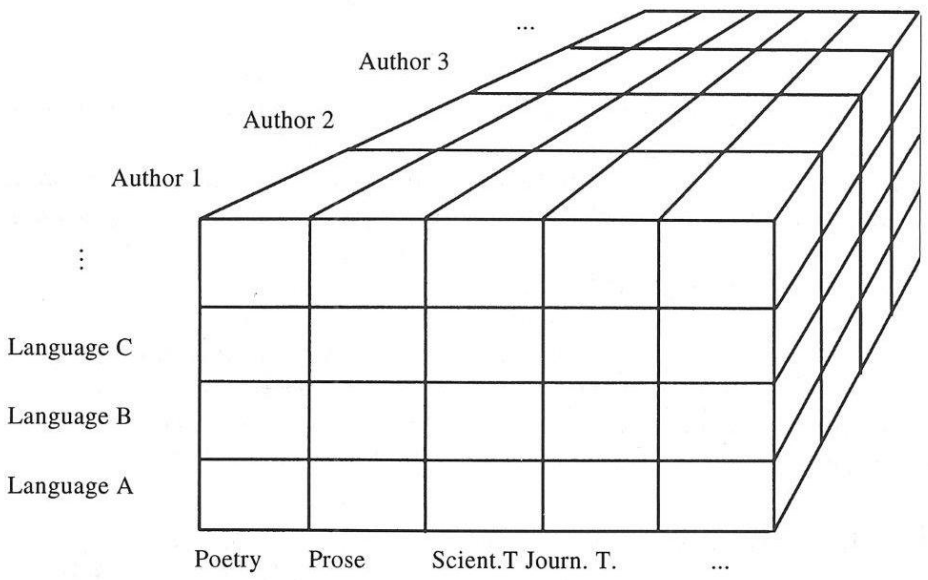


Figure 1: Theory of word length - some boundary conditions (Wimmer et al. 1994)

As can easily be seen, in addition to language, two of the factors, authorship and genre, are of relevance for possible intra-lingual differences, i.e. language-intrinsic word length heterogeneity. This approach has provoked numerous studies on word length, which need not be mentioned here in detail (cf., e.g., Best 1997, 2001). As compared to earlier research, and with regard to the boundary conditions mentioned, these studies are characterized by an interesting tendency: whereas (at least the titles of) part of the papers continue to speak about word length in some language *L* as a whole, others are more specific (or careful) and refer to word length in individual texts of a given author *A* and/or written in some genre *G*.[13]

[13] In this respect, a specific strategy has been to concentrate on the genre of letters, which have been considered to be a prototypical textual representation of a given language, due to its intermediate location between oral and written communication, on

What we are likely to have in the first case, is – given there are intra-lingual differences – some overgeneralized model, based on heterogeneous data. In the second case, we are likely to have more specific models, based on (more) homogeneous data; what we usually do not have, however, are systematic studies of how such specific models (if they differ) or the parameter values of a given model (if there is only one model) relate to each other within a given language. This is, however, a major problem: if there are indeed systematic intra-lingual differences, attention must be paid to them also in inter-lingual comparisons, if one does not want to compare chalk to cheese. Moreover, it is an open question, if and to what degree there are systematic differences not only within a given genre, but also within a given text, which in turn, may well be composed of heterogeneous components.

In the remaining sections, we will first systematically study the problem of language-intrinsic differences, then shifting our attention to text-intrinsic heterogeneities, in order to finally study how such heterogeneities can be dealt with in the process of theoretical modeling.

3. Intra-lingual heterogeneities

3.1. Linguistic data and test material

In order to find out, if systematic word length differences exist within a given language, we need linguistic data as test material which is (or, eventually, can be prepared in such as way that it turns out to be) appropriate for the study of this question. Therefore, the data should not only consist of individual elements (e.g., texts), but also should these elements lend themselves to some kind of systematic grouping (i.e., some text typology).

These groups can either emerge as the result of quantitative analysis – in this case the individual elements are *a posteriori* shown to belong to categories which are heterogeneous with regard to the criterion studied –, or they can represent the starting point, when the individual elements are *a priori* attributed to higher-order classes, or types, which then are tested for systematic differences – in this case, classes which differ with regard to the criterion studied, are considered to be heterogenous, those which do not, as homogeneous. Both approaches are not mutually exclusive – the results may even coincide, although one should not expect this, at least not fully, since we are concerned with one variable only (i.e., word length) here; they simply ask for different methods which will be described and applied further below.

---

the one hand, and the assumption that they represent the outcome of a single, homogeneous (“undisturbed”, “non-interrupted”) process of text generation.

In any case, to pursue these options, we need a corpus[14] of texts and some text typology serving as the higher-order system to which the individual texts can be attributed. From a practical point of view, it is necessary (or at least desirable) to cover the whole textual spectrum of a given language, in order to arrive at systematic results; in that case, at least some implicit knowledge of the textual spectrum and the variety of text types is necessary.

With regard to our objective, it seems reasonable to choose and apply two different text typologies, in order to at least minimially control influences of authoritative decisions in this respect: one of them with maximal abstraction, resulting in minimal specifity and a minimum number of categories (text types), the other one with maximal specifity and, as a consequence, a maximal number of categories. Again, both approaches do not exclude each other, but are to be seen complementary, since the more specified typology should eventually allow for an attribution of its elements to the more general one:

1. As to a maximally specified typology, reference can be made to results from extensive text type research (German: "Textsortenforschung"), where lists with more than 4000 different text types have been provided; these text types[15] are distinguished according to specific communicative-situational functions, which tend to be interpreted in terms of differences in their thematic-propositional or illocutive characteristics (cf., e.g., Adamczik 1995: 255ff.).
2. A minimally specified (and thus maximally reduced) text typology can be seen in the concept of functional styles. This approach originates mainly in Czech functionalism and structuralist positions from the 1930s and 1940s (e.g., Havránek, and others), and it has later been elaborated by Russian scholars (as, e.g., Vinogradov, and others), too.[16] Generally speaking, the concept of functional styles is characterized by the attempt to relate specific stylistic features to extra-linguistic pragmatic or social functions, assuming that specific purposes of language usage influences

[14] It should be emphasized here that within this "corpus" all texts keep their individuality and are not merged into one corpus in the usual understanding of this term. As Orlov (1982) has pointed out some decades ago, any kind of textual combination is, from a theoretical point of view of quantitative linguistics, not an increasingly better approximation to some abstract norm, but a mixture of heterogeneous components – a "pseudo-text", in other words. In this context, the qualitative attribution of individual texts to text types is but tentative, as is the combination of more than one text in some kind of sub-corpus, as long the homogeneity of these texts is not tested and, eventually, proven by adequate statistical methods.

[15] The term 'text type' may be used differently in other contexts; here, it serves as a translation of the term 'Textsorte' as it has become commonly used in German scholarly discourse.

[16] For an informative survey on functional stylistics, including Russian research, see Ohnheiser (1999).

linguistic form.[17] Functional styles have been successfully submitted to quantitative and probabilistic approaches[18], e.g., by Doležel (1964), or Mistrík (1973), who used the "traditional" schema with five major categories of discourse: everyday (colloquial), scholarly (scientific), administrative, journalistic and artistic, the latter inturn being subdivided into literary prose, poetry, and dramatic language (cf. Figure 2).

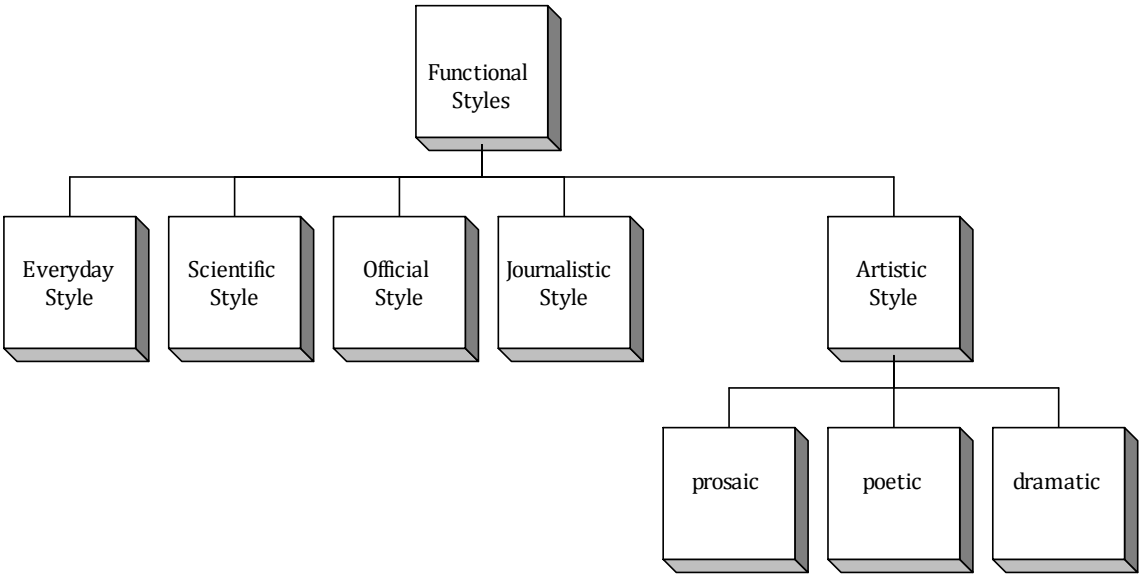


Figure 2: Functional styles (Mistrík 1973: 23ff.)

Although more specific approaches are favored in contemporary text typology (cf. Blühdorn 1990: 218), functional styles continue to play an important role still today, as e.g., in the recently published statistical analyses of the representative Czech National Corpus (Bartoň et al. 2009).[19] For our purposes, it thus seems reasonable and sufficient to refer to these two concepts of text typology. More-

[17] In this respect, the concept of functional style has very much anticipated of what has more recently been discussed in contemporary approaches favoring the notion of 'register' (cf., e.g., Biber 1988, 1995).

[18] In this respect, modern approaches like the ones mentioned before are much more elaborate, both with regard to the number of linguistic variables taken into consideration, and to the amount of quantitative methods applied. For Biber, for example, 'text types' are quantitatively determined on the basis of linguistic similarities, and as the result of extensive statistical analyses; these analyses remain, however, on a descriptive level only, and do not touch upon the question of modeling with regard to a theory, as this is relevant and considered to be crucial in a quantitative linguistics framework.

[19] Detailed results are offered, among others, for word length, separately calculated for three functional styles: scientific, journalistic, and literary prose; word length is counted both in the number of syllables and the number of phonemes per word; additionally, data are given separately for both type and token occurrences.

over, with regard to the compilation of our text data base for the analyses, it seems plausible to choose such text types from the vast amount available, that each category of the less specified text typology (i.e., of the functional styles) is "filled", each by at least ca. 30 texts from at least one text type.[20]

For the sake of illustration, we will use material from a Slovenian text data base described and analyzed in detail elsewhere – cf. Grzybek/Kelih (2005a,b), Grzybek et al. (2006), Kelih et al. (2005): 398 texts from seven different text types were tentatively attributed to functional styles, with four of the seven functional styles being represented by either more than one text type or texts from more than one author, thus representing allegedly homogeneous subgroups of ca. 30 texts each.

It should be emphasized once more that the 398 texts were not merged into a corpus. With regard to the above-mentioned assumption that the genre of letters is a prototypical textual representation of a given language, particular emphasis was laid on different kinds of letters in compiling the text data base. These letters were tentatively attributed to different kinds of text types and, by way of that, to different functional styles: private letters as one instance of everyday communication, open letters as instances of administrative/public and letters to the editor of journalistic communication, chapters from epistolary novels as belonging to literary prose. Additionally, to complement the schema of functional styles, journalistic comments, poems, dramatic texts, and short novels were included, the sum of texts thus summing up to a corpus of 398 items. These texts were not, however, fused into one large corpus – rather, each text was treated in its individuality, thus allowing for tests of the classificatory principles according to the two methods described above. Table 1 represents the text data base in detail.

Table 1
Text basis of 398 Slovene texts

Functional style	**Authors**	**Text types**	**N**
Colloquial	Cankar, Jurčič	Private letters	61

[20] This procedure seems reasonable because the authoritative attribution of text types to functional styles involves the possible methodological problem that it is based on some kind of *a priori* decision only. As such, we are concerned with a qualitative decision, which may well bias the overall result. In fact, as has been shown elsewhere in more detail, such text type attributions to functional styles may be highly subject-dependent: in a study involving 24 experts in text typology, there was high agreement as to some text types, but large disagreement as to others, when subjects attributed specific text types to either more than one functional style at a time, or to different functional styles – for details, see Grzybek/Kelih (2005a,b).

Administration	various	Open letters	29
Journalistic	various	Letters to the editor, Comments	65
Prose	Cankar	Chapters from long stories (‘povest’)	68
	Švigelj-Mérat / Kolšek	Letters from an epistolary novel	93
Poetic	Gregorčič	Versified poems	40
Dramatic	Jančar	Single acts from dramas	42
Total			**398**

3.2. Methods

For the purposes outlined, various statistical methods may be applied. Generally speaking, there are three commonly used approaches, which may be termed quantitative (i), or quantitative-qualitative (ii), respectively. Specifically, we are concerned with:

i. *Clustering methods*, which introduce no qualitative information into the process of classification; rather, they represent some kind of *tabula rasa* principle, introducing specific quantitative information only (such as, in our example, mean word length of a given sample), and aiming at the distinction of sub-groups which in the end will have to be interpreted qualitatively;
ii. *Post hoc* and *discrimination methods*, which are to be understood as specific combinations of *a priori* and *a posteriori* (qualitative and quantitative) principles, which are both based on tentative attributions of the individual samples to groups:
    a. in post hoc analyses, more often than not (but not necessarily) based on the means of the observations, the major question is if specific homogeneous subgroups can be detected among the groups tentatively distinguished *a priori*,
    b. in discriminant methods, the adequacy of tentative *a priori* attributions is tested by first mathematically transforming the variables in order to arrive at a maximal distinction of occurrences, and then calculating the percentage of “correct” attributions – the higher the percentage of “correct” attributions, the better the discrimination is interpreted to be.

Strictly speaking, we need no text typology, if we want to apply method (i) only. Explicit recourse to some text typology is necessary, however, for methods (iia) and (iib); it goes without saying that, in this case, the results to be obtained may at least partly depend on the concrete typology chosen.

3.2.1. Cluster analysis

In a first approach, clustering methods were applied, where no qualitative information is introduced. A usual procedure to determine the optimal number of clusters is the so-called elbow technique, which is based on the mean of the squared errors of an analysis of variance for a give number of clusters (which can be stepwise varied). Table 2 contains the values obtained for 3-8 clusters, which are graphically represented in Figure 3: in the two-dimensional graphic, the number of clusters and the sum of the squared errors are depicted on the *x* and *y* scales; the "best" number of clusters can be seen from that point of the curve, where a salient descent (the 'ellbow') can be observed.

As compared to this, two-step analyses represent an explorative procedure to identify groups within a given data set, various distance measures being used to calculate the (dis)similarities between clusters.

Table 2
Mean of squared errors for different numbers of clusters

Number of clusters	**Mean of squared errors**
2	0.017
3	0.009
4	0.006
5	0.004
6	0.003
7	0.003
8	0.002

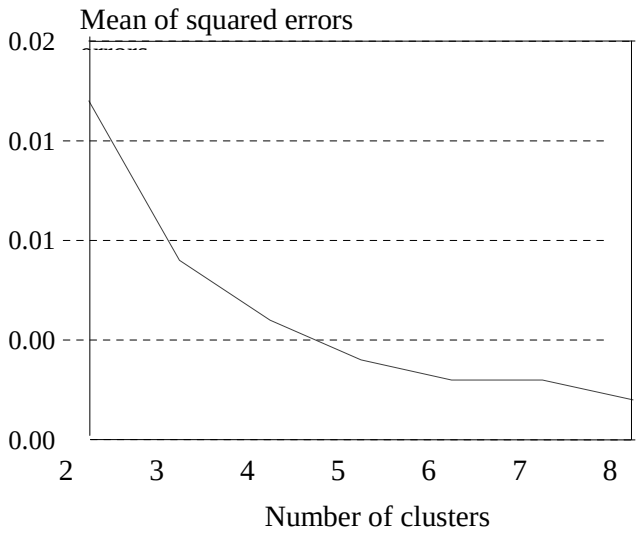


Figure 3: Optimal number of clusters (ellbow technique)

Table 3 shows the results, based on log-likelihood distances.

Table 3
Two-step cluster analyses

Centroids			
		$\bar{x}$	$\bar{s}$
Clusters	**1**	2.4020	0.1293
	2	1.8114	0.0885
	3	2.0450	0.0794
	Combined	1.9889	0.2379

As can be seen, the same result is obtained by both procedures, the visual elbow technique as well as two-step cluster analyses based on log-likelihood distances: accordingly, the "optimal" number of clusters to be distinguished for the material under study turns out to be three. This result is most surprising, since the number of three categories corresponds neither to the number of text types studied, nor to the number of functional styles. In other words, there do seem to be specific textual sub-categories – what is a clear indication of systematic intra-lingual heterogeneity –, but not in agreement with either of the text typologies applied.

3.2.2. Post hoc analysis

Approaching the problem from a different side, post hoc comparisons of means can be run, based on a priori attributions to text types, on the one hand, and quantitative information (in our case: word length averages per text) on the other, in order to identify possibly existing homogeneous subgroups (i.e., without significant differences within the groups, but with significant distinctions between the groups. Table 4 represents the result of these analyses.

Table 4
Post hoc comparisons of means (8 text types, 398 texts)

		Homogeneous subgroups ($\alpha = 0.05$)				
Text Type	*N*	1	2	3	4	5
Poems	40	1.7127				
Short stories	68		1.8258			
Private letters	61		1.8798			
Drama	42		1.8977			
Epistolary Novel	93			2.0026		
Letters to the Editor	30				2.2622	
Comments	35				2.2883	
Open Letters	29					2.4268
Significance		≈1.00	0.37	≈1.00	0.99	≈1.00

As can be seen from Table 4, five homogeneous sub-groups do indeed exist, to which the texts from the eight text types chosen can be attributed. At closer sight, however, some more specific observations raise interpretative problems:

1. With five sub-groups, the number of identified homogeneous sub-groups is different from the number of clusters, obtained in cluster analyses.
2. There is no consistent attribution of text types to functional styles as predicted.
3. The four different letter types fall into four different categories.

In sum, we seem to have homogeneous sub-groups, but neither of the two qualitative typologies applied corresponds to these five subgroups.

3.2.3. Discriminant analysis

In discriminant analyses, individual cases (here: texts) are first attributed to groups (here: text types, or functional styles, respectively) on the basis of specific predictor variables (here: average word length and statistical characteristics derived therefrom[21]), these variables then being submitted to linear transformations, in order to arrive at an optimal discrimination of the cases. However, running discriminant analyses with text types, thus testing the hypothesis "Word length is a variable, which is characteristic of text types", we arrive at the poor result of only 56.3% correct attributions of the texts – what causes us to reject this hypothesis.

A better – though still far from satisfying – result is obtained for discriminant analyses on the basis of functional styles: in this case, in contrast to the assumption of homogeneity of word length within functional styles, we arrive at a still overall poor percentage of 73% correct discriminations.

There are various possible explanations at hand for the overall poor results of the discriminant analyses, e.g.:

a. the tentative *a priori* attributions of the individual texts to text types and/or the attribution of text types to functional styles have been wrong or inconsistent;
b. none of the typologies (i.e., neither the text types distinguished nor the functional styles) provides an adequate (basis for) text typology;
c. there are no consistent subgroups to be distinguished on the basis of word length, which thus turns out to be no good indicator for the demonstration of systematic intra-lingual heterogeneity: the property of word length, used as a basis of classification, is not adequate for the given purpose.

[21] Statistical characteristics derived from the word length frequency distribution, are measures such as variance, dispersion coefficient, skewness, kurtosis, etc., in addition to the mean.

By stepwise (progressive) re-grouping, it can be shown, however, that indeed three general categories can be distinguished, i.e., a number of categories which corresponds to the initial result of our cluster analyses. According to the results, we are concerned with *three discourse types* (as they shall be termed here), which can be distinguished rather clearly on the basis of word length: juxtaposing (i) poetic texts vs. (ii) public (written) speech vs. (iii) private (or oral speech[22]), the outcome is a remarkable percentage of 92.7% correct discriminations. Table 5 represents the results.

Table 5
Three discourse types as a results of discriminant analyses

	Predicted group			Total
	Oral /Private	Written /Public	Verse	
Oral / Private	260	3	1	264
Written / Public	19	75	0	94
Verse	6	0	40	40

Figure 4 offers a graphical illustration of these results.

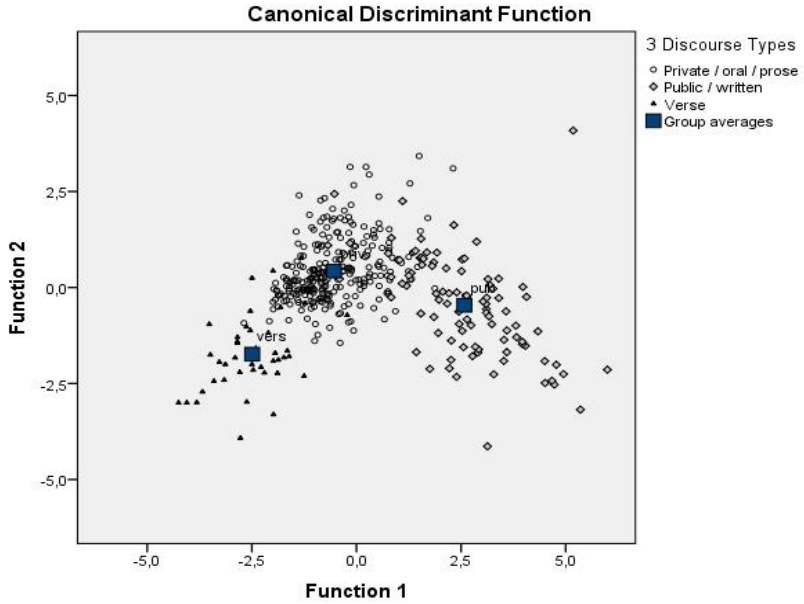


Figure 4: Discrimination of three discourse types

[22] It should be mentioned that the 19the century literary stories analyzed here not only include many dialogues (i.e., fictitious oral speech), but that the whole *ductus* of these texts aims at the illusion of the narrator using oral speech, a phenomenon known as "skaz" in literary theory; both factors might explain why these texts might rather be classified as oral speech (what by no means must be characteristic of literary prose in general).

Summarizing the major results of this section, we can thus conclude that language turns out to be no homogeneous whole, at least not with regard to word length.[23] Rather, there is a large portion of systematic heterogeneity immanent to a given language[24], beyond (or rather below) extra-linguistically motivated categories.[25] Functional styles, albeit the most "radical" kind of intra-lingual text typology, seem to reflect this heterogeneity inadequately; rather, there seem to exist a limited number of more general discourse types which can be distinguished on the basis of word length, obviously even more "trivial" than the maximally reduced functional styles. This, in turn, might be a hint at the conclusion that, due to the "triviality" of these categories, concrete texts from either one of the functional styles or one of the text types, might be composed of mixtures of these categories, what will have to be tested further below, when intra-textual heterogeneity will be at stake.

With this in mind, let us now turn to the questions of variation within a given text type and within one and the same text, again with regard to word length.

3.4. Heterogeneity within text type and texts

As to the study of variation within a given text type, let us analyze, by way of an example, two Russian texts: both belong to literary prose, both are written by one and the same author, Aleksej N. Tolstoj (1882-1945), and both were published within a time span of six years and thus approximately at one and the same period of the author's life. One text, *Гадюка* [The Adder], is a story from 1928, the other one is *Золотой ключик* [*The Little Golden Key*], a story for children from 1936.

Starting with a comparison of average word (x) length for these two texts[26], it turns out that words in the children's text are shorter on the average, with $\overline{x} = 2.31$ ($s = 1.10$), than those in the adults' text, with $\overline{x} = 2.42$ ($s = 1.21$).

[23] Quite similar results have been obtained with studies based on sentence length, or on a combination of word and sentence length jointly (cf. Kelih et al. 2006).

[24] Very similar results have also been obtained with analogical studies on Russian (Friedl 2006).

[25] The observed heterogeneity cannot be explained, by the way, by individual authors' style: in a detailed study on authorship, letters and poems by three Russian poets (A.S. Puškin, A. Achmatova, D. Charms) were analyzed (ca. 30 texts per author and genre, summing up to a total of 190 texts); as a result, it turned out that, with authorship as the discriminant variable, there was a percentage of only 38.4%, as compared to 89.5%, with genre as the discriminating variable (cf. Kelih et al. 2005).

[26] Merging both texts into one "corpus" results in a mean word length of $\overline{x} = 2.35$ ($s = 1.15$). As compared to the idea that corpus construction is an appropriate procedure to "smoothen" heterogeneities and to illuminate a language's "norm", it can easily be seen that actually, such kind of "norm" is but an artificial construct, the corpus in fact turning out to be but a mixed pseudo-text in Orlov's sense (see above).

Since word length frequencies are known not to be normally distributed (see above), a Mann-Whitney *U*-test is in order to test the differences for significance. As the result shows, the differences are highly significant (z = -5.23, $p < 0.001$). In other words: mean word length clearly differs for these two texts, written by one and the same author, belonging to one and the same text type, literary prose.[27]

Given this finding, we can go one step further, showing that heterogeneity may characterize not only relations between two texts of one and the same text type, but also characterizing specific textual subgroups within these texts. To demonstrate this, let us separately analyze the narrative and the dialogical passages of both texts (combined), with regard to average word length, and compare the results for differences between both sub-groups.

Calculating average word length yields in an interesting – though, at second sight, not really astonishing – result: the (combined) narrative passages of both texts are characterized by clearly shorter word length (with $\overline{x} = 2.41$, $s = 1.16$) as compared to the (combined) dialogical passages ($\overline{x} = 2.15$, $s = 1.11$), the difference being highly significant ($z = –16.60$, $p < 0.001$).

Given this observation, it is obvious that average word length of a given text is heavily biased by specifics of text composition, and it is likely to be influenced by the proportion of narrative and dialogical passages contained. Taking this finding into account, it seems reasonable to pay attention to the percentages of these two constituting elements in the children's and the adults' texts, and to compare differences in proportions: of the overall 10590 words in the adults' text, 8120 are represented by narrative[28] passages, and 1527 by dialogues; this corresponds to 76.68% narrative passages and 14.42% dialogues. As compared to this, the children's text contains 10085 words from narrative, and 5291 from dialogical passages, from a total sum of 17470 words; in percentages, this corresponds to only 57.73% of narrative passages and 30.29% dialogues. As a chi square test shows, these differences are highly significant ($X^2 = 1032.55$, $p < 0.001$). Thus, the quantity distribution of narrative and dialogical passages – which, as has been shown above, clearly differ with regard to word length – turns out to heavily influence the overall result.

Yet, the observation of intrinsic heterogeneity is not at its end here. Comparing average word length in narrative passages and in dialogues of the adults' and the children's text, separately for each of the two texts individually, it turns out that these again clearly differ across texts: the narrative passages in the adults' text are significantly ($z = –2.98$, $p < 0.005$) longer ($\overline{x} = 2.46$) than those in

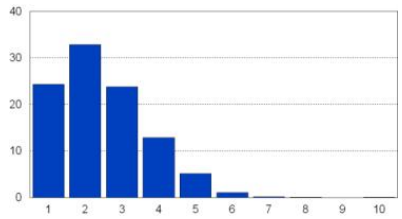
[27] It is a well-known statistical fact that differences in large samples generally tend to be more likely to be significant; however, in case of the non-parametric *U*-test, sample size plays no crucial role, thus indicating the results to be reliable.

[28] No distinction is made here between narrative and descriptive passages; moreover, auctorial narrative sequences preceding (i.e., introducing) or following figures' direct speech, are ignored here.

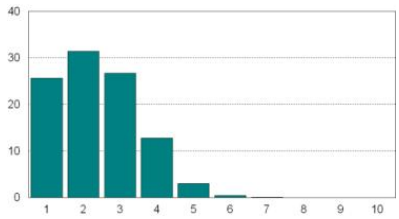
the children's text ($\bar{x}$ = 2.37); furthermore, word length in the dialogical sequences of the adults' text ($\bar{x}$ = 2.21) is significantly shorter than in the children's text ($\bar{x}$ = 2.13), the difference in this case not being significant, however (z = –1.21, p = 0.22).

The observed differences are, of course, a result of differences in frequency distribution – after all, the mean is but one central measure of central tendency. Figures 4a-d illustrate the observed word length frequencies for the four subsamples.

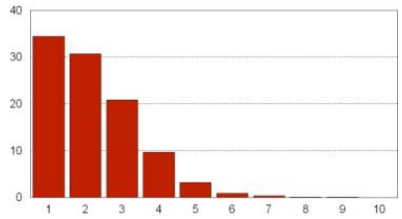
As can easily be seen, the distribution profiles for the narrative and for the dialogical sequences clearly differ, word length for the narrative passages having a peak at two-syllable words, whereas monotonously decreasing for the dialogues.



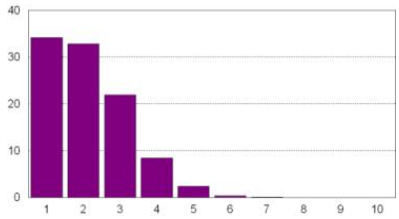
(a) Narrative adults



(b) Narrative children



(c) Dialogues adults



(d) Dialogue children

Figure 4: Empirical word length frequencies in four sub-samples

In any case, despite the seemingly similar profiles of the two dialogical as well as the two narrative sequences across the children's and the adult's texts, differences in average word length are significant, as has been shown above. This is confirmed by the non-parametric Kruskal-Wallis *H*-test for between-group differences with all four groups, which in our case, with the variable 'word length' not following a normal distribution, has to be used for the analysis of variance (ANOVA). As a result, the differences between the four sub-groups are highly significant (X^2 = 285.78, d.f. = 3, p < 0.001). This test result can only indicate the existence of differences, but it cannot identify which of the groups is (or are) responsible for the differences. Therefore additionally computing post hoc tests with all four samples, in order to identify possibly existing homogeneous subgroups, yields the insight that there are no homogeneous subgroups; rather they all fall into a separate category of their own, the three most common post hoc tests (Student-Newman-Keuls, Tukey-B and Scheffé) all equally yielding

high significance, with $p \approx 1.00$. Table 6 represents the results of the post hoc tests.

Table 6
Post hoc tests with mean word length

		Subset for α = 0.05			
Group	*N*	**1**	**2**	**3**	**4**
Dialogue (children)	5291	2.13			
Dialogue (adults)	1527		2.21		
Narrative (children)	10085			2.37	
Narrative (adults)	8120				1.46
Significance		≈1.00	≈1.00	≈1.00	≈1.00

As a result, it thus turns out that not only the two texts are heterogeneous with regard to word length, but that, in addition to this, each of these texts is heterogeneous in itself. Eventually, even more specific sub-samples might be thought of, as e.g., differences between individual speakers, between neutral utterances, questions, and imperatives, and so on and so forth… But instead of further complicating matters, let us draw some preliminary conclusions from the foregoing observations.

3.5. In-Between-Conclusion

Summarizing the results obtained from cluster, post hoc, and discriminant analyses, we may thus say that the principle of heterogeneity goes indeed much farther as is often assumed or taken into account. It seems, any linguistic attempt at describing general "norms" of a language, must ask itself in how far this fact is relevant for the given question and eventually pay due attention to it.

At first sight, the insights gained may seem to be most important for corpus linguistics, particularly when the latter is concerned with theoretical generalizations of empirical results obtained. With corpus linguistics abandoning its "the-more-the-better-principle" – and, by way of that, changing its orientation from establishing the norm of a given language to that of specific domains of it – only a first step seems to be done. Ultimately, any linguistic attempt at (re)-constructing generally valid norms must take into account a major conclusion to be drawn from the observations above, namely that such norms seem to vanish, the deeper one goes into details. In trying to provide homogeneous material – as was said earlier in this text, a necessary pre-condition for statistical testing and reasoning –, the ice gets getting increasingly thinner under the linguists' feet: after the illusion of finding a norm of language as a whole, attention was directed to corpora considered to be "domain-specific", or of "context-related relevance", and it seems attention may, or must further be turned towards "balanced" corpora of specific text types, eventually restricted to specific individual authors, maybe

even from a clearly defined period of time, and so on, and so forth… In the end, we have nothing but the text itself; but even a text is heterogeneous in itself, as could be seen above.

This phenomenon is far from being specific for linguistic objects, and well-known to scientist from many other fields. It seems that under these circumstances, no generalization beyond the object observed is possible any longer. Ultimately, this disillusioning result would make it impossible to do scientific research. The idea to base linguistic research on allegedly prototypical texts, may seem to be a way out; but as could be seen, a single prototype, does not exist, and it has to be chosen, or rather defined, anew, with any question to be pursued. And, what is even more important, if we do not want to restrict ourselves to authoritative qualitative decisions, we tend to know only post hoc, what an adequate prototype is for problem under study.

This raises the final question, how one can deal with these problems in the "everyday practice" of quantitative linguistics, for which the establishment of theoretical models is a sine qua non condition in its research paradigm.

4. Modeling heterogeneities

As has been emphasized above, linguistic objects tend to be principally characterized by heterogeneity, being essential to any kind of linguistic material under study. Yet, with regard to an abstract model, adequate to theoretically describe and eventually explain these data, it is just homogeneity which is needed: data homogeneity is necessary as soon as forming and testing a hypothesis is at stake, which refers to a mechanism one assumes to exist „behind" or "beyond" the data observed.[29]

Data acquisition, in a quantitative linguistics framework, has to be functionally seen as the foundation of theoretical conclusions, with the aim to develop stochastic laws, and quantification is but a necessary step in the logical sequence of scientific steps (cf. Altmann 1993) which generally comprise:

1. Qualitative formulation of a hypothesis, which relates to language(s) or text(s), are of empirical relevance and testable;
2. Statistical formulation („translation") of the hypothesis;
3. Empirical testing (retaining / rejecting) the hypothesis;

[29] In this respect, it is important to pay attention to the conceptual distinction of (a) linguostatistics, or statistics of language(s), on the one hand, and (b) quantitative linguistics, or quantitative text analysis, on the other: whereas linguostatistics aims at the description of language(s) and texts, including the number of languages, of speakers, etc., and refers to any statistical description of linguistic phenomena, (b) primarily aims at the formulation of linguistics laws, including theoretical hypotheses to be tested, and thus is, among others, characterized by a different status and function of both data gathering and quantification.

4. Statistical interpretation of the result with regard to the initially formulated hypothesis;
5. Qualitative interpretation.

Quantification thus is not the aim or the outcome of quantitative linguistics, but one necessary step in the course of scientific study.[30] Anyway, the need to base any generalization on homogeneous data, has been explicitly pointed out at the very beginning of this text. Given the theoretical and empirical insights reported above, it turns out, however, that realistically, obtaining (perfectly) homogeneous data is almost impossible (as in other research fields, too).

Generally speaking, in case heterogeneity is observed in (a set) of data, there are two options to deal with it (cf. Altmann 1992):

1. strive for a diversification of the data, aiming at homogeneous subsets, in order to guarantee the *ceteris paribus* condition.
2. integrate heterogeneity into the model, what results, among others, in mixture or composite models;

Let us illustrate this problem, once more using the word length data presented above. Figure 5 represents the data of both texts mentioned above, merged into one "corpus"; since not only narrative and dialogical passages are included, but also auctorial speech accompanying direct speech, the total of words sums up to N = 28060. The second column of Table 7 represents the observed frequencies (f_x) of the individual word length classes (x), graphically represented in Figure 5 by dark grey bars; for the time being, the values in the third column (nP_x) and the light grey bars can be ignored here (see below).

Table 7
Word length frequencies

x	f_x	nP_x
1	7538	7205.73
2	9017	9795.96
3	6982	6658.65
4	3327	3017.41
5	1000	1025.52
6	161	278.83
7	27	63.18
8	6	12.27
9	1	2.09
10	1	0.36

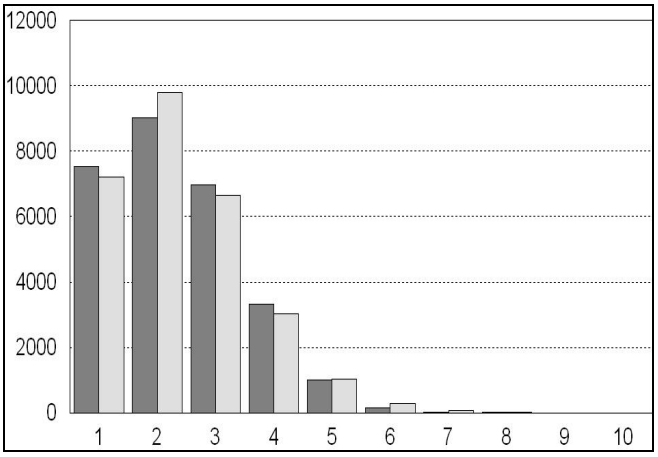


Figure 5: Word length frequencies of two combined texts – observed data and fit of the Poisson distribution (2b)

[30] If Bailey (1991) had been familiar with these principles, and with the theoretical and methodological basics of quantitative linguistics, he could easily give a positive answer to his provocative question "Variation in the data: Can linguistics ever become a science?"

An attempt to find an adequate frequency distribution model for these data may be theoretically based on the general approach discussed above,

$$(1) \qquad P_x = \left(1 + a_0 + \frac{a_1}{x} + \frac{a_2}{x^2} + \ldots\right) P_{x-1}.$$

For $a_0 = -1$, $a_1 > 0$, and $a_i = 0$ for $i = 2,3,\ldots$ we obtain the well-known Poisson distribution (2a):

$$(2a) \qquad P_x = \frac{a^x e^{-a}}{x!} \qquad x = 0,1,2,\ldots$$

Testing the goodness of this model for our word length data, it is reasonable to use this model in its 1-shifted form since, according to our word definition there are no zero-syllable words. We thus arrive at

$$(2b) \qquad P_x = \frac{a^{x-1} e^{-a}}{(x-1!)} \qquad x = 1,2,3,\ldots$$

In fact, with parameter value $a = 1.36$, which is estimated from the data, the fit turns out to yield very good results, as indicated by the discrepancy coefficient[31] $C = X^2/N = 0.0071$. The third column of Table 7 (see above) contains the theoretical values (nP_x) obtained, which are graphically represented by the light grey bars in Figure 5.

Testing the same model for the two texts separately, it turns out, however, that the fit is, in fact, excellent for the adults' text ($C = 0.0057$, with $a = 1.43$), but less good for the children's text ($C = 0.0144$, with $a = 1.33$). Moreover, with regard to the four subgroups (i.e., separately for the narrative and dialogical passages of each text), we see that the model is not only less good for the narrative passages in the children's text, but even has to be rejected for the dialogical passages of the adults' text. The fitting results are represented in detail in Table 8.

[31] The usual goodness of fit test would be the well-known chi square test. Since the X^2 value increases linearly with an increase of sample size, it tends to yield significant result the sooner, the larger the sample is. Since this is the standard case in linguistics, the discrepancy coefficient is preferred, with $C < 0.02$ being interpreted as a good, $C < 0.01$ as a very good fit.

Table 8
Fitting the Poisson distribution to narrative and dialogical sequences

	Adult		Child	
	narrative	dialogue	narrative	dialogue
N	8120	1527	10085	5291
a	1.46	1.20	1.39	1.13
C	0.0039	0.0231	0.0170	0.0070

Thus, with regard to the four subgroups, which we tentatively assume to be homogeneous, and attempting to find a common model for all of them, a usual procedure would include one of the following options – cf. Wimmer/Altmann (1996), Wimmer et al. (1999):

1. To test some 'local' modification.– Usually, in word length studies, it is just the first frequency (f_1), which is modified in one way or another, i.e. for some reason (to be explained), there are "too many" or "not enough" words in this class, thus worsening the overall fit of the model. In such a case, the first probability class (P_1) is modified, i.e. modeled separately, usually being estimated from the observed frequency (f_1). In our case, the Singh-Poisson (4) distribution – which, for α = 1 corresponds to the ordinary (1-displaced) Poisson distribution (2a/b) – might be an adequate model (cf. Wimmer/Altmann 1999: 605f.), in case the assumption above should turn out to be correct.

(4) $$Px = \begin{cases} 1-\alpha+\alpha\cdot e^{-a} & x=1 \\ \dfrac{\alpha\cdot a^{x-1}e^{-a}}{(x-1)!} & x=2,3,4,\ldots \end{cases}$$

2. To test some composite (mixture) model.– Since it cannot be excluded that an allegedly homogeneous subgroup is in fact composed of further heterogeneous components, a mixture of either two different distribution models, or of one and the same with two different weighting factors, might be appropriate. In our case, given the overall adequacy of the Poisson distribution (see above), it seems reasonable to test the Mixed Poisson distribution (cf. Wimmer/Altmann 1999: 417f.) which, for α = 0 or α = 1, again results in the ordinary (1-displaced) Poisson distribution (2a/b):

(5) $$P_x = \frac{\alpha\cdot a^{x-1}e^{-a}}{(x-1)!} + \frac{(1-\alpha)\cdot b^{x-1}e^{-b}}{(x-1)!}, \quad x = 1,2,3,\ldots$$

3. To search for some generalization.– A generalization is a more general model, of which particular sub-models turn out to be special cases, usually with one or more of the general model's parameters being equal to or approximating some limit (0, 1, ∞). In quantitative linguistics in general, and with regard to word length frequency particularly, a well-known generalization of (2a/b) is the (1-displaced) hyper-Poisson distribution (6)

$$(6) \qquad P_x = \frac{a^{x-1}}{{}_1F_1(1;b;a)\,b^{(x-1)}} \quad x = 1,2,3,\ldots$$

which for $b = 1$, results in the ordinary (1-displaced) Poisson distribution (2a/b).

Table 9 summarizes the fitting results for all three data options: (i) the whole corpus, (ii) both texts separately, and (iii) the narrative and dialogical passages in each of the two texts. For all three models, the values of both the parameters and the discrepancy coefficient *C* are given.

As can be seen from Table 9, only for the dialogical passages of the adults' text an improvement can be observed with any of the three modifications, as compared to the ordinary Poisson distribution (cf. Table 7 above). Although the overall results are far from being bad, the relatively worse fit for the children's text, particularly for its narrative passages, is obvious. Interestingly enough, none of the modifications yields crucial improvements for these two sub-samples. This is also reflected in the parameter behavior of the models; for both samples we have parameter $\alpha \to 1$ for the Singh-Poisson distribution, $a = b$ and $\alpha \to 0$ for the Mixed Poisson distribution, and $b \to 1$ for the Hyperpoisson distribution, thus all of them having the ordinary Poisson distribution as special or limiting case, the modifications consequently being of no substantial benefit.

Table 9

Fitting modifications and generalizations of the Poisson distribution

Corpus		**Adults'**	**Children's**	**Narrative**		**Dialogical**	
		(Г.)	**(З. к.)**	**Adults'**	**Children's**	**Adults'**	**Children's**
Singh Poisson							
a	1.40	1.48	1.36	1.50	1.42	1.39	1.21
α	0.97	0.96	0.98	0.98	0.98	0.88	0.94
C	0.0058	0.0031	0.0136	0.0031	0.0166	0.0017	0.0032
Mixed Poisson							
a	1.36	2.27	1.33	1.90	1.39	1.43	1.50
b	1.36	1.42	1.33	1.46	1.39	0.16	1.13
α	0.01	0.01	0.01	0.01	0.01	0.83	0.01
C	0.0071	0.0057	0.0144	0.0043	0.0170	0.0015	0.0070

Hyperpoisson							
a	1.41	1.64	1.32	1.60	1.33	1.88	1.31
b	1.07	1.29	1.00	1.18	0.93	2.01	1.25
C	0.0069	0.0036	0.0144	0.0036	0.0168	0.0021	0.0049

Given these findings, it seems reasonable to tackle the problem differently, starting "from the bottom", i.e., searching for an adequate model covering the sub-samples, first, and only then extending the findings to the complete texts, and to the corpus. Thus, re-analyzing the data, it turns out that a specific modification of the well-known binomial distribution (7)

$$(7) \qquad P_x = \binom{n}{x} p^x q^{n-x} \qquad x = 0,1,...,n; \qquad 0 \le p \le 1,\ q = 1-p$$

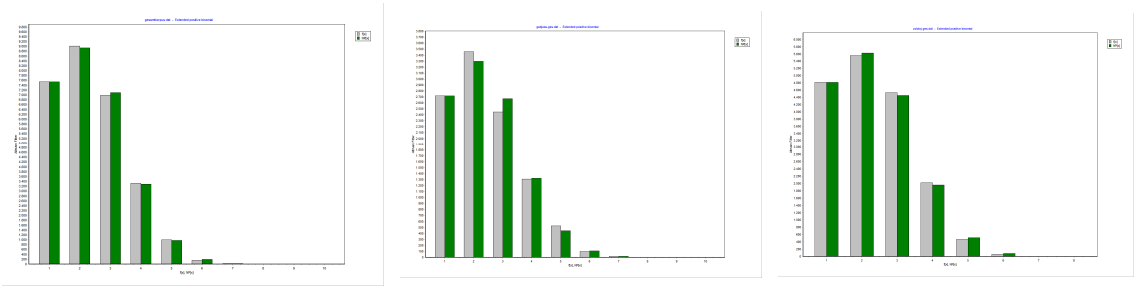
is an excellent model for each of the four narrative and dialogical sub-groups. The binomial distribution (7) can be derived from (1) with $a_0 < -1$ and $i = 2,3,\ldots$ Its modification consists of (a) a left-truncation (which is reasonable, since there are no 0-syllable words, according to our word definition), and (b) a special treatment of the first frequency class P_1 (which would be a hint that it is just the 1-syllable words, which tend to be used in specific ways, asking for some qualitative interpretation). We are thus concerned with the extended positive binomial distribution (cf. Wimmer/Altmann 1999: 148)

$$(8a) \qquad P_x = \begin{cases} 1-\alpha & x = 0 \\ \dfrac{\alpha \binom{n}{x} p^x q^{n-x}}{1-q^n} & x = 1,2,3...,n \end{cases}$$

which in our case is to be used in its 1-displaced form:

$$(8b) \qquad P_x = \begin{cases} 1-\alpha & x = 1 \\ \dfrac{\alpha \binom{n}{x-1} p^{x-1} q^{n-x+1}}{1-q^n} & x = 2,3,4,...n+1 \end{cases}$$

It yields excellent fitting results not only for the four sub-groups, but also for the two texts, and for the complete corpus (with $C < 0.005$ in all cases). This can clearly be seen from the graphical illustrations in Figures 6a-c which show the results for the two text and the combined corpus



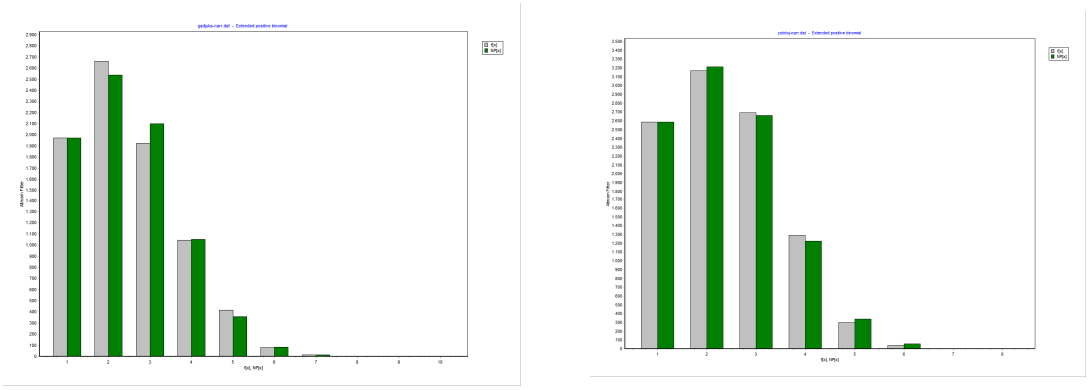
(a) Corpus
$N = 9$, $p = 0.17$, $\alpha = 0.73$
$C = 0.0008$

(b) *Гадюка*
$N = 13$, $p = 0.12$,
$\alpha = 0.74$, $C = 0.0045$

(c) *Золотой ключик*
$N = 7$, $p = 0.21$, $\alpha = 0.72$
$C = 0.0007$

Figure 6: Fitting the extended positive Poisson distribution (texts and corpus)

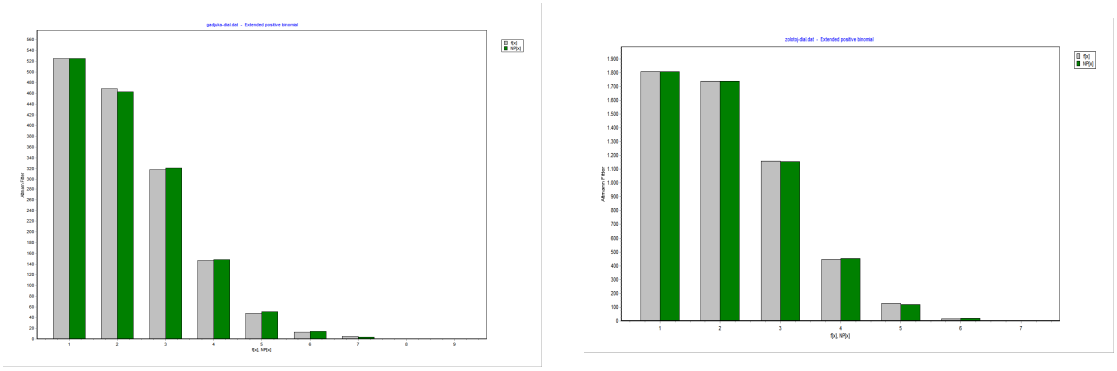
Figures 7a-d present the results for the four narrative and dialogical subgroups. As can be seen, this model is able to grasp all samples equally well, despite the obviously different profile of narrative and dialogical sequences: with $\alpha \approx f_1$ in all cases, and $\alpha \approx 0.75$ for the narrative and $\alpha \approx 0.65$ for the dialogical passages, parameter p of this modified binomial model ranges from $0.12 \leq p \leq 0.21$. Interestingly enough, the model for the dialogical passages in the adults' text slightly deviates from all others, with $n \to \infty$ and $p \to 0$, thus converging to the (extended positive) Poisson distribution.



(a) / (b) Narrative

(b) *Гадюка*
$N = 12$, $p = 0.13$, $\alpha = 0.75$, $C = 0.0043$

(c) *Золотой ключик*
$N = 7$, $p = 0.22$, $\alpha = 0.74$, $C = 0.0014$



(c) / (d) Dialogical

(c) *Гадюка*
$N = 940$, $p = 0.0015$, $\alpha = 0.66$,
$C = 0.0017$

(d) *Золотой ключик*
$N = 10$, $p = 0.13$, $\alpha = 0.66$
$C = 0.0003$

Figure 7: Fitting the extended positive Poisson distribution (four subsamples)

4. Summary and Perspectives

This contribution has started from the assumption that not only are languages different, but that also languages are principally characterized by intrinsic heterogeneity; homogeneity and heterogeneity can only be obtained by way of abstract reduction to specific features under observation, and with reference to some super-ordinate system, or model, concentrating on these features.

Both deductive methods in the Saussure-Chomsky tradition and contemporary approaches favoring inductive methods are doomed to failure in their attempts to arrive at a theory of language, adequately taking into account, among others, variation within language(s), as long as they do not integrate both procedures in an abductive approach, including the formulation of testable hypotheses.

From a quantitative linguistics point of view, linguistic variation is in important object to be studied, which cannot be reduced to extralinguistic factors, but must be understood as the effect of boundary conditions of more general laws, which thus are local specifications, or modifications, of more general language regulations, and which today can already be deduced from a general theoretical concept. Much empirical evidence has been gathered over the last decades, corroborating hypotheses deduced from the “Unified derivation of some linguistic laws”, developed by Wimmer and Altmann (2005, 2006).

By way of an illustrative example, the present contribution demonstrates these principles and procedures with regard to word length, for which systematic varieties are been proven to exist not only within language as a whole, but within specific text types and individual texts, and for which the establishment of the-

oretical frequency distributions are discussed, which attempts to pay due attention to the problems outlined.

References

Adamczik, Kirsten (1995): *Textsorten – Texttypologie. Eine kommentierte Bibliographie.* Münster: Nodus.

Altmann, Gabriel (1985): "On the dynamic approach to language." In: Ballmer, Thomas T. (ed.), *Linguistic Dynamics. Discourses, Procedures, and Evolution*. Berlin: de Gruyter; 181-189.

Altmann, Gabriel (1987): "The levels of linguistic investigation", in: *Theoretical Linguistics, 14; 227-239.*

Altmann, Gabriel (1992): „Das Problem der Datenhomogenität." In: Rieger, Burghard (ed.), *Glottometrika 13*. Bochum: Brockmeyer; 287-298.

Altmann, Gabriel (1993): "Science and linguistics." In: Köhler, Reinhard; Rieger, Burkhart B. (eds.), *Contributions to Quantitative Linguistics*. Dordrecht, NL: Kluwer Academic Publishers; 3-10.

Altmann, Gabriel (2008): „Methodologische Probleme der Sprachtypologie." In: Altmann, Gabriel; Zadorožna, Iryna; Matskulyak, Yuri (eds.), *Проблеми загального, германського та слов'ясьнкого мовознавства до 70-риччя профессора В.В. Левицького. Problems of General, Germanic and Slavic Linguistics. Papers for 70-th anniversary of Professor V. Levic'kij*. Černivci: Books–XXI; 98-105.

Altmann, Gabriel; Lehfeldt, Werner (1973): *Allgemeine Sprachtypologie. Prinzipien und Messverfahren* München: Fink.

Antić, Gordana; Stadlober, Ernst; Grzybek, Peter; Kelih, Emmerich (2006): "Word Length and Frequency Distributions in Different Text Genres." In: Spiliopoulou, Myra; Kruse, Rudolf; Nürnberger, Andreas; Borgelt, Christian; Gaul, Wolfgang (eds.), *From Data and Information Analysis to Knowledge Engineering*. Heidelberg, Berlin: Springer, 310-317.

Auroux, Sylvain; Koerner, Ernst F.K.; Niederehe, Hans-Josef (eds.) (2006): *History of the Language Sciences.* Berlin, New York: de Gruyter. [= Handbücher zur Sprach- und Kommunikationswissenschaft; 18/3]

Bailey, Charles-Jaimes N. (1991): *Variation in the data: Can linguistics ever become a science?* Kea'au, HI: Orchid Land.

Best, Karl-Heinz (ed.) (1997): *Glottometrika 16. The Distribution of Word and Sentence Length*. Trier: wvt.

Best, Karl-Heinz (ed.) (2001): *Häufigkeitsverteilungen in Texten*. Göttingen: Peust & Gutschmidt.

Biber, Douglas (1988): *Variation across speech and writing*. Cambridge etc.: Cambridge University Press.

Biber, Douglas (1995): *Dimensions of register variation: a cross-linguistic comparison*. Cambridge etc.: Cambridge University Press.

Blühdorn, Hardarik (1990): „Korpuslinguistische Befunde als Ausgangspunkt für eine modifizierte Funktionalstilistik – Anregungen zu einer Neuaufnahme der Diskussion", In: *Linguistische Berichte, 127*; 217-231.

Čebanov, Sergej G. (1947): "O podčinenii rečevych ukladov 'indoevropejskoj' gruppy zakonu Puassona" [= On Conformity of Language Structures within the Indo-European Familiy to Poisson's Law], in: *Doklady Akademii Nauk SSSR / Comptes Rendus (Doklady) de l'Adadémie des Sciences de l'URS*, 55/2; 99-102.

Bartoň, Tomaš; Cvrček, Václav; Čermák, František; Jelínek, Tomáš; Petkevič, Vladimír (2009): *Statistyky čestiny*. Praha: *Lidové Noviny*.

Doležel, Lubomír (1964): „Verojatnostnyj podchod k teorii chudožestvennogo stilja", in: *Voprosy jazykoznanija 1*; 19-29.

Elderton, William P. (1949): "A Few Statistics on the Length of English Words", in: *Journal of the Royal Statistical Society, series A (general)*, 112; 436-445.

Friedl, Alexander (2006): *Untersuchungen zur Texttypologie im Russischen*. M.A. Thesis, Graz University.

Fucks, Wilhelm (1955a): *Mathematische Analyse von Sprachelementen, Sprachstil und Sprachen.* Köln/Opladen. [= Arbeitsgemeinschaft für Forschung des Landes Nordrhein-Westfalen; 34a]

Fucks, Wilhelm (1955b): „Theorie der Wortbildung", in: *Mathematisch-Physikalische Semesterberichte zur Pflege des Zusammenhangs von Schule und Universität, 4*; 195-212.

Greenberg, Joseph H. (1960): „A Quantitative Approach to the morphological typology of language", in: *International Journal of American Linguistics*, 26; 178-194.

Grotjahn, Rüdiger (1982): „Ein statistisches Modell für die Verteilung der Wortlänge"', in: *Zeitschrift für Sprachwissenschaft, 1*; 44-75.

Grotjahn, Rüdiger; Altmann, Gabriel (1993): „Modelling the Distribution of Word Length: Some Methodological Problems." In: Köhler, Reinhard; Rieger, Burghard (eds.), *Contributions to Quantitative Linguistics.* Dordrecht, NL: Kluwer Academic Publishers; 141-153.

Grzybek, Peter; Stadlober, Ernst; Kelih, Emmerich; Antić, Gordana (2006): "Quantitative Text Typology: The Impact of Word Length." In: Weihs, Claus; Gaul, Wolfgang (eds.), *Classification. The Ubiquitous Challenge*. Heidelberg, New York: Springer; 53-64.

Grzybek, Peter (2005): "History and Methodology of Word Length Studies. The State of the Art." In: Grzybek, Peter (ed.), *Contributions to the Science of Text and Language. Word Length Studies and Related Issues*. Dordrecht, NL: Springer; 15-90. [Text, Speech and Language Technology; 31]

Grzybek, Peter; Kelih, Emmerich (2005a): „Empirische Textsemiotik und quantitative Text-Typologie." In: Bernard, Jeff; Fikfak, Jurij; Grzybek,

Peter (eds.), *Text & Reality. Text & Wirklichkeit.* Ljubljana, Wien, Graz: ZRC; 95-120.

Grzybek, Peter; Kelih, Emmerich (2005b): „Textforschung: Empirisch!“ In: Banke, Julia K.; Dumont, Björn; Schröter, Anke (eds.), *Die Leipziger Text-Tage.* Leipzig: FSR; 13-34.

Humboldt, Karl Wilhelm von (1836): *Über die Verschiedenheit des menschlichen Sprachbaues und ihren Einfluß auf die geistige Entwickelung des Menschengeschlechts.* Berlin: Königliche Akademie der Wissenschaften.

Kelih, Emmerich (2011): „Zum Analytismus und Synthetismus in den slawischen Sprachen: Morphologische Wortstrukturen in Paralleltexten.“ In: *Polyslav 14*. [In print]

Kelih, Emmerich; Antić, Gordana; Grzybek, Peter; Stadlober, Ernst (2005): “Classification of Author and/or Genre? The Impact of Word Length.” In: Weihs, Claus; Gaul, Wolfgang (eds.), *Classification. The Ubiquitous Challenge.* Heidelberg, New York: Springer; 498-505.

Kelih, Emmerich; Grzybek, Peter; Antić, Gordana; Stadlober, Ernst (2006): “Quantitative Text Typology. The Impact of Sentence Length.” In: Spiliopoulou, Myra; Kruse, Rudolf; Nürnberger, Andreas; Borgelt, Christian; Gaul, Wolfgang (eds.), *From Data and Information Analysis to Knowledge Engineering*. Heidelberg, Berlin: Springer; 382-389.

Kempgen, Sebastian; Lehfeldt, Werner (2004): „Quantitative morphologische Typologie.“ In: Booij, Geert E.; Lehmann, Christian; Mugdan, Joachim (eds.), *Morphologie. Ein internationales Handbuch zur Flexion und Wortbildung.* 2. Halbband. Berlin, New York; 1235-1246.

Köhler, Reinhard (1986): *Zur synergetischen Linguistik: Struktur und Dynamik der Lexik.* Bochum: Brockmeyer.

Köhler, Reinhard (2005): „Properties of lexical units and systems.“ In: Köhler, Reinhard; Altmann, Gabriel; Piotrowski, Rajmund G. (eds.), *Quantitative Linguistik · Quantitative Linguistics. Ein internationales Handbuch · An International Handbook.*. Berlin, New York: de Gruyter; 305-312.

Krupa, Viktor (1965): „On quantification of typology“, in: *Linguistics*, 3/12; 31-36

Krupa, Viktor; Altmann, Gabriel (1966): “Relations between typological indices”, in: *Linguistics* 4/24; 29-37.

Mistrík, Jozef (1973): *Exakte Typologie von Texten.* München: Sagner.

Ohnheiser, Ingeborg (1999): „Funktionale Stilistik.“ In: Jachnow, Helmut (ed.), *Handbuch der sprachwissenschaftlichen Russistik*. Wiesbaden: Harrassowitz; 660-686.

Orlov, Jurij K. (1982): „Linguostatistik: Aufstellung von Sprachnormen oder Analyse des Redeprozesses? (Die Antinomie ‚Sprache–Rede‘ in der statistischen Linguistik)”. In: Orlov, Jurij K; Boroda, Moisej G.; Nadarejšvili, I.Š., *Sprache, Text, Kunst. Quantitative Analysen*. Brockmeyer: **Bochum: Brockmeyer; 1-55.**

Saussure, Ferdinand de (1916): *Course in General Linguistics*. New York: The Philosophical Library, 1959.

Skalička, Vladimír (1966): „Ein «typologisches Konstrukt».“ In: *Travaux linguistiques de Prague,* 2; 157-163.

Wimmer, Gejza; Altmann, Gabriel (1996): “The theory of word length: some results and generalizations. In: Schmidt, Peter (ed.), *Glottometrika 15. Issues in General Linguistic Theory and The Theory of Word Length.* Trier: wvt; 112-133.

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

Wimmer, Gejza; Altmann, Gabriel (2005): “Unified derivation of some linguistic laws.” In: Köhler, Reinhard; Altmann, Gabriel; Piotrowski, Rajmund G. (eds.), *Quantitative Linguistik · Quantitative Linguistics. Ein internationales Handbuch · An International Handbook.* Berlin, New York: de Gruyter; 791-807.

Wimmer, Gejza; Altmann, Gabriel (2006): “Towards a Unified Derivation of Some Linguistic Laws.” In: Grzybek, Peter (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“', in: *Journal of Quantitative Linguistics, 1/1*; 98-106.

Wimmer, Gejza; Witkovský, Viktor; Altmann, Gabriel (1999): “Modification of Probability Distributions Applied to Word Length Research“, in: *Journal of Quantitative Linguistics, 6/3*; 257-268.

The influx rate of Turkic glosses in Hungarian and Polish post-mediaeval texts

Kamil Stachowski, Jagiellonian University

Abstract. The paper analyzes Turkic glosses in Hungarian and Polish post/mediaeval texts from the point of view of their correlation with historical events, and of their compatibility with the Piotrovskij-Altmann law. The correspondence is found to be very good in both cases. A slight modification is proposed to the equation to lend more linguistic significance to one of the coefficients.

0 Rationale

The goal of the present paper is twofold. On one hand, it continues the work pioneered by Karl-Heinz Best (Best/Kohlhase 1983, Best 2003, 2006, 2008, 2010 and others), of collecting empirical evidence for the so-called Piotrovskij-Altmann law. By providing Hungarian and Polish data, it also adds to the issue of Turkic influence in Europe, first discussed quantitatively in Best (2005) using the example of German.

On the other hand, it attempts to show how the quantitative and qualitative analyses complement rather than oppose or exclude each other. Luděk Hřebíček worked to demonstrate this to a more traditionalistic audience (1990: 371). My aim here is to illustrate how the quantitative approach can reveal a general tendency in a collection of detailed observations gathered and explained with the philological method.

I will: **1.** explain how I prepared the data for analysis, **2.** conduct a qualitative (philological; 2.1) and quantitative (Piotrovskij-Altmann law; 2.2) analysis, and **3.** summarize the conclusions.

1 Data

The data have been collected from two historical dictionaries (Kakuk 1973 and Stachowski S. 2007), and perhaps the majority of items are no longer in use. See 1.1 and 1.2 below for language-specific details.

Here, two general assumptions need to be justified.

The first is that I trust entirely the authors of the dictionaries in the semantic separation of items, that is to say, I always count one entry as one gloss. A precise analysis of polysemy is possible (see Levyc'kyj 2003 for an overview of methods) but beyond the scope of the present work. Since polysemous entries are rare in the material used here, I chose to assume that all such occurrences have been identified by the lexicographers and broken up into separate entries. The

second assumption is that I count each item only once, without attempting to establish whether there existed a gap in its appearance in texts. This limitation is enforced by the available data and by uncertainty as to for how long exactly it would have to be missing from the sources to be counted separately. The fact, however, that even so prepared material correlates well with the historical data on one hand, and the Piotrovskij-Altmann law on the other, appears to confirm that this disadvantage is in no way critical.

1.1 Hungarian

The Hungarian data contains Ottoman (literary and dialectal) glosses from the period 1500–1698, extracted from Kakuk 1973, which is presently the only catholic study of the subject. From a purely etymological point of view, the material is mixed in three ways and thus less heterogeneous than the Polish data (see 1.2 below).

Firstly, it is only rarely established whether an item has come directly from Ottoman. Serbo-Croatian mediation is explicitly indicated in 20 cases out of 1220 in total. Other possible intermediaries are also only sporadically mentioned. This is improbable, given that, especially in the early period of contact, neither Hungarians nor the Ottomans had good translators, and were forced to rely on foreigners, most frequently Serbians (see e.g. Hazai 1977). In the strict, etymological sense, an unknown but likely a considerable part of the material are not actually *Ottoman* glosses. In what seems to be the great majority of cases, however, their appearance in Hungarian sources can only be attributed to the cultural and administrative impact of the Ottoman rule. Thus, from the general point of view of influence, they have to be considered Ottoman glosses after all, regardless of what the immediate donor was.

Secondly, the exact Turkic source within the linguistic melting pot of the Ottoman Empire is almost never established. Etymologically, this is a significant deficiency. From the point of view of the Hungarian population, however, it would have been quite irrelevant whether a particular word is known in the entire Empire or limited to just one dialect which will, in the remote future, become Turkish or Azerbaijani, or whether it ultimately stems from Proto-Turkic or Arabic or Persian. Rather, it would have been just another word used by the occupying soldiers, settlers and tax collectors, and as such, an Ottoman gloss.

Thirdly, significant parts of the material are personal and place names which only ever occur in a Hungarian context as *Fremdwörter*. What is important for me here, however, is not how much and for how long Hungarian has made them its own, but the fact that they do appear in it at all. Even a hapax legomenon is a proof of interest and, indirectly, of influence. (In an extreme case, perhaps even better a proof than the longevity of a specific borrowing. Else, in the particular case of Hungarian, the European influence would have to be concluded

nearly purged by the end of the 19th c.)

The above does not mean in any way that more detailed studies of Ottoman influence on Hungarian are unnecessary. It merely attempts to justify why greater precision is not a sine qua non in this case.

The total size of the dataset was 1220 items. From these, I removed in the following order those cases where the dating was:

1. imprecise because the source was published over a period of several years (32 items),
2. uncertain because Kakuk (apparently) only used a later copy of the original document (4 items, see Kakuk 1973: 11),
3. imprecise (a century given rather than a specific year; 11 items) or
4. unrepresentative of the given period because it lay outside of the primary scope of my source (16th–17th c., see Kakuk 1973: 8) (11 items in years 1405–94 + 61 items in years 1701–1873, but see 2.1.1 below).

Points 1. and 2. could perhaps be saved for our purpose by some arbitrary method, as e.g. taking the earlier date, or the appropriate fraction for each year in the range. But this would not make a significant difference: for the items in point 1., the mean of glosses per year ($\overline{p}$) = 2.46, standard deviation (σ) = 2.54; for those in point 2., $\overline{p}$ = 1, σ = 0. Points 3. and 4. cannot be included in any meaningful way.

The pruned dataset contains 1101 glosses from 169 unique years from the period 1500–1698. There are 1–56 glosses per year with $\overline{p}$ = 6.515, σ = 8.227.

1.2 Polish

The Polish data contains Turkic (Ottoman and Tatar) glosses from the period 1388–1791, extracted from Stachowski S. (2007), which also is presently the only catholic study of the subject. Etymologically, the material is more homogeneous than the Hungarian data (see 1.1 above), albeit still mixed.

Firstly, it contains items not only from Ottoman and its dialects but also from Tatar. Separating these two groups from each other is often difficult or effectively impossible because of relatively high phonetic similarity of the two languages, especially when forced through the Polish phonological filter. Hence, the data must be viewed as reflecting the linguistic influence of the *Turkic element* on Polish, rather than that of Ottoman or, the more so, literary Ottoman.

The other two limitations of the Hungarian material, namely the unknown route and a high number of personal and place names, are only marginally valid for the Polish data.

Probably, the great majority of borrowings from Ottoman (the Ottoman Empire, see 1.1 above) are direct. In the case of words from Tatar, perhaps Ukrainian or other mediation might have been more frequent. As for personal and place names, these are very rare in S. Stachowski's dictionary.

There is one more major difference between the two datasets analyzed here. Unlike the Hungarian data, the Polish material has been extracted in a greater part from works which deal very specifically with Turkic matters, such as accounts of legations to the Ottoman Empire. This has resulted in a few large skips. I will consider here both the entire dataset, and its trimmed down, more continuous subset.

The total size of the dataset was 1204 items. From these, I removed in the following order those cases where the dating was:

1. imprecise because the source was published over a period of several years (117 items),
2. uncertain because the source was published 35 years after it had been written (11 items from one source, see Stachowski S. 2007: XXV s.v. *LeszD*),
3. imprecise (half a century given rather than a specific year; 4 items) or
4. unrepresentative of the given period because it lays outside of the primary scope of my source (14^{th}–18^{th} c., see Stachowski S. 2007: XVII) (176 items in years 1812–1899).

As opposed to the Hungarian data in 1.1 above, here including the items from point 1. would make a significant difference: out of the 114 glosses, 83 come from one source, eleven from another, and only the remaining twenty are distributed evenly among 15 different sources. The former of these two sources has been written in years 1496–1501 (see Stachowski S. 2007: XXV s.v. *KT-Z*), the latter – in years 1582–84 (see p. XXIX s.v. *RadzPZŚ*). Unfortunately, there seems to be no objective and good way to interpret these cases.

Items from point 2. would not make a significant difference. Items from points 3. and 4. cannot be included in any meaningful way.

The pruned dataset contains 896 glosses from 130 unique years from the period 1388–1791. There are 1–203 glosses per year with $\overline{p}$ = 6.9, σ = 20.59 and kurtosis (Γ_2) = 65.56.

As the distribution of items per year is extremely uneven, I will also use here a trimmed down subset, from which the eight most bountiful years (sources) have been removed. (They supply, in order, 203, 78, 58, 51, 44, 43, 29 and 24 glosses. The following years all bring less than 17.) This subset contains 366 glosses from 122 unique years from the same period as the full set (1388–1791), with $\overline{p}$ = 3, σ = 3.29 and Γ_2 = 4.34.

2 Analysis

I will give a qualitative (philological) and a quantitative (Piotrovskij-Altmann law) analysis. The former suggests that there exists a correlation between the influx rate of unique glosses and historical events. The latter, that this influx can be modelled by the equations given below.

2.1 Qualitative (philological)

2.1.1 Hungarian

Fig. 1 shows the influx of Turkic items in Hungarian texts in years 1500–1757. This is a slightly longer period than the focus of Kakuk 1973 (see 1.1 above), and the data from the final 58 years must be considered unrepresentative. I chose to include them here nonetheless because of one important remark that needs to be made about them.

The period of Ottoman rule in Hungary, 1541–1699, is hatched in the figure. Vertical lines represent what appear to be the most important and/or relevant historical events of the time (based mostly on Kálmán 19s89 and Molnár 2001). They are:

1521: Belgrade is conquered by the Ottomans.

1526: The Hungarian army is defeated at Mohács. The Ottomans enter and pillage Buda but retreat soon afterwards, holding however central Hungary and suzerainty over Transylvania.

1541: Buda is conquered; central and southern Hungary is annexed.

1547: A five-year armistice is concluded in Edirne.

1570: The semi-independent Principality of Transylvania is established and will last until 1711.

1683: The Ottoman army is defeated at Vienna, leading to the liberation of Hungary during the next sixteen years.

1686: Buda is reconquered by the Holy League's army.

1699: A peace treaty is signed in Karlowitz, effectively transferring Hungary to the Habsburgs.

1718: A peace treaty is signed in Passarowitz, transferring the Banat of Temeswar and much of present-day Serbia to the Habsburgs.

1739: A peace treaty is signed in Belgrade, restoring Ottoman administration in Serbia, the southern part of the Banat of Temeswar and northern Bosnia.

Rather surprisingly, the conquest of the country, the greater part of which had happened till around 1550, appears to have had next to no impact on the influx of Turkic vocabulary. A pronounced rise can be seen shortly after 1550, probably due to a more wide-spread application of Ottoman administration. Another surge occurs after 1570, perhaps following the establishment of the Principality of Transylvania which remained through most of its history under Ottoman suzerainty.

The small fluctuations around 1615 and 1650 are difficult to explain by historical events. Given the modest scale, they can probably be attributed to the random factor.

After 1660, the influx rate visibly drops. Like the two aberrations above, it, too, does not seem to coincide with any particular historical event, but unlike them, this is a long-term phenomenon. Perhaps then, it is to be concluded that

after well more than a hundred years of close contact, Hungarian has finally reached a (temporary) point of saturation.

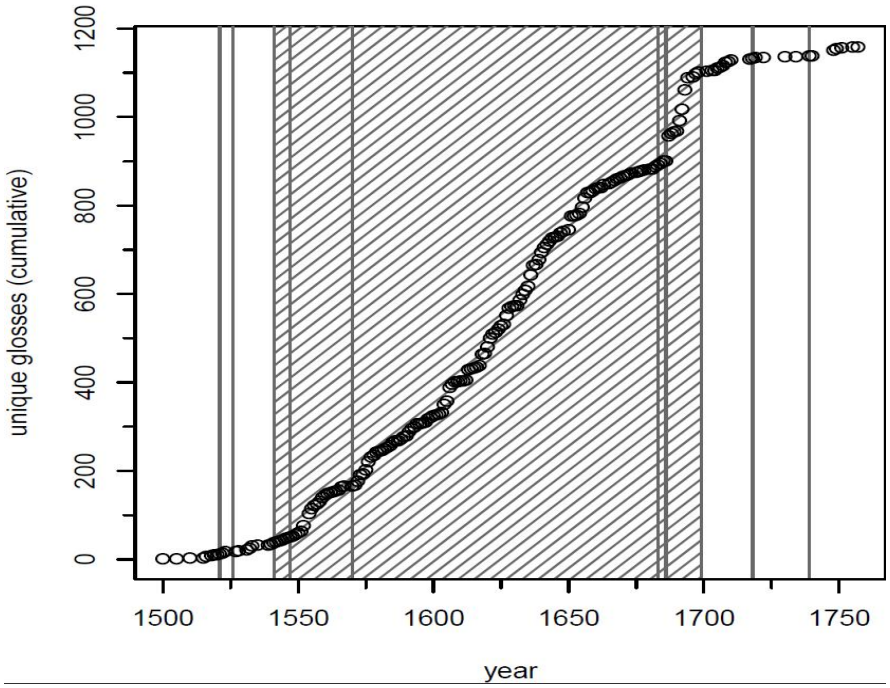


Figure 1. A cumulative sum of unique Ottoman glosses in Hungarian texts in years 1500–1757. Hatched is the period 1541–1699; see 2.1.1 for the meaning of vertical lines.

The situation changes dramatically during the War of the Holy League (1683–1698), and specifically after the reconquest of Buda in 1686. Unsurprisingly, the hope for regaining the country must have renewed the interest in Turkic matters which had lasted almost throughout the war, until the Peace of Karlowitz (1699); see also 2.1.2 below.

Subsequent events of the war between the Habsburgs and the Ottomans in present-day Croatia, Bosnia and Transylvania apparently left Hungarians indifferent, but see below. It is to be noted, however, that even after the end of the Ottoman occupation, new Ottoman items still continued to appear in Hungarian texts. This might seem counter-intuitive at first but can be easily explained. First, these items are glosses, not necessarily actual loanwords. Though weakened, the Ottoman Empire did not cease to play the role of a superpower and as such, to attract attention. Also, it is natural to expect historical treaties dealing with the period of occupation, to begin to appear after it had ended, and to contain at least a small number of previously unattested lexical items.

To sum up, the change of the influx rate can, in most cases, be explained by historical events. Interestingly, however, the coincidence does not hold equally well the other way round. 18th c. data are uncertain because this period lies out-

side of the focus of Kakuk 1973 and the numbers might be too low. Particularly surprising, however, is the early period of conquest until around 1550, especially if contrasted with the striking surge after the liberation of Buda in 1686. One conceivable explanation is that Hungarian authors would have been more eager to write about what they probably saw as a success than what obviously was a failure. Further study is necessary to provide a more certain answer.

2.1.2 Polish

Fig. 2a shows the influx of Turkic (Ottoman and Tatar) items in Polish texts in years 1388–1791. Hatched are the periods of big wars; vertical lines represent what appear to be the most important and/or relevant historical events of the time (based mostly on Davies 2005 and Markiewicz 2002). They are:

1444: The Hungarian-Polish army is defeated at Varna. The consequences are much more significant for Hungary than for Poland.

1485–1503: The First Polish-Ottoman War. The contact with the actual Turkic element, however, is limited.

1533: A peace treaty is signed in Istanbul, valid for the life of both rulers.

1569: The Union of Lublin replaces the previous personal union and creates a single state, the Polish-Lithuanian Commonwealth.

1591–93, 1594–96: The Kosiński and Nalyvaiko Uprisings, which both begin more as private quarrels but soon transform into civil wars between the local nobility and the Cossacks.

1620–21: Another conflict with the Ottoman Empire, which begins with a defeat of the Polish-Lithuanian army at Cecora in 1620 and ends with the indecisive battle of Khotyn in 1621 and the subsequent, equally ambivalent peace treaty.

1633–34: Another conflict, concluded eventually by a peace treaty in 1635 which effectively extends the previous, 1621 peace.

1648–55: Khmelnytsky Uprising, which begins more as a peasant revolt but soon transforms into a Ukrainian war of liberation of a sort; the offending party are primarily Zaporozhian Cossacks, supported by Crimean Tatars who, however, change sides twice.

1672–76: The Second Polish-Ottoman War. In 1672, the Commonwealth signs an unfavourable peace treaty in Buczacz. The war concludes with the Treaty of Żurawno in 1676 which only revises a few points of the 1672 peace.

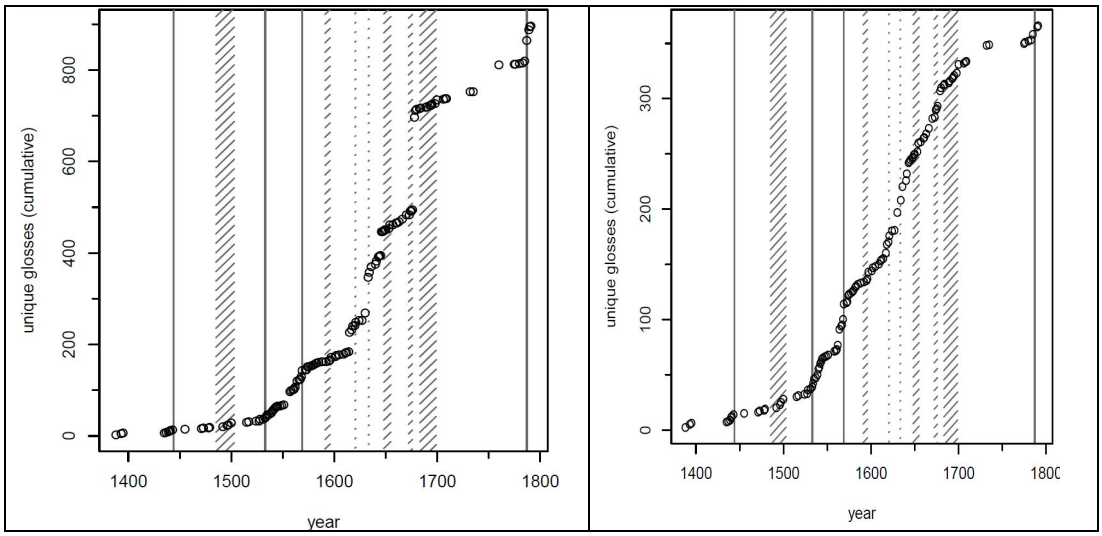
1683–99: The Third Polish-Ottoman War, which starts with the epic victory at Vienna in 1683 and ends with the Treaty of Karlowitz in 1699.

1787: The Russo-Turkish War breaks out, raising Poland's hopes of loosening the Russian protectorate.

Fig. 2b shows the same historical events as fig. 2a but superimposed on it is the trimmed down, more continuous subset of the Polish data (see 1.2 above).

The decimation of the Hungarian-Polish forces at Varna does not seem to have had much impact on the influx of Turkic glosses into Polish sources. A

slight rise is visible before the actual event. This is only mildly surprising. Overall, the aftermath of the battle was very significant, as it in fact paved the way to the Constantinople for the Ottoman Empire, but it was not critical for Poland.



2a. The entire dataset. 2b. The trimmed subset.

Figure 2. A cumulative sum of unique Turkic glosses in Polish texts in years 1388–1791. Hatched are the periods of big wars (1485–1503, 1591–96, 1620–21, 1633–34, 1648–55, 1672–76 and 1683–99); see 2.1.2 for the meaning of vertical lines.

More unexpectedly, the First Polish-Ottoman War, too, has only left a small imprint on the number of glosses. It is true that the war was relatively short and, in essence, consisted of two battles, but both were big and important ones. Perhaps it was the Lithuanian-Russian War (1512–22) and the Polish-Teutonic War (1519–21) that had drawn the attention away from the conflict in the South.

The influx rate appears to finally rise after 1533 when a long-term treaty had been signed between Poland and the Ottoman Empire, guaranteeing many years of peace. It might be suspected that this encouraged trade and other forms of pacific contacts, which caused the said rise.

Around 1550, a fluctuation can be seen, which does not seem to coincide with any significant historical event. Possibly a random variation? The clear slowdown after 1569 should probably be attributed, again, to the shift of public attention away from Turkic matters.

The first bigger difference between figures 2a and 2b turns up shortly before 1600. Inflated by a single source (Rycaut/Kłokocki 1678 with 203 items), the year 1678 has visually dominated the entire plot. The momentary increase at that time, possibly due to the two subsequent Cossack uprisings and the general tension in contacts with the Crimean Khanate, is barely visible in fig. 2a. Therefore, as far as a qualitative analysis is concerned, I believe that a somewhat

random occurrence, which the publication of one description of a country in 1678 surely is, is better disregarded if it is to eclipse a valid coincidence with historical data.

In 1620, the first of a series of consecutive wars breaks out. Shortly before it, the influx visibly regains momentum, which it is to maintain throughout the wartime, till the end of the century. It needs to be noted that in the figures, only the most important wars are marked, while in reality the entire 17th c. was defined by almost unceasing conflicts with the Ottomans, raids to and from the Crimean Khanate and Cossack uprisings (not to mention the hostilities with Sweden, Russia, Moldova and others).

Although somewhat random in nature, the skips in fig. 2a coincide surprisingly well with the periods of (big) war. In years 1645–76 and 1679–1709, however, this figure shows a slowdown whereas in fig. 2b, the rate is almost constant. A similar situation happens in year 1760 where the number of glosses rapidly increases but, as can be seen from the two figures, inflated by a single source. Again, the culprits here are the few unusually bountiful sources which distort the general picture.

Finally, in 1787 the last surge begins. In opposition to the previous two, this one has been brought about by multiple sources. Most probably, it is to be connected with the breakout of the Russo-Turkish war in that same year, which had kindled a hope in Poland that some loosening of the Russian protectorate over the country would soon follow.

To sum up, most changes in the influx rate can be explained by historical events which, in turn, are rather well reflected in the linguistic material significantly better than was the case with Hungarian in 2.1.1 above. An interesting observation is that the peace of 1533 appears to have caused an equally sharp boost, if less long-lived, as a century of almost permanent conflict.

It is not clear to me, which of the figures 2a and 2b gives a more appropriate overview of the course of influence. The former is jagged, as would be expected to result from a series of individual contacts. But on the other hand, these contacts were very numerous and followed very closely one after the other, almost blending into one, long period of war. It is perhaps debatable, what *random* actually means in our context.

2.1.3 Conclusions

In the two cases analysed here, the coincidence between historical events and linguistic data has proven to be quite accurate and mutual, that is, to be a *correlation*. This is key, as it not only provides empirical grounds for intuitive explanations resorted to in linguistics, but also it paves the way for quantitative linguistic data to support historical interpretations.

Two interesting conjectures can be made: 1. a historical success is more likely to cause a surge in the number of glosses than a failure (based on 2.1.1),

and 2. an increase in the number of glosses is equally likely to be caused by hostilities as by friendship (base on 2.1.2). Both still require further study.

2.2 Quantitative (Piotrovskij-Altmann law)

The Piotrovskij-Altmann law states that change in language can be modelled with the logistic equation (1):

$$(1) \qquad p(t) = \frac{c}{1 + ae^{-bt+Ct^2}}$$

where *p(t)* is the number of forms in question at time *t*, and *a*, *b*, *c* and *C* are coefficients.

The law has three variants which describe: 1. a complete change (the replacement of an obsolete form by the new one, where $c = 1$ and $C = 0$), 2. a partial change (where $C = 0$, see below), and 3. a reversible change (one that ended before it could dominate the language). On the history of the law, see e.g. Altmann (1983: 59–60), Lehfeldt/Altmann (2003: 142–44) and Leopold (2005: 627–28, 631–32), but also Vulanović (2007: 112f, 116f, 120).

Our point of interest here is the middle variant (there is no limit to how many glosses can possibly appear in texts, and once they appear, they become attested and cannot be taken back), which is given by

$$(2) \qquad p(t) = \frac{c}{1 + ae^{-bt}}$$

The function is sigmoid (has the shape of a stretched letter *s*) and fine-tuned by the three coefficients which control: a (> 0) – the horizontal displacement of the centre of the slope, b ($\neq 0$) – the steepness, and c ($\neq 0$) – the height. In other words, it states that a change begins slowly, then accelerates in its middle phase and later slows down again as it nears its end. It will be shown below that the intuitiveness of this observation is, to some degree, illusory.

As can be seen, the only independent variable is time (t). Additional, process-specific parameters, such as the duration or intensity of linguistic influence, are encoded in the coefficients and calculated anew for every analyzed change. Thus, the coefficients can be viewed as indices of a kind: a describing the point in time at which the process intensifies, b – its intensity, and c – its strength (the degree of influence). This is illustrated in fig. 3. Note that a and b depend on the employed time scale and are meaningless if the unit of time is not specified (year, half a century &c.).

The point of the analysis is to find how well the curve given by eq. 2 can approximate the empirical data. The goodness of fit is typically measured by the R^2 coefficient of determination, with the maximum value of 1 which defines a perfect fit.

As advised by Best/Beöthy/Altmann (1990: 117), the data are usually grouped into periods, typically of twenty-five, fifty or a hundred years. Randomness is thus partially removed from the picture and the resulting plot clearer. However, some potentially interesting details may also be lost. I chose to be as precise here as the data allow, which is one year.

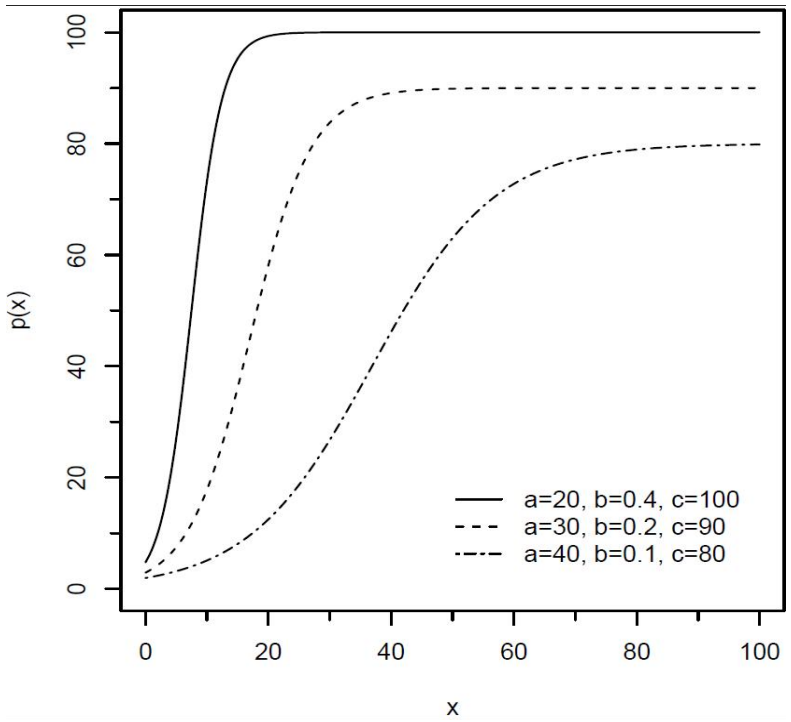


Figure 3. A comparison of function *p* (eq. 2) with different coefficients.

The numbering of years and centuries is usually transformed to begin with 1. It is, after all, purely contractual, and large numbers tend to hinder the fitting because of coefficient *a*. However, my goal here is to demonstrate how the qualitative and quantitative approaches can be used together. Leopold (2005: 631) classifies our law as an *intuitive Heuristik, die allerdings der empirischen Überprüfung standhält*. The parameter *a* does not represent any particular linguistic, historical or other value. Hence, it seems potentially beneficial to transform the equation to

$$p(t) = \frac{c}{1+e^{-b(t-A)}} \tag{3}$$

where $A = ln(a)/b$ and, unlike *a*, remains within sane limits even with higher year numbers (e.g. with the Hungarian data below, $A = 1633.038$ which corresponds to $a = e^{47.298} = 3.476 \times 10^{20}$ in eq. 2). Also, *A* is much more meaningful linguistically, as it marks the precise centre of the sigmoid, i.e. the turning point of the influx rate.

2.2.1 Hungarian

Fig. 4 shows the Hungarian empirical data as points and the fitted curve (eq. 3, fitted with *R*). The fit is quite good, with $R^2 = 0.9924$.

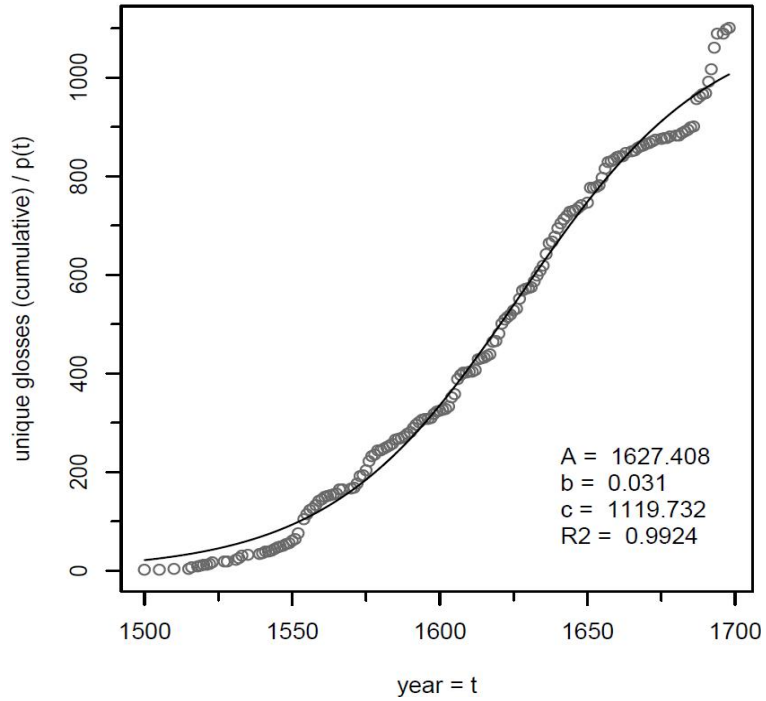


Figure 4. Fitting eq. 3 to the Hungarian data.

Perhaps in most linguists' intuition, the slow-quick-slow scheme associates with a single period of intensified contact, such as a war or occupation, and every another such period would repeat the same pattern from the start, in complete detachment from the previous one or ones. The Hungarian data show that this is not necessarily true.

Firstly, there are in fact two sigmoids visible in the empirical data, and the greater parts of both fall entirely within the one period of occupation (fig. 1; see also 2.2.2 below). Secondly, the two separate sigmoids can still be approximated quite accurately with just a single curve (fig. 4). Neither does one period necessarily cause one sigmoidal influx, nor are the subsequent sigmoids in complete detachment from each other. This seems to prove that the Piotrovskij-Altmann law is less intuitive than it might seem at face value.

These observations would be impossible if years were grouped into intervals of e.g. 25, as has often been the custom in quantitative studies. The appropriate data are given and plotted in fig. 5.

The question, then, is how much precision is optimal. The second sigmoid in fig. 4 might be an irrelevant, random fluctuation, which grouping only irons out – or a perfectly valid occurrence, which would have contributed a deeper

insight, had grouping not obscured it. But greater precision might also be misleading. Typically, sources used for historical dictionaries are prints. In each case, the time that elapsed between the moment when the original text had been written, and the year when it was printed, is different. Grouping helps remove this randomness from the picture. On the other hand, the surge in 1686 can be rather believably explained by the reconquest of Buda. Without sound empirical evidence, this discussion appears to be a battle of beliefs.

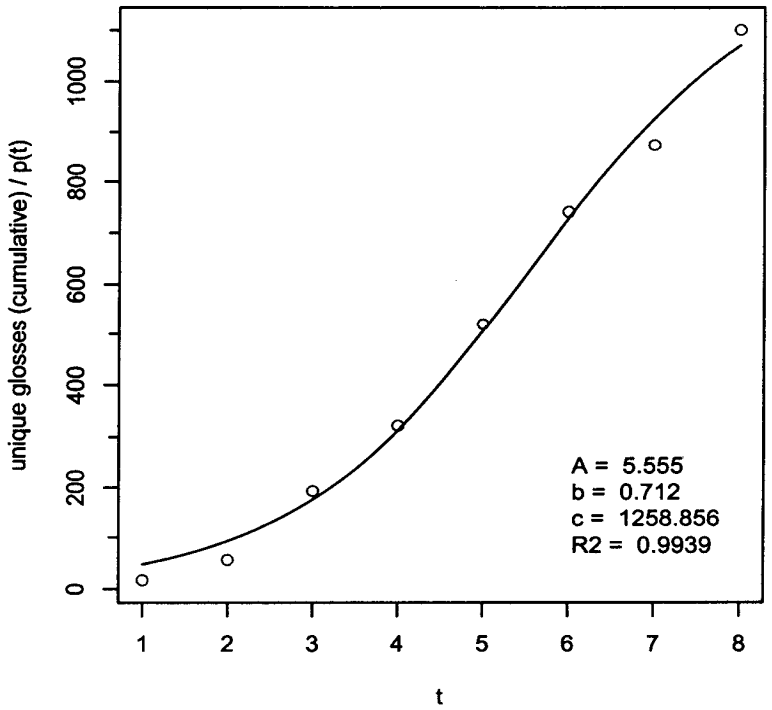


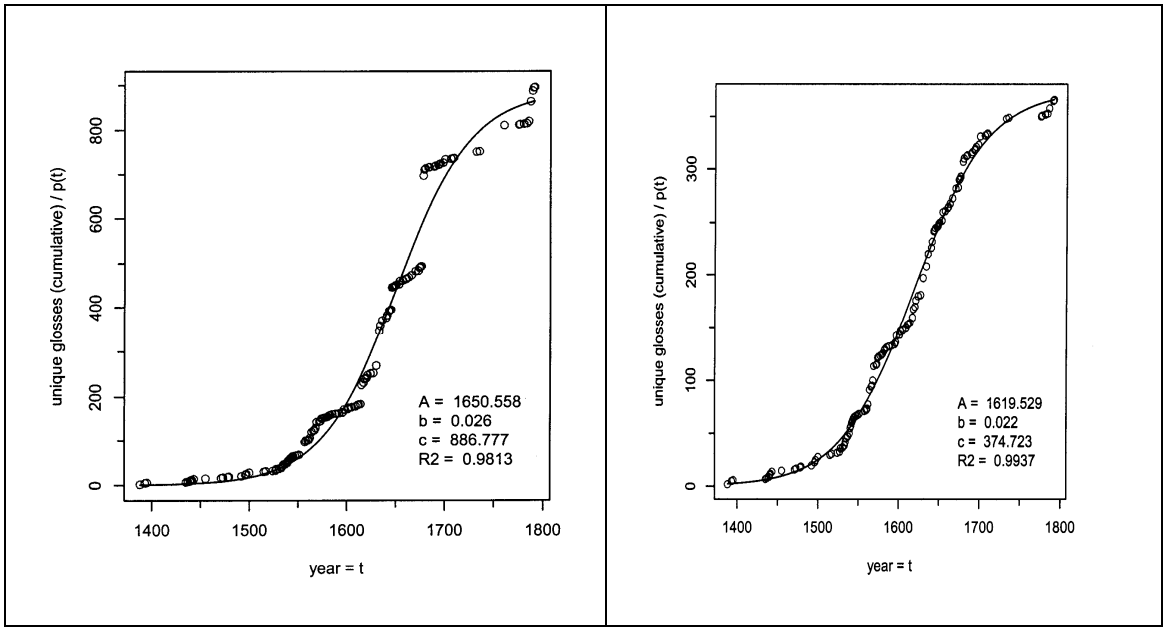
Figure 5. Fitting eq. 3 to grouped Hungarian data

Table 1
Hungarian data

century	*t*	observed	cumulative	predicted
16/1	1	17	17	47.349
16/2	2	39	56	92.851
16/3	3	137	193	175.735
16/4	4	129	322	312.762
17/1	5	198	520	506.654
17/2	6	221	741	728.225
17/3	7	133	874	927.248
17/4	8	227	1101	1070.89
$A = 5.555$ $b = 0.712$ $c = 1258.856$ $R^2 = 0.9939$				

2.2.2 Polish

Figures 6a and 6b show the Polish empirical data as points and the fitted curves (eq. 3, fitted with *R*). The fits are quite good, with R^2 = 0.9813 and 0.9937, respectively. The difference between the two sets is negligible, which appears to be a visually convincing confirmation of the validity of the Piotrovskij-Altmann law.



6a. The entire dataset 6b. The trimmed subset

Figure 6. Fitting eq. 3 to the Polish data

As opposed to the Hungarian data in 2.2.1, here there had not been a single, long period of intensified contact within which more than one sigmoid could be found. Rather, there had been a long sequence of consecutive shorter ones. The fractal-like course of the influx is perhaps less striking but visible nonetheless. At least five component sigmoids can be easily discerned and approximated rather accurately with a multi-logistic curve (with R^2 = 0.9987 versus 0.9813 with a single one), as illustrated in fig. 7. This is hardly surprising. A much more insightful observation, which, incidentally, would have been lost had the years been grouped, is that the entire dataset can still be modelled with quite high fidelity, by just one single sigmoid.

It might also be remarked that the plots are somewhat reminiscent of the punctuated equilibrium theory (Dixon 1997), only with (considerably) shorter periods of stasis. Possibly, some renewal of interest in the concept, combined perhaps with an effort to formulate for it a more precise definition, is desirable.

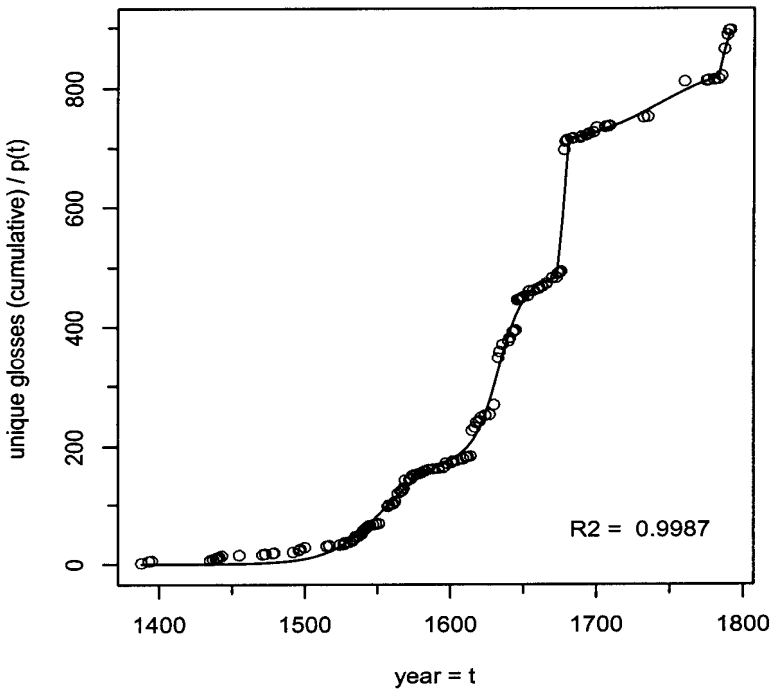


Figure 7. Fitting a multi-logistic curve to the Polish data

Table 2
Coefficients of the five component sigmoids in Fig. 7

$p(t) = \Sigma \frac{c_i}{1 + e^{-b_i(t-A_i)}}$			
i	*Ai*	*bi*	*ci*
1	1555	0.055	185.369
2	1634	0.119	292.535
3	1677	2.820	227.452
4	1742	0.039	135.793
5	1787	2.031	70.472

2.2.3 Conclusions

The three datasets analyzed here can all be approximated very closely by the curve given in equations 2 and 3, which strengthens the empirical support for the

Piotrovskij-Altmann law.

The inconvenient and linguistically meaningless coefficient *a* has been replaced here with an equivalent coefficient $A = ln(a)/b$ which allows the fit to be performed on a timescale in actual years AD, and which is considerably more interpretable from the historical and linguistic point of view.

Also, it has been shown that the Piotrovskij-Altmann law is non-trivial in that it goes in fact against what I believe to be the standard linguistic intuition.

3 Final conclusions

In the present paper, I conducted a qualitative (philological) and a quantitative (Piotrovskij-Altmann law) analysis and attempted to show that the two complement rather than oppose or exclude each other.

It has been suggested (2.1.3) that the coincidence between historical events and linguistic data can probably be considered a correlation, which would enable both linguistics and history to benefit from the results of similar studies. Also, two conjectures have been made on the finer points of this relation.

The two (three) analyzed datasets proved to conform to the Piotrovskij-Altmann law (2.2.3). A slightly modified version of the equation has been proposed, which replaces one of the coefficients with its linguistically and historically more meaningful equivalent. Also, it has been suggested that the purported self-ecidence of the law is, at least to some degree, illusory.

References

Altmann, G. (1983). *Das Piotrovski-Gesetz und seine Verallgemeinerungen.* In: Best/Kohlhase (1983), 59–90.

Best, K.-H. (2003). Spracherwerb, Sprachwandel und Wortschatzwachstum in Texten. Zur Reichweite des Piotrowski-Gesetzes. *Glottometrics 6, 9–34.*

Best, K.-H. (2005). Turzismen im Deutschen. *Glottometrics 11, 56–63.*

Best, K.-H. (2006). Deutsche Entlehnungen im Englischen. *Glottometrics 13, 66–72.*

Best, K.-H. (2008). Sinismen im Deutschen und Englischen. *Glottometrics 17, 87–93.*

Best, K.-H. (2010). Zur Entwicklung des Wortschatzes der deutschen Umgangssprache. *Glottometrics 20, 34–37.*

Best, K.-H., Beöthy E., Altmann G. (1990). Ein methodischer Beitrag zum Piotrowski-Gesetz. *Glottometrika 12, 115–24.*

Best, K.-H., Kohlhase, J. (eds.) (1983). *Exakte Sprachwandelforschung. Theoretische Beiträge, statistische Analysen und Arbeitsberichte* (= *Göttinger Schriften zur Sprach- und Literaturwissenschaft* 2). Göttingen: edition

herodot.

Davies, N. (2005). *God's Playground. A History of Poland*, 1: *The Origins to 1795*, Oxford – New York: Oxford University Press [revised edition].

Dixon, R.M.W. (1997). *The rise and fall of languages*, Cambridge: Cambridge University Press.

Hazai, Gy. (1977). Zur Rolle des Serbischen im Verkehr des Osmanischen Reiches mit Osteuropa im 15.–16. Jahrhundert. In: Décsy Gy., Dimov-Bogoev H.D. (ed.), *Eurasia Nostratica 2, 82–88*, Wiesbaden: Eurolingua.

Hřebíček, L. (1990). *Quantitative studies*. In: Hazai Gy. (ed.), *Handbuch der türkischen Sprachwissenschaft 1, 371–387*. Budapest: Akadémiai Kiadó.

Kakuk, S. (1973). *Recherches sur l'histoire de la langue osmanlie des XVI*e *et XVII*e *siècles. Les éléments osmanlis de la langue hongroise.* Budapest: Akadémiai Kiadó.

Kálmán, B. (ed.) ([3]1989). *Magyarország történeti kronológiája*, 2: *1526–1848*, Budapest: Akadémiai Kiadó.

Lehfeldt, W., Altmann, G. (2003). Протекание падения редуцированных в древнерусском языке в свете закона Пиотровских. *Russian Linguistics 27, 141–49.*

Leopold, E. (2005). Das Piotrowski-Gesetz (Piotrowski's law). In: Köhler R., Altmann G., Piotrowski R.G. (eds.), *Quantitative Linguistik. Quantitative Linguistics. Ein internationales Handbuch. An International Handbook: 627–633.* Berlin – New York: Walter de Gruyter.

Levyc'kyj, V.V. (2003). Лексична полісемія та квантитативні методи її дослідження. *Мовознавство 4, 17–25*.

Markiewicz, M. (2002). *Historia Polski 1492–1795*. Kraków: Wydawnictwo Literackie.

Molnár, M. (2001). *A Concise History of Hungary*. Cambridge: Cambridge University Press.

R Development Core Team (2011). *R: A Language and Environment for Statistical Computing*, Vienna: R Foundation for Statistical Computing.

Rycaut/Kłokocki 1678 = **Rycaut P.**, *Monarchia Turecka, Opisana. Przez Ricota Sekretarza Posła Angielskiego. V Porty Ottomanskiey Residviącego. Z Francvskiego Ięzyka Na Polski Przetłvmaczona. Przez Szlachcica Polskiego. Y do Druku podána w Roku 1678.* Słuck: Druk. Radziwiłłowska [transl. H. Kłokocki].

Stachowski, S. (2007). *Słownik historyczny turcyzmów w języku polskim.* Kraków: Księgarnia Akademicka.

Vulanović, R. (2007). Fitting Periphrastic *do* in Affirmative Declaratives. *Journal of Quantitative Linguistics 14(2-3), 111-126.*

Distances between words of equal length in a text

Peter Zörnig, Brasília

1. Introduction

While there exist countless articles to describe the frequency distributions of words or other text units, which represent static properties of text or language, the study of dynamical properties of text generation, e.g. the study of the regularities underlying the repetition of text units, is a relative new research topic, see e.g. Corral et al. (2009, p. 2). There are e.g. some approaches in natural language processing, studying the distribution of distances in a text, where usually the distances between occurrences of *the same* word are considered, see Altmann et al. (2009, p.3). In the present article we are engaged in some other dynamical property, i.e. we consider the distances between words of equal length occurring in the course of text deployment.

Whatever linguistic unit or its property we take into account, each of them has the chance to be repeated in a text. This fact is given by several circumstances: (a) the finiteness of inventories of whatever kind (e.g. of phonemes, morphemes, words, clause patterns, etc.) that forces some units to be repeated, (b) Skinner's effect of reinforcing, increasing the probability of the occurrence of a unit in a short distance after its previous appearance, and (c) the existence of (short) synsemantics which play different roles in different languages and may depend on the writers or the text sorts.

Since these circumstances reach from necessity down to free use, they may evoke the rise of different patterns which characterize either the given text or are valid for all texts. The patterns may be static, e.g. follow the frequency distribution of an entity, where different indicators may be valid for the whole text (golden section, lambda indicator, etc.), or dynamic, e.g. following the Hurst exponent related to the autocorrelation in forming word and sentence lengths.

Because there is a very restricted inventory of word lengths and the grammar forces some lengths to be repeated, the distribution cannot follow the random model. However, the question is which law governs the repetitions of word lengths. In the present article we show that the distribution of distances between words of equal length can be well described by means of the mixed negative binomial distribution.

2. Basic concepts and previous results

As in Zörnig (2010), we interpret a real text as a sequence $S = (s_1,\ldots,s_n)$ of length n, consisting of elements chosen from the set $\{1,\ldots,m\}$, where the element r occurs exactly k_r times for $r = 1,\ldots,m$ $(k_1 + \ldots + k_m = n)$.

For such a sequence, $f_d^{(r)}$ denotes the number of occurrences of the distance d between two consecutive elements of type r. The distance between two consecutive elements of type r is defined as the number of elements $\neq r$, lying between them. Furthermore we consider the total frequency of the distance d, that is

$$f_d = f_d^{(1)} + \ldots + f_d^{(m)}. \tag{2.1}$$

We are interested in the distribution of the distances $f_0, f_1, \ldots, f_{n-2}$ in a given sequence S.

In the present application, the element s_j of the sequence S represents the length of the j-th word of the real text, measured by the number of syllables.

Example 2.1: Consider the title of the article Zörnig (2010) „Statistical simulation and the distribution of distances between identical elements in a random sequence". Writing down the length of the $n = 14$ words we obtain the sequence

$$S = (4,4,1,1,4,1,3,2,4,3,1,1,2,2). \tag{2.2}$$

The frequencies of the text elements are

$$k_1 = 5,\ k_2 = 3,\ k_3 = 2,\ k_4 = 4 \quad (m = 4).$$

Between the consecutive elements of type 1 we encounter the distances 0, 1, 4 and 0. Thus we have two occurrences of the distance 0, one occurrence of the distance 1, and one occurrence of the distance 4, i.e. $f_0^{(1)} = 2$, $f_1^{(1)} = f_4^{(1)} = 1$.

In the same way we find for the other text elements $f_0^{(2)} = f_4^{(2)} = 1$, $f_2^{(3)} = 1$, $f_0^{(4)} = f_2^{(4)} = f_3^{(4)} = 1$. Thus the overall distance frequencies are, see (2.1):

$$f_0 = 4,\ f_1 = 1,\ f_2 = 2,\ f_3 = 1,\ f_4 = 2.$$

In general it holds

$$\sum_{d=0}^{n-2} f_d = \sum_{d=0}^{n-2}\sum_{r=1}^{m} f_r^{(d)} = \sum_{r=1}^{m}\sum_{d=0}^{n-2} f_r^{(d)} = \sum_{r=1}^{m}(k_r - 1) = n - m. \tag{2.3}$$

The distribution of distances has been intensively studied for randomly constructed sequences, see Zörnig (1984, 1987, 2010). In this case the expected frequencies between elements of type r are given by

$$\bar{f}_d^{(r)} = E(f_d^{(r)}) = \frac{k_r(k_r - 1)(n - k_r)_{(d)}}{n_{(d+1)}} \qquad \text{for } d = 0,\ldots,n\text{-}2, \tag{2.4a}$$

and by summing up we obtain

$$\bar{f}_d = E(f_d) = \frac{1}{n_{(d+1)}} \sum_{r=1}^{m} k_r (k_r - 1)(n - k_r)_{(d)} \quad \text{for } d = 0,\ldots,n\text{-}2 \tag{2.4b}$$

where $n_{(d)} = n(n\text{-}1)\ldots(n\text{-}d\text{+}1)$ denotes the decreasing factorial, see e.g. Zörnig (2010, Theorem 2.2)).

The sequence $\bar{f}_0, \bar{f}_1, \ldots, \bar{f}_{n-2}$ is decreasing with falling rate of descent and for small distances d, the expected frequencies can be very well approximated by the formulas

$$\bar{f}_d^{(r)} \approx (k_r - 1) f_r (1 - f_r)^d \quad \text{for } d = 0,\ldots,n\text{-}2, \tag{2.5a}$$

$$\bar{f}_d \approx \sum_{r=1}^{m} (k_r - 1) f_r (1 - f_r)^d \quad \text{for } d = 0,\ldots,n\text{-}2, \tag{2.5b}$$

where $f_r = k_r/n$.

The expression (2.5b) represents a mixture (convex combination) of geometric distributions with weights $k_r - 1$, see Zörnig (2010, p. 12).

From (2.4a) we obtain the recurrence relation

$$\bar{f}_d^{(r)} = \frac{n - k_r - d - 1}{n - d} \bar{f}_{d-1}^{(r)} \tag{2.6}$$

for the distance frequencies between elements of the type r. In particular, for small d we obtain the approximate relation

$$\bar{f}_d^{(r)} \approx \frac{n - k_r}{n} \bar{f}_{d-1}^{(r)} \tag{2.7}$$

which is the recurrence relation corresponding to (2.5a).

The above models can not be directly applied to real texts which are not randomly generated. However, in the probabilistic model studied in the next section we will make use of the fact that the distance frequencies f_d are naturally composed of the frequencies $f_d^{(r)}$.

In previous studies of natural language processing the distance between occurrences of the same word has been modeled by the stretched exponential distribution, see Altmann et al. (2009, p.3). The so-called dependency distances, i.e. distances between dependent and governor of a sentence have been adjusted by the right truncated Zeta distribution, see Liu (2007). For fitting the distribution of distances in various information strings (e.g. specific natural texts, sequences of amino acids, exe-files) the exponential, lognormal, Weibull and

negative binomial distributions could be fitted with different precision, see e.g. Kunz and Rádl (1998) and the site kunzmilan.20.m.com.distrib.pdf.

3. Fitting data by means of the mixed negative binomial distribution

We study the distribution of the distances between words of equal length in 22 texts of different languages with length between 280 and 3140 words, presented in the following tables. For each text given in form of a sequence (2.2), the observed distance frequencies $f_0, f_1,\ldots, f_{20}$ have been determined with the aid of a MAPLE program (see the respective columns f_d in the following tables). Then the software Altmann-Fitter (1997) has been applied to determine automatically (out of a stock of about 200 distributions) the model which is best suited to fit the observed values $f_0, f_1,\ldots, f_{20}$. In all of the 22 cases the mixed negative binomial distribution (MNB) was suitable for adjusting the data and in most of the cases this distribution provided the best fit among all models available by this software.

The results for the fitting the MNB to the 22 texts are presented in tables 1-6. For each text the following information is given:

n: length of the sequence S
m: number of different text elements of S
k_i: frequency of the element i
f_d : observed frequency of the distance d
$\bar{f}_d$: calculated frequency of the distance d

The frequencies of distances greater than 20 have been pooled; the line "rest" presents the sums $\sum_{d>20} f_d$ and $\sum_{d>20} \bar{f}_d$, respectively. The lower boxes in the tables contain the results of the adjustment:

k, p_1 , p_2 and α are the optimal parameter values, χ^2 is the chi square value, $P(\chi^2)$ is the p-value (probability to exceed the observed chi square value)
DF represents the number of degrees of freedom.
$C = \chi^2/N$ is a discrepancy coefficient, where $N = n - m$ denotes the sample size, see (2.3).

As a common criterion of quantitative linguistics the fit can be considered as good or satisfactory, if the p-value exceeds 0.05 or 0.01, respectively. According to this criterion the fitting of the MNB was good for 18 of 22 texts and satisfactory in the remaining 4 cases.

Table 1:

	Bulgarian N. Ostrovskij, Kak se kaljavaše stomanata, Chapter 1 n = 926, m = 6 k_1 = 336 k_2 = 269 k_3 = 213 k_4 = 78 k_5 = 27 k_6 = 3		**Hungarian** press: A nomina-lizmus forradalma n = 1314 , m = 9 k_1 = 392 k_7 = 9 k_2 = 304 k_8 = 8 k_3 = 266 k_9 = 2 k_4 = 159 k_5 = 128 k_6 = 46		**Hungarian** press: Kunczekolbász n = 458, m = 9 k_1 = 122 k_7 = 8 k_2 = 129 k_8 = 1 k_3 = 81 k_9 = 1 k_4 = 68 k_5 = 34 k_6 = 14		**Macedonian** N. Ostrovskij, Kako se kaleše čelkiot, Chapter 1 n = 1123, m = 6 k_1 = 426 k_2 = 280 k_3 = 217 k_4 = 123 k_5 = 56 k_6 = 21	
d	f_d	$\bar{f}_d$	f_d	$\bar{f}_d$	f_d	$\bar{f}_d$	f_d	$\bar{f}_d$
0	201	194.76	232	245.91	79	75.53	202	228.62
1	223	199.25	230	206.44	73	79.49	251	223.27
2	140	151.34	182	166.48	70	65.47	189	172.73
3	88	102.85	129	132.47	55	49.32	114	122.80
4	63	66.92	95	104.75	31	35.77	80	84.32
5	44	43.40	84	82.56	24	25.65	51	57.22
6	31	28.93	55	64.95	21	18.52	35	39.23
7	27	20.29	41	51.07	15	13.66	34	27.59
8	16	15.15	41	40.17	12	10.39	26	20.16
9	12	12.02	34	31.62	5	8.20	20	15.42
10	8	10.02	26	24.94	5	6.71	9	12.36
11	9	8.64	20	19.72	5	5.68	11	10.34
12	9	7.60	16	15.65	8	4.94	11	8.95
13	5	6.76	14	12.47	6	4.39	3	7.95
14	0	6.05	15	9.99	6	3.96	9	7.17
15	2	5.41	4	8.06	0	3.60	9	6.55
16	2	4.84	5	6.56	4	3.29	5	6.01
17	4	4.32	7	5.38	2	3.02	7	5.54
18	1	3.84	2	4.46	3	2.77	3	5.11
19	5	3.41	4	3.74	3	2.55	1	4.72
20	2	3.02	6	3.18	1	2.34	2	4.35
rest	28	21.16	63	64.42	21	23.78	45	46.67
	k = 2,1265 p_1 = 0,1609 p_2 = 0,5254 α = 0,1828 χ^2 = 25,10 P(χ^2) = 0,0925 DF = 17 C = 0,0273		k = 1.0865 p_1 = 0.2287 p_2 = 0.0244 α = 0.9299 χ^2 = 19.22 P(χ^2) = 0.3163 DF = 17 C = 0.0147		k = 1.8079 p_1 = 0.1143 p_2 = 0.4266 α = 0.2371 χ^2 = 13.70 P(χ^2) = 0.6884 DF = 17 C = 0.0305		k = 1.7413 p_1 = 0.4452 p_2 = 0.1081 α = 0.8224 χ^2 = 26.17 P(χ^2) = 0.0714 DF = 17 C = 0.0234	

Table 2:

	Romanian O, Paler, Aventuri solitare, excerpt n = 891 , m=7 k_1 = 392 k_2 = 220 k_3 = 151 k_4 = 92 k_5 = 22 k_6 = 13 k_7 = 1		**Romanian** N, Steinhardt, Jurnalul fericirii, Trei soluții n = 1511, m=7 k_1 = 706 k_2 = 375 k_3 = 220 k_4 = 142 k_5 = 51 k_6 = 13 k_7 = 4		**Russian** Ostrovskij , Kak zakaljalas stal' n = 792, m=7 k_1 = 264 k_2 = 265 k_3 = 168 k_4 = 70 k_5 = 17 k_6 = 7 k_7 = 1		**Serbian** N. Ostrovskij, *Kako se kalio čelik*, Chapter 1 n = 1001, m=6 k_0 = 7 k_1 = 359 k_2 = 328 k_3 = 198 k_4 = 81 k_5 = 28	
d	f_d	$\bar{f}_d$	f_d	$\bar{f}_d$	f_d	$\bar{f}_d$	f_d	$\bar{f}_d$
0	200	225.95	408	432.30	208	194.70	260	251.77
1	240	193.76	367	316.70	152	156.40	185	201.73
2	118	129.83	200	208.71	111	113.56	152	147.41
3	75	80.94	117	133.94	79	80.04	123	104.41
4	53	50.41	80	86.08	63	55.97	67	72.91
5	29	32.87	63	56.41	37	39.28	48	50.61
6	26	23.11	38	38.28	24	27.89	26	35.09
7	15	17.66	25	27.24	24	20.16	29	24.41
8	17	14.48	31	20.47	11	14.93	20	17.12
9	12	12.45	18	16.24	8	11.36	10	12.15
10	11	11.01	14	13.53	6	8.91	9	8.78
11	7	9.88	11	11.71	9	7.18	8	6.49
12	7	8.91	16	10.42	3	5.95	5	4.94
13	8	8.05	12	9.44	6	5.04	3	3.88
14	6	7.26	7	8.67	2	4.35	1	3.16
15	8	6.53	7	8.01	6	3.81	3	2.65
16	6	5.86	3	7.44	2	3.38	2	2.30
17	2	5.25	7	6.92	3	3.02	1	2.05
18	6	4.69	8	6.45	0	2.71	1	1.86
19	2	4.18	2	6.01	2	2.45	2	1.72
20	0	3.71	2	5.61	3	2.21	4	1.61
rest	36	27.20	68	73.43	26	21.69	36	37.94
	k = 1,8831 p_1 = 0,5555 p_2 = 0,1484 α = 0,7528 χ^2 = 29,21 P(χ^2) = 0,0326 DF = 17 C = 0,0330		k = 1.2797 p_1 = 0.4369 p_2 = 0.0793 α = 0.8078 χ^2 = 31.47 P(χ^2) = 0.0175 DF = 17 C = 0.0209		k = 1.2563 p_1 = 0.0994 p_2 = 0.3708 α = 0.1700 χ^2 = 15.49 P(χ^2) = 0.5602 DF = 17 C = 0.0197		k = 1.2174 p_1 = 0.3439 p_2 = 0.0438 α = 0.9215 χ^2 = 16.55 P(χ^2) = 0.4854 DF = 17 C = 0.0166	

Table 3:

	Slovak Bachletová, Moja Dolná zem n = 873, m = 9 k_1 = 232 k_6 = 3 k_2 = 325 k_7 = 0 k_3 = 204 k_8 = 0 k_4 = 87 k_9 = 1 k_5 = 21		**Slovak** Bachletová, Riadok v tlačive: nezamestnaný n = 924, m = 7 k_1 = 258 k_6 = 11 k_2 = 258 k_7 = 1 k_3 = 233 k_4 = 120 k_5 = 43		**Slovenian** N. Ostrovskij, Kako se je kalilo jeklo, Chapter 1 n = 977, m = 6 k_1 = 426 k_2 = 300 k_3 = 172 k_4 = 61 k_5 = 17 k_6 = 1	
d	f_d	$\bar{f}_d$	f_d	$\bar{f}_d$	f_d	$\bar{f}_d$
0	217	232.02	210	207.60	300	301.67
1	169	168.92	157	161.67	203	187.83
2	129	118.53	119	123.56	141	126.82
3	86	82.78	114	93.87	73	88.01
4	58	58.15	73	71.15	52	62.03
5	42	41.35	45	53.88	46	44.22
6	38	29.92	46	40.80	33	31.84
7	24	22.11	24	30.93	30	23.16
8	13	16.74	19	23.48	13	17.03
9	13	13.01	17	17.87	6	12.68
10	18	10.37	11	13.64	10	9.58
11	7	8.48	14	10.47	10	7.35
12	7	7.08	7	8.07	2	5.74
13	3	6.02	6	6.27	2	4.57
14	4	5.20	6	4.91	8	3.72
15	6	4.55	3	3.89	5	3.09
16	3	4.02	2	3.12	1	2.62
17	2	3.58	4	2.53	3	2.27
18	4	3.20	4	2.09	1	2.00
19	0	2.87	2	1.76	3	1.79
20	0	2.58	1	1.50	0	1.62
rest	23	24.52	33	33.94	29	31.35
	k = 1.0965 p_1 = 0.3512 p_2 = 0.0969 α = 0.7936 χ^2 = 20.04 P(χ^2) = 0.2724 DF = 17 C = 0.0231		k = 1.0398 p_1 = 0.2523 p_2 = 0.0251 α = 0.9427 χ^2 = 14.79 P(χ^2) = 0.6107 DF = 17 C = 0.0161		k = 0.8585 p_1 = 0.2788 p_2 = 0.0389 α = 0.9142 χ^2 = 28.92 P(χ^2) = 0.0352 DF = 17 C = 0.0298	

Table 4:

	Sundanese Aki Satimi (Online) n = 1283, m = 5 k_1 = 308 k_2 = 593 k_3 = 284 k_4 = 81 k_5 = 17		**Sundanese** Agustusan (Salaka Online) n = 416, m = 6 k_1 = 97 k_2 = 203 k_3 = 74 k_4 = 36 k_5 = 5 k_6 = 1		**Indonesian** Pengurus PSM terbelah (press) n = 345, m = 6 k_1 = 35 k_2 = 139 k_3 = 109 k_4 = 56 k_5 = 5 k_6 = 1		**Indonesian** Sekolah ditutup (press) n = 280, m = 6 k_1 = 40 k_2 = 94 k_3 = 105 k_4 = 33 k_5 = 5 k_6 = 3	
d	f_d	$\bar{f}_d$	f_d	$\bar{f}_d$	f_d	$\bar{f}_d$	f_d	$\bar{f}_d$
0	342	341.74	121	119.46	110	106.88	73	74.16
1	303	280.16	93	75.26	59	60.97	72	56.88
2	179	193.21	52	52.67	43	41.12	28	39.30
3	126	126.48	38	38.15	32	29.20	28	26.36
4	77	81.91	15	28.13	25	21.28	14	17.57
5	56	53.79	20	20.99	9	15.78	12	11.80
6	37	36.48	13	15.81	13	11.85	7	8.08
7	24	25.88	8	12.00	6	8.99	9	5.69
8	17	19.32	11	9.18	11	6.89	3	4.16
9	21	15.16	7	7.07	6	5.34	2	3.17
10	17	12.41	8	5.49	0	4.17	5	2.52
11	9	10.49	0	4.30	2	3.29	3	2.09
12	11	9.07	3	3.39	3	2.63	0	1.79
13	9	7.96	4	2.70	1	2.12	2	1.57
14	5	7.05	3	2.17	1	1.74	2	1.41
15	5	6.27	2	1.76	3	1.44	1	1.28
16	5	5.59	0	1.44	1	1.21	0	1.17
17	2	5.00	0	1.20	2	1.03	1	1.08
18	0	4.46	2	1.00	0	0.88	0	1.00
19	1	3.98	0	0.85	0	0.77	1	0.93
20	0	3.56	2	0.72	1	0.68	2	0.86
rest	32	28.01	8	6.26	11	10.74	9	11.15
	k = 1,5081 p_1 = 0,4669 p_2 = 0,1289 α = 0,8170 χ^2 = 21,95 P(χ^2) = 0,1866 DF = 17 C = 0,0172		k = 0.8207 p_1 = 0.0729 p_2 = 0.2381 α = 0.0868 χ^2 = 21.11 P(χ^2) = 0.1334 DF = 15 C = 0.0515		k = 0.7350 p_1 = 0.2289 p_2 = 0.0377 α = 0.9073 χ^2 = 15.23 P(χ^2) = 0.4350 DF = 15 C = 0.0449		k = 1.2658 p_1 = 0.4020 p_2 = 0.0795 α = 0.8369 χ^2 = 18.96 P(χ^2) = 0.2155 DF = 15 C = 0.0692	

Table 5:

	Bamana Masadennin n = 2616, m = 8 k_1 = 1680 k_2 = 535 k_3 = 231 k_4 = 100 k_5 = 50 k_6 = 10 k_7 = 9 k_8 = 1		**Bamana** Sonsanin n = 2393, m = 7 k_1 = 1515 k_2 = 575 k_3 = 159 k_4 = 89 k_5 = 43 k_6 = 11 k_7 = 1		**Bamana** Namakɔrɔba halakilen n = 1407, m = 5 k_1 = 893 k_2 = 384 k_3 = 97 k_4 = 24 k_5 = 9		**Bamana** Bamak' sigicoya n = 1138, m=6 k_1 = 695 k_2 = 255 k_3 = 126 k_4 = 43 k_5 = 18 k_6 = 1	
d	f_d	$\bar{f}_d$	f_d	$\bar{f}_d$	f_d	$\bar{f}_d$	f_d	$\bar{f}_d$
0	1227	1210.61	1086	1093.61	706	703.61	461	454.80
1	501	501.37	508	477.39	266	248.75	248	231.19
2	230	232.72	223	238.89	121	132.99	116	127.05
3	105	120.14	124	129.04	76	79.75	63	73.50
4	76	70.74	73	75.38	48	50.82	49	44.95
5	52	48.00	56	48.14	39	33.81	26	29.34
6	48	36.79	34	33.77	22	23.34	19	20.59
7	44	30.67	22	25.82	16	16.68	21	15.51
8	24	26.87	21	21.14	15	12.34	23	12.42
9	29	24.18	20	18.15	6	9.44	10	10.44
10	18	22.07	24	16.09	11	7.47	5	9.08
11	22	20.28	12	14.53	8	6.10	10	8.07
12	22	18.71	15	13.28	2	5.12	10	7.28
13	22	17.30	10	12.22	3	4.40	4	6.64
14	18	16.02	9	11.29	5	3.87	8	6.08
15	12	14.85	5	10.47	1	3.45	3	5.59
16	10	13.77	6	9.72	2	3.13	3	5.15
17	10	12.77	5	9.03	1	2.87	3	4.75
18	4	11.85	6	8.40	6	2.65	1	4.39
19	5	11.01	8	7.82	2	2.46	4	4.06
20	5	10.22	7	7.29	1	2.30	4	3.75
rest	124	137.04	112	104.53	45	46.64	41	47.36
	k = 0,8857 p_1 = 0,5571 p_2 = 0,0659 α = 0,7400 χ^2 = 30,92 P(χ^2) = 0,0204 DF = 17 C = 0,0119		k = 0.8116 p_1 = 0.4816 p_2 = 0.0595 α = 0.7910 χ^2 = 18.12 P(χ^2) = 0.3815 DF = 17 C = 0.0076		k = 0.4991 p_1 = 0.3051 p_2 = 0.0324 α = 0.8628 χ^2 = 18.78 P(χ^2) = 0.3413 DF = 17 C = 0.0134		k = 0.8969 p_1 = 0.4530 p_2 = 0.0702 α = 0.7751 χ^2 = 26.79 P(χ^2) = 0.0612 DF = 17 C = 0.0237	

Table 6:

	Vai Mu ja vaa I (T. Sherman) n = 3140, m=5 k_1 = 1893 k_2 = 1033 k_3 = 186 k_4 = 86 k_5 = 2		**Vai** Sa'bu Mu'a'… n = 495, m=4 k_1 = 281 k_2 = 189 k_3 = 21 k_4 = 4		**Vai** Vande bɛ Wu'u n = 426, m=4 k_1 = 270 k_2 = 124 k_3 = 29 k_4 = 3	
d	f_d	$\bar{f}_d$	f_d	$\bar{f}_d$	f_d	$\bar{f}_d$
0	1496	1489.93	233	219.40	176	173.50
1	670	668.30	104	120.00	124	111.26
2	369	350.00	65	58.67	44	51.85
3	183	193.79	33	28.55	20	22.35
4	101	111.50	13	14.67	13	10.46
5	55	66.59	5	8.38	10	6.15
6	43	41.53	6	5.49	4	4.64
7	39	27.28	2	4.08	2	4.07
8	20	19.02	5	3.32	5	3.77
9	15	14.13	0	2.85	0	3.53
10	11	11.14	4	2.50	1	3.28
11	6	9.25	2	2.23	4	3.04
12	13	8.00	0	1.99	6	2.78
13	6	7.12	1	1.78	2	2.53
14	4	6.47	3	1.59	2	2.28
15	3	5.96	1	1.42	0	2.05
16	8	5.54	1	1.26	0	1.83
17	7	5.18	1	1.12	2	1.63
18	8	4.86	0	1.00	0	1.44
19	3	4.57	0	0.89	0	1.27
20	3	4.31	0	0.79	0	1.12
rest	72	80.53	9	6.03	7	7.18
	k = 0,7580 p_1 = 0,4149 p_2 = 0,0439 α = 0,9094 χ^2 = 22,33 P(χ^2) = 0,1725 DF = 17 C = 0,0071		k = 1.3369 p_1 = 0.5991 p_2 = 0.1266 α = 0.8764 χ^2 = 19.12 P(χ^2) = 0.2082 DF = 15 C = 0.0392		k = 2.3902 p_1 = 0.7350 p_2 = 0.1784 α = 0.8534 χ^2 = 24.59 P(χ^2) = 0.1042 DF = 17 C = 0.0583	

4. Justification of the MNB model

In order to justify the adequacy of the MNB linguistically, we assume that the distances are controlled by two forces (parameters), say *a* and *b*, where *a* represents a constant property of language, controlled by the grammar, and *b* expresses an individual text characteristics, depending on type and style of the text and diverse author characteristics.

Since the distances can not grow infinitely, we assume the variable proportionality

$$f_d = \frac{a + bd}{d} f_{d\text{-}1} \tag{4.1}$$

between the distance frequencies f_d and $f_{d\text{-}1}$. This relation is an analogy to the relation $P_x = g(x)P_{x\text{-}1}$ for word length distributions, where x is the word length and $g(x)$ may assume several specific forms, see Wimmer/Altmann (1996, 11; 131f.). Now we can write (4.1) as

$$f_d = \frac{a/b + d}{d} b\, f_{d\text{-}1}$$

and the substitution $a/b = k - 1$, $b = 1 - p$ yields

$$f_d = \frac{k + d - 1}{d} (1\text{-}p)\, f_{d\text{-}1} \tag{4.2}$$

which differs essentially from the corresponding formula (2.6) in the random case and represents the recurrence formula for the negative binomial distribution (NB):

$$P_d = \binom{k + d - 1}{d} p^k (1 - p)^d \qquad \text{for } d = 0,1,2,\ldots \tag{4.3}$$

$$(k > 0,\ 0 < p < 1).$$

However, it was not possible to fit the simple NB to the data of section 3. Due to the facts that some distance frequencies $f_d^{(r)}$ related to the texts in section 3 could be fitted by the NB and that f_d is given as $f_d = f_d^{(1)} + \ldots + f_d^{(m)}$, one could try to fit the f_d by a convex combination of m NB´s, i.e. by a distribution

$$\sum_{i=1}^{m} \alpha_i \binom{k_i + d - 1}{d} p_i^{k_i} (1 - p_i)^d \tag{4.4}$$

with $\alpha_1 + \ldots + \alpha_m = 1$, $\alpha_i > 0$ for i = 1,…,m.

However, this is not practicable, since the distribution (4.4) has 3*m*-1 parameters, where *m* is the number of different elements in the sequence of type (2.2), i.e. the number of different word lengths in the real text.

Instead we used the MNB which is a mixture (convex combination) of only two NB´s (where the parameters of the "type *k*" are assumed to be identical), i.e. the observed distance frequencies have been fitted by the expression

$$\begin{aligned}&(n\text{-}m)\,[\alpha \binom{k+d-1}{d} p_1^k (1-p_1)^d + (1\text{-}\alpha)\binom{k+d-1}{d} p_2^k (1-p_2)^d] \\ &= (n\text{-}m)\binom{k+d-1}{d}[\alpha\ p_1^k (1-p_1)^d + (1\text{-}\alpha)\, p_2^k (1-p_2)^d], \qquad (4.5)\end{aligned}$$

where (*n*-*m*) is the sample size (see (2.3)) and the four parameters k, p_1, p_2, α satisfy $k > 0$ and $0 < \alpha$, p_1, $p_2 < 1$, respectively.

This procedure may be justified, assuming that the forming of distances in text is essentially determined by only two types of text elements in a sequence of type (2.2) which might correspond, for example, to two strata in the text, consisting of short synsemantics and longer autosemantics.

Acknowledgements

I would like to thank Gabriel Altmann, Emmerich Kelih, Sven Naumann, Anja Overbeck, Ioan-Iovitz Popescu and Andrij Rovenchak for providing me with files containing the word lengths of the analyzed texts.

References

Altmann, E.G.; Pierrehumber, J.B.; Motter, A.E. (2009). Beyond word frequency: Bursts, lulls, and scaling in the temporal distribution of words, PloS ONE, 4(11): e7678.

Corral, A.; Ferrer-i-Cancho, R.; Díaz-Guilera, A. (2009). Universal Complex Structures in Written Language, arXiv: 0901.2924v1[physics.soc-ph], 19 Jan. 2009.

Liu, H. (2007). Probability distribution of dependency distance. *Glottometrics 15*, 1-13.

Kunz, M./Rádl Z. (1998). Distribution of Distances in Information Strings. *Journal of Chemical Information and Computer Sciences* 38(3), 374-378.

Wimmer, G.; Altmann, G. (1996). The Theory of Word Length Distribution. In: Schmidt, P. (ed.), *Glottometrika 15, 112-133*, Wissenschaftlicher Verlag Trier.

Zörnig, P. (1984). The distribution of distances between like elements in a sequence, part I. *Glottometrika 6, 1-15;* part II. *Glottometrika 7, 1-14.* (= Quantitative Linguistics, Vol. 25 and 26). Brockmeyer, Bochum.

Zörnig, P. (1987). A theory of distances between like elements in a sequence. *Glottometrika 8, 1-22* (= Quantitative Linguistics, Vol. 32). Brockmeyer, Bochum.

Zörnig, P. (2010). Statistical simulation and the distribution of distances between identical elements in a random sequence. *Computational Statistics & Data Analysis*, Vol. 54, 2317-2327.

Software

Altmann-Fitter (1997). *Iterative Fitting of Probability Distributions.* Lüdenscheid, RAM-Verlag

From translational cross-analysis to ecotranslatology

Irma Sorvali, Oulu

Abstract. In the present study, an attempt is made to characterize the concept of 'ecotranslatology' with a model of cross-translation as a starting point. In cross-translation the process of translation is studied quantitatively in both directions. A sample of proper names is used in order to illustrate the model. By means of the cross-analysis, applied in interlingual and intersemiotic translation, a proposal is made to find a way from cross-translation to ecotranslatology.

Keywords: cross-analysis; ecotranslatology; quantitative studies

1. Starting point

When studying the process of translating, it is not enough to analyze translations from one language into another and from one culture into another, but an analysis of the process in both directions is needed. It is also of relevance to use authentic materials in real linguistic-cultural and other contexts. In my earlier studies a model of cross-analysis was presented. This interlingual model, where the verbal signs are translated into another language, is further developed by means of intersemiotic translation of proper names. In this type of translation the verbal signs are translated into another, nonverbal system of signs (Jakobson 1966). The intersemiotic cross-translation is a kind of introduction to ecotranslatology, seen as a field between translation studies and ecology. Here, ecological questions are focused in the process of translation. Finally, the role of translators is discussed in relation to ecotranslatology. In the present study only a simple quantitative method is used.

2. Interlingual cross-analysis

In translational cross-analysis, a number of Finnish and Swedish texts and translations, as well as authors as translators, are emphasized (Sorvali 1986, 1987). The interlingual cross-analysis deals with source language and target language in the following way: A Finnish text translated into Swedish, and a Swedish text, translated into Finnish, are compared with each other. This double translational process is supposed to result in a mirroring effect: if x in original, then y in translation, and respectively: if y in original, then x in translation.

It has often been stated that the Finnish language makes frequent use of non-finite verb phrases, which are sentence equivalences, while the Swedish language prefers complete clauses and sentences in corresponding positions. The analysis of non-finite verb phrases shows a mirroring effect with following tend-

encies: In translation from Finnish into Swedish the ratio of non-finite verb phrases is diminishing from 2.6 in Finnish to 1.0 in Swedish, whereas a growth of non-finite verb phrases from 1.0 to 2.2 is observed in the opposite direction. The way of translating non-finite verb phrases is seemingly constant, regardless of the direction of translation. The texts in the above study are written by totally ten well-known Finnish resp. Swedish authors, and there are about 17.000–18.000 lexical words in each group. (Sorvali 1987: 357–363.)

A mirroring effect is also found in a study, aiming at measuring of information in the translation process. For the purpose, the lexical words and finite verbs are used in order to measure information by the help of a model called 'inforeme'. The inforeme is defined as the smallest syntactic unit with lexical meaning, and it is counted in original texts as well as in their translations in the following way: the number of 'lexical words' minus the number of 'finite verbs' divided by the number of 'finite verbs'. In translating from Finnish into Swedish, the inforeme is varying from -0.4 to -0.1, while the variation in translating in the opposite direction is between +0.5 and 0.0. (Sorvali 1986: 58–63.) Independent on the direction in the translation process, the result can be interpreted as preserved information.

The above presented cross-analyses have been made manually, a fact that does not allow any greater amounts of material. The quality of the texts in the analyzed corpuses is, however, guaranteed by the qualified authors and translators engaged in the process. Since then, greater corpuses have been analyzed non-manually (see Olohan 2004: 36–37). A corpus of ten million words, consisting of texts written in Finnish and of translations from different languages into Finnish has been studied by Anna Mauranen (2000, 2006). As far as I can see, the translated texts in this great corpus are not necessarily equivalents of the original texts, as is the case in the cross-translation, where the translators and the translational environment are focused.

3. From intersemiotic cross-analysis to ecotranslatology

After the above interlingual cross-analyses, an intersemiotic cross-translation is now presented in order to widen the perspective. For this purpose, a corpus of product names in ecological environments is used as illustrations. This study is further aimed to throw light on the way from the intersemiotic cross-analysis to ecotranslatology. The product names are linguistically expressed, and they are used in commercial purposes; in one way or another, they refer to the idea of the product (Ainiala et al. 2008: 276–278). The product names are proper names, in general considered as universals (Steiner 1976: 97). Here, the ecological names, i.e. the verbal signs, are translated into nonverbal signs and vice versa (cf. Jakobson1966: 233). The diversification of proper names has been analyzed by Karl-Heinz Best, who presents the following regularities: "Als Ergebnis kann betrachtet werden, dass Eigennamen offenbar ungeachtet der Tatsache, dass es sich

bei ihnen um eine ganz eigene Kategorie von Zeichen handelt, den gleichen Gesetzmäßigkeiten folgen wie auch andere sprachliche Erscheinungen" (Best 2007: 27). His analysis of German family names has also resulted in the law statements, as follows: "Man kann nun feststellen, dass die Hypothese, die Rangordnung der Familiennamen unterliege einem Sprachgesetz, sich in mehreren Fällen bewährt hat" (Best 2008: 459).

In the present study, some universal features are aimed to be emphasized, even though the study relies entirely on simple quantitative analyses and does not allow any comprehensive conclusions. The focus is lying on the relation between the ecological bread name and the ecological ingredients of the bread, which are verbally and/or visually given on the bread packages. The main point is, however, that the actual entity is connected with real-life activities.

The quantitative analysis is based on a sample of 140 Finnish bread packages (collected 1990–2000). The verbally mentioned cereal ingredients are found in totally 68 per cent of the names of the packages, while the cereal ingredients are more seldom pictured on them, i.e. in 49 per cent of the cases. In this respect, the verbal signs are more usual than the pictured ones.

The advertiser wants to draw the consumer's attention to the product, and for this purpose he makes use of indexical signs (see Nöth 1990: 138, 480). The modern consumer is moreover considered to be interested in the nutritive ingredients of the bread. In the present study, the cereals named in the list of the ingredients are ecological indexes of the packages, and are therefore marked with [ecological]. Any cereal, e.g. rye, wheat, oat etc., used as an ingredient in the baking process, is marked with [RYE].

The product names are considered as signs which can be interpreted in a way, somewhat comparable with intersemiotic translation, i.e. from verbal signs into another, nonverbal system of symbols (cf. Jakobson 1966: 233; see above). When eating rye bread, the consumer is transforming the concept of [RYE] into [ecological] nutrition, or in other words: if [RYE], then [ecological], and in the reversed form: if [ecological], then [RYE].

The concepts of [RYE] and [ecological] are representing the ecological and natural product. As stated above, the ingredient [RYE] is included in verbal expressions as well as in visual pictures. [RYE] takes the role of an 'original version' of the bread. At the earliest stage, this is 'translated' into [ecological] in the baking process, and further by the producer in the process of giving a name to the product. In the opposite direction, the 'original version' of the consumer is [ecological], which s/he 'translates' into [RYE]. In marketing, this is aimed to result in buying of bread with [RYE].

As to the variation in the use of the bread names in the corpus, a kind of rotation is observable: a name is used again and again with smaller variations only, e.g. 'Rye bread', 'RyeRye', and 'Ecologically baked bread'. In this way the rotated names are examples of linguistic economy, but another question is how

they work in marketing, where the product must be identified by means of its name.

Furthermore, a modern example in everyday life is the gluten-free bread, which has increasing commercial functions (see *Commission Regulation (EC) No 41/2009*). The special gluten-free diet is verbally marked in a sample of gluten-free packages. Together with the linguistic expressions, a visual symbol of 'gluten-free' is regularly found on the packages. The Finnish product name *Luontaisesti gluteeniton tumma leipä* is translated into English ('*Naturally Gluten-Free Dark Bread*') and into a couple of other languages, while the product names of the 'normal' packages are shorter and usually expressed in Finnish only (Sorvali 2008: 464).

Due to the vital importance of bread, the relation between the names and the ingredients could be studied from a wider perspective and by means of more complicated concepts than is the case in the present study. It is worth reminding that the cereals in the above presentation are chosen at random, while any other entity, for instance the lexical features of the texts, or the visual design of packaging, could be relevant future objects. Interesting is the concept of 'hreb', introduced by Luděk Hřebíček as "aggregate" and further developed and re-named by Gabriel Altmann. Together with Arne Ziegler, Altmann analyzed the hrebs in the poem "Der Erlkönig" by Goethe, and published the results in the work *Denotative Textanalyse* (Ziegler & Altmann 2002). The method, which is later applied by Reinhard Köhler and Sven Naumann (2007: 317–329), enables semantic analyses of large adherent entities, giving possibilities to interdisciplinary studies (Ziegler & Altmann 2002: 21–28, 135).

As to the interdisciplinarity of ecotranslatology, the study of the bread names and the ingredients in cross-translation refers to the branch of ecology, which has had a great influence on other fields of study, since scholars have been aware of the worldwide ecological crisis, as pointed out by Winfried Nöth (1998: 332). The ecologically produced bread and the recycled packages also belong to the 'green' environment, and a special 'green' vocabulary is said to be on the increase (see Mühlhäusler 2001: 163). By means of theoretical, methodological and empirical linguistics, ecolinguistics opens new perspectives for the linguists interested in ecology (see Fill 2001: 43–53). Consequently, studies of ecological problems could be intensified by means of translation studies.

4. Role of translators in ecotranslatology

Ecotranslatology deals with the whole translational enviroment, where the human translators and other persons engaged in the process and ecological questions as well stand in focus. Ecotranslatological problems are illustrated by the help of a carousel diagram, originally created in order to describe the whole process of translating with special respect to the persons (Sorvali 1990: 149, 1996: 83). The carousel diagram, divided in four sectors, is now completed with ecological material, as follows: (1) the creation of the original, i.e. ecological text, by the

author(s), (2) the reception of this original by the receptors, readers and consumers, (3) the translation process, i.e. the original translated by the translator(s), and (4) the reception of the translated version by the receptors, readers and consumers. The ecological environment is included in each of the sectors. The crucial point is that the carousel goes both clockwise and counterclockwise. In practice, the analysis can be made from the original to the translation, respectively from the translation back to the original. The authors and the translators, respectively the receivers of the originals and those of the translations, can be compared to each other. The whole process of translation and the persons engaged in this processes are in this way emphasized.

The translator and the environment are said to have a central role also in Chinese ecotranslatology (see Dollerup 2010; Stolze 2011). The Chinese ecotranslatology considers the translation as a holistic translational ecosystem, where ancient Chinese philosophies are taken into account, and the relationship between the translator and the translational ecoenvironment is focused. In short, it is based on the Chinese life and culture. According to Cay Dollerup the Chinese school of ecotranslatology "has been inspired by nature, by our environments which are greater and last longer than the individual human being" (Dollerup 2010: 4). The field of the Chinese ecotranslatology, as developed by the Chinese Hu Gengshen, approaches translation as a process of the translator's adaptations and selections, and the translator's subjectivity and the translational ecoenvironment are emphasized (Liu 2011: 87). The Chinese concept of ecotranslatology is also compatible with certain western hermeneutical and phenomenological theories of translation (cf. Stolze 2011: 14). Moreover, the underlying grammar of human speech is a mapping of the world, as stated by George Steiner (1976: 98).

Ecotranslatology has been discussed especially from semiotic points of view, and a special field of ecosemiotics has also been presented. Ecosemiotics is among other things considered as "the study of the *semiotic* interrelations between organisms and their environment" (Nöth 1998: 333). Ecological approach is not alien to semiotics, as Timo Maran puts it (Maran 2007: 289). In seeking for an integrated methodology of ecosemiotics, he uses nature-text as a concept of ecosemiotics. His quadripartite model of analysis consists of textual natural environment, written text, author of the text, and reader. According to this model the reading experience of a nature essay may cause the reader to visit the depicted natural environment, or other way round. The meanings of nature need mediation by human semiotic processes. (Maran 2007: 280–282.)

To conclude, researchers in many fields have been engaged in ecological problems dealing with ecotranslatology (e.g. Dollerup 2010), and even other fields have been mentioned, such as ecosemiotics (e.g. Nöth 1998, Maran 2007) and ecolinguistics (e.g. Fill 2001). However, it must be admitted that it is not possible at the moment to give any overwhelming or at least appropriate definition(s) of ecotranslatology. Anyway, the process of translating deals with words,

languages and cultures in world-wide systems, where the ecotranslatology could give great possibilities to translators and researchers in their work for the future.

5. Summary

In the above study, an attempt has been made in order to throw light on the concept of 'ecotranslatology'. First, a model of cross-analysis is shortly presented. Inspired of this model, a simple quantitative analysis of bread names with features [RYE] and [ecological] in intersemiotic cross-translation is made. The discussion is mainly based on methodological issues, articles and other publications dealing with features on the area of translation studies and ecological thinking as well.

References

Ainiala, Terhi & Saarelma, Minna & Sjöblom, Paula (2008). *Nimistöntutkimuksen perusteet* [Basics of Onomastics]. Tietolipas 221. Helsinki: Suomalaisen Kirjallisuuden Seura.

Best, Karl-Heinz (2007). Diversifikation bei Eigennamen. In: Grzybek, Peter & Köhler, Reinhard (eds.), *Exact Methods in the Study of Language and Text.* Quantitative Linguistics 62. Berlin, New York: Mouton de Gruyter. P. 21–31.

Best, Karl-Heinz (2008). Rangordnungen deutscher Eigennamen. In: Altmann, Gabriel & Zadorozhna, Iryna & Matskulyak, Yuliya (eds.), *Problems of General, Germanic and Slavic Linguistics*. *Papers for 70-th Anniversary of Professor V. Levickij*. Chernivtsi: Books – XXI. P. 454–460.

Commission Regulation (EC) No 41/2009 = Commission Regulation (EC) No 41/2009 of 20 January 2009 concerning the composition and labelling of foodstuffs suitable for people intolerant to gluten. http://eur-lex.europa.eu [visited 24.10.2012]

Dollerup, Cay (2010). *Chinese eco-translatology in translation theory contexts*. Paper delivered at the First International Conference on Eco-translatology held in Macau, China, in 2010. P. 1–7. http://www.cay-dollerup.dk/publications.asp [visited 24.10.2012]

Fill, Alwin (2001). Ecolinguistics: State of art 1998. In: Fill, Alwin & Mühlhäusler, Peter (eds.), *The Ecolinguistics Reader*. Language, Ecology and Environment. London & New York: Continuum. P. 43–53.

Jakobson, Roman (1966). On Linguistic Aspects of Translation. In: Brower, Reuben A. (ed.), *On Translation*. New York: Oxford University Press. P. 232−239.

Köhler, Reinhard & Naumann, Sven (2007). Quantitative analysis of co-reference structures in texts. In: Grzybek, Peter & Köhler, Reinhard (eds.), *Exact Methods in the Study of Language and Text.* Quantitative Linguistics 62. Berlin, New York: Mouton de Gruyter. P.317–329.

Liu, Aihua (2011). An Eco-translatological Perspective to Translator: A Case Study of Xu Chi. In: *Theory and Practice in Language Studies*, Vol. 1, No. 1, January 2011. Academy Publisher. 87–90.

Maran, Timo (2007). Towards an integrated methodology of ecosemiotics: The concept of nature-text. *Sign Systems Studies* 35. 1/2, 2007. P. 269–294.

Mauranen, Anna (2000). Strange strings in translated language. In: Olohan, Maeve (ed.), *Intercultural Faultlines. Research Methods in Translation Studies I: Textual and Cognitive Aspects*. Manchester: St Jerome. P. 119–141.

Mauranen, Anna (2006). Genre, käännös ja korpus [Genre, translation and corpus]. In: Mäntynen, Anna & Shore, Susanna & Solin, Anna (eds.), *Genre ja tekstilaji*. Tietolipas 213. Helsinki: Suomalaisen Kirjallisuuden Seura. P. 214–239.

Mühlhäusler, Peter (2001). Babel Revisited. In: Fill, Alwin & Mühlhäusler, Peter (eds.), *The Ecolinguistics Reader*. Language, Ecology and Environment. London & New York: Continuum. P. 159–171.

Nöth, Winfried (1990). *Handbook of Semiotics*. Bloomington and Indianapolis: Indiana University Press.

Nöth, Winfried (1998). Ecosemiotics. *Sign System Studies* 26. P. 332–343.

Olohan, Maeve (2004). *Introducing Corpora in Translation Studies*. London & New York: Routledge.

Sorvali, Irma (1986). Inforeme – How to Measure Information-Content. *Babel* Vol. XXXII, No. 1/1986. P. 58–63.

Sorvali, Irma (1987). On translating non-finite verb phrases from Finnish into Swedish and vice versa. In: Lilius, Pirkko & Saari, Mirja (eds.), *Proceedings of the Sixth International Conference of Nordic and General Linguistics in Helsinki, August 18–22, 1986. The Nordic Languages and Modern Linguistics 6.* Helsinki: Helsinki University Press. P. 357–367.

Sorvali, Irma (1990). *Studier i översättningsvetenskap* [Translation Studies]. Oulu: University of Oulu. Department of Scandinavian Languages.

Sorvali, Irma (1996). *Translation Studies in a New Perspective*. Frankfurt am Main, Berlin, Bern, New York, Paris, Wien. Frankfurt am Main: Peter Lang. Europäischer Verlag der Wissenschaften.

Sorvali, Irma (2008). Analysis of Everyday Texts. In: Altmann, Gabriel & Zadorozhna, Iryna & Matskulyak, Yuliya (eds.), *Problems of General, Germanic and Slavic Linguistics*. Papers for 70-th Anniversary of Professor V. Levickij. Chernivtsi: Books – XXI. 2008. P. 461-471.

Steiner, George (1976). *After Babel. Aspects of Language and Translation*. London, Oxford, New York: Oxford University Press.

Stolze, Radegundis (2011). Eco-translatology in China. *The EST Newsletter*, No. 39, November 2011. European Society for Translation Studies. P.13-14.

Ziegler, Arne & Altmann, Gabriel (2002). *Denotative Textanalyse. Ein Textlinguistisches Arbeitsbuch.* Wien: Edition Praesens.

Subjectival position and syntactic complexity in English sentences

Fan Fengxiang, Yu Yang, Wang Hua
School of Foreign Languages, Dalian Maritime University

Abstract. This article examines the relationship between sentential subjectival position and sentential syntactic complexity using the written section of the ICE-GB as the data source. Results show that sentential syntactic complexity is a function of sentential subjectival position and can be captured with a linear regression model.

Key words: *subjectival position, syntactic complexity, synergetic linguistics, linear regression model*

1. Introduction

The sentence is a pivotal linguistic unit in the study of language. It is regarded as the largest stretch of language forming a syntactic construction (Huddleston, 1984:19), the highest-ranking unit of grammar (Quirk, 1985:47), and the key unit of syntax (Biber et al, 1999:46). Linguists deal with sentential components and their interrelationships with a variety of descriptive apparatus such as rules, categories, functions, algebra, set theory, formal logics etc within certain theoretical frames, i.e., structuralist, transformational-generative, systemic-functional, to name just a few.

Apart from the above mentioned approaches, there is another way for sentential syntactic examinations within the theoretical frame of quantitative-synergetic linguistics, which is concerned with the quantitative aspects of the sentential components and their law-like interrelations, such as inventory size, frequency, length measured in different linguistic units, complexity, position/ distances of components etc.

Take length as an example. For quantitative linguists, length of linguistic constructs is a very important measurement revealing, at least partially, the self-organising and self-regulating nature of language systems. Literature on length of linguistic constructs abounds. For example, Köhler (1982) examined the relationship between sentence length and the length of its immediate constituent, the clause, and found the dependence of mean clause length on sentence length. Altmann (1988) studied sentence length distributions and concluded that the probabilities of neighbouring length classes is a function of the probability of the preceding length classes and modelled this dependency using the 1-displaced negative hyper-Pascal distribution. More examples of research on sentence length are provided by Best (2005), who has also delved into the length distributions of the sub-components of the sentence — the morph and the syllable, in addition to sentence length (Best, 1977a, 1977b, 1998, 2011, 2012). Another researcher who

examined the relationship between length, NP distribution and NP complexity is Wang (2012) within the quantitative-synergetic theoretical frame; she used math models to describe these interrelationships.

Another prominent quantitative concept in quantitative syntactic analyses is the position of sentential components. Köhler (1999) investigated the relationship between the positions of syntactic units and their structural complexity in number of immediate constituents, and the result shows that the structural complexity is a function of the position of the corresponding unit. Fan, Grzybek and Altmann (2010) studied word length in number of syllables and the sentential positions of the corresponding words, and found that sentence-final words tend to be longer, hence more complex, than words in other positions.

In the above position-related sentential component examinations, the position and the complexity of the linguistic units examined are sub-sentential, i.e., from clause to words. It would be of interest to search for a sub-sentential component whose position has a relationship with sentential complexity.

The sentential components consist of primary syntactic classes such as noun, verb, adjective, adverb, preposition and so on; the next higher classes are the different phrases: noun phrase, verb phrase, adjective phrase, adverbial phrase, prepositional phrase etc. These higher syntactic classes have different syntactic functions: subject, predicator, object, complement etc, which are known as functional classes. Semantically, the subject is the most important element of the sentence, since it is normally the theme, the topic, around which the sentence develops. As for information processing within the sentence, it is common to process information so as to achieve a linear presentation from low to high information values, with the verb as the transitional point between a thematic low communicative dynamism and a focal high (Quirk, 1985:1357). The information the subject bears is generally regarded as given, with new information provided by other functional classes after it. This means that generally the part of the sentence after the verb would be more elaborate, hence longer, than the one before it, manifesting the principle of end weight. Since the English language is basically an SVO language, this seems to suggest that the position of the subject in the sentence would have a certain relationship with sentence length, hence sentence complexity. The following two sentences illustrate this point:

a. On a cold morning of December 4th, 1970, the year in which Peter's father set up his 20th restaurant, **Peter** *was born.*
b. On a cold morning of December 4th, 1970, the year in which Peter's father set up his 20th restaurant, **Peter** *was born in the very hospital where, 45 years before, his father came into the world, with a silver spoon in his mouth, as the saying goes.*

In normal circumstances, sentences like *a* are avoided because the pre-verbal part is relatively too long, caused by the proximity of the subject to the end of the sentence; such a long pre-verb part needs a longer post-verbal part to make the sentence well balanced, both from the point of view of sentential structure and information dynamism. So *b* is better because of the heavier post-verbal part.

However, this increases the overall length of the sentence, hence its complexity.

In view of the importance of the subject and the linearity of information processing within the sentence, this contribution intends to study the relationship between subjectival positions and sentential complexity. Sentential subjectival position refers to the position of the first subject of the entire sentence measured from the beginning of the sentence, whether it is a simple, complex or compound sentence. As to the measurement of complexity of the sentence, there are three ways to do it: a. one based on the number of its immediate constituents (Köhler, 2012), the clause in this case; b. one by counting the number of phrases; c. one by the total number of words. In this study, we use the equivalent of the second measurement, the total number of the phrasal syntactic functional elements such as subject, verbal, object, etc, instead of the total number of the corresponding phrases because the subject is a syntactic function of a phrase. Hereafter we shall refer to functions performed by a phrase, i.e. subject, determiner, object etc, as phrasal functional elements.

2. Data and methods

The data source is the ICE-GB corpus, which is a 1,000,000-word corpus of contemporary British English consisting of the spoken English section and the written English section, hereafter referred to as ICE-GBW. The former contains 300 2,000-texts totalling about 600,000 words, while the latter 200 texts totalling about 400,000. These texts are syntactically tagged. In the present study only ICE-GBW was used. Table 1 displays the tags representing the major syntactic functions and syntactic classes in ICE-GBW.

Table 1
Major syntactic function and syntactic class tags

A	adverbial	CL	clause
ADJ	adjective	CLEFTIT	cleft it
ADV	adverb	CLOP	cleft operator
AJHD	adjective phrase head	CO	object complement
AJP	adjective phrase	COAP	appositive connector
AJPO	adjective phrase postmodifier	CONJUNC	conjunction
AJPR	adjective phrase premodifier	CONNEC	connective
ART	article	COOR	coordinator
AUX	auxiliary	CS	subject complement
AVB	auxiliary verb	CT	transitive complement
AVHD	adverb phrase head	DEFUNC	detached function
AVP	adverb phrase	DISMK	discourse marker
AVPO	adverb phrase postmodifier	DISP	disparate
AVPR	adverb phrase premodifier	DT	determiner
CF	focus complement	DTCE	central determiner
CJ	conjoin	DTP	determiner phrase

DTPE	predeterminer
DTPO	determiner postmodifier
DTPR	determiner premodifier
DTPS	postdeterminer
EMPTY	empty
EXOP	existential operator
EXTHERE	existential there
FNPPO	floating NP postmodifier
FOC	focus
FRM	formulaic expression
GENF	genitive function
GENM	genitive marker
IMPOP	imperative operator
INDET	indeterminate
INTERJEC	interjection
INTOP	interrogative operator
INVOP	inverted operator
MVB	main verb
NADJ	nominal adjective
NONCL	nonclause
NOOD	notional direct object
NOSU	notional subject
NP	noun phrase
NPHD	noun phrase head
NPPO	noun phrase postmodifier
NPPR	noun phrase premodifier
NUM	numeral
OD	direct object
OI	indirect object
OP	operator
PARA	parataxis
PC	prepositional complement
PMOD	prepositional modifier
PP	prepositional phrase
PREDEL	predicate element
PREDGP	predicate group
PREP	preposition
PROD	provisional direct object
PROFM	proform
PRSU	provisional subject
PRTCL	particle
PS	stranded preposition
PU	parsing unit
PUNC	punctuation
REACT	reaction signal
SBHD	subordinator phrase head
SBMO	subordinator phrase modifier
SU	subject
SUB	subordinator
SUBP	subordinator phrase
TAGQ	tag question
UNTAG	untag
VB	verbal
VP	verb phrase

The syntactic tags for all the syntactic functions from the phrase level upwards (inclusive), including the nested ones, of each of the sentences were extracted, keeping their original positional order in the sentence, and arranged linearly in a sentential syntactic function element vector for the measurement of the sentential subjectival position and for the calculation of sentential syntactic complexity in terms of the total number of the phrasal syntactic function elements (hereafter referred to as PSFE) of the sentence. Position is measured conventionally from the beginning of the sentential syntactic function vector, with the first element being in position 1, the second 2, the third 3, etc, until the end. For sentences without a subject, the subjectival positional value is assigned 1, instead of 0. The following is a syntactically parsed sentence with tags for syntactic functions and syntactic classes, taken from file W1A-001.COR of ICE-GBW:

```
PU,CL(main,intr,pres)
 A,PP()
  P,PREP(ge) {In}
  PC,NP()
```

```
    NPPR,AJP(attru)
     AJHD,ADJ(ge) {recent}
    NPHD,N(com,plu) {years}
SU,NP()
 DT,DTP()
  DTPS,PRON(quant,plu) {several}
 NPHD,N(com,plu) {schools}
 NPPO,PP()
  P,PREP(ge) {of}
  PC,NP()
   NPHD,N(com,sing) {thought}
VB,VP(intr,pres,perf)
 OP,AUX(perf,pres) {have}
 MVB,V(intr,edp) {emerged}
PUNC,PUNC(comma) {,}
FNPPO,CL(depend,zrel,montr,pass,edp)
 SU,NP()
  NPHD,PRON(univ,sing) {each}
 VB,VP(montr,edp,pass)
  MVB,V(montr,edp) {championed}
 A,PP()
  P,PREP(ge) {by}
  PC,NP()
   NPPR,AJP(attru)
    AJHD,ADJ(ingp) {leading}
   NPHD,N(com,plu) {exponents}
   NPPO,PP()
    P,PREP(ge) {of}
    PC,NP()
     DT,DTP()
      DTCE,ART(def) {the}
     NPHD,N(com,sing) {period}
PUNC,PUNC(per) {.}
```

In this syntactically parsed sentence, there are altogether 17 phrases (represented by the syntactic class tags after the comma) excluding the initial parsing unit tag (*PU,CL(main,intr,pres)*, standing for *parsing unit, clause* placed at the beginning of every sentence, which is excluded in position measurement), performing 17 syntactic functions (represented by the syntactic function tags before the comma). They are as follows (the notes in the brackets are removed to save space):

A,PP(); PC,NP(); NPPR,AJP(); SU,NP(); DT,DTP(); NPPO,PP(); PC,NP(); VB,VP(); FNPPO,CL(); SU,NP(); VB,VP(); A,PP(); PC,NP(); NPPR,AJP(); NPPO,PP(); PC,NP(); DT,DTP().

For example, *A,PP()* stands for adverbial (syntactic function) performed by prepositional phrase (syntactic class); *PC,NP()*, prepositional complement (syntactic function), performed by noun phrase (syntactic class); *NPPR,AJP()*, noun phrase premodifier (syntactic function), performed by adjective phrase (syntactic class); *SU,NP()*, subject (syntactic function), performed by noun phrase (syntactic class), etc. The sentential subject (*SU,NP()*) is the seventh element in the sentential syntactic function element vector, so its position is 7, and the syntactic complexity of the sentence is 17.

3. Result and analysis

The total number of sentences of ICE-GBW is 27,647. Of these sentences, 7,117 are non-clause sentences, such as *IIV, section 1.4, the history of Japan, the man who escaped* etc; these non-clause sentences were excluded from the present study. The actual number of sentences examined is 20,530. These sentences consist of 381,819 PSFEs with 36 different syntactic functions. Figure 1 displays the distribution of these PSFEs.

The ten most frequent PSFEs are as follows (with their frequency and percentage): DT, 52,692, 13.8%; VB, 52,389, 13.72%; PC, 45,654, 11.96%; SU, 37,375, 9.76%; NPPO, 31,089, 8.14%; CJ, 30,518, 7.99%; NPPR, 22,439, 5.88%; OD, 21,185, 5.55%; CS, 10,456, 2.74%. The total number of these PSFEs is 351,013, accounting for 91.93%. The number of the remaining 26 PSFEs is only 30,806, accounting for 8.07%.

The mean sentential syntactic complexity is 18.5981, obtained by dividing the total number of the PSFEs, 381,819, with the total number of sentences, 20,530. The median of the sentential syntactic complexity is 17, and the mode 14. The range of the sentential syntactic complexity is between 1 and 95. However, sentences with syntactic complexity between 1 and 10 constitute the first quartile of the total number of sentences: 54.71; those with syntactic complexity between 11 and 25 constitute the inter-quartile, totaling 10,318, while those with syntactic complexity of 26—95 constitute the last quartile, totaling 4,741. Of the 20,530

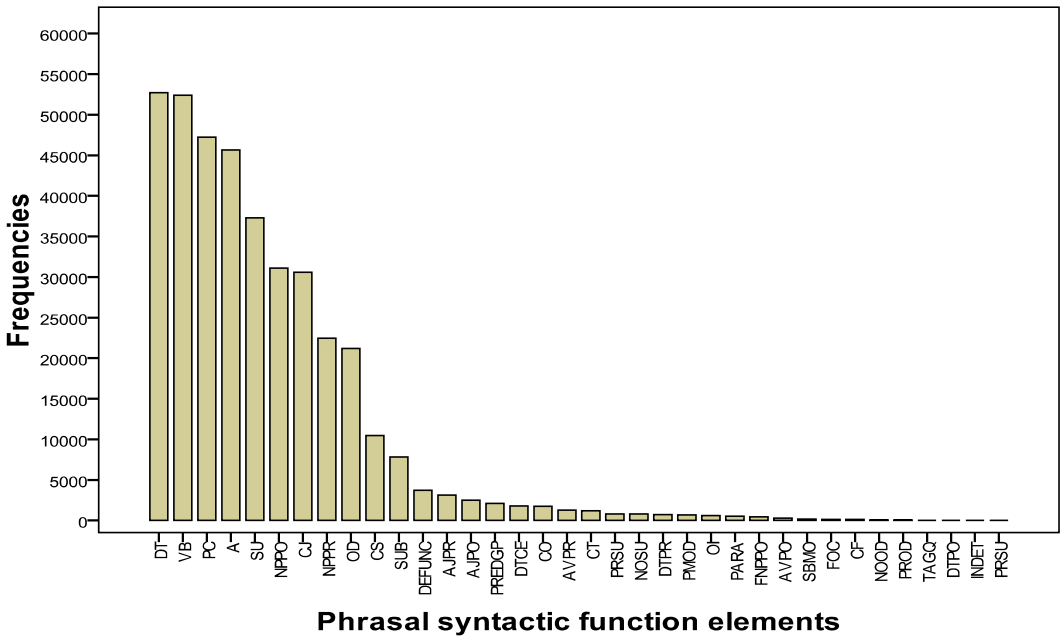


Figure 1. The distribution of PSFEs.

sentences, only 33 have the syntactic complexity of 1, and two have the complexity of 95. The following is one of the sentences with a complexity of 95:

A PC DT NPPR NPPO VB A PC NPPO NPPO NPPO SU DT NPPO PC CJ DT CJ NPPO PC DT VB CT CJ DEFUNC SU DT A PC A PC PREDGP CJ VB A PC NPPO PC DT OD SU DT A PC DT VB A PC NPPR CJ VB A PC DT NPPO PC DT OD DT NPPO PC CJ NPPO CJ NPPO PC DT CJ DEFUNC SU DT NPPO PC CJ DT CJ NPPO PC CJ CJ NPPO PC DT A PC DT NPPO PC CJ DT CJ VB A PC DT

There is a certain relationship between sentential syntactic complexity and sentential structural variation in a number of different structures. The number of sentences with different sentential structures increases along with the sentential syntactic complexity, until reaching to the peak of 803 at a complexity of 14, then begins to decrease. The sentential syntactic complexity and the corresponding sentential structural variation are shown in Table 1.

Table 1
Sentential syntactic complexity and structural variation. SC: syntactic complexity, SV: structural variation in number of sentences with different structures.

SC	SV	SC	SV	SC	SV	SC	SV
1	3	4	132	7	566	10	721
2	15	5	256	8	633	11	773
3	55	6	431	9	754	12	725

13	756	31	280	49	46	67	5
14	803	32	249	50	31	68	2
15	773	33	245	51	28	69	2
16	799	34	217	52	29	70	5
17	725	35	201	53	28	71	5
18	713	36	162	54	22	72	4
19	746	37	156	55	21	73	4
20	701	38	137	56	14	75	2
21	592	39	121	57	12	76	2
22	587	40	117	58	14	77	1
23	573	41	101	59	14	78	3
24	542	42	98	60	12	82	1
25	509	43	87	61	15	87	1
26	428	44	69	62	5	89	1
27	423	45	61	63	5	93	1
28	374	46	52	64	11	95	2
29	349	47	50	65	4		
30	355	48	55	66	12		

The following displays the structures of the 55 sentences whose syntactic complexity is 3:

1. A SU CS
2. A SU DT
3. A SU VB
4. A VB A
5. A VB OD
6. A VB SU
7. CS AVPO SU
8. CS DT DEFUNC
9. CS DT SU
10. CS SU A
11. CS SU DT
12. CS SU NPPO
13. CS SU VB
14. CS VB SU
15. DEFUNC SU VB
16. DEFUNC VB DEFUNC
17. ELE SU VB
18. OD SU VB
19. OI SU VB
20. PARA SU VB
21. SU A VB
22. SU CS A
23. SU CS AJPR
24. SU CS DEFUNC
25. SU CS DT
26. SU CS PC
27. SU DT CS
28. SU DT VB
29. SU NPPO VB
30. SU NPPR VB
31. SU VB A
32. SU VB CS
33. SU VB DEFUNC
34. SU VB INDET
35. SU VB OD
36. SU VB OI
37. SU VB PARA
38. SUB SU VB
39. VB A DEFUNC
40. VB A OD
41. VB A PC
42. VB A SU
43. VB CJ CJ
44. VB CS A
45. VB CS AJPR
46. VB CS DT
47. VB CS PC
48. VB OD A
49. VB OD CO
50. VB OD DEFUNC
51. VB OD DT
52. VB OD NPPO
53. VB OI OD
54. VB SU DT
55. VB SU NPPR

The relationship between sentential syntactic complexity and sentential structural variation can be captured by Nemcová and Serdelová's (2005) model describing

the relationship between the number of synonyms *y* of a word and the length of the word in syllables *x*:

(1) $$y = ax^b e^{cx} + 1$$

(1) is a special case of Wimmer & Altmann (2005). This relationship also holds for sentential syntactic complexity *SC* and structural variation *SV*:

(2) $$SV = aSC^b e^{cSC} + 1$$

The fit is very good, with R^2 = 0.992, a = 14.333, b = 2.459 and c = −0.176. Figure 2 is the model fit.

The 20,530 sentences have 32 different sentential subjectival positions, the rightmost, the "largest" being 46. Generally, as the sentential subjectival position increases, the number of the sentences with the corresponding subjectival position decreases. The number of sentences whose sentential subject is in position 1 is predominantly large, totalling 12,795, accounting for 62%. The number of the sentences whose sentential subjectival position is larger than 10 is only 318, accounting for only 1.55%. Table 2 is the sentential subjectival position and the number of sentences with such subjectival position.

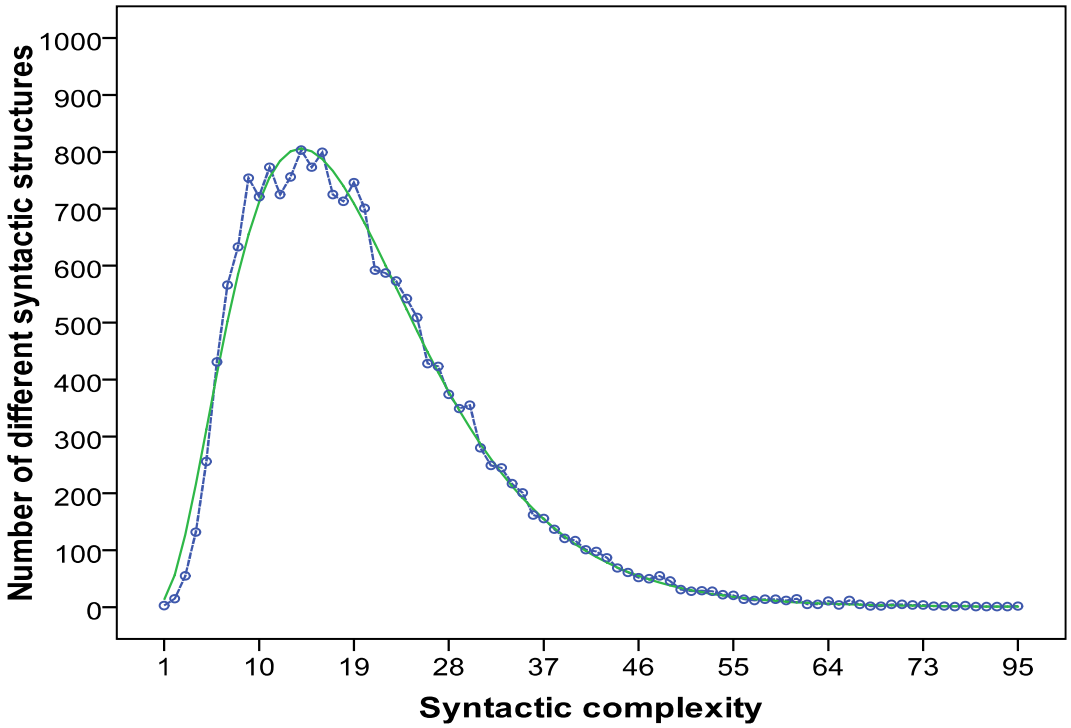


Figure 2. The relationship between syntactic complexity and structural variations in number of different structures. The solid line: the model fit, the small circles: the empirical values.

Table 2

Sentential subjectival position and the number of sentences with such subjectival position. SP: sentential subject position, NS: number of sentences

SP	NS	SP	NS	SP	NS	SP	NS
1	12795	9	127	17	8	25	2
2	3177	10	98	18	12	26	2
3	1772	11	66	19	12	27	3
4	969	12	56	20	9	30	1
5	518	13	48	21	2	31	2
6	332	14	30	22	6	34	1
7	249	15	24	23	2	44	1
8	175	16	28	24	2	46	1

The sentential syntactic complexity of the sentences in relation to their sentential subjectival position was examined. Figure 3 shows the distribution of sentential complexity with sentential subjectival position from 1 to 10.

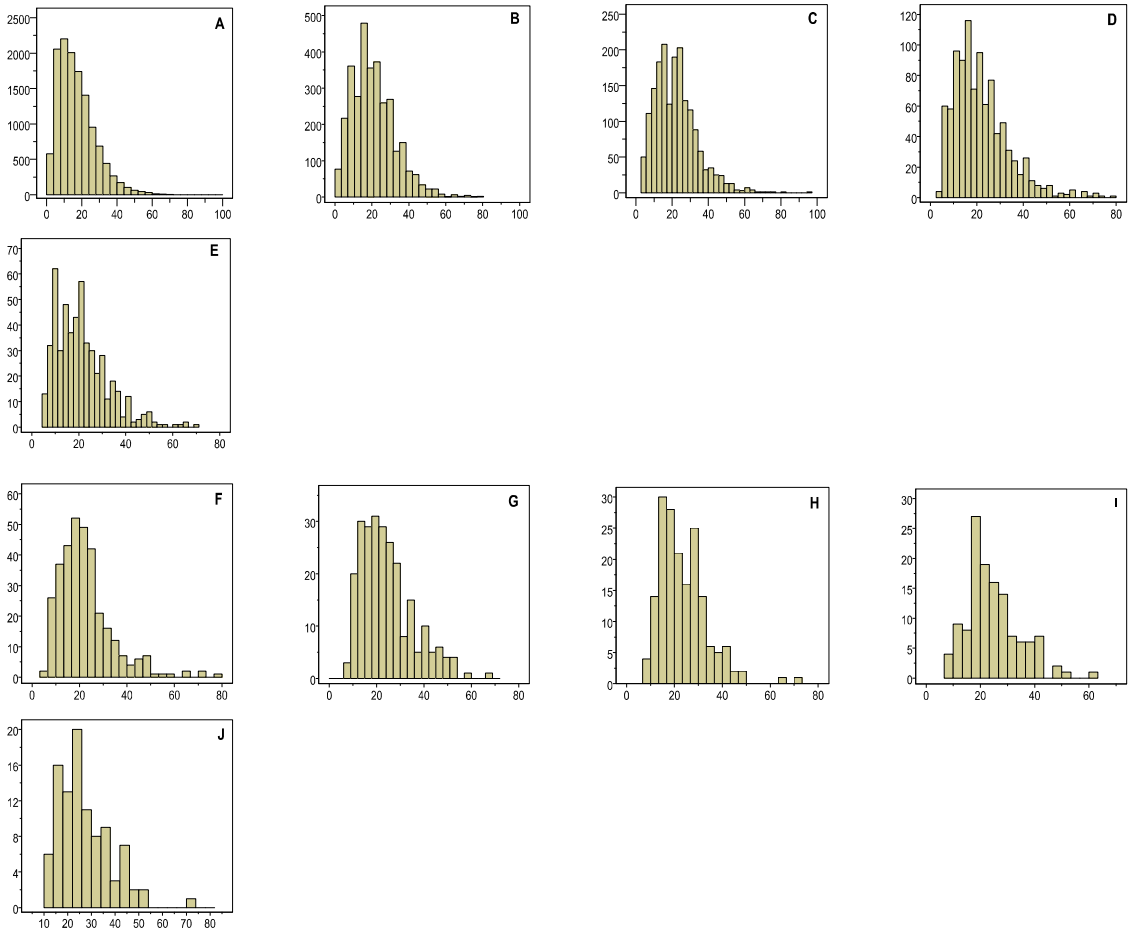


Figure 3. The distribution of sentential syntactic complexity with sentential subjectival position from 1 (panel A) to 10 (panel J). The *X* axis: sentential syntactic complexity in number of PSFEs, the *Y* axis: number of sentences with such complexity.

The general distribution patterns of the sentential syntactic complexity with sentential subjectival positions from 1 to 10 are similar, which are skewed, tailing out to the right. The mean sentential syntactic complexity of the sentences with the same sentential subjectival position was computed in order to study the relationship between sentential syntactic complexity and sentential subjectival position. The result is in Table 3.

Table 3
Sentential subjectival position and the mean sentential syntactic complexity.
SP: sentential subjectival position,
MSSC: mean sentential syntactic complexity in number of PSFEs

SP	MSSC	SP	MSSC	SP	MSSC
1	16.5782	12	26.2857	23	34.5000
2	20.6308	13	26.8542	24	41.0000
3	22.0418	14	30.4000	25	48.5000
4	21.5697	15	31.2083	26	42.5000
5	21.2857	16	27.7500	27	36.0000
6	22.4247	17	29.5000	30	38.0000
7	23.8153	18	34.7500	31	37.0000
8	23.7086	19	31.5000	34	50.0000
9	24.6614	20	38.2222	44	51.0000
10	26.7551	21	32.5000	46	59.0000
11	24.7879	22	36.6667		

The mean sentential syntactic complexity (MSSC) is a function of the sentential subjectival position (SP). This relationship is linear and can be expressed with the following linear regression model:

(3) $MSSC = \alpha + \beta SP$

The fit is good, with $R^2 = 0.910$, $\alpha = 17.546$, $\beta = 0.835$. Figure 4 displays the model fit.

4. Conclusion

The PSFEs in ICE-GBW perform 36 different sentential syntactic functions and the ten most frequently used are (in order of descending frequency): DT, VB, PC, SU, NPPO, CJ, NPPR, OD, CS, accounting for 91.93% of the total. Despite the importance of the sentential subject, it ranks only the fourth. This is possibly because of the structural complexity of the subject itself—its substructures formed with other PSFEs, such as DT, VB, PC, etc.

The relationship between sentential syntactic complexity and sentential structural variation in number of sentences with different structures is bell-shaped,

which can be described with Nemcová and Serdelová's synonyms and word length model.

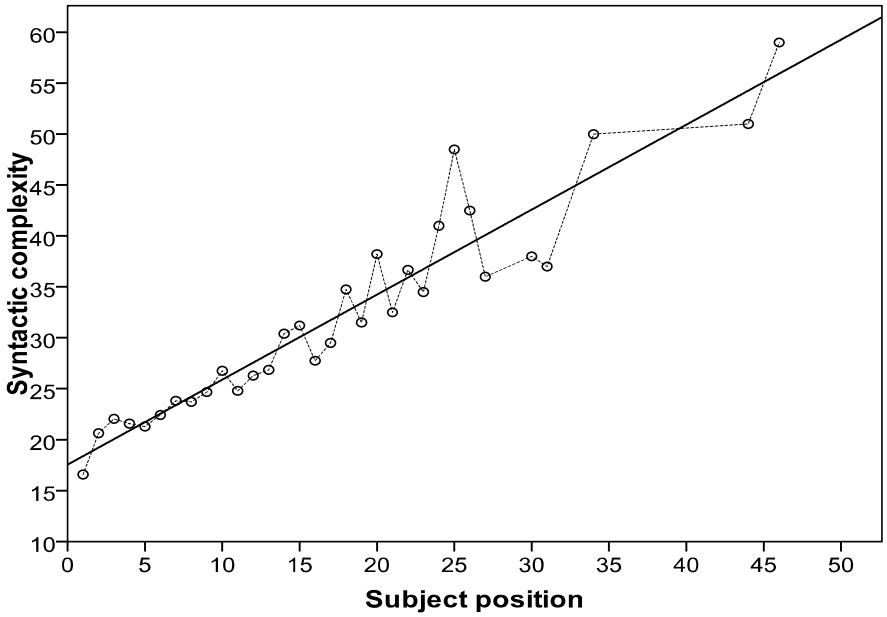


Figure 4. The linear regression model for the relationship between the subjectival position and sentential syntactic complexity. The small circles: the observed sentential syntactic complexity.

The sentential subjects appear in 46 different positions in the sentence, but the predominant position is sentence initial. Generally, the sentential subjectival position is an indicator of sentential syntactic complexity; the larger the sentential subjectival position, the more syntactically complex the sentence. This relationship can be captured with a linear regression model. This phenomenon, apart from rhetorical and stylistic reasons, is due to the principle of end weight and communication dynamism in the sentence.

References

Altmann, G. (1988). *Verteilungen der Satzlängen.* In: Schulz, K.-P. (ed.), *Glottometrika 9: 147-169*. Bochum: Brockmeyer.

Best, K.-H. (1997). Zum Stand der Untersuchungen zu Wort- und Satzlägen. In: *Third International Conference on Quantitative Linguistics.* Helsinki, 172–176.

Best, K.-H. (1997). Zur Wortartenhäufigkeit in Texten deutscher Kurzprosa der Gegenwart. In: Best, Karl-Heinz (ed.), *Glottometrika 16.* Trier: Wiss. Verlag Trier, 276–285.

Best, K.-H. (1998). Results and perspectives of the Göttingen Project on Quan-

titative Linguistics. *Journal of Quantitative Linguistics 5(3), 155-162.*

Best, K.-H. (2005). Satzläge. In: Köhler, R.; Altmann, G.; Piotrowski, G. (eds.), *Quantitative Linguistics. An International Handbook.* Berlin, New York: de Gruyter, 298–304.

Best, K.-H. (2011). Silben-, Wort- und Morphlängen bei Lichtenberg. *Glottometrics 24,* 1-13.

Best, K.-H. (2012). How many words are in a verse? An exploration. In: Naumann, S.; Grzybek, P.; Vulanović, R.; Altmann, G. (eds.). *Synergetic Linguistics, Text and Language as Dynamic Systems*. Praesens Verlag: Wien. 13-22

Biber, D., Johansson, S., Leech, G., Conrad, S. & Finegan, E. (1999). *Longman Grammar of Spoken and Written English*. Pearson Education Limited: Harlow.

Fan, F., Grzybek, P., Altmann, G. (2010). Dynamics of word length in sentence. *Glottometrics 20, 70-109.*

Huddleston, R. (1984). *Introduction to the Grammar of English*. Cambridge University Press: Cambridge 1984

Köhler R. (1982) Das Menzerathsche Gesetz auf Satzebene. In: Lehfeldt, Werner; Strauss, Udo (eds.), *Glottometrika 4.* Bochum: Brockmeyer, 103–113.

Köhler R. (1999).Syntactic structures: properties and interrelations. *Journal of Quantitative Linguistics*, 6(1). 46–57.

Köhler R. (2012). *Quantitative syntax analysis*. Berlin/Boston: de Gruyter.

Nemcová, E, & Serdelová, K. (2005). On synonymy of Slovak. In: Altmann, G, Levickij, V & Perebyinis, V. (eds.), *Problems of Quantitative Linguistics*: *194-209*. Chernivtsi: Ruta.

Quirk, R., Greenbaum, S., Leech, G. & Svartvik, J. (1985). *A comprehensive grammar of the English language*. Longman Group Limited: New York.

Wang, H. (2012). Length and complexity of NPs in Written English. *Glottometrics 24, 79-87.*

Wimmer, G. & Altmann, G. (2005). Towards a unified derivation of some linguistic laws. In: Grzybek, P. (ed.). *Contributions to the science of language: Word length and related issues*: *93-117*. Boston: Kluver.

Efficiency of Word Order in Flexible Parts-of-Speech Systems

Relja Vulanović, Kent State University at Stark

Abstract. A simplified version of the previously developed formula for calculating relative grammar efficiency is applied to flexible parts-of-speech systems. Only fixed word order is modeled, enabling the simplification of the formula, as well as a more detailed analysis. All theoretically possible parts-of-speech systems are considered, regardless of whether they are attested or not. It is shown that the most efficient systems are mainly unattested.

1. INTRODUCTION

In this paper, a formal mathematical approach to relative grammar efficiency is presented. It is applied to parts-of-speech (PoS) systems in the sense of Hengeveld (1992) and Hengeveld, Rijkhoff, and Siewierska (2004). In a PoS system, four propositional functions are considered, viz. the obligatory heads of predicate and referential phrases and their optional modifiers. What word classes are used to fulfill the existing propositional functions is what defines a PoS system. Based on the PoS systems found in natural languages, Hengeveld (1992) proposes seven main PoS system types. Six additional intermediate types are introduced in Hengeveld et al. (2004). Only main PoS systems are considered here, but their number is extended to all theoretically possible ones, regardless of whether these systems are attested or not. Of primary interest are flexible PoS systems, i.e., those that have fewer word classes than propositional functions. This is because such systems show various degrees of efficiency, as opposed to the rigid PoS systems which have the same measure of efficiency (in a rigid PoS system, the number of word classes is equal to the number of propositional functions).

As for the way relative grammar efficiency is measured, the paper follows the approach developed in Vulanović (2003, 2007), where grammar efficiency is calculated using an equation which is based on a formal way of analyzing sentences. These two papers are not concerned with PoS systems, but the same approach has been applied to PoS systems in Vulanović (2008b, 2009, 2012) and Vulanović and Miller (2010). The present paper is similar to Vulanović (2008b) and Vulanović and Miller (2010) in the sense of discussing theoretically possible, unattested PoS systems alongside the attested ones, but the novelty here is the focus on fixed word order. On the one hand, this enables a simplification of the relative-efficiency formula, and, on the other hand, more details regarding the possible orders of propositional functions can be provided. The interest in fixed word order is also what

distinguishes the present paper from Vulanović (2009, 2012), in addition to the fact that these two papers only consider PoS systems found in the Hengeveld et al. (2004) linguistic sample.

One of the results presented here is that it is mainly the unattested PoS systems which are the most efficient ones. This confirms the findings in Vulanović (2008b) and can be partly explained by the need of natural languages for a certain amount of redundancy (greater redundancy corresponds to smaller efficiency).

Hengeveld's word classes and PoS systems are described in section 2. Then, in section 3, a gradual derivation of the grammar-efficiency formula is presented. The formula is applied to flexible PoS systems in section 4 and the results are discussed there. Finally, section 5 contains some concluding remarks.

2. WORD CLASSES AND PoS SYSTEMS

In Hengeveld's (1992) approach, word classes are defined according to what propositional functions they can fulfill. The four propositional functions (or syntactic slots) considered are: P – the head of the predicate phrase, R – the head of the referential (noun) phrase, r – the modifier of the referential-phrase head, and p – the modifier of the predicate-phrase head. Whereas *P* and *R* are obligatory, r or p, or both, may be missing in some languages. For instance, Tagalog has no slot for the modifier of the predicate phrase (Hengeveld et al. 2004). In the present paper, PoS systems with 4 or 3 propositional functions are mainly considered. The word classes they use are defined in Tables 1 and 2. The tables also introduce the notation which is used for different word classes. The label X in Table 1 is used generically – it does not stand for the same word class in each row in which it occurs; note also that these word classes are left unnamed. Whereas the only propositional function verbs can have is the head of the predicate phrase, other word classes are defined by their predominant usage and may have functions different from those indicated in the tables. Adverbs other than manner adverbs are not considered because they typically modify the whole sentence and not just the head of the predicate phrase.

Word classes can be divided into two groups, rigid and flexible. A rigid word class is specialized for one and only one propositional function and a flexible word class can have two or more different functions. Therefore, verbs, nouns, adjectives, and manner adverbs are the only rigid word classes and all other word classes are flexible.

Table 1
Word classes in the presence of 4 syntactic slots

Word class	P	R	r	p
Verbs	V	-	-	-
Nouns	-	N	-	-
Adjectives	-	-	a	-
Manner adverbs	-	-	-	m
Heads	H	H	-	-
Predicatives	$\mathcal{P}$	-	-	$\mathcal{P}$
Nominals	-	$\mathcal{N}$	$\mathcal{N}$	-
Modifiers	-	-	M	M
-	X	-	X	-
-	-	X	-	X
Non-verbs	-	Λ	Λ	Λ
Non-nouns	Z	-	Z	Z
-	X	X	X	-
-	X	X	-	X
Contentives	C	C	C	C

Table 2
Word classes in the presence of 3 syntactic slots
(x is either r or p)

Word class	P	R	x
Verbs	V	-	-
Nouns	-	N	-
Adjectives (if x = r)	-	-	a
Manner adverbs (if x = p)	-	-	m
Heads	H	H	-
Non-verbs	-	Λ	Λ
Non-nouns	Z	-	Z
Contentives	C	C	C

The above word classes are combined to create different PoS systems. Hengeveld (1992) identifies the 7 PoS system types presented in Table 3; see (Hengeveld et al. 2004) as well. The table shows what word classes are used in each PoS system type. Types 1–3 are flexible and 4–7 are rigid. A PoS system is rigid if all of

its word classes are rigid, while a flexible PoS system has at least one flexible word class.

Table 3
PoS systems types in Hengeveld (1992)

Type	P	R	r	p
1	C	C	C	C
2	V	Λ	Λ	Λ
3	V	N	M	M
4	V	N	a	m
5	V	N	a	-
6	V	N	-	-
7	V	-	-	-

The PoS system types in Table 3 represent ideal structures that natural languages only approximate (Rijkhoff 2007: 718). For instance, strictly speaking, type 7 is not attested, but there are languages, like Tuscarora, which come close to this type (Hengeveld 1992, Hengeveld et al. 2004). The PoS system of Tuscarora can be viewed as intermediate between types 6 and 7, having a small and closed class of nouns. Therefore, it can be said that type 7 is attested indirectly, through this intermediate PoS system type. Moreover, 5 other intermediate types are considered in Hengeveld et al. (2004). Each intermediate type falls in between two consecutive main types. Any rigid intermediate type, like Tuscarora's, is characterized by the presence of one small, closed word class. On the other hand, in a flexible intermediate type, some propositional functions can be fulfilled by two word classes. This feature is not shared by the main types or rigid intermediate types.

The way word classes can be combined to form a PoS system can be even more intricate, as a detailed description of sample languages in Hengeveld and van Lier (2010b) reveals. The focus of the present paper is on the main PoS system types and the intermediate ones are not analyzed. From this point on, when a PoS system is mentioned, it is understood that it is a main system, i.e., all its word classes are large and open and no existing propositional function is fulfilled by more than one word class.

The classification in Table 3 does not cover all possible PoS system types. Six other types are reported in Hengeveld and van Lier (2010a) as attested. All theoretically possible PoS system types, 51 of them in all, are listed there with the purpose of finding constraints that eliminate most of the unattested types. Vulanović (2008b) also considers all theoretically possible types and calculates and compares their grammar-efficiency values. This is done under the assumption that P and R are

obligatory syntactic slots, except in the case when P is the only existing slot. Because of this assumption, the number of possible PoS systems is reduced from 51 to only 28. The 28 PoS systems are presented in Table 4.

Table 4
All theoretically possible PoS systems
(*k* = number of word classes, *n* = number of propositional functions,
+ = attested system, ? = indirectly attested system, * = unattested system)

k.n type	Attestation	P	R	r	p
1.4	+	C	C	C	C
2.4	+	V	Λ	Λ	Λ
	*	Z	N	Z	Z
	*	X	X	a	X
	*	X	X	X	m
	*	H	H	M	M
	?	$\mathcal{P}$	$\mathcal{N}$	$\mathcal{N}$	$\mathcal{P}$
	*	X_1	X_2	X_1	X_2
3.4	*	H	H	a	m
	*	$\mathcal{P}$	N	a	$\mathcal{P}$
	+	V	$\mathcal{N}$	$\mathcal{N}$	m
	*	X	N	X	m
	*	V	X	a	X
	+	V	N	M	M
4.4	+	V	N	a	m
1.3	+	C	C	C	-
	*	C	C	-	C
2.3	*	H	H	a	-
	*	H	H	-	m
	?	V	Λ	Λ	-
	*	Z	N	-	Z
	*	Z	N	Z	-
	*	V	Λ	-	Λ
3.3	+	V	N	a	-
	+	V	N	-	m
1.2	?	C	C	-	-
2.2	+	V	N	-	-
1.1	?	V	-	-	-

The attestation status of a PoS system in Table 4 is based on the findings in Hengeveld et al. (2004) and Hengeveld and van Lier (2010a). The PoS systems with a question mark are attested only indirectly, i.e., as subsystems or components of other attested systems, such as attested intermediate PoS systems. There are 13 directly or indirectly attested systems in the table and they are the exact same ones as in Hengeveld and van Lier (2010a). Any system using word classes H, Z, or X is unattested.

A PoS system is going to be referred to by listing its word classes as they fulfill the propositional functions ordered as PRrp. Thus, there are CCCC, VΛΛΛ,..., VNam, CCCØ, CCØC,..., VNØm, HHØØ, VNØØ, and VØØØ systems. The following languages are given in Hengeveld et al. (2004) or Hengeveld and van Lier (2010a) as examples of the directly or indirectly attested PoS systems: Kharia (CCCC), Warao (VΛΛΛ), Kayardild (𝒫𝒩𝒩𝒫), Hungarian (V𝒩𝒩m), Dutch (VNMM), Goergian (VNam), Tagalog (CCCØ), Nhanda (VΛΛØ), Pipil (VNaØ), Garo (VNØm), Nivkh (CCØØ), Krongo (VNØØ), and Tuscarora (VØØØ).

3. RELATIVE GRAMMAR EFFICIENCY

A string of word classes used in the PoS system represents a sentence if this string can be interpreted as at least one of the following 18 orders of propositional functions:

(1a) PR, RP,

(1b) PRr, PrR, RrP, rRP,

(1c) PpR, pPR, RPp, RpP,

(1d) PpRr, PprR, pPRr, pPrR, RrPp, RrpP, rRPp, rRpP.

It is assumed for simplicity that both predicate and referential phrases are continuous and this is how the above 18 strings are formed.

Word order is not a defining characteristic of a PoS system. Any PoS system can be combined with different word orders, which are described here by the corresponding orders of propositional functions. A PoS system, together with the permitted fixed word order, constitutes a grammar in which sentences are interpreted according to the rules of the system, i.e., according to what *k* word classes are used to fulfill what *n* propositional functions. Examples of four grammars are presented in Table 5. Three of them belong to the same PoS system type and only differ by the

permitted fixed order of propositional functions. The examples are used in this section to derive the grammar-efficiency formula step by step. This derivation is similar to the one used in Vulanović (2007). Although grammar complexity, not efficiency, is discussed in Vulanović (2007),[1] the difference is inessential since the measures of grammar complexity and efficiency are taken to be the reciprocals of one another. The approach taken here introduces some simplifications of the formula since only fixed word orders are of interest in this paper.

Similar to Vulanović (2007), a formula for measuring absolute grammar efficiency, denoted by *AE*, is derived first. It is assumed that each sentence, as a string of word classes, is analyzed from left to right, one word-class symbol at a time.[2] The result of the analysis has to be at least one of the orders in (1); otherwise, the string of word classes is not a sentence. For instance MMVN is not a sentence in grammar G_1 since this string cannot be interpreted as any of the orders in (1). The only sentences permitted by G_1 are NV, MNV, NVM, and MNVM. All these sentences are unambiguous, since each can be interpreted in only one way: NV as RP, MNV as rRP, NVM as RPp, and MNVM as rRPp. Grammar G_2, on the other hand, permits only three sentences: NV, NMMV, and NMV, which is ambiguous since it can be analyzed as both RrP and RpP.

Table 5
Grammars used to illustrate how *AE* is calculated

Grammar	PoS system	Order of propositional functions
G_1	VNMM	rRPp
G_2	VNMM	RrpP
G_3	HHMM	rRPp
G_4	VNMM	rRpP

Ambiguity has to be considered when grammar efficiency is discussed (Frazier 1985: 135, Hawkins 2004: 38). In general, greater amount of ambiguity should imply smaller efficiency, all other things being equal. If the order of propositional functions is fixed and no sentence is ambiguous, the number of sentences permitted by a grammar is easy to determine. It is 4 in PoS systems with four propositional functions ($n = 4$), each sentence being interpreted as one string in each (1a), (1b), (1c), and (1d). If $n = 3$, one sentence results in one string from (1a) and another in one string from either (1b) or (1c), so the total number of unambiguous sentences is

[1] (Vulanović 2003) is a more formal, mathematical predecessor of (Vulanović 2007).
[2] There is no intention here to suggest that this is how the human mind works.

2. There is just one sentence, either PR or RP, if $n = 2$, and finally, $n = 1$ is a special case in which V (interpreted as P) is the only sentence. These counts of unambiguous sentences are reduced if the grammar permits some ambiguous sentences. Since such a grammar should be less efficient, the following principle should be satisfied:

(2) *AE* is directly proportional to the total number of unambiguous sentences, denoted by *US*.

An immediate consequence of principle (2) is that grammars only permitting ambiguous sentences have efficiency equal to 0, which is an indication that they cannot be used for successful communication. This, however, will never be the case with the grammars considered in this paper. Because of the fixed order of propositional functions, the sentence which conveys all existing propositional functions cannot be ambiguous, regardless of the word classes used in the system.

Consider now the grammar denoted as G_3. This grammar also admits four unambiguous sentences: HH, MHH, HHM, and MHHM. Judging solely by this, it follows that G_1 and G_3 are equally efficient grammars. However, G_3 has one word class less to fulfill the same four propositional functions. In order to explore the possibility that G_1 and G_3 may have different efficiency because of their different structures, the approach of regulated rewriting (Dassow and Păun 1989) is taken. Each word-class symbol is rewritten as any propositional function it can fulfill, keeping the continuity of both predicate and referential phrases (1) in mind. The information about the permitted orders of propositional functions is only used *after* all possible sentence-analyses are obtained. This information serves as a regulation that may eliminate some of the analyses. Thus, in G_3, each sentence has two possible interpretations:

(3) HH → PR | RP MHH → rRP | pPR HHM → PRr | RPp MHHM → rRPp | pPRr

However, no sentence is ambiguous because the assumed rRPp order eliminates one of the analyses of each sentence.

In general, some analyses may be just attempted and may be impossible to complete successfully. This happens when some of the required propositional functions cannot be identified. However, each sentence has at least one successful analysis. Let the number of all possible analysis attempts (whether they are successful or not) be denoted by *AA*. In G_3, $AA = 8$, as can be seen in (3). This total is smaller in G_1, where $AA = 6$:

NV → RP MNV → rRP | p- NVM → RPp MNVM → rRPp | p-

Here, the second and fourth sentences have one unsuccessful analysis attempt, indicated by the hyphen. The analysis is abandoned when it is realized that the initial M cannot be interpreted as p since the next propositional function is not P.

Since smaller values of *AA* indicate simpler analyses of sentences, the following principle should apply:

(4) *AE* is inversely proportional to the total number of analysis attempts, *AA*.

The simplest formula for defining *AE* according to principles (2) and (4) is given in (5),

$$(5) \quad AE = \frac{US}{AA}.$$

This formula becomes the definition of *AE*. The corresponding formula in Vulanović (2007) is more complicated because it is intended for measuring complexity of grammars with any order of propositional functions, whereas the present interest is just in fixed orders. For instance, the total of all analysis attempts in Vulanović (2007) involves *all permutations* of each sentence admitted.

According to (5), the absolute efficiency of G_1 is 4/6 = 2/3 = 0.667 and that of G_3 is 4/8 = 1/2 = 0.5. G_1 is more efficient than G_3 because of the smaller *AA* count. Instead of calculating the *AA* values, it would be simpler to compare the number of word classes in the two grammars, but the resulting measure of absolute grammar efficiency would not discover any difference between G_1 and G_4, for instance. However, the formula proposed in (5) differentiates between G_1 and G_4 because G_4 has *AA* = 7:

NV → RP NMV → RrP | RpP MNV → rRP | p- MNMV → rRpP | p-

Therefore, *AE* = 4/7 = 0.571 for G_4 and this grammar is less efficient than G_1.

To complete the discussion of absolute grammar efficiency, consider G_2. The sentences permitted by this grammar have the following analysis attempts:

NV → RP NMV → RrP | RpP NMMV → RrpP | Rp-

Here, *AA* = 5, but there are only two unambiguous sentences and therefore *AE* = 2/5 = 0.4, which is the smallest measure of all four grammars considered.

However, the results for G_1 and G_3 should be revisited because these two grammars belong to two different classes, viz. Γ(3.4) and Γ(2.4) respectively, where Γ(*k*.*n*) contains all grammars with *k* word classes and *n* propositional functions, as

well as with fixed orders of propositional functions. It turns out that both 2/3 and 1/2 are the greatest possible values of *AE* in the respective grammar classes. Therefore, G_1 and G_3 are optimal in their respective classes and they should be equally efficient in this relative sense. This can be achieved by dividing the absolute efficiency of the grammar by the greatest possible *AE* value of all the grammars in the same class. The rescaled value represents a relative measure of efficiency. In general, the relative efficiency, *RE*, of a grammar G_0 in $\Gamma(k.n)$ can be measured as

$$(6) \quad RE(G_0) = \frac{AE(G_0)}{\max_{G \in \Gamma(k.n)} AE(G)}, \quad G_0 \in \Gamma(k.n),$$

where *AE*(*G*) indicates the absolute efficiency of grammar *G*, as calculated in (5). Thus, $RE = 1$ for both G_1 and G_3.

It should be pointed out that there is something relative about the measure in (5) as well. This can be illustrated by considering rigid PoS systems. If types 4.4 and 2.2, for instance, are compared, it can be argued that type 4.4 is more complex (and therefore less efficient) because it has more word classes. However, the greater number of word classes enables more propositional functions and the formula in (5) gives $AE = 1$ for both grammars. This is so because each sentence in any rigid PoS system only has one analysis attempt (which is successful) and therefore $AA = US$. In other words, the value of *AE* also indicates how close the PoS system is to the one-to-one correspondence[3] between word classes and propositional functions. In rigid PoS systems, where $k = n$, this one-to-one correspondence exists and the value of $AE = 1$. On the other hand, in flexible PoS systems, $k < n$ and $AE > 1$. If *k* is less, *AE* is greater, cf. G_1 and G_3 above.

The preceding paragraph also shows that $RE = 1$ for any rigid PoS system. This is why it is only interesting to find *RE* values of flexible PoS system types. Of those, 1.4 and 1.2 trivially have $RE = 1$ because these types contain no subtypes and because each order of propositional functions produces the same value of *AE*. Type 1.3 has two subtypes, but they are equivalent; formally speaking, it does not matter if the only propositional function present is r or p. These subtypes may have some different linguistic features, but those are not represented in the model and do not influence the calculations. This is why $RE = 1$ for both subtypes of type 1.3. Therefore, only types 2.4, 3.4, and 2.3 are left for discussion.

[3] This is the One-Meaning–One-Form principle of (Miestamo 2008).

4. RELATIVE EFFICIENCY OF PoS SYSTEM TYPES 2.4, 3.4, AND 2.3

The following list presents the results of calculations, using formula (6), for all subtypes of PoS system types 2.4, 3.4, and 2.3, as well as for all possible fixed orders of propositional functions.

List 1. *RE* values (in decreasing order) for all subtypes of PoS system types 2.4, 3.4, and 2.3, and for all fixed orders of propositional functions

RE = 1

2.4: HHMM (pPRr, rRPp), *AE* = 1/2

3.4: HHam (pPRr, pPrR, rRPp, rRpP), VNMM (PpRr, pPRr, RrPp, rRPp), VXaX (rRPp, rRpP), XNXm (pPRr, pPrR), *AE* = 2/3

2.3: HHaØ (rRP), HHØm (pPR), *AE* = 2/3

$RE = \frac{6}{7} = 0.857$

3.4: VNMM (pPrR, rRpP), VXaX (PprR, pPrR), XNXm (RrpP, rRpP)

$RE = \frac{4}{5} = 0.8$

2.4: HHMM (pPrR, rRpP)

$RE = \frac{3}{4} = 0.75$

3.4: V$\mathfrak{N}\mathfrak{N}$m (all 8 orders), $\mathcal{P}$Na$\mathcal{P}$ (all 8 orders), HHam (PpRr, PprR, RrPp, RrpP), VXaX (PpRr, pPRr, RrPp, RrpP), XNXm (PpRr, PprR, RrPp, rRPp)

2.3: HHaØ (PRr, PrR, RrP), HHØm (PpR, RPp, RpP), VΛΛØ (all 4 orders), ZNØZ (all 4 orders), ZNZØ (all 4 orders), VΛØΛ (all 4 orders)

$RE = \frac{8}{11} = 0.727$

2.4: $X_1X_2X_1X_2$ (PprR, pPrR, RrpP, rRpP)

$RE = \frac{2}{3} = 0.667$

2.4: HHMM (PpRr, RrPp), $X_1X_2X_1X_2$ (PpRr, pPRr, RrPp, rRPp)

$RE = \frac{3}{5} = 0.6$

3.4: VNMM (PprR, RrpP)†

$RE = \frac{4}{7} = 0.571$

2.4: $\mathcal{PNNP}$ (all 8 orders), XXaX (rRPp, rRpP), XXXm (pPRr, pPrR)

$RE = \frac{8}{15} = 0.533$

2.4: VΛΛΛ (pPRr, pPrR, RrPp, rRPp), ZNZZ (PpRr, pPRr, rRPp, rRpP)

$RE = \frac{4}{9} = 0.444$

2.4: XXaX (RrPp, RrpP), XXXm (PpRr, PprR)

$RE = \frac{8}{19} = 0.421$

2.4: XXaX (PprR, pPrR), XXXm (RrpP, rRpP)

$RE = \frac{2}{5} = 0.4$

2.4: HHMM (PprR, RrpP)†, XXaX (PpRr, pPRr), XXXm (RrPp, rRPp)

$RE = \frac{4}{11} = 0.364$

2.4: VΛΛΛ (PpRr, PprR, RrpP, rRpP)†, ZNZZ (PprR, pPrR, RrPp, RrpP)†

The list contains 128 grammars in all: 56 of type 2.4 (7 subtypes combined with 8 possible orders of propositional functions), 48 of type 3.4 (6 subtypes, 8 orders of propositional functions), and 24 of type 2.4 (6 subtypes, each having 4 possible orders of propositional functions). For each calculated *RE* value, the PoS system type is listed first, followed by the subtype and parenthesized orders of propositional functions. *AE* values are given for the maximally efficient grammars, which are listed under *RE* = 1. Using these maximal *AE* values and the presented *RE* values, it is possible to find the measure of absolute efficiency for any of the grammars, $AE = RE \cdot \max AE$.

Ambiguity is present in some grammars of type 2.4 and in the VNMM subtype of type 3.4; such grammars are marked by a dagger. The number of unambiguous sentences in these grammars is reduced down to 2, from the regular number of 4 that can be achieved with other orders of propositional functions. This is why these grammars can be found at the bottom of the list, with the exception of VNMM with

orders PprR or RrpP, for which the *RE* value is the middle one. However, this is the smallest efficiency of all 3.4 grammars.

There is a considerable symmetry in the values and structures presented. Many grammars preserve the same *RE* when the strings describing the order of propositional functions are transformed by interchanging P and R, and, at the same time, p and r. The HHam subtype is an example of this. In some other cases, this transformation has to be accompanied by the corresponding change of the subtype. For instance, *RE* values are invariable when all of the following interchanges take place: VΛΛΛ ↔ ZNZZ, P ↔ R, and p ↔ r. In this sense, it can be said that the PoS systems VΛΛΛ and ZNZZ are mathematically or theoretically equivalent,[4] both having a specialized word class for one head and another word class for the three remaining propositional functions. There are six more equivalent pairs of this kind: XXaX and XXXm, V𝔑𝔑m and 𝔓Na𝔓, VXaX and XNXm, HHaØ and HHØm, VΛΛØ and ZNØZ, and finally, VΛØΛ and ZNZØ. Table 4 contains two more such pairs: CCCØ and CCØC (which have already been discussed at the end of section 3) and VNaØ and VNØm.

There are 13 values of *RE* in all, spanning from 4/11 to 1. Grammars of type 2.4 are spread over this whole interval, whereas the span is much narrower for the other two types: from 3/5 to 1 in the case of type 3.4 and from 3/4 to 1 for type 2.3. In fact, the bottom six *RE* values belong to type 2.4 grammars and all but two grammars with the bottom nine values are of the 2.4 type. That the 2.4 type has the widest span is not surprising since it also has the greatest difference between *k* and *n*. The distribution is presented graphically in Figure 1, where the percentage of grammar types 2.4, 3.4, and 2.3 is given for each value of *RE*. The mean *RE* value is 0.689, with the standard deviation of 0.182.

[4] It should not be forgotten that in any mathematical model, features outside the model are not of interest.

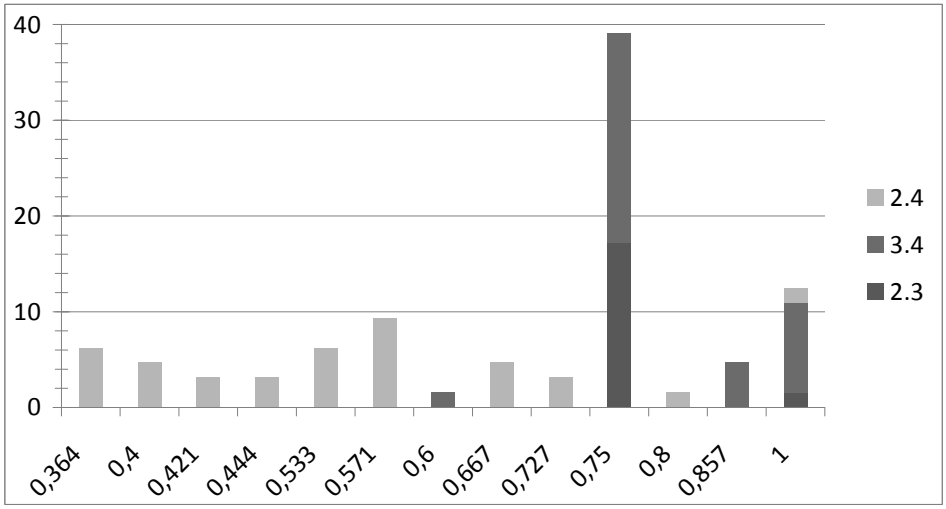


Figure 1. Percentage of grammar types for each value of relative efficiency

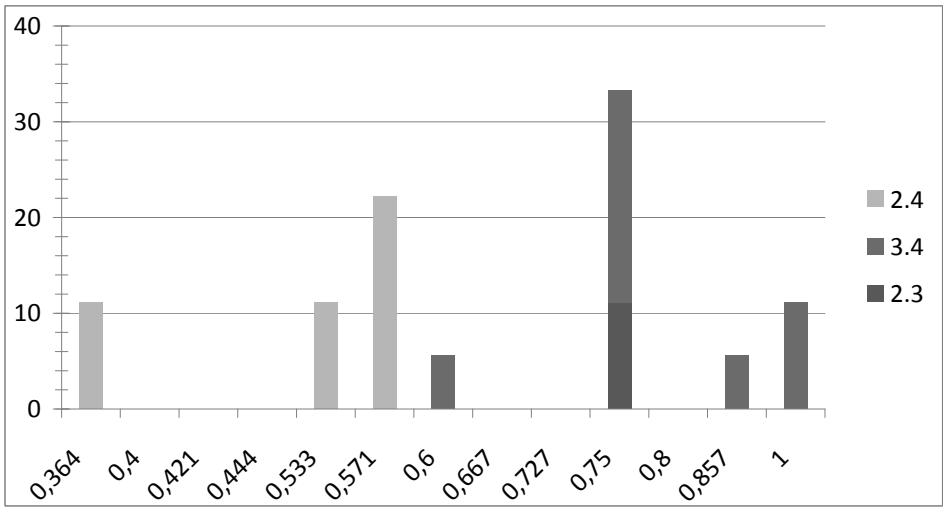


Figure 2. Percentage of grammars of attested subtypes for each value of relative efficiency

The only attested subtypes in List 1, whether they are attested directly or indirectly, are VΛΛΛ, 𝒫𝒩𝒩𝒫, V𝒩𝒩m, VNMM, and VΛΛØ. There are 36 grammars that belong to these five subtypes. Their distribution regarding the *RE* values is presented in Figure 2. The mean and standard deviation remain close to those for all 128 grammars: 0.669 and 0.179, respectively.

In an agreement with the findings in Vulanović (2008b), the greatest *RE* values mainly belong to unattested PoS systems. The presence of the word class of heads (H) is noticeable in the list of grammars with *RE* = 1. All PoS systems using H are

unattested, but each type considered has a grammar with H and maximal relative efficiency. There must be some other reason, not the tendency to maximize grammar efficiency, why word classes H, Z, and X are absent from attested PoS systems. The VNMM subtype is the only attested system with the maximal relative efficiency, and the only one within the span of *RE* values from 0.8 to 1.

The values of grammar efficiency and redundancy are complementary (Vulanović 1991, 1993). Therefore, greater grammar efficency corresponds to smaller redundancy. This is one possible reason why attested PoS systems are less efficient – they contain a greater amount of redundancy, which is necessary for successful communication.

That natural languages do not strive to maximize their efficiency can be illustrated by comparing the findings of this section to the relevant languages from the 50-language sample in (Hengeveld et al. 2004). The sample contains 9 languges[5] of types 2.4 and 3.4 and none of type 2.3.

The PoS systems of Mundari, Hurrian, Imbabura Quechua, Warao, and Turkish can be described as subtype VΛΛΛ of type 2.4. In the case of Mundari and Turkish, this is only partly so because these languages have intermediate PoS systems – Mundari is between types 1.4 and 2.4 and Turkish between 2.4 and 3.4. The basic word order in Warao corresponds to the RrpP order of propositional functions. Without additional markers, this structure would permit ambiguous sentences, but Warao has a postposition *tane* to mark p. Formally speaking, Λ + *tane* can be treated as m, thus, when the marker is taken into account, the PoS system of Warao is V𝔑𝔑m. This system is neither maximally efficient, nor does it permit ambiguous sentences. The other four languages do not have markers to resolve ambiguity in spite of the ambiguity-permitting rRpP order they use. Of course, languages may use other disambiguation means, beyond syntax and morphology, such as prosodic, semantic, pragmatic, or visual means (Hengeveld et al. 2004). Imbabura Quechua is the only language of the five, which permits a variation of its basic word order. The variation is pPrR and is accompanied with a referential-phrase marker, quite unnecessarily, since this order does not permit ambiguous sentences.[6] Because of the marker, this part of Imbabura Quechua can be considered a VNMM (pPrR) system, which has a relatively high, but not the highest, efficiency.

The 3.4 type languages in the Hengeveld et al. (2004) sample are closer to the theoretical ideal. They are Turkish (again partly), Ket, Miao, Tidore, and Lango (as an intermediate language between types 3.4 and 4.4). The corresponding PoS system subtype is VNMM. Turkish and Ket share the rRpP order, which does not create

[5] Ngiti has to be excluded since its structure is not modeled here – its predicate phrase does not have to be continuous.

[6] It is observed in Vulanović *2008a) that natural languages often do not use markers in the most efficient way.

ambiguity, but is not the most efficient one either. The remaining three languages, however, all have the RrPp basic order, which means that their structures are maximally efficient. In spite of this, Lango has a marker, which is redundant, to indicate both r and p, thus the marker does not change the VNMM subtype of Lango. In conclusion, only two languages (Miao and Tidore) of the nine have optimal structures – maximal efficiency and no redundant markers.

5. CONCLUSION

According to Miestamo (2008), the overall language complexity, which can be referred to as *global complexity*, is impossible to measure using our current, limited linguistic and mathematical tools. We can only analyze *local complexity*, i.e., the complexity of a restricted local area of grammar. Since the measures of grammar efficiency and complexity are defined as the reciprocals of each other, this means that only local grammar efficiency can be considered. This kind of grammar efficiency is the topic of the preceding sections.

The local area of grammar dealt with consists of simple intransitive sentences which describe the essence of parts-of-speech systems in Henegveld's sense. The sentences only carry the information about up to four propositional functions that are of interest in Hengeveld's approach. All theoretically possible parts-of-speech systems are considered and they are combined with all possible fixed word orders. There are 128 combinations in all.

Parts-of-speech systems and their corresponding sentences are represented by a formal grammatical model which is suitable for measuring relative efficiency using the formula developed in Vulanović (2003, 2007). However, because of the focus on fixed word orders, a simpler formula is derived and applied.

Grammar efficiency of parts-of-speech systems has been analyzed in Vulanović (2008b, 2009) and Vulanović and Miller (2010), but not in such detail. In these papers, the only fixed word orders of interest are those which produce the smallest possible grammar-efficiency values. Here, efficiency is calculated for all 128 grammars and 13 different values are obtained. It is shown that, in this local area of grammar, attested grammars are typically less efficient than what is theoretically possible. Natural languages generally do not have the tendency to maximize grammar efficiency. This is confirmed by the 9 relevant languages from the Hengeveld et al. (2004) sample – only 2 of them have optimal structures in the sense of the model considered here.

References

Dassow, J., Păun, G. (1989). *Regulated Rewriting in Formal Language Theory*. New York: Springer.

Frazier, L. (1985). Syntactic complexity. In: Dowty, D. R., Karttunen, L., and Zwicky, A. M. (eds.), *Natural Language Parsing: Psychological, Computational, and Theoretical Perspectives*. Cambridge: Cambridge University Press, 129-189.

Harrison, M.A. (1978). *Introduction to Formal Language Theory***.** Reading, Massachusetts: Addison-Wesley.

Hawkins, J.A. (2004). *Efficiency and Complexity in Grammars*. Oxford/New York: Oxford University Press.

Hengeveld, K. (1992). Parts of speech. In: Fortescue, M., Harder, P., Kristoffersen, L. (eds.), *Layered Structure and Reference in Functional Perspective*. Amsterdam/Philadelphia: John Benjamins, 29-55.

Hengeveld, K., Rijkhoff, J., Siewierska, A. (2004). Parts-of-speech systems and word order. *Journal of Linguistics 40, 527–570.*

Hengeveld, K., Lier, E. van (2010a). An implicational map of parts of speech. *Linguistic Discovery 8, 129-156.*

Hengeveld, K., Lier, E.van (2010b). Parts of speech and dependent clauses in functional discourse grammar. In: Ansaldo, U., Don, J., Pfau, R. (eds.), *Parts of Speech: Empirical and Theoretical Advances.* Amsterdam/Philadelphia: John Benjamins, 253-285.

Miestamo, M. (2008). Grammatical complexity in a cross-linguistic perspective. In Miestamo, M., Sinnemäki, K., Karlsson, F. (eds.), *Language Complexity: Typology, Contact, Change.* Studies in Language Companion Series. Amsterdam: Benjamins, 23-41.

Rijkhoff, J. (2007). Word classes. *Language and Linguistics Compass 1, 709–726.*

Vulanović, R. (1991). On measuring grammar efficiency and redundancy. *Linguistic Analysis 21, 201-211.*

Vulanović, R. (1993). Word order and grammar efficiency. *Theoretical Linguistics 19, 201-222.*

Vulanović, R. (2003). Grammar efficiency and complexity. *Grammars 6, 127–144.*

Vulanović, R. (2007). On measuring language complexity as relative to the conveyed linguistic information. *SKY Journal of Linguistics 20, 399-427.*

Vulanović, R. (2008a). The combinatorics of word order in flexible parts-of-speech systems. *Glottotheory 1, 74-84.*

Vulanović, R. (2008b). A mathematical analysis of parts-of-speech systems. *Glottometrics 17, 51-65.*

Vulanović, R. (2009). Efficiency of flexible parts-of-speech systems. In Köhler, R. (ed.), *Issues in Quantitative Linguistics.* Studies in Quantitative Linguistics 5. Lüdenscheid: RAM-Verlag, 136-157.

Vulanović, R., Miller, B. (2010). Grammar efficiency of parts-of-speech systems. *Glottotheory 3/2, 65–80.*

Vulanović, R. (2012). Efficiency of grammatical markers in flexible parts-of-speech systems. In: Naumann, S., Grzybek, P., Vulanović, R., and Altmann, G. (eds.), *Synergetic Linguistics: Text and Language as Dynamic Systems*. Wien: Praesens, 241-256.

The relationship of dependency relations and parts-of-speech in Hungarian

Veronika Vincze, Szeged

1. Introduction

The distribution of parts-of-speech and morphological features has been investigated from a quantitative viewpoint in several languages (Tuzzi, Popescu and Altmann 2010, Best 1994, Ziegler 1998, 2001, Vulanović and Köhler 2009). Furthermore, frequencies of syntactic relations have also been analyzed within the framework of quantitative syntax (Liu 2007, 2009, Čech, Pajas and Mačutek 2010, Köhler 1999, 2012).

Strongly agglutinative languages abound in morphemes which are responsible for encoding syntactic relations (among other grammatical functions) and as such, they constitute a rich soil for morphological and syntactic investigations. For instance, Köhler (2012) investigated the distribution of grammatical roles in one part the Hungarian Szeged Treebank and Väyrynen, Noponen and Seppänen (2008) analyze semantic roles in Finnish.

In this work, we investigate the relationship between dependency relations and parts-of-speech in Hungarian. We make use of the Szeged Dependency Treebank, which was manually POS-tagged and annotated for dependency relations.

2. The Szeged Dependency Treebank

We utilize the Szeged Dependency Treebank (Vincze et al., 2010) as the basis of our experiments. It contains 82,000 sentences, 1.5 million words and 230,000 punctuation marks from six domains (student compositions, computer texts, literature, legal texts, newspaper texts and short business news). The corpus contains morphological (MSD) and syntactic (dependency) annotation too, both of which were carried out manually.

Table 1
Basic statistical data on the Szeged Dependency Treebank

	com-position	computer	litera-ture	law	news-paper	newsml	**total**
Sentences	24,720	9,627	18,558	9,278	10,210	9,574	81,967
Punctuation marks	59,419	31,241	47,990	33,515	32,880	25,712	230,757
Words	283,591	183,562	189,751	225,207	190,406	201,527	1,504,801

The main parts-of-speech are as follows:

N – noun
V – verb
A – adjective
R – adverb
C – conjunction
T – article
P – pronoun
M – numeral
S – postposition
I – interjection
Y – abbreviation
O – open class (e.g. identifiers, web addresses etc.)
Z – misspelled words
X – unknown

Possible dependency relations are as follows:

APPEND – relation between the root and non-integral parts of sentences (e.g. addressing terms)
ATT – relation between noun and adjective, postposition and noun, noun/nominal modifier and noun
AUX – relation between verb and auxiliary
CONJ – relation between the conjunction and the first member of the coordination/subordination
COORD – relation between the first and second member of coordination (if there is no conjunction) or between the conjunction and the second member of coordination
DAT – relation between a dative noun and its parent
DET – relation between noun and determiner
FROM – relation between an adverb or postpositional phrase answering for the question „from where?" and its parent
INF – relation between infinitive and verb
LOCY – relation between an adverb or postpositional phrase answering for the question „where?" and its parent
MODE – relation between other adverbs or postpositional phrases and its parent
NE – relation between members of multiword named entities (e.g. *Coca Cola Ltd.)*
NEG – relation between negative adverbs and the verb
NUM – relation between members of multi-token numbers (e.g. *3 million*)
OBJ – relation between verb and object
OBL – relation between verb and its other nominal argument
PRED – relation between verb and nominal predicate

PREVERB – relation between verb and preverb (verbal prefix)
PUNCT – relation between punctuation mark and its parent
QUE – relation between question word and verb
ROOT – main element of the sentence
SUBJ – relation between verb and subject
TFROM – relation between an adverb or postpositional phrase answering for the question „from when?" and its parent
TLOCY – relation between an adverb or postpositional phrase answering for the question „when?" and its parent
TO – relation between an adverb or postpositional phrase answering for the question „where to?" and its parent
TTO – relation between an adverb or postpositional phrase answering for the question „till when or by when?" and its parent

3. Dependency relations compared to parts-of-speech tags

We examined the frequency distributions of dependency relations compared to the parts-of-speech in the corpus. Data are shown in Table 2.

We carried out an ANOVA test and found that the data is not significant ($p = 0.0685$). However, disregarding some obvious annotation errors, there are several interesting tendencies that can be observed in the data. First, we discuss cases where there is one-to-one correspondence between the dependency relation and the part-of-speech, then we turn to dependency relations that can be valid for all (or the main) parts-of-speech. We also pay special attention to nominal and adverbial relations.

3.1. Part-of-speech-specific relations

Among the relations, there are some which (almost) exclusively occur together with a given part-of-speech. Such verbal relations are AUX (auxiliary) and INF (infinitive): in the MSD coding system, auxiliaries and infinitives are subtypes of verbs, thus, these relations mark subtypes of verbs. For instance, the two words in the phrase *jött volna* '(he) would have come' are attached by the relation AUX since *volna* is an auxiliary or in *kell mennie* '(he) must go', *mennie* is an infinitive thus it bears the relation INF. The relation PREVERB connects preverbs and verbs, thus it is a relation paired with adverbs (preverbs are a subtype of adverbs in the MSD coding system) as in *ment el* '(he) went away'. The relation CONJ denotes conjunctions while the relation DET denotes determiners. There are two relations that connect members of multiword entities: NUM connects members of multi-token numbers (e.g. *3 million*), hence it is related to numbers and NE connects members of multiword named entities, thus it is related to (proper) nouns (e.g. *Coca Cola Ltd.*).

Table 2
Dependency relations compared to part-of-speech tags in the Szeged Dependency Treebank

	Adjective	Conjunction	Interjection	Numeral	Noun	Pronoun	Adverb	Postposition	Article	Verb
APPEND	608	762	248	121	4643	91	404	294	5	1017
ATT	127448	94	133	33264	89034	18418	953	53	26	45451
AUX	0	0	0	0	0	0	0	0	0	1049
CONJ	0	95459	1	0	0	1	151	0	0	0
COORD	5944	350	67	687	27834	955	796	404	2	32136
DAT	967	1	0	51	4945	925	1276	6	0	0
DET	0	0	2	6	1	3026	4	0	162866	0
FROM	0	0	0	0	4	5	525	417	0	0
INF	1	0	0	0	2	0	2	4	0	17988
LOCY	30	0	0	2	19	3	5313	1503	0	2
MODE	11864	42	815	1116	145	988	51337	12407	1	54
NE	0	0	0	49	24941	0	0	0	0	0
NEG	0	13	12	0	0	0	17461	1	0	0
NUM	0	0	0	3339	0	0	0	0	0	0
OBJ	333	0	1	650	45559	12042	15	0	2	10
OBL	2186	0	0	1975	102093	10819	4059	127	0	8
PRED	13913	4	55	424	8703	2161	128	5	2	7
PREVERB	3	0	1	0	3	2	18112	16	0	0
QUE	2	0	34	0	0	16	757	0	0	0
ROOT	311	140	440	297	2720	238	484	53	0	67951
SUBJ	640	1	3	621	70019	16946	113	0	20	13
TFROM	4	0	0	3	185	0	394	374	0	0
TLOCY	300	9	0	31	200	12	26279	4861	0	0
TO	4	0	0	0	9	6	1699	947	0	0
TTO	3	0	0	2	17	63	1700	21	0	0

There are annotation errors in the data, unfortunately: either the dependency relation is mislabeled (e.g. nouns cannot function as preverbs but there are three instances of nouns labeled as PREVERB) or the MSD code was wrongly selected (this can be the case with adverbs labeled as CONJ since in Hungarian there are quite a few words that can act as conjunctions and adverbs, depending on the context (e.g. *viszont* 'but/however') and they were probably mistagged, still, the correct dependency label was chosen for them.

3.2. General relations

There are three relations that can be applied for all parts-of-speech, namely, the attributive (ATT), the coordination (COORD), and the appenditive (APPEND) relations. Table 3 shows these data.

Table 3
General dependency relations compared to part-of-speech tags
in the Szeged Dependency Treebank

	A	C	I	M	N	P	R	S	T	V
APPEND	608	762	248	121	4643	91	404	294	5	1017
ATT	127448	94	133	33264	89034	18418	953	53	26	45451
COORD	5944	350	67	687	27834	955	796	404	2	32136

We applied an ANOVA test for these data and found that the results are significant ($p = 0.0375$). It can be also observed from the data that it is mostly nouns that function as extra elements in the sentence (APPEND) but conjunctions and interjection also often introduce appenditive parts of the sentence. On the other hand, it is mostly pairs of nouns and verbs (clauses) that are coordinated.

For the main parts-of-speech, we examined the distribution of four dependency relations, beside the three above mentioned ones, ROOT (the main element of the sentences) was also investigated. Table 4 shows the frequency distributions.

Table 4
General dependency relations compared to main part-of-speech tags
in the Szeged Dependency Treebank

	Adjective	Numeral	Noun	Pronoun	Adverb	Verb
APPEND	608	121	4643	91	404	1017
ATT	127448	33264	89034	18418	953	45451
COORD	5944	687	27834	955	796	32136
ROOT	311	297	2720	238	484	67951

The ANOVA test indicates that results are significant ($p = 0.0252$).

The attributive relation is primarily connected to adjectives and numerals as they modify the following noun. In the case of nouns, the attributive relation mostly denotes a possessive relation, that is, the possessor is in an ATT relation with the possessed noun as in *a kutya farka* the dog tail-3SGPOSS 'the dog's tail', there is an ATT relation between *dog* and *tail*.

In dependency grammars, verbs play the central role of the clause. The main element of each sentence is marked with ROOT. In sentences with more than one clauses, the verb in the coordinated clause is marked with COORD, while the verb in a subordinated clause is marked with ATT. 41% of the verbs bear the ROOT relation, 27% the ATT relation and 19% of the COORD relation, which reflects that the subordination of clauses is more frequent in the corpus than their coordination.

3.3. Nominal relations

We investigated the relations that are connected to nominal case suffixes. SUBJ denotes subjects (nominals in the nominative), OBJ denotes objects (nominal in the accusative), DAT denotes dative and OBL denotes all the other nominal cases (e.g. illative, ablative etc.). Nominal case suffixes can be attached to nouns, adjectives, numerals and pronouns in Hungarian (these parts-of-speech are commonly called nominals), thus we investigate the frequency distribution of these relations in connection with nominals.

We ranked the relations as occurred in connection with the given part-of-speech. The results can be seen in Table 5.

Table 5
Ranked nominal dependency relations compared to nominals
in the Szeged Dependency Treebank

	Adjective	Numeral	Noun	Pronoun
DAT	2	4	4	4
OBJ	4	2	3	2
OBL	1	1	1	3
SUBJ	3	3	2	1

We calculated Kendall's coefficient for these relations and found that there are differences among the distributions ($W = 0.425$, results are not significant). Except for pronouns, the most frequent relation is OBL, which can be explained by the fact that it is a comprehensive relation involving many different case suffixes. As for pronouns, they mostly occur as subjects, which is probably due to personal pronouns.

3.4. Adverbial relations

Among the dependency relations, there were some that were not only related to the syntactic role of the given word but a deeper semantic level also influenced the labeling. Such relations denote adverbial modifiers in the sentence, which can manifest either as postpositions or adverbs. Locative and temporal modifiers were classified according to the tridirectionality typical of Hungarian adverbs and case suffixes: where, from where and to where (or when, from what time and till what time) the action has taken place. For instance, *onnan* 'from that point' denotes the source of the action, *ott* 'there' denotes the location of the action and *oda* 'to that point' denotes the endpoint of the action. Thus, there are six dependency relations dedicated to these aspects and the other adverbials that cannot be described in this way are grouped under the relation MODE. Table 6 shows the ranking of these dependency relations according to the part-of-speech of the word.

We calculated Kendall's coefficient for these relations, too, and found that although there are differences between the distributions, the data are concordant ($W = 0.8929$, results are not significant).

In this way, it can be seen that among the three directions, it is most typically the locative one which is expressed in the language (i.e. where and when): TLOCY being ranked as second and LOCY as third. The least frequent direction is the source (from where and from when) whereas the goal-oriented direction is situated in the middle. The fact that this distribution is typical in Hungarian is also shown by the frequency distribution of nominal case suffixes: they can also be grouped according to tridirectionality, and they exhibit the same tendency, i.e. locative suffixes are the most frequent, followed by goal-oriented suffixes and finally source-oriented suffixes.

Table 6
Adverbial dependency relations compared to adverbs and postpositions in the Szeged Dependency Treebank

	Adverb	Postposition
FROM	6	5
LOCY	3	3
MODE	1	1
TFROM	7	6
TLOCY	2	2
TO	5	4
TTO	4	7

4. Conclusions

In this paper, we investigated the relationship between dependency relations and parts-of-speech in Hungarian on the basis of corpus data. We found that there are dependency relations that are characteristic of one (or a few) part(s) of speech, on the other hand, there are more general relations that can belong to more parts-of-speech. Concerning the tridirectionality of Hungarian adverbs and case suffixes, it was revealed that they primarily encode locative relations and source-oriented directions are the least frequent ones.

In the future, it would be interesting to examine the above tendencies in other types of texts or maybe in other languages which also encode tridirectionality.

References

Best, Karl-Heinz (1994). Word class frequencies in contemporary German short prose texts. *Journal of Quantitative Linguistics 1, 144–147.*

Čech, Radek; Pajas, Petr; Mačutek, Ján (2010). Full valency. Verb valency without distinguishing complements and adjuncts. *Journal of Quantitative Linguistics 17(4), 291–302.*

Köhler, Reinhard (1999). *Syntactic Structures. Properties and Interrelations*. *Journal of Quantitative Linguistics 6(1), 46–57.*

Köhler, Reinhard (2012). *Quantitative Syntax Analysis*. Berlin, New York: de Gruyter.

Liu, Haitao (2007). Probability distribution of dependency distance. *Glottometrics 15, 1–12*.

Liu, Haitao (2009). Probability distribution of dependencies based on Chinese dependency treebank. *Journal of Quantitative Linguistics 16(3), 256–273.*

Tuzzi, Arjuna; Popescu, Ioan-Iovitz; Altmann, Gabriel (2010). *Quantitative analysis of Italian texts*. Lüdenscheid: RAM.

Väyrynen, Pertti Alvar; Noponen, Kai; Seppänen, Tapio (2008). Preliminaries to Finnish word prediction. *Glottotheory 1, 65–73*.

Vincze, Veronika; Szauter, Dóra; Almási, Attila; Móra, György; Alexin, Zoltán; Csirik, János (2010). Hungarian Dependency Treebank. In *Proceedings of the Seventh Conference on International Language Resources and Evaluation (LREC'10).*

Vulanović, Relja; Köhler, Reinhard (2009). Word order, marking, and Parts-of-Speech Systems. *Journal of Quantitative Linguistics 16(4), 289–306.*

Ziegler, Arne (1998). Word class frequencies in Brazilian-Portuguese press texts. *Journal of Quantitative Linguistics 5, 269–280*.

Ziegler, Arne (2001). Word class frequencies in Portuguese press texts. In: Uhlířová, Ludmila; Wimmer, Gejza; Altmann, Gabriel; Köhler, Reinhard (eds.): *Text as a linguistic paradigm: levels, constituents, constructs. Fest-*

schrift in honour of Luděk Hřebíček: 295–312. Trier: Wissenschaftlicher Verlag Trier.

Statistical analysis of perception adjectives ‘soft’ – ‘hard’ in English

Olexandra Nuzban, Sergej Kantemir

Abstract. The article focuses on the syntagmatic characteristics of antonymous perception adjectives ‘soft’ – ‘hard’. The application of statistical methods enabled us by means of formal criteria to objectively reveal similarities and dissimilarities of syntagmatic peculiarities of polysemous perception adjectives. The research yielded the following findings:

1. The value of the χ^2 criterion (which indicates the presence or absence of a relation) is not necessarily directly proportional to the indices of the frequency of occurrence.
2. Wide discrepancies are found between the significant standard relations of the dominant lexemes “**soft**” and “**hard**”. Statistically, they are characterized by a contrasting distribution.

Keywords: *perception adjectives, syntagmatic relations, polysemy, dominant lexemes, semantic combinability, statistical methods, correlation analysis.*

1. Introduction

Recent decades have seen a steady growth of interest in the phenomenon of perception (sensory) vocabulary in different languages sparked by the works of Rakhilina (2000), Laenko (2005), Bons (2008), Tribushinina (2008), Shindo (2009), and others. Among various parts of speech under study, perception adjectives move centre stage and hold one of the leading positions. This can be accounted for the ability of adjectives to reflect the speaker’s way of construing the situation, because they increase the precision of description, and their gradable characteristics tend to imply the speaker’s attitude to or evaluation of the situation.

There has been a huge bulk of research on perception adjectives that denote such vital qualities as smell, colour (sight), hearing, taste, and tactile phenomena. The verbalization of tactile senses, in particular, adjectives with the semantic component “**soft**” – “**hard**”, have, nonetheless, received little focus from researchers studying the English language, and, accordingly, calls forth the need for in-depth research in this area. The growing interest in the studies of tactile lexis is particularly stipulated by the fact that tactile senses are one of a human’s first “points of contact” with the real world. Touch perception is considered to have the utmost importance for a human. It precedes all other types of perception, which can be corroborated by the conceptualization of all other types of senses via touch (since touch markers transfer to other perceptual domains). The English tactile lexis is abundant, and is best represented in the form of a system,

or rather a structure. Figure 1 demonstrates the classification of tactile lexis. (by way of illustration, see Appendix A).

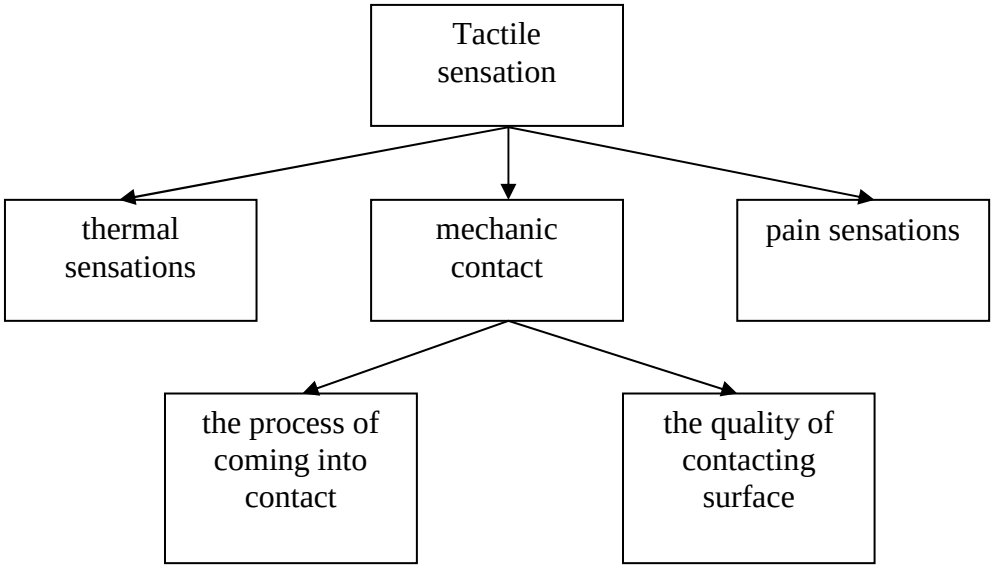


Figure 1. Semantic classification of tactile lexis

Modern investigations of perception adjectives are generally conducted from three main perspectives: *psycholinguistic*, *cognitive*, and *structural*. Nonetheless, researchers point out that these approaches are not exclusive, and often complement one another contributing to more specific and all-embracing findings of a research.

In the domain of psychology and cognitive linguistics scholars have already identified metaphorical relation across sensory modalities, known as *synaesthesia*. Ullmann (1951) defines synaesthesia as a linguistic universal. The cognitive approach is characterized by the assertion that perception and cognition are closely interrelated. Scientists are unanimous that people utilize perceptive experiences for conceptualizing abstract domains.

From cognitive perspective perception adjectives with the common semantic component "**soft**" – "**hard**" have been investigated by Russian scholars Laenko (2005) and Rakhilina (2000) among others. Laenko presents a contrastive analysis of the lexical-semantic group of adjectives with the meaning "**soft**" – "**hard**" on the basis of their lexical combinability in English and Russian, and examines their semantic structure. The author suggests the hypothesis that evaluation of the nominative potential of qualitative perception adjectives, as well as their lexical combinability, shed light on the way perceptual information about the world is processed in the consciousness of the English and Russian linguocultures. On the whole, the study itself has a conceptual framework.

In the field of structural linguistics perception adjectives with the meaning "**soft**" – "**hard**" have been analyzed by Agalakova (2003). The scholar explores

the semantic field of synaesthetic adjectives (*hard*, *soft* and 10 others) in the history of the English language: examines their semantic change within language dynamics and analyzes the interdependence of these changes in the language system in general.

From a different perspective, Bons (2008) presents a corpus-driven usage-based research of tactile adjectives 'hart', 'weich', 'sanft' and 'grob' in the German language and sets out to explore their polysemy and distribution. The author posits the view that the semantics of tactile adjectives can be best represented as a complex network of metonymical and metaphorical usages. Additionally, a thorough differential analysis of the spheres of application (tactile / visual / auditory / gustatory / olfactory) enables the scholar to witness a distinct affinity of all four lexemes that enter into antonymic and quasi-synonymous relations.

Syntagmatic relations of perception adjectives "**soft**" – "**hard**", in particular, their semantic combinability as "the ability of a lexical unit to combine with other words at the level of a group of words" (Levickij 2012: 245) has been the subject of less intensive and systematic study. Appearance of new denotations that eventually results in the expansion of denotative spheres of adjectives calls for the need of checking their syntagmatic characteristics. Determining the peculiarities of this line of evidence is the focus of this paper. We do this by performing a noun-based co-occurrence experiment with the application of statistical methods (chi-square test, and Phi coefficient).

Thus, our ultimate goal is not only to describe the preference of perception adjectives to combine with specific subclasses of nouns in discourse; it is rather to uncover the mechanisms of semantic combinability of lexical units.

Data for the present study was extracted from 25 English and American novels of the 20th to 21st centuries (D.Lawrence, I.Murdoch, G.Orwell, D.Brown, E.Gilbert, J.Rowling, and others).

As our research aim is broadly-based and requires linguistic conclusions to be drawn from the data, it is important to begin by reviewing previous research on syntagmatic relations and types of combinability before proceeding to undertake the statistical analysis.

2. Syntagmatic relations and types of combinability

The advent of Structuralism, with its view of language as a structured, organized network of relations, brought an epochal change, the consequences of which have affected the later developments in all fields of linguistic investigation. The work of the Swiss linguist Ferdinand de Saussure provided innovative means of describing the internal organization of language. Saussure's main contribution is a view of language as a unique relational structure, or system, in which the units identified as basic theoretical constructs acquire essence from their relationship with other units in the system (Martelli 2007: 4). Among the dichotomies used by

Saussure in referring to the structure and nature of language, it is the distinction between the paradigmatic and the syntagmatic aspects of language that has revealed itself particularly important for developments in lexical studies.

Interest in syntagmatic relations emerges from the observation that words are not independent from each other and that their combinatorial potential relies on a variety of factors. According to Lipka (1990: 160), as far back as the 1930s, when studies in lexical and structural semantics were exclusively paradigmatic, the German linguist Porzig provided one of the first attempts to explain the combinatorial properties of words. In his view, lexical items which co-occur are bound together by what he defined as *an essential meaning relation* (*wesenhafte Bedeutungsbeziehung*). Porzig's work, therefore, represented a starting point for the development of a systematic description of syntagmatic relations.

More recent studies on syntagmatic relations have been represented by two main approaches: the generative approach and the structuralist one. The former trend is characterized by the terms *selection restriction* and *transfer feature*, while the latter is represented by the concept of *lexical solidarities.* Selection restrictions (Katz and Fodor 1963; Katz 1972) provide a new way to avoid violating combinatory rules and at the same time are responsible for the disambiguation of certain lexical items. According to transfer feature theory (Weinrich 1966) combinability of items is ensured by the transference of semantic markers or 'features' from one word to the other. Taking up Porzig's view Coseriu coined the term "lexical solidarity" (E. Coseriu 1967: 296) which he defines as a class (of words), an archilexeme, or a lexeme functioning as a distinguishing feature in the meaning constitution of a word. Coseriu is genuinely concerned with the nature and structural properties of the relations between the constituents of collocations. Thus, his focus is on the constitution of the meaning of individual lexemes, not on collocations as linguistic structures. An overview of both generative and structuralist approaches was provided by Kastovsky (1980). After a comparison of selection restrictions and transfer features on the one hand and lexical solidarities on the other, he concludes that the two approaches are not exclusive or in contradiction with one another, but rather that they describe the same phenomenon from two different points of view. Lexical solidarities are described as positive semantic implications as they explain why some lexical items have a tendency to co-occur. Selection restrictions are described as constraints expressed in terms of semantic features which limit the combinability of certain items. On the whole, both approaches focus their analysis on the description of meaning relationships that exist between lexical items combined by syntagmatic relations.

It has already become a point of general agreement to identify syntagmatic relations with lexical combinability. The classification of lexical combinability introduced by Amosova and Apresyan is suggestive in this context. The scholars differentiate between *syntactic*, *semantic*, and *lexical* types of combinability (Apresyan 1974: 233). In light of this approach, Koteleva defines syntactic com-

binability as an ability of a word to establish syntagmatic relations with other words at the level of some grammatical category (e.g. "parts of speech"), that is to say, at the level of a class of words (Koteleva 1975: 81). According to Levickij, "semantic combinability implies the ability of a lexical unit to combine with other words at the level of a group of words" (Levickij 2012: 245). Lexical combinability (in the narrow sense) is a combination of two words: a lexical unit and any other word. Thus, a lexical unit may combine with: 1) a class of words; 2) a group of words; 3) individual words.

This approach differs from the previous two in the way it views syntagmatic relations, based on the types of possible syntagmatic partners that a word may have. In dealing with lexical combinability (in the broad sense), Levickij and his followers concentrate on statistical rather than exclusively semantic grounds.

In our study the typology introduced by Amosova and Apresyan will be applied as it is instructive for our present purposes. The current study addresses issues of semantic combinability of the adjectives "**soft**" – "**hard**" in the pattern "*a word + a group*".

3. Quantitative characteristics of the syntagmatic partners of "soft" – "hard"

The analysis of lexical combinability of adjectives that the present study addresses requires data of the combinatorial observations of the adjectives under investigation with nouns (in AN-combinations) in the pattern "*a word + a group of nouns*". All instances were classified in terms of their head-nouns. Within the collected data both attributive and predicative uses of the adjectives were analysed. After a meticulous review of noun classifications introduced by some scholars (Levickij 1989: 135-136; Levickij / Ogui et al. 2001: 52-53), we worked out our own semantic classification of nouns which, according to the empirical datasets, most comprehensively reflects the character of the context-based selection of perception adjectives "**soft**" – "**hard**" in the English language. In the case of predicative uses, subjects were taken as head-nouns. Instances of pronominal subjects were left out of the scope of our study. Each head-noun was afterwards coded as belonging to one of the following lexical-semantic subclass:

1) **Human appearance and body parts**: eye, shoulder, lip, bruise, palm;
2) **Names of humans**: baby, man, girl, hussy, youth;
3) **Social status**: master, worker, lieutenant, outcast;
4) **Proper nouns, names**: Rita Skeeter, Barker and Morel, Snickers;
5) **Flora (plants) and its constitution**: *grass, bloom, leave, branch, petal;*
6) **Fauna (birds are animals) and its constitution**: *leopard cub, wing, paw;*

7) **Bodies of nature, natural phenomena and space notions**: *flood, cloud, sky, air, snow;*
8) **Items of clothing and footwear**: *shirt, hat, slipper, coat, sarong;*
9) **Edifices, their elements, and premises**: *room, floor, hospital, wall, exterior;*
10) **Interior and its elements**: *mattress, couch, painting, chair, pillow;*
11) **Inanimate objects**: *disc, ball, money, sac, balloon;*
12) **Substances and materials:** *wood, coal, sand, cement, logs;*
13) **Food and beverages**: *loaf, candy, drink, vegetable, pizza dough;*
14) **Time notions**: *time, afternoon, night, day, century;*
15) **Character traits, features and characteristics of humans**: *kindness, look, smile, humour, violence;*
16) **Feelings, emotions, and relationships of humans:** *feeling, malignancy, anguish, exultance, thrill;*
17) **Abstractions**: *understanding, promise, energy, reality, magic;*
18) **Actions, processes, and arrangements**: *touch, kiss, nip, divorce,*
19) **Acoustic phenomena**: *voice, tone, thud, noise, whisper;*
20) **Olfactory phenomena**: *fragrance, odor;*
21) **Light phenomena and colour**: *light, sheen, blackness, glitter, hue;*
22) **Motion and movement**: *flow, tread, tug, travel, swerve;*
23) **Language and speech units**: *accent, word, litany, question, reiteration;*
24) **Shapes, figures and their elements**: *spot, side, tip, line, edge;*
25) **Other notions**: *drugs, yoga, juncture, thing, array.*

Table 1
Syntagmatic relations of the adjectives "**soft**" – "**hard**"

№	Subclasses of nouns	Adjectives under study		
		soft	hard	Total
1.	Human appearance	87	57	144
2.	Names of humans	12	4	16
3.	Social status	1	5	6
4.	Proper nouns	2	1	3
5.	Flora	13	2	15
6.	Fauna	3	0	3
7.	Nature, space	35	6	41
8.	Clothes, footwear	14	6	20
9.	Edifices, premises	3	13	16
10.	Interior objects	24	4	28
11.	Inanimate objects	7	7	14

12.	Substance, materials	19	17	36
13.	Food, beverages	6	5	11
14.	Time notions	7	26	33
15.	Character traits and humans' features	14	18	32
16.	Feelings, emotions, relationships	6	18	24
17.	Abstractions	16	42	58
18.	Actions, arrangements	12	20	32
19.	Acoustic phenomena	94	16	110
20.	Olfactory phenomena	2	1	3
21.	Light phenomena	46	3	49
22.	Motion, movement	15	9	24
23.	Language and speech units	5	6	11
24.	Shapes, figures	12	15	27
25.	Other notions	4	8	12
	Total	459	309	768

As exemplified in Table 1, the frequency of occurrence (hereafter "FO") of the adjective "**soft**" is somewhat higher than that of the adjective "**hard**". For example, "**soft**" is characterized by superior numbers with the following subclasses: "*human appearance*", "*names of humans*", "*flora*", "*fauna*", "*nature and space*", "*clothes and footwear*", "*interior objects*", "*light phenomena*", "*acoustic phenomena*", "*motion and movement*". In sum, there are 14 such subclasses (where "**soft**" is a quantitative leader).

In addition, there is a noticeable discrepancy in the distribution in terms of the FO of the adjectives under study. For example, the FO of the adjective "**soft**" with noun subclasses "*time notions*" and "*light phenomena*" are in relation 7: 46, respectively; while "**hard**" with the same subclasses has a relation 26: 3. On the whole, the ability of a word to establish syntagmatic relations with other words, as well as a differing distribution of these relations is grounded upon extralinguistic and intralinguistic factors. The underlying property of combinability is a semantic coordination (congruence) or, in other words, compatibility of two words. Therefore, the characteristic feature of combinability is that with some words a certain word can combine practically without any restrictions, with others – with a few restrictions, and with the third – cannot combine at all. In this connection, Levickij suggested differentiating among three types of compatibility of words – *denotative*, *pragmatic*, and *lexical* (Levickij 2012: 243). *Denotative compatibility* of two words is characterized by the presence of some objective relations between the denotations of these words in the real world, whereas *pragmatic compatibility* implies coordination of communicative events in which words can be used. *Lexical compatibility*, accordingly, originates as a result of selection and entrenchment in a language of certain word combinations under the influence of language usage.

However, while analyzing the reasons of quantitative variation of syntag-

matic relations between words, it is possible to assume that the frequency of synchronous emergence of two words in a text may depend on the FO of every separate word. For example, the combinability of the pattern "**soft** + the noun subclass *human appearance*" is predetermined both by relatively high frequencies of the adjective "**soft**" and the subclass of nouns denoting human appearance. The lexeme "**soft**" with the frequency *n* = 459 indeed has the highest figure, while the noun subclass "*human appearance*" (*n* = 144) is dominant among other subclasses (by way of illustration, see Table 1). A similar pattern can be observed in the relation of the following subclasses with the given adjective: "**soft** + *acoustic phenomena*" (*n* = 110), "**soft** + *light phenomena*" (*n* = 49) and so on. Nonetheless, such regularity is not absolute, since a less frequent adjective "**hard**" (*n* = 309) is combined with a subclass "*abstractions*" (*n* = 58), which is the third in the FO.

These findings appear to suggest an irregularity in syntagmatic relations of the adjectives under investigation. Henceforth, we will suppose that the reason of quantitative variation of syntagmatic relations at the present stage of investigation is not yet clarified, and requires the application of additional research methods. Therefore, we find it necessary to introduce one more parameter – *the range of combinability* (hereafter 'RC'). This parameter here refers to the differences in combinability of every subclass with the adjectives under study. The RC is expressed by some relative value (from 0 to 1), namely by the ratio between the number of recorded word combinations and the total number of units that constitute the database under investigation.

Figures from Table 2 indicate that in all noun subclasses the RC equals 1, with the exception of the subclass "*fauna*" (the RC equals 0.5). Apparently, almost homogenous RC in all the noun subclasses is called forth by a relatively small number of units which constitute the database under study (only 2 adjectives). Therefore, in order to differentiate between noun subclasses in a more explicit way as well as to characterize the degree of their relation with the adjectives "**soft**" – "**hard**" we introduce such a relative value as *the utilization* of noun subclasses. In order to get this parameter, it is necessary to divide the total number of all word occurrences by the span of a subclass. Specifically, with the words "**soft**" – "**hard**" 24 nouns with the meaning of "*nature and space*" were used by the authors 41 times. The utilization value of this subclass with the adjectives "**soft**" – "**hard**" equals 41: 24 = 1.7. The findings of utilization value of all the noun subclasses with the adjectives under study in English fiction are shown in Table 2.

Table 2
Quantitative characteristics of noun subclasses

№	Noun subclasses	Frequency of occurrence	Range of combin-ability	Utilization value	Span
1.	Human appearance	144	1	4	36
2.	Names of humans	16	1	1.3	11
3.	Social status	6	1	1.5	4
4.	Proper names	3	1	1	3
5.	Flora	15	1	1.2	13
6.	Fauna	3	0.5	1	3
7.	Nature, space	41	1	1.7	24
8.	Clothes, footwear	20	1	1.8	11
9.	Edifices, premises	16	1	2	8
10.	Interior objects	28	1	1.5	18
11.	Inanimate objects	14	1	1.4	10
12.	Substances, materials	36	1	2.1	17
13.	Food, beverages	11	1	1.1	10
14.	Time notions	33	1	4.1	8
15.	Character traits and humans' features	22	1	1.2	18
16.	Feelings, emotions, relationships	24	1	1.4	17
17.	Abstractions	58	1	2.3	25
18.	Actions, arrangements	32	1	1.3	24
19.	Acoustic phenomena	110	1	3.1	35
20.	Olfactory phenomena	3	1	1.5	2
21.	Light phenomena	49	1	2.6	19
22.	Motion, movement	24	1	1.1	21
23.	Language and speech units	11	1	1.8	6
24.	Shapes, figures	27	1	2.7	10
25.	Other notions	12	1	1.5	8

As exemplified in the table above, the subclasses of the most frequently

occurring syntagmatic partners have a rather high utilization coefficient: "*human appearance*" – 144: 4; "*acoustic phenomena*" – 110: 3.1; "*light phenomena*" – 49: 2.6. This raises the question whether there is a directly proportional dependency between the recorded parameters. Yet, such correlation is not absolute: "*time notions*" – 33: 4.1; "*shapes, figures*" – 27: 2.7. Taking into account these highly contrastive patterns, we refer to statistical methods to check the correlation between the abovementioned values. Since the number of observations *n* for each pair of adjectives equals 25 (the number of syntagmatic relations according to Table 1), then the number of degrees of freedom for correlation analysis makes up df = 25 - 2 = 23.

According to statistical table the statistically significant coefficient equals: r = 0.40 at the significance level α = 0.05 (5%) or r = 0.51 at the significance level α = 0.01 (1%). Thus, the relations between the values of frequency and utilization lower than 0.40 (≤ 0.40) are considered weak; from 0.40 to 0.51 (> 0.40; < 0.51) – medium; higher than 0.51 (0.51 ≥) – strong. Table 3 illustrates the data of statistical analysis.

Table 3
Coefficient values of the noun subclasses correlation

	Range	Utilization	Span
Frequency	0.17	0.67	0.86
Range		0.21	0.25
Utilization			0.33

It is apparent from the table above that there is a rather high correlation (*r* = 0.67) between the frequency and the utilization of noun subclasses with the adjectives "**soft**" **–** "**hard**"**.** As it can be observed, the utilization depends on the frequency of occurrence of the whole subclass or its separate elements. Again, this signifies that certain contextual partners have considerable syntagmatic potency. Utilization also depends on the span of a subclass (*r* = 0.33), but such ratio is not significant.

On the other hand, we observe a very high correlation between the frequency of occurrence of noun subclasses and the number of elements that constitute a certain subclass (*r* = 0.86). It is, therefore, obvious that the extension of span results in the increase of its frequency.

As regards the range parameter, it negligibly depends on frequency (*r* = 0.17) and largely – on utilization (*r* = 0.21) and span (*r* = 0.25). In other words, the higher is the subclass utilization, the higher will be the RC of these nouns with the adjectives under study. It can be concluded here that the larger is the number of reiterating words embedded in a text, the larger will be the number of adjectives they combine with. However, the correlation coefficient in this case is insignificant. Therefore, we can speak of a tendency rather than regularity.

4. The range of combinability and the utilization value of "soft" – "hard"

The RC of adjectival lexemes is analyzed on the basis of their combinability with different subclasses of nouns. Specifically, the adjective "**soft**" enters into syntagmatic relations with 25 out of 25 subclasses, while its opposite, "**hard**", combines with 24 out of 25 subclasses. In the first case the RC equals 25:25 = 1, and in the second the RC equals 24:25 = 0.96. As it can be observed, the RC value of both adjectives is quite high and almost the same. Apparently, this result is called forth by certain factors. On the one hand, the adjectives "**soft**" and "**hard**" are centre, dominant lexemes of semantic groups that denote "softness" and "hardness", therefore, it is quite logical that their combinability has to be the widest.

Notwithstanding a rather wide, almost similar RC of the adjectives "**soft**" **–** "**hard**" with their contextual datasets ("**soft**" – 1, "**hard**" – 0.96), the intrinsic character of syntagmatic relations of the lexemes under study with the noun subclasses will vary each time. This can be ascertained, however, by means of estimating the utilization value of every adjective with noun subclasses. To obtain this value, it is necessary to divide the number of all word occurrences by the span of a subclass. For example, with the word "**hard**" 7 nouns denoting "*edifices and premises*" were used by the authors 13 times. Henceforth, the utilization value of the adjective "**hard**" with the given subclass equals 13:7 = 1.8. Observations of utilization value of the adjectives under study are made in Table 4.

Table 4
Utilization value of "**soft**" – "**hard**" with the noun subclasses

№	Subclasses of nouns	Adjectives under study	
		soft	hard
1.	Human appearance	3.3	3
2.	Names of humans	1.5	1
3.	Social status	1	1.7
4.	Proper names	1	1
5.	Flora	1.2	1
6.	Fauna	1	-
7.	Nature, space	1.7	1
8.	Clothes, footwear	1.3	6
9.	Edifices, premises	1	1.8
10.	Interior objects	1.5	1.3
11.	Inanimate objects	1	2.3
12.	Substances, materials	1.7	2.8
13.	Food, beverages	1	1
14.	Time notions	2.3	4.3
15.	Character traits and humans' features	1.8	1.4

16.	Feelings, emotions, relationships	1.5	1.3
17.	Abstractions	1.3	2.8
18.	Actions, arrangements	1.7	1.1
19.	Acoustic phenomena	3.03	2
20.	Olfactory phenomena	2	1
21.	Colour and light phenomena	2.4	1
22.	Movement, motion	1.3	1
23.	Language and speech units	1.3	2
24.	Shapes, figures	2.4	2.5
25.	Other notions	1	2

5. The assessment of semantic relations in the pattern "*Adjective + a group of nouns*"

The investigation of syntagmatic relations between lexical items should not be confined to registering the range and frequency of occurrence in a text. An important feature of combinability is the strength of relations, that is, their *intensity*. The data on the intensity of relations of adjectives with different subclasses of nouns is obtained by means of statistical methods, Chi-square test, and contingency coefficient *Φ*.

$$\chi^2 = \sum \frac{(O-E)^2}{E}, \tag{1}$$

where

$\chi 2$ – the criterion of congruence or correlation;
Σ – total;
O – practically researched values;
E – theoretically projected values.

$$\Phi = \frac{ad-bc}{\sqrt{(a+b)(c+d)(a+c)(b+d)}}, \tag{2}$$

where *Φ* – the coefficient of mutual contingency;

a, b, c, d – empirical values in a four-field table.

While the χ^2 value indicates the presence or absence of a relation, the coefficient Φ points to the intensity of this relation. Since the critical value of χ^2 is 3.84, the combinations in which the degree of a relation equals or exceeds 3.84 are considered standard. All the instances when χ^2 is smaller than 3.84, are statistically insignificant. The measure of the relation between markers was established by means of the coefficient of mutual contingency (*Φ*) that can range

from 0 to 1 depending on the presence or absence of a relation.

Instead of these methods it is possible to use the ***u*** criterion defined as

$$u = \frac{n_{ij} - E_{ij}}{\sqrt{\frac{n_i \cdot n_j \cdot (N - n_i) \cdot (N - n_j)}{N^2 \cdot (N - 1)}}}, \quad (3)$$

where n_i –marginal sums on the right side of a table

n_j –marginal sums at the bottom of a table

N – the sum of all the rates in a table.

As the practice of statistical studies shows, it is reasonable to present the data for analysis in the form of four-field tables set up by two columns and two rows. Such tables can be compiled by means of reducing or combining the number of columns and rows in a many-field table. Accordingly, an alternative table by means of which it would be possible to determine the relation (in pairs) between the markers under study will look as follows:

Table 5
Alternative distribution of a word's frequencies

a	b	(a+b)
c	d	(c+d)
(a+c)	(b+d)	N

a, b, c, d – empirical values in a four-field table;

N – the total number of observations.

The corresponding calculations showed that the highest values of χ^2 criterion and Φ are observed in the following cases of combinability of the adjectives "**soft**" – "**hard**":

- "**soft** + *acoustic phenomena*" (χ^2 = 35.23; Φ = 0.21);
- "**hard** + *abstractions*" (χ^2 = 27.02; Φ = 0.19);

When comparing the observations made in Tables 6 and 1, we can make the following assumptions: while the FO of "**soft**" with the subclass "*acoustic phenomena*" considerably exceeds these indices in the rest of the subclasses (94 out of 459 units), the FO of "**hard**" with the subclass "*abstractions*" does not constitute an absolute majority (42 out of 309 units) (see Table 1). Instead, the adjective "**hard**" most frequently occurs with the subclass of nouns "*human appearance*" (57 out of 309 units). However, with this subclass no standard relations were detected. These findings appear to suggest that the value of χ^2 criterion is not necessarily directly proportional to the indices of the frequency of occurrence. It follows that a high FO does not imply a close connection between

the signified (a subclass of substantive lexemes) and the signifying (an adjectival lexeme). It rather indicates a speaker's individual peculiarities of perception of the categories "softness" and "hardness", which are reflected in a language through certain regularities of usage of the adjectives "**soft**" – "**hard**" in a text's structure.

The analysis of χ^2 and Φ values of each adjective under study with the noun subclasses has produced the following findings. The adjective "**soft**" has standard syntagmatic relations with such subclasses of nouns:

1) **soft +** "*acoustic phenomena*" (χ^2 = 35.23; Φ = 0.21). For example, "*The cries of hopelessness against the howling wind of the Pyrenees and the **soft sobs** of forgotten men.*" (Brown 2003: 46).
2) **soft** + "*colour and light phenomena*" (χ^2 = 25.33; Φ = 0.18). For example, "*Illuminated in the **soft lights** of the deserted entresol, the two pyramids pointed at one another, their bodies perfectly aligned, their tips almost touching*" (Brown 2003: 382).
3) **soft** + "*nature, space*" (χ^2 = 11.80; Φ = 0.12). For example, "*The metal looks delicate, and marble is a **soft rock*** (Brown 2003:169).
4) **soft** + "*interior objects*" (χ^2 = 8.14; Φ = 0.10). For example, "*His **bed** was **soft** like a cloud, and the air around him smelled sweet with candles*" (Brown 2003: 48).
5) **soft** + "*flora*" (χ^2 = 4.60; Φ = 0.08). For example, "*She stayed away for the rest of the afternoon, stopping at a shady creek to water the horse and let him graze on the **soft**, fresh **grasses** nearby*" (Howard 1997: 13).

The adjective "**hard**" has the following standard syntagmatic relations:

1) **hard** + "*abstractions*" (χ^2 = 27.02; Φ = 0.19). For example, "*Do you think the Church would actually be capable of uncovering **hard evidence** of the Priory's date?*" (Brown 2003: 226).
2) **hard** + "*time notions*" (χ^2 = 21.31; Φ = 0.17). For example, "*They had had a **hard year**, and…*" (Orwell 1945: 28).
3) **hard** + "*feelings, emotions, relationships"* (χ^2 = 12.45; Φ = 0.12). For example, "*Even though Webb had come back for Lucinda's sake, there were **hard feelings** between them that needed to be settled*" (Howard 1997: 9).
4) **hard** + "*edifices, their elements, and premises*" (χ^2 = 11.43; Φ = 0.12). For example, "*Langdon shouted, raising his arm and suspending the cryptex precariously over the **hard stone floor**.*" (Brown 2003: 355).
5) **hard** + "*actions, arrangements*" (χ^2 = 6.88; Φ = 0.19). For example, "*Langdon gave it a **hard kick** and sensed he might be able to break it off entirely*" (Brown 2003: 180).
6) **hard** + "*social status*" (χ^2 = 4.67; Φ = 0.08). For example, "*In past years Mr. Jones, although a **hard master**, had been a capable farmer, but of late he had fallen on evil days*" (Orwell 1945: 8).

The investigation has shown that the adjectives "**soft**" and "**hard**" have a similar number of standard relations: 5 and 6, respectively. Yet, all the discovered relations are marked by a noticeable contrast in the types of syntagmatic partners (subclasses of nouns) – in those instances where "**soft**" establishes standard relations, the adjective "**hard**" exhibits no relations, and vice versa. These findings clearly confirm the above-mentioned hypothesis that a high FO of lexemes does not necessarily imply a close relation between them.

The biggest number of standard syntagmatic relations for the adjective "**soft**" has been recorded in combinations with the subclasses of nouns that denote: flora, natural phenomena, space notions, interior objects, light and acoustic phenomena. Apparently, these are the very contextual datasets with which the adjective with the meaning "**soft**" in the English language establishes steady combinations, acquiring both direct and figurative meanings.

Table 6
The most significant standard relations of the adjectives "**soft**" – "**hard**" with the semantic subclasses of nouns

Subclasses of nouns	**The adjectives under study**	
	soft	hard
Human appearance	-	-
Names of humans	-	-
Social status	-	$\chi^2 = 4.67$ $\Phi = 0.08$
Proper names	-	-
Flora	$\chi^2 = 4.60$ $\Phi = 0.08$	-
Fauna	-	-
Nature, space	$\chi^2 = 11.80$ $\Phi = 0.12$	-
Clothes, footwear	-	-
Edifices, premises	-	$\chi^2 = 11.43$ $\Phi = 0.12$
Interior objects	$\chi^2 = 8.14$ $\Phi = 0.10$	-
Inanimate objects	-	-
Substances, materials	-	-
Food, beverages	-	-
Time notions	-	$\chi^2 = 21.31$ $\Phi = 0.17$
Character traits and humans' features	-	-

Feelings, emotions, relationships	-	$\chi^2 = 12.45$ $\Phi = 0.12$
Abstractions	-	$\chi^2 = 27.02$ $\Phi = 0.19$
Actions, arrangements	-	$\chi^2 = 6.88$ $\Phi = 0.19$
Acoustic phenomena	$\chi^2 = 35.23$ $\Phi = 0.21$	-
Olfactory phenomena	-	-
Colour and light phenomena	$\chi^2 = 25.33$ $\Phi = 0.18$	-
Movement, motion	-	-
Language and speech units	-	-
Shapes, figures	-	-
Other notions	-	-

The adjective "**hard**", as well as "**soft**", possess a polysemantic status. The semantic boundaries between certain lexical-semantic variants are indistinct, which fosters reinterpretation of physical properties of an object into diverse types of abstract relations. Henceforth, the closest semantic relation with the subclass of abstractions seems quite unequivocal. We can conclude here that a set of human's concepts about the category of "hardness" has been expanding in due course, and new derivative meanings of the adjective "**hard**" became entrenched in language. As a result of abstraction they came to denote human feelings, emotions, relationships, time notions, various actions and processes.

Given the above-mentioned facts, the categories of "softness" and "hardness" can be viewed as dynamic structures which gradually acquire in language newer forms of expression.

6. Conclusion

On a final note, it has been one of the goals of this study to research the syntagmatic characteristics of dominant adjectival lexemes that denote "softness" and "hardness" in the English language with the application of statistical methods. In so doing, the investigation has shown standard elements of combinability of the adjectives "**soft**" – "**hard**" at the level of semantic subclasses.

Through the unique examination of the adjectives "**soft**" – "**hard**" in the fiction discourse, the present study has also provided an example of how lexical items "interact" within the frameworks of language. The analysis of the empirical data has demonstrated that the adjectives under study with opposite denotational meanings, have a similar range of combinability. Yet, neither a wide range of combinability nor the frequency of occurrence are the factors which condition a

close relation between the markers. Therefore, the χ^2 value is not necessarily directly proportional to the indices of occurrence. Findings from the present study have revealed that wide discrepancies are found between the significant standard relations of the dominant lexemes "**soft**" and "**hard**". Statistically, they are characterized by contrasting distribution.

Based on the findings of this preliminary study, several outlines of future research may be useful to gain further insight into the relationships among other lexical items that belong to the reviewed microsystems with the common semantic components "**soft**", "**hard**". The study of paradigmatic relations of words will, thus, be a logical continuation to the present research, inasmuch as any relations of syntagmatic nature have context dependence on paradigmatic relations. In so doing, we will be able to explore a separate fragment of "linguistic image of the world" of the English linguoculture.

References

Agalakova, T.B. (2003). *Stanovleniye leksiko-semanticheskogo polya sinesteticheskikh prilagatelnykh v anglijskom jazyke:* Avtoref.dis...kand.filol.nauk.: www.dissercat.com

Apresyan, Y.D. (1974). *Leksicheskaya semantika. Sinonimicheskie sredstva yazyka*. Moskva: Nauka.

Bons, I. (2009). *Polysemie und Distribution : Zur Theorie und Methode einer korpusbasierten Semantik deutscher Adjektive*: Gießen : Gießener Elektronische Bibliothek: http://geb.uni-giessen.de/geb/volltexte/2009/7356/

Brown, D. (2003). *The Da Vinci Code*: http://pictoumasons.org/libraby/Brown, Dan~The Da Vinci Code [pdf].pdf

Coseriu, E. (1967). Lexikalische Solidaritäten. *Poetica, 1, 293-303.*

Howard, L. (1997). *Shades of Twilight*: http://freebooks2u.org/romance/Shades_Of_Twilight/15282.html.

Kastovsky, D. (1980). Selectional restrictions and lexical solidarities. In *Kastovsky, D. (Ed.), Perspektiven der Lexikalischen Semantik. Beiträge zum Wuppertaler Semantikkolloqium vom 2.-3. Dezember 1977 (70-92).* Bonn: Bouvier.

Katz, J. and Fodor, J. (1963). The structure of semantic theory. *Language, 39, 170-210.*

Kotelova, N.Z. (1975). *Znachenie slova i ego sochetaimost.* Leningrad: Nauka.

Laenko, L.V. (2005). *Perzeptivnyj priznak kak objekt nominazii*: Avtorefer .dis. dokt. filol.nauk.: www.dissercat.com

Levickij, V.V., Ogui, O.D. u.a. (2000). *Aproksimatyvni metody vyvchennja leksychnogo skladu. Černovcy*: Ruta.

Levickij, V.V. (2012). *Semasiologiya: monografiya dlya molodykh issledovatelej*. Vinniza: Nova knyga.

Lipka, L. (1990). *An outline of English Lexicology*. Tübingen: Niemeyer.

Martelli, A. (2007). *Lexical collocations in Learner English: A Corpus-based Approach.* Alessandria: Edizioni dell'Orso.

Orwell, G. (1945). *Animal Farm*. http://msxnet.org/orwell/print/animal_farm.pdf.

Rakhilina, E.V. (2000). *Kognitivnyj analiz predmetnykh imen: semantika i sochetaimost.* Moskva: Russkie slovari.

Shindo, M. (2009). *Semantic Extension, Subjectification, and Verbalization.* Maryland: University Press of America.

Tribushinina, **E**. (2008). Cognitive reference points. Semantics beyond the prototypes in adjectives of space and colour. https://openaccess.leidenuniv.nl/handle/1887/13224

Ullmann, S. (1951). *The Principles of Semantics*. Glasgow: University of Glasgow.

Weinrich, U. (1966). Explorations in semantic theory. In *Sebeok, T.A. (Ed.), Current Trends in Linguistics (395-477).* The Hague: Mouton.

Distribution of the Depth of Argumentation Relations

Andrei Beliankou, Reinhard Köhler, Sven Naumann, Trier

Abstract. The logical structure of an argument can be represented as a tree, whose elements are argumentation relations such as justification, elaboration, concession, circumstance etc. The present paper is a study on the depths of the individual argumentation relations in the corresponding tree. Depth will be considered as a random variable. Its probability distribution will be derived using Altmann's proportionality approach, and the model will be tested on data from an annotated German newspaper corpus. Statistical characteristics (mean and skewness) of the empirical frequency distributions will be used to describe the typical behaviour of the relations types. Finally, the order of the argumentation relations in the logical structure will be compared with the order of the corresponding elements in the linguistic surface. It is observed that the distribution of the degree of agreement between these orders can be modelled with the (U-shaped) binomial arcus sinus distribution.

The Data

For this study we used data extracted from the Potsdam Commentary Corpus (Stede, 2003). This resource is an ongoing project combining multilayered linguistic annotations and some interesting aspects of the regional language. This corpus is the most mature corpus of German with an annotation layer using the formalism called RST (Mann & Thomson, 1987). Furthermore, the tree structure of rhetorical relation was encoded in a flexible xml-based format baptized URML (Reitter & Stede, 2003). The corpus was produced at the Potsdam University. It is based on short commentaries from a regional daily newspaper "Märkische Allgemeine Zeitung". The source for the texts and the genre were chosen guided by the following considerations: (1) The texts are short; (2) the lexical richness is less than in a national daily and (3) every article is rather opinionated and thus can be annotated unambiguously.

The version of the PCC we took for our analysis consists of 172 texts with an average text length of 10.9 sentences per document. All 1876 sentences are split into 2771 units produced by human annotators. The length of a discourse unit averages to 1.5 discourse units per sentence. The average length of a discourse unit is 11.7 tokens.

Due to our focus on RST, in this study we want to pay particular attention to the quantitative evaluation of this annotation level. The corpus annotation scheme contains 23 different relation types (18 mononuclear vs. 5 multinuclear). In the evaluated part of the corpus, we observed the following absolute frequencies, e.g. for some mononuclear relations: elaboration – 730, evaluation – 268, evidence – 264, summary – 1. In case of multinuclear relations these frequencies are: list – 166; contrast – 54; joint – 42; sequence – 27; conjunction – 1.

The whole corpus contains 220 RST Trees spread over 124 single-tree documents and 48 documents with two trees. The majority of the annotated relations (2075 occurrences) are mononuclear, and only a few relations (290 occurrences) have multinuclear nature. The average relation tree contains 32.4 nodes. Figure 1 shows a small argumentation tree as an illustration of the data.

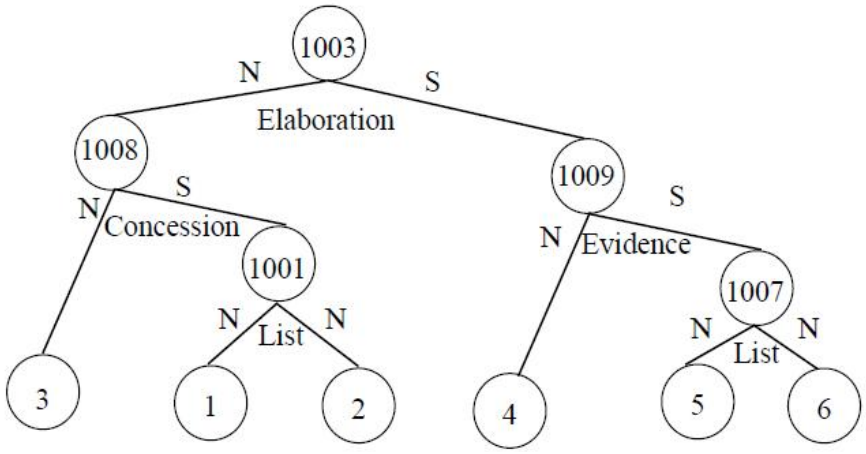


Fig. 1: Example of a text abstract, represented using RST: [**1** *Wie schwierig es ist, in dieser Region einen Ausbildungsplatz zu finden,*] [**2** *haben wir an dieser und anderer Stelle oft und ausführlich bewertet.*] [**3** *Trotzdem bemühen sich Unternehmen sowie die Industrie- und Handelskammer Potsdam den Schulabgängern Wege in die Ausbildung aufzuzeigen.*] [**4** *Der Berufemarkt – erstmals in der Aula der Gesamtschule "Am Weinberg" angeboten war ein guter und sinnvoller Beitrag.*] [**5** *Das hat allein das Interesse der Schülerinnen und Schüler bewiesen.*][**6** *Noch etwas hat der "Berufemarkt ans Licht gebracht.*]

[1 How difficult it is to find an apprenticeship in this region] [2 have we mentioned several times and commented on extensively.] [3 Even so are companies and the *Industrie- und Handelskammer Potsdam* trying to show graduates a way to find an apprenticeship.] [4 The 'Berufemarkt' – arranged for the first time in the auditorium of the comprehensive school "Am Weinberg" - was a good an meaningful contribution.] [5 That was proven already by the interest of the pupils.] [6 And something else was revealed by the 'Berufemarkt'.]

The Hypothesis

The variable we want to study here is the depth of a node representing an argumentation relation in the tree. In the example in figure 1, 'elaboration' forms the root of the tree and is thus assigned depth 0 while 'concession' and 'evidence' have depth 2 etc.

With regard to the theoretical distribution of the depth of a relation, we can set up the following hypothesis: The more complex the argument the deeper the tree will be nested. A given relation will, on the average, appear on deeper levels if it is mostly used in complex argumentations. This property of a relation

will be represented by the parameter *C* in our hypothesis. A second factor introduced here is the tendency to limit the depth of an argumentation, a constant factor which reflects the cognitive limitations both of the speaker/writer and the hearer/reader. The interdependence of these two factors and their influence on depth can be modelled in the form of Köhler/Altmann's approach (Köhler/ Altmann 1996): The probability of a given depth *x* is proportional to the probability of depth *x*-1, where the proportionality is a linear function:

$$P_x = \frac{C}{x + T - 1} P_{x-1}$$

C has an increasing influence on this relation whereas *T* has a decreasing one. The probability class *x* itself has also a decreasing influence, which reflects the fact that the probability of deeply nested relations decreases with the already given depth. This equation leads to the hyper-Poisson distribution (Wimmer/ Altmann 1999, 281):

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

where ${}_1F_1(1;b;a)$ is the confluent hypergeometric function

$${}_1F_1(1;b;a) = \sum_{j=0}^{\infty} \frac{a^j}{b^{(j)}}.$$

and $b^{(x)} = b(b+1)\ldots(b+x-1)$. According to this derivation, the hyper-Poisson distribution should be a good model of the depth distribution of a relation in argumentation structures. In fact, fitting this distribution to the data from the Potsdam corpus yields good and some acceptable results (cf. Table 1 and Fig. 2). In Table 1, the columns 2, 3, and 4 give the results of the goodness-of-fit tests, column 5 specifies the degrees of freedom, the values of the coefficient of determination R2 are shown in column 6 although it is defined for linear functions only; it may help to evaluate a fit nevertheless. Columns 7 and 8 give the parameter estimations and column 9 the sample size.

Skewness

The skewness of a distribution, defined as $\gamma_1 = \mu_3 / \sigma^3$, shows how unsymmetrical the aggregation of the instances of a relation is. We can see that volitional results and causes are given on levels which are symmetrically distributed around the mean; circumstances are concentrated on high levels cf. Fig. 2b.

Table 1
Results of fitting the hyper-Poisson distribution to the depth data of the most important argumentation relations.

	Distribution: Hyper-Poisson (a,b)							
Relation:	X^2	$P(X^2)$	C	DF	R^2	a	b	N
nonvolitional-cause	1.09	0.8963	0.0131	4	0.9676	2.1651	0.5851	83
circumstance	0.39	0.8224	0.0156	2	0.9073	3.5359	4.0411	25
contrast	2.3	0.6814	0.0765	4	0.8129	4.3906	3.3156	30
volitional-cause	1.61	0.6579	0.0423	3	0.8375	2.0292	1.2259	38
condition	3.09	0.5423	0.0967	4	0.7906	2.7249	0.2193	32
background	2.5	0.4754	0.0219	3	0.9566	1.8518	1.3611	114
volitional-result	0.57	0.4503	0.0475	1	0.8545	1.0524	0.2631	12
preparation	6.99	0.2216	0.0564	5	0.9475	2.0299	0.1104	124
concession	7.57	0.1816	0.0772	5	0.806	2.7765	0.8085	98
elaboration	10.47	0.1632	0.0254	7	0.9583	3.1462	0.7397	413
evidence	7.42	0.1155	0.0412	4	0.9443	2.0466	0.2646	180
evaluation	10.36	0.1103	0.054	6	0.8838	13.1137	15.6186	192
nonvolitional-result	8.1	0.0881	0.2249	4	0.5252	3.0058	1.2099	36

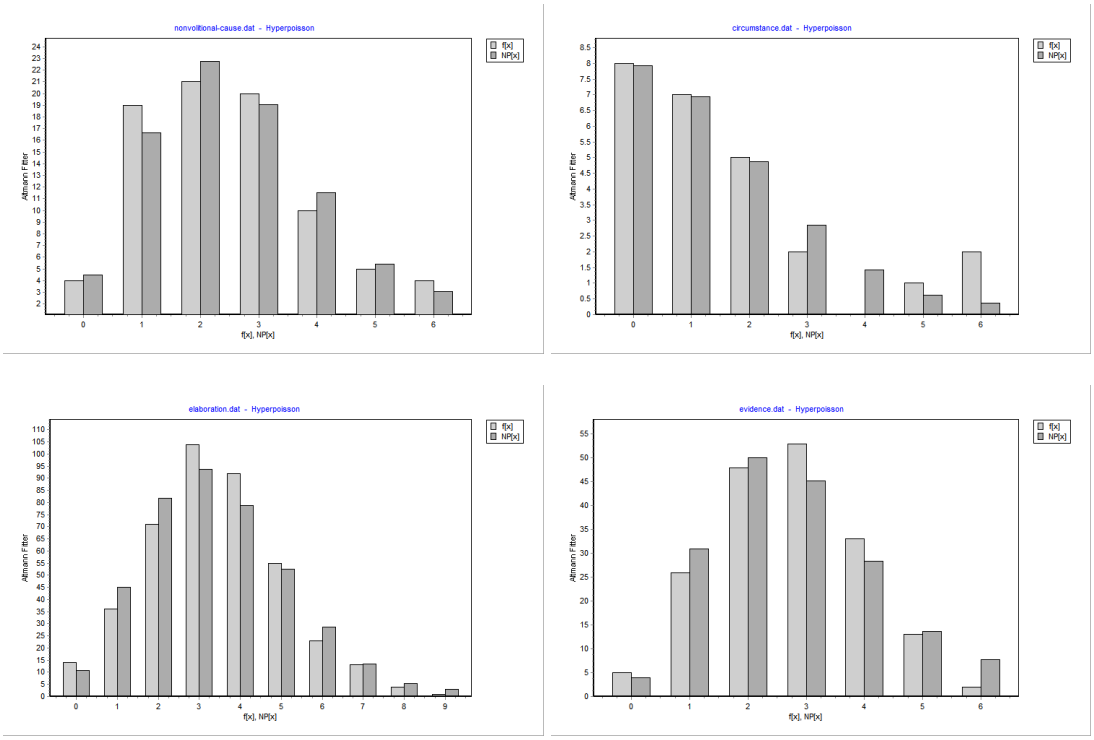


Figure 2 (a-d): Some examples of the fitting results

Table 2
Mean values of the depth distributions
of the argumentation relations

Relation	Mean value
background	1.5877
circumstance	1.6000
evaluation	2.2552
contrast	2.4000
nonvolitional-cause	2.5301
volitional-result	2.6667
evidence	2.7222
volitional-cause	2.8421
preparation	2.8548
nonvolitional-result	2.8611
concession	2.9082
elaboration	3.3971
condition	3.4375

Table 3
Skewness values of the depth distributions of the argumentation relations

Relation	Skewness
volitional-result	0.2708
volitional-cause	0.5838
evidence	0.5839
preparation	0.6819
condition	0.7080
elaboration	0.8117
nonvolitional-cause	0.8604
concession	0.9690
nonvolitional-result	1.1098
background	1.1471
contrast	1.5444
evaluation	1.8071
circumstance	1.9500

Ord's criteria

Another way to compare statistical characteristics of the argumentation relations are the well-known Ord criteria (Ord 1972: 98f; 133ff) *I* and *S*. These two quantities are functions of three moments of a distribution the mean (m_1), the variance (m_2) and the skewness (m_3). The definitions are $I = m_2/ m_1$ and $S = m_3/ m_2$. These characteristics can be used to locate an empirical distribution as a point in a two-dimensional plane, where *I* and *S* are the co-ordinates. Every theoretical distribution which has the three moments can be assigned a point, a line, or an area of possible *I/S*-values in the plane, and empirical distributions can be described with respect to the theoretical and with respect to other empirical distributions. The values calculated from the moments of the depth distributions of the argumentation relations are as follows and as shown in Figure 3.

Table 4
Ord's criteria of the depth distributions of the argumentation relations

I	S
0.1172	2.2237
-0.1207	-1.54
0.5263	5.8463
0.2957	2.3467
0.3074	11.1854
1.1199	8.0852
0.4866	9.459
1.3152	18.3864
0.5202	8.8385
0.7764	8.3362
0.713	20.7587
0.74	11.8639
0.7095	20.1068

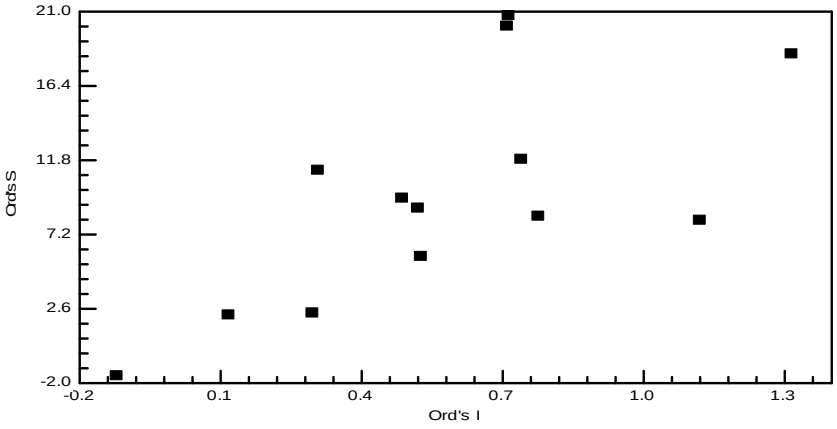


Figure 3: Plot of the values of Ord's *S* as a function of *I* for the depth distribution of the relations

The criteria show, numerically and graphically, that the relations form an inhomogeneous group without a clear structure – although a tendency can be observed.

Order of segments

From the illustration in Figure 1, it can be seen that the order in which the argumentation relations appear in the logical argumentation structure, the tree, (in this case [1]concession, [2]elaboration1, [3]elaboration2, [4]evidence ...) does not agree with the order of the corresponding elements of the linguistic surface, which are clauses in most cases (in this case [2]elaboration1 "*Wie schwierig es ist, in dieser Region einen Ausbildungsplatz zu finden*", [3]elaboration2 "*haben wir an dieser und anderer Stelle oft und ausführlich bewertet*", [1]concession "*Trotzdem bemühen sich Unternehmen sowie die Industrie- und Handelskammer Potsdam den Schulabgängern Wege in die Ausbildung aufzuzeigen*" ...). While the logical order is 1, 2, 3,..., the corresponding linguistic elements have the order 2, 3, 1,... The illustration in Table 5 and Figure 4 gives an example of two diverging orders.

Table 5
Linguistic and logical order of arguments in a text

Linguistic order	**Logical order**
1	8
2	9
3	10
4	11
5	12
6	13
7	14
8	7
9	5
10	6
11	3
12	4
13	2
14	1

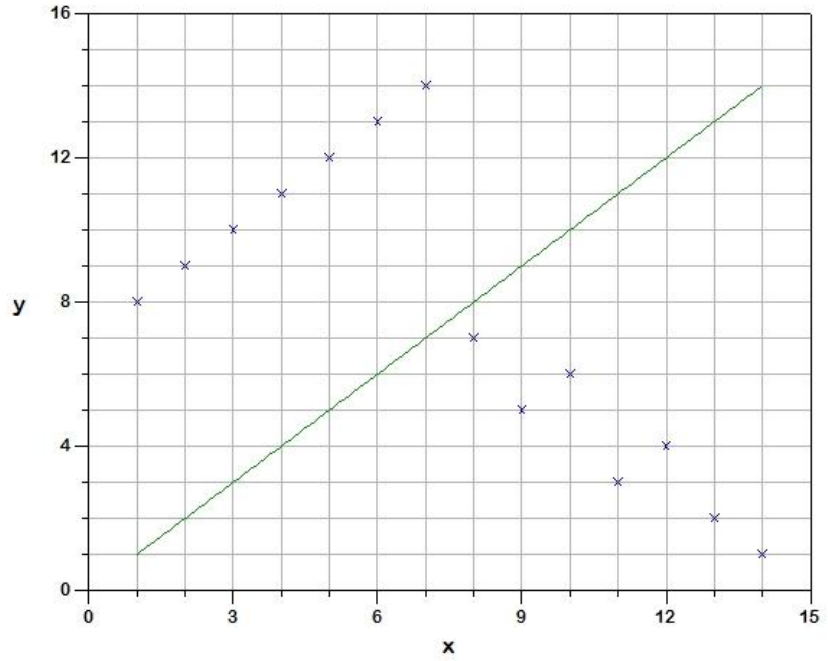


Figure 4: The two orders of segments (according to the linguistic and the logical order) are represented by the coefficients of the points.
The straight line shows the hypothetic 1:1 correspondence.

It is, of course, interesting to determine the deviation or correspondence between the two orders in general or with respect to a language or text type etc. A measure of correspondence can be based on Spearman's rank correlation coefficient ρ, which we can apply to each individual pair of orders. The closer the values to 1.0 the better the correspondence between the two sequences. Values close to zero indicate that there is no correspondence at all while negative values are obtained in cases where the two orders are more or less inverse.

As we are not so much interested in individual order-pairs of relations we will determine the distribution of the ρ-values of the 123 texts in the corpus (cf. Table 6), i.e. we determine the rank correlation coefficient for each of the pairs of orders. We have to know the distribution of the *ρ values* also because otherwise a test for significance is not possible. The Shapiro-Wilks test for normality yields $W = 0.9206$, $p = 2.208 \cdot 10^{-06}$, hence, the hypothesis that the values are normally distributed in our text sample must be rejected. Therefore, we cannot apply any test which is based on the assumption of normality and will use the distribution only to characterise the behaviour of the deviations of the two orders in the corpus.

Scrutinizing the distribution of the data after pooling them into 10 classes yields a good fit of the binomial arcus sinus distribution with parameters a =

0.4554, n = 9 and P(X^2) = 0.4787 (cf. Fig. 5). This distribution is a mixture of the binomial and the beta distributions. It is given as (cf. Wimmer/Altmann 1999: 28f)

$$P(X = x) = \int_0^1 p^x q^{n-x} \frac{1}{B(1-\alpha,\alpha)} p^{-\alpha}(1-p)^{\alpha-1} dp$$

$$= \binom{n}{x} \frac{\sin \alpha\pi}{\pi} B(x-\alpha+1, n-x+\alpha)$$

$$= \binom{\alpha+n-x-1}{n-x} \binom{x-\alpha}{x}, \; x = 0,1,2,..,n; \; n \in \mathbb{N}_0; 0 \le \alpha \le 1$$

where B(a,b) is the beta function.

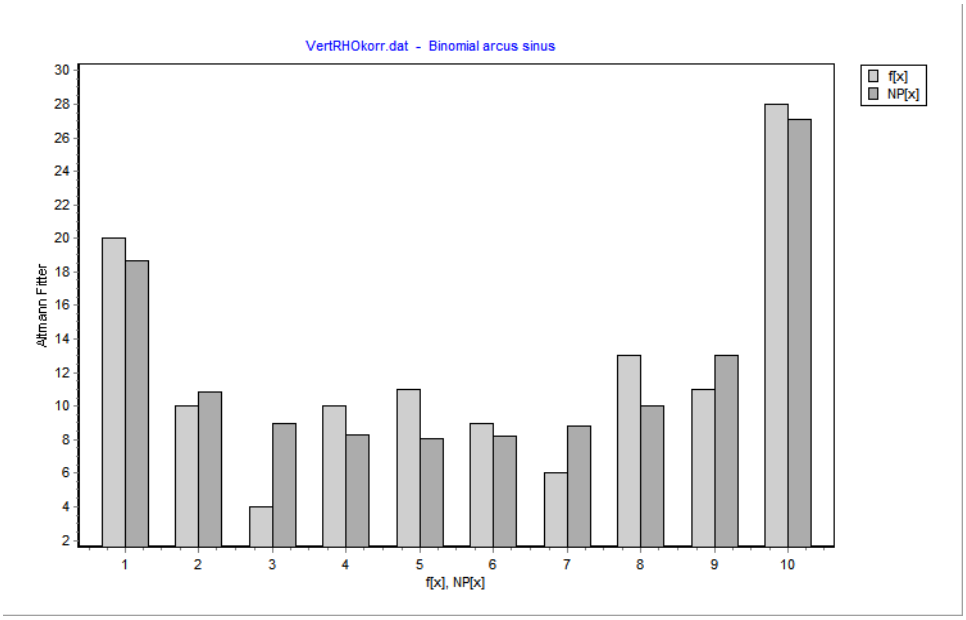


Figure 5: Fit of the binomial arcus sinus distribution to the rank correlation coefficients.

As can be seen in this figure, there is a tendency towards extremes, i.e. either direct or inverse correspondences between the linguistic and the logical order whereas moderate divergences are less frequent. This preliminary result raises new problems. In particular, no theoretical model which would lead to the binomial arcus sinus distribution can be presented by now. Therefore, questions as to whether order correspondence is characteristic of text types and similar hypotheses must be postponed until clarification of the statistical properties of our measure.

Table 6
The Spearman rank correlation coefficient for 122 texts

Text No.	ρ	**Text No.**	ρ	**Text No.**	ρ
1	0.7934	42	0.8800	83	0.6754
2	0.5486	43	0.7058	84	0.5745
3	-0.5705	44	-0.1715	85	0.2479
4	-0.1111	45	-0.3518	86	0.8255
5	-0.6800	46	0.6603	87	0.8519
6	-10.135	47	-0.1500	88	0.1731
7	0.9033	48	-0.1733	89	-0.2325
8	-0.2933	49	-0.9792	90	-0.9936
9	-0.3897	50	0.9544	91	0.8460
10	0.5733	51	0.5417	92	0.9933
11	-11.538	52	0.2036	93	0.9900
12	0.5486	53	0.6095	94	-10.833
13	-0.7093	54	-11.796	95	10.000
14	0.1429	55	-0.1615	96	0.5577
15	0.9848	56	-0.1259	97	0.4231
16	0.8600	57	0.5041	98	0.8500
17	-0.7836	58	0.1074	99	-0.6627
18	-0.9800	59	-0.1111	100	-0.3125
19	0.1986	60	-0.7396	101	-0.8245
20	-10.693	61	0.0625	102	0.9766
21	0.6388	62	0.9567	103	-0.2593
22	-0.8133	63	0.9808	104	0.8654
23	0.8582	64	-0.4295	105	-0.6529
24	0.5755	65	-0.1157	106	0.9421
25	0.3269	66	0.0577	107	-0.2223
26	-0.8293	67	0.2537	108	-0.3959
27	0.9407	68	10.000	109	-10.248
28	0.8898	69	-10.513	110	-0.4735
29	0.6945	70	0.7438	111	0.5067
30	-0.6116	71	-0.9592	112	0.7769
31	0.5704	72	-0.1260	113	0.5064
32	0.5111	73	-0.5513	114	0.9885
33	-0.7853	47	-10.980	115	0.9446
34	-12.367	75	0.7093	116	0.1714
35	-0.1385	76	-0.9338	117	0.2231
36	-0.0473	77	-14.490	118	-0.7708
37	0.6033	78	-0.3878	119	0.9556
38	-0.3673	79	0.9848	120	0.8486
39	0.0321	80	-0.8109	121	-0.9577
40	0.3052	81	0.9236	122	0.8533
41	0.1633	82	-0.7701		

Conclusion

The hypothesis that the depths of argumentation relations in their respective tree structures follow the hyper-Poisson distribution was supported by the data from the German newspaper corpus. This is, of course, only a first empirical test; far-reaching generalisations are not yet possible. Another finding, viz. the U-shaped distribution of the degree of correspondence between the logical and the linguistic orders of segments (argumentation elements) is perhaps language-specific. Follow-up studies will be conducted as soon as appropriate material is available.

References

Köhler, R., Altmann, G. (1996): "Language Forces" and synergetic modelling of language phenomena. In: *Glottometrika 15.* Trier: WVT, 62-76.

Mann W. C., Thompson S. A. (1987): Rhetorical structure theory: description and construction of text structures. In: Kempen, G. (Ed.): *Natural Language Generation*. Dordrecht: Martinus Nijhoff Publishers, 85-96.

Ord, J. K. (1972): *Families of frequency distributions.* London: Griffin.

Stede. M. (2003): Surfaces and depths in text understanding: the case of newspaper commentary. In: *Proceedings of the HLT/NAACL Workshop on Text Meaning*, Edmonton/AL.

Stede, M. (2004): The Potsdam commentary corpus. In: *Proceedings of the Workshop on Discourse Annotation, 42nd Meeting of the ACL.*

Reitter, D., Stede, M. (2003): Step by step: underspecified markup in incremental rhetorical analysis. In: *Proceedings of the 4th International Workshop on Linguistically Interpreted Corpora (LINC-03)*, Budapest.

Wimmer, G., Altmann, G. (1999): *Thesaurus of univariate discrete probability distributions*. Essen: Stamm.

Systematic stylometric differences in men and women authors: a corpus-based study

George K. Mikros, Athens

Abstract. The aim of this paper is to explore the differences existing in written texts between male and female authors. In particular we will investigate a number of well-known stylometric variables which have been used previously in authorship attribution with considerable success and we will try to expand their discriminatory power in author's gender classification.

Keywords: author profiling, stylometry, MANOVA, Discriminant Function Analysis, gender, Modern Greek.

1 Introduction

The gender differences in language understanding and production is a widely researched issue which has been investigated by a number of different disciplines, from neurophysiology and cognitive science to sociolinguistics. Our main research hypothesis in this article is that the author's gender influences the stylometric profile of the text in a systematic way. In order to explore this hypothesis we measure a wide variety of stylometric textual features in a balanced corpus of news texts written form both men and women. More specifically in the first part of this study we present the most important research findings of neurobiology related to cross-gender differentiation of linguistic ability. In the second part we examine whether author's gender leave a stylometric trace in the text. We use multivariate techniques in order to explore whether specific stylometric variables groups differ systematically across authors' gender.

2 Brain diversity between men and women

2.1 Anatomical differences in the brain

For a long time diverse behavior patterns between men and women were attributed to the unequal socio-cultural conditions existing in western societies.

However, recently a large number of studies have controlled the impact of social structure concluding that a significant amount of cross-gender variation can

be explained due to the biological differences existing in the brain. We know already that babies in their first months already present gender variation, reflecting the later differentiations observed in male and female brains (Moir & Jessel, 1992).

One of the most important gender-based brain anatomical differences is related to the corpus callosum, the tissue that connects the left and right cerebral hemispheres and facilitates interhemispheric communication. Different studies have connected this specific difference with female "intuition" (Gorman & Nash, 1992) and musical "talent" (Levitin, 2006). Recent research (Clarke et al., 2007) has confirmed its difference between sexes but its exact cognitive role has not been determined.

Another important brain difference can be found in the inferior-parietal lobule (Frederikse, Lu, Aylward, Barta, & Pearlson, 1999). This area is located over the ears and at the height of the temple and has been found to be significantly bigger in men than women. Also in men the right lobe is larger than the left, while in women this asymmetry is reversed. The right lobe is also connected to the temporary memory, a function that the brain needs to understand and manipulate spatial relationships and the ability to understand the relationships that exist between different parts of the body. It is also associated with the perception of our own emotions. The left lobe, on the other hand, is involved in the perception of time and speed and the ability of mental rotation of three-dimensional images.

2.2 Functional Magnetic Resonance Imaging – fMRI findings

One of the major developments in the field of diagnostic imaging is the development of Functional Magnetic Resonance Imaging (fMRI). This technique when applied to the brain can show in real time which parts of the cerebral cortex show increased activity by measuring the amount of circulating blood. The use of fMRI in experimental conditions with controlled stimuli can reveal which regions of the brain are associated with specific skills.

The main finding in language tests is the functional lateralization observed in fMRI of men, i.e. the exploitation of only the left lobe for processing linguistic data (Shaywitz et al., 1995). Instead, women seem to use both hemispheres of the cerebral cortex when they produce, as well as when they hear human speech.

The simultaneous use of both hemispheres in the female brain is partially explained by the larger corpus callosum, which has the female brain (see above in section 2.1) and is currently the most important biological interpretation of female superiority in language processing.

The ability of distributed language processing allows faster and more accurate processing of linguistic data (Kimura, 2000; Linn & Petersen, 1985). Instead, as a result of functional lateralization, men have twice the rates in dyslexia

(Flannery, Liederman, Daly, & Schultz, 2000) and significantly higher rates of aphasia in stroke (McGlone, 1980).

Surveys from Kimura and her associates (Kimura, 1993; Kimura & Hampson, 1994) show that from the total number of patients who suffer some kind of damage to the left hemisphere of the brain, more men (48.5%) than women (30%) show signs of aphasia. The relationship between affected brain area and gender has now been determined more accurately and we now know that hits in the left anterior cerebral cortex affect more the linguistic ability of women, while hits in the back left portion of the frontal cortex produce more frequently symptoms of aphasia in men.

fMRI studies conducted in recent years have not produced consensus related to the existence of functional lateralization in men. One of the largest meta-analysis of fMRI data (Sommer, Aleman, Bouma, & Kahn, 2004) concluded that the hypothesis of functional lateralization cannot be accepted with certainty for the general population. However, recent research results (Harrington & Farias, 2008) show that men and women activate different regions of the brain when they face specific linguistic tests, supporting the hypothesis of biological diversity of linguistic competence between sexes.

3 Text profiling studies predicting the author's gender

Information Retrieval and Text Mining research was among the first fields that tried to profile the author of a text using stylometric features and machine learning algorithms. Author profiling falls into the standard paradigm of text classification with class labels the author's gender (Argamon, Koppel, Pennebaker, & Schler, 2007; Koppel, Argamon, & Shimoni, 2002; Schler, Koppel, Argamon, & Pennebaker, 2006), age (Argamon, Koppel, Fine, & Shimoni, 2003) or psychological type (Argamon, Dhawle, Koppel, & Pennebaker, 2005; Luyckx & Daelemans, 2008a, 2008b).

One of the first studies which tried to use stylometric features to predict author's gender was from Koppel et al. (2002). They compiled a sub corpus controlled for genre from British National Corpus (BNC) which contained 566 texts written from equal number of men and women authors. They counted a wide variety of topic-neutral stylometric features including the 405 most frequent function words and the most frequent Part of Speech n-grams. The total vector size contained 1081 features which trained a variant of the Exponential Gradient algorithm. The accuracy of the author's gender prediction ranged from 79.5% in literary texts to 82.6% in non-literary texts. One of the most interesting finding was that literary texts used different features from non-literary text to mark gender. Moreover, previous findings that women and men use more frequently different Parts of Speech (pronoun and definite article correspondingly) were confirmed.

The same research group used a large corpus from blogs (37,475 blog posts totalling 300 million words) and tried to predict both the authors' gender and age (Schler et al., 2006). The specific study used 1,502 features including specific content words, selected morphological categories, function words and blogs specific features such as "blog words" - lol, haha, ur, etc. - and hyperlinks. The machine learning algorithm used was Multi-Class Real Winnow and the prediction accuracy for the author's gender reached 80.1%. Interestingly, the authors noted that despite the great diversity found among stereotyped word content usage between men and women, the most important gender distinctive features were semantically neutral (such as frequent functional words and Parts of Speech).

In another study Corney (2003) analyzed an email corpus and tried to predict the email sender's gender. He used a wide variety of stylometric features including the most frequent function words, the word and sentence length, etc. The prediction accuracy reached 70.1% and the most important gender predictors were the most frequent function words, the average word length and the letter frequencies.

Hota et al. (2006) studied the linguistic usage of men and women characters in 34 Shakespeare plays. The main research question posed was whether a male author could effectively approximate features of woman's language producing real characters and natural dialogues. The researchers used semantic neutral features (frequent functional words, numbers, prepositions, contracted word forms) and content words with high frequency (greater than 10) in the corpus. Prediction accuracy ranged from 60% to 75% depending on the features used. The authors interpret the somewhat less precise gender identification, in the fact that Shakespeare although intuitively approached the language of women characters, failed to deliver it in its entirety.

4 Methodology

4.1 Corpus description

A serious problem related with the corpora used in the authorship attribution studies is the lack of their homogeneity. As Rudman (1997) states the most striking deficiencies are:

- The improper selection, unavailability or fragmentation of the texts.
- The text normalization that often applies from the editor or the publisher causing serious distortion in the writer's style.
- The differences observed in many cases between training and cross-validated texts in terms of genre, topic, date and medium.

Linguistic variation extends across text genre, topic and medium. The linguistic boundaries between these categories are obscure and the linguistic structures exhibit frequencies which co-vary with topic and genre, or medium and topic. Even the most abstract stylometric variables exhibit significant correlation with text metadata such as topic and genre. Examples of this correlation can be found in Mikros & Argiri (2007) who examined the effect of the topic in the authorship information carried by several stylometric variables widely used in authorship research. The study demonstrated that many stylometric variables can be used with success in topic classification. This characteristic is highly undesirable especially in cases where the researcher attempts authorship attribution in corpora where topic and other textual metadata have not been taken into account.

For the needs of our research we developed a corpus which was controlled simultaneously for the author's gender, text topic, genre and medium. More specifically the corpus design was based on the following premises:

- Equal number of texts (50) written by male and female authors.
- Each text from a male author with specific topic and genre should be matched by a text in the same topic and genre from a female author.
- All texts should be published in the same newspaper (Eleftherotypia) in a brief time span (1 year).
- The collected texts should belong to many different and distinct topics and genres in order to represent a wide socio-pragmatic space of language usage.

The resulting corpus contains 700 texts equally divided in 7 male and 7 female authors. Although, there are some small differences between specific topic and genre categories in spite of the strict sampling restrictions described above, the corpus should be considered balanced. Its size in words (W) and number of texts (N) is displayed in Table 1.

The specific corpus aims to form a difficult challenge for stylometric analysis. It is highly homogeneous regarding its textual metadata and additionally contains small size texts which is untypical of most corpora used in stylometry. In particular, 84% of the texts have less than 1,000 words and this poses a further difficulty since most stylometric variables exhibit authorship quantitative patterns in larger text sizes (Baillie, 1974; Ledger & Merriam, 1994).

The texts were obtained using "Minotauros" a tool for creating corpora from web sources (Koutsis, Kouklakis, Mikros, & Markopoulos, 2005). Tokenization and Part of Speech tagging was performed by 'Ellogon' a multi-lingual, cross-platform, general-purpose language engineering environment, developed from the Institute of Informatics and Telecommunications, NCSR "Demokritos" (Petasis, Karkaletsis, Paliouras, Androutsopoulos, & Spyropoulos, 2002). Measurements of various stylo-

metric variables were made using "Corpus Manager" (Kouklakis, Mikros, Markopoulos, & Koutsis, 2007) as well as specialized PERL scripts.

Table 1
Corpus size breakdown by author's gender, text topic and genre

	Topic / **Genre**	**Science**		**Society**		**Economy**		**Art**		**Total**	
		N	W	N	W	N	W	N	W	N	W
Female	Opinion	7	3,169	84	60,489	14	5,748	17	14,568	**122**	**83,974**
	News	11	6,308	111	77,811	31	16,982	15	10,865	**168**	**111,966**
	Discourse	8	4,999	33	23,071	2	1,646	17	21,710	**60**	**51,426**
	Subtotal	*26*	*14,476*	*228*	*161,371*	*47*	*24,376*	*49*	*47,143*	***350***	***247,366***
Male	Opinion	8	3,847	88	60,283	14	8,453	17	11,301	**127**	**83,884**
	News	17	8,353	117	68,793	32	20,471	16	11,023	**182**	**108,640**
	Discourse			22	20,595	2	1,736	17	17,218	**41**	**39,549**
	Subtotal	*25*	*12,200*	*227*	*149,671*	*48*	*30,660*	*50*	*39,542*	***350***	***232,073***
	Total	**51**	**26,676**	**455**	**311,042**	**95**	**55,036**	**99**	**86,685**	**700**	**479,439**

4.2 Feature sets

In this study we measured six broad sets of stylometric features which contain both lexical and sublexical units. Each set groups a number of variables which function complementarily and all together approximate a specific textual construct. Although the listing is not exhaustive, it contains most of the variables that have been employed in modern stylometric research and we consider them as socio-linguistically neutral. All the features used in this study are the following:

1. *Lexical "richness"*
   a) Yule's K: Vocabulary richness index that exhibits stability in different text sizes (Tweedie & Baayen, 1998).
   b) Lexical Density: The ratio of functional to content words frequencies in the text, also known as Functional Density (Miranda & Calle, 2007).
   c) % of Hapax- and Dis-legomena: The percentage of words with frequency 1 and 2 in the text segment.
   d) Dis-/Hapax-legomena: The ratio of dis-legomena to hapax-legomena in the text segment, indicative of authorship style (Hoover, 2003).

e) Relative entropy: Is defined as the ratio between the text entropy and its maximum entropy multiplied by 100. Maximum entropy for a text is calculated if we assume that every word appears with frequency 1 (Oakes, 1998, p. 62).

f) Word rareness: Percentage of words in each text which do not belong to the 5,000 and the 10,000 most frequent words of Modern Greek.

2) *Word length*
   a) Average word length (per text) measured in letters.
   b) Word length distribution: The frequency of words of 1, 2, 3 … 14 letters long normalized in 1,000 words sample.
3) *Sentence length*
   a) The average sentence length measured in words.
   b) The percentage of long sentences (>18 words) in each text.
4) *Character frequencies*
   The frequency of each letter in the text segment normalized in 1,000 words sample. We measured in total 31 letters (we calculated separately the frequencies of the stressed and the unstressed vowels since in Modern Greek spelling the stressed vowels have stress marked orthographically, thus representing different grapheme).
5) *Part of Speech frequencies*
   The frequency of each Part of Speech tag, expressed as percentage of the text size.
6) *Frequent Function Words (FFW)*
   The frequency of the 50 most frequent function words of Modern Greek normalized in 1,000 words sample.

4.3 Statistical analysis

Gender effect analysis in linguistic production requires multivariate methods. In order to examine in detail the way each of the six variable sets relates to author's gender we used Multiple Analysis of Variance (MANOVA). MANOVA is a multivariate statistical analysis which is specifically designed to analyze the effect of one or more categorical independent variables on two or more continuous dependent variables. Although the problem could be tackled with multiple univariate tests, the overall Type I error will be inflated and the probability to reject the null hypothesis when it is true is increased. MANOVA controls against Type I error and offers an omnibus test of significance that takes into account the effect of the independent variable(s) to all the dependent variables simultaneously (Weinfurth, 1995). Furthermore, MANOVA is particularly useful if the dependent variables are conceptually related and there is a moderate inter-correlation between them. Since each variable

group contains stylometric variables that attempt to measure the quantitative expression of a specific textual construct, we expect a certain amount of redundancy. MANOVA takes into account this shared common information and tests the effect of the independent variable(s) in a multivariate way (i.e., taking all dependent variables at once). Another reason why a multivariate approach is preferable in our data is that it can detect differences when groups differ on a system of variables (Huberty & Morris, 1989). MANOVA finds a linear composite of the dependent variables that maximizes the separation of the categories that form the independent variable.

A non-significant MANOVA result means that the specific set of dependent variables examined simultaneously do not differ across the categories of the independent variable and no further analysis should be made. A significant MANOVA however indicates that at least one dependent variable differs significantly across the categories of the independent variable. In the relevant literature most researchers perform univariate tests (t-tests in our case) with adjusted alpha level (Bonferroni correction) for each of the dependent variables in order to detect which variable is different between the categories of the independent variable (Hair Jr, Anderson, Tatham, & Black, 1995; Stevens, 2002). This procedure however has been criticized (Bray & Maxwell, 1982; Huberty & Morris, 1989) among others for confusing the univariate with the multivariate research questions. Since gender and language structure interact in complex and multilevel ways we chose to further explore significant MANOVAs with Discriminant Function Analysis (DFA). Conducting DFA following a significant multivariate effect allows the researcher to investigate in detail the linear composites of the dependent variables and to determine their structure as well as the weights of each dependent variable (Meyers, Gamst, & Guarino, 2006).

5 Results

5.1 Feature group importance

In our data we performed separated MANOVAs for each one of the six stylometric groups with independent variable the author's gender. The multivariate statistic we calculated was Hotelling T^2 which is the multivariate counterpart of the univariate t statistic. Furthermore, partial η^2 was calculated indicating the percentage of the variance explained by the combined dependent variables.

Table 2 summarizes the MANOVA results in the six variable groups.

Table 2
Ranking of the feature groups based on their explanatory power (Partial η^2) in the author's gender

Feature Groups	**Hotelling T^2**	**p**	**Partial η^2**
Frequent Function	416.008	0.000	0.374
Character Frequencies	252.676	0.000	0.266
Word Length	101.908	0.000	0.128
Part of Speech	59.33	0.000	0.078
Lexical "richness"	17.45	0.032	0.024
Sentence length	2.792	0.211	0.004

As can be seen all stylometric groups had a multivariate statistically significant effect in author's gender except sentence length. The group that accounts for the biggest amount of variance is Frequent Function Words (37%) followed by Character Frequencies (27%) and Word Length (13%). Small (< 10%) but statistically significant amount of explained variance in author's gender have the Part of Speech frequencies and the Lexical "richness". For each of the five variable groups that Hotelling T^2 was found statistically significant we performed DFA in order to further explore the linear composite structure and to assess each dependent's variable contribution to author's gender discrimination.

5.2 Frequent function words DFA analysis

The DFA with Frequent function words variables as independent and author's gender as dependent showed that 28 variables differ significantly between the male and the female authors.

Table 3 and all subsequent tables in next sections summarize the DFA results. In particular each table presents the Wilk's λ, of each statistically significant predictor, mean (M) and standard deviation (SD) in male ($_M$) and female ($_F$) authors as well as the within-groups correlations between the predictors and the discriminant function. Furthermore, standardized weights for each variable are reported in order to assess their relative importance in author's gender discrimination.

In Table 3 we see that male authors use more frequently the words (with decreasing importance in the author's gender discrimination) όμως [instead], αλλά [but], στην [in], σ'[contracted form of 'in'], ο [male singular article], απ' [contracted form of 'from'] , τη [female singular article], της [female singular article in genitive

case], την [female singular article or female personal pronoun], με [with], που [where], η [female singular article] while female authors present higher percentages in the use of μας [us], των [article in genitive case], το [neutral singular article], δεν [not], σε [in], οι [plural article], μόνο [just], μέσα [inside], πως [how], σου [personal pronoun in genitive], τους [them], γιατί [why], τα [neutral plural article], πάνω [on], στα [in], από [from]. Among the words that males use more frequently we can group two distinct categories: a) coordinated conjunctions (αλλά, όμως) and b) contracted forms of prepositions (σ', απ'). The former category characterizes the syntactic structure of the text and previous research has revealed that can be used as a potential gender discriminator (Mulac, Bradac, & Mann, 1985; Mulac, Studley, & Blau, 1990). The latter grouping (contracted forms) has also been described by many linguists as a typical male marker in text production (Baron, 2004).

Table 3
Ranking of frequent function words based on their overall usefulness (absolute value of the standardized coefficient) in the authors' gender differentiation

Predictors	Wilk's λ	p	MM	SDM	MF	SDF	Correlation Coeff.	Stand. Coeff.
μας	0.953	0	0.143	0.247	0.306	0.459	0.286	0.443
όμως	0.908	0	0.282	0.271	0.136	0.178	-0.411	-0.427
των	0.973	0	1.002	0.689	1.274	0.917	0.217	0.381
το	0.975	0	2.055	0.88	2.335	0.87	0.208	0.370
δεν	0.986	0.002	0.775	0.526	0.912	0.615	0.156	0.347
σε	0.973	0	0.731	0.47	0.89	0.48	0.216	0.346
οι	0.977	0	0.809	0.525	0.995	0.68	0.198	0.291
αλλά	0.993	0.03	0.293	0.264	0.251	0.244	-0.106	-0.214
στην	0.968	0	0.887	0.521	0.708	0.462	-0.235	-0.192
μόνο	0.986	0.002	0.104	0.15	0.142	0.172	0.152	0.161
μέσα	0.985	0.001	0.076	0.131	0.112	0.158	0.159	0.159
σ'	0.994	0.045	0.074	0.136	0.049	0.179	-0.099	-0.155
ο	0.988	0.004	1.372	0.774	1.206	0.743	-0.142	-0.130
πως	0.991	0.013	0.081	0.179	0.122	0.248	0.122	0.107
σου	0.994	0.041	0.014	0.059	0.027	0.112	0.100	0.096
απ'	0.992	0.021	0.041	0.101	0.025	0.074	-0.113	-0.082
τους	0.976	0	0.678	0.532	0.854	0.594	0.202	0.075
τη	0.991	0.011	0.824	0.439	0.74	0.441	-0.125	-0.075
της	0.956	0	2.204	0.977	1.794	0.945	-0.276	-0.059
γιατί	0.994	0.045	0.105	0.178	0.132	0.178	0.098	0.052
τα	0.981	0	0.908	0.575	1.092	0.737	0.18	0.046

πάνω	0.99	0.008	0.043	0.099	0.067	0.136	0.129	0.046
την	0.972	0	1.778	0.763	1.534	0.676	-0.219	-0.044
με	0.991	0.01	1.54	0.764	1.407	0.591	-0.127	-0.039
που	0.982	0	1.768	0.61	1.603	0.604	-0.176	-0.033
στα	0.992	0.015	0.285	0.289	0.345	0.351	0.119	0.024
η	0.958	0	1.792	0.882	1.459	0.703	-0.271	0.021
από	0.989	0.006	1.313	0.605	1.444	0.645	0.136	0.004

Instead, in the words that characterize women authors we can distinguish the presence of personal pronouns (μας, σε, σου). The preference of personal pronoun usage has been confirmed by previous corpus-based studies (Argamon et al., 2007; Holmes, 1990; Preisler, 1986; Rayson, Leech, & Hodges, 1997) and is related to the fact that female discourse is characterized by interpersonal involvement. This has also been described by Tannen (1991) as the "report vs. rapport" distinction, i.e. the female speaker/author's tendency to produce texts that concentrate on interaction with her readers/listeners and maintain their relationship while males focus on the information transmission.

5.3 Character frequencies DFA analysis

The DFA with character frequencies as independent variables and author's gender as dependent showed that 12 variables differ statistically significant between the male and the female authors. Table 4 presents the analysis findings.

Table 4
Ranking of character frequencies based on their overall usefulness (standardized coefficient) in the authors' gender differentiation

Predictors	Wilk's λ	p	M_M	SD_M	M_F	SD_F	Correlation Coeff.	Stand. Coeff.
η	0.961	0.000	3.937	0.731	3.643	0.741	0.333	1.241
κ	0.967	0.000	4.038	0.556	3.833	0.563	0.305	1.043
ρ	0.938	0.000	4.473	0.508	4.197	0.567	0.426	1.033
λ	0.979	0.000	2.765	0.482	2.627	0.473	0.240	0.870
ί	0.987	0.003	2.501	0.387	2.419	0.346	0.187	0.816
ό	0.993	0.028	2.091	0.402	2.029	0.348	0.138	0.777
ο	0.977	0.000	7.871	0.859	8.139	0.922	-0.250	0.749
α	0.983	0.000	9.246	0.871	9.476	0.887	-0.218	0.558
ζ	0.988	0.004	0.358	0.197	0.318	0.168	0.181	0.475

δ	0.990	0.011	1.712	0.364	1.788	0.423	-0.159	0.359
γ	0.973	0.000	1.703	0.402	1.829	0.369	-0.273	0.320
ω	0.992	0.020	1.430	0.338	1.492	0.366	-0.146	0.274
φ	0.978	0.000	0.801	0.238	0.882	0.307	-0.245	0.144

In Table 4 we can see that male authors use more frequently (with decreasing importance in the author's gender discrimination) the characters η, κ, ρ, λ, ί, ό, ζ. On the other hand female authors use more frequently the characters ο, α, δ, γ, ω, φ. The relative importance of each character was determined by its standardized coefficient.

5.4 Word length DFA analysis

The DFA with Word length as independent variable and author's gender as dependent showed that seven variables differ significantly between the male and the female authors. Table 5 summarized the analysis findings.

Table 5
Ranking of word lengths based on their overall usefulness (absolute value of the standardized coefficient) in the authors' gender differentiation

Predictors	Wilk's λ	p	M_M	SD_M	M_F	SD_F	Correlation Coeff.	Stand. Coeff.
2 letter words	0.967	0.000	11.361	2.079	12.131	2.073	0.487	0.869
3 letter words	0.994	0.046	23.156	2.208	23.498	2.331	0.198	0.720
13 letter words	0.983	0.000	1.100	0.613	1.270	0.667	0.349	0.442
14 letter words	0.991	0.011	0.704	0.471	0.799	0.517	0.252	0.367
8 letter words	0.976	0.000	7.402	1.488	6.932	1.488	-0.414	-0.086
4 letter words	0.984	0.000	10.585	2.045	10.050	2.163	-0.334	-0.040
9 letter words	0.987	0.003	6.188	1.562	5.842	1.513	-0.295	0.009
10 letter words	0.991	0.012	5.317	1.400	5.052	1.408	-0.247	0.003

In Table 5 we see that female authors use greater percentage of 2, 3, 13 and 14 letter words while male authors present higher percentages in the use of 4, 8, 9 and 10 letter words. The examination of the standardized coefficients shows that the most useful markers for the detection of female writing is the percentage of 2 and 3 letter words followed by the percentage of 13 and 14 letter words. Correspondingly, the most useful markers for the detection of the male writing is the percentage of 8 and

4 letter words followed by the percentage of 9 and 10 letter words. These results give us a relative clear picture regarding female writing in news and word length. Female authors use more than males the lower and upper boundary of the word length spectrum. They use smaller words (2-3 letter words) which in their majority in Modern Greek belong to the group of function words. They use also many words which have many letters (13 and 14 letter words) and they are related inversely to function word usage. Since the 13 and 14 letter words have relatively small effect on gender discrimination compared to the 2 and 3 letter words, we can hypothesize that they reflect inversely the major trend of the small words to characterize women's writing. This hypothesis is further supported by examining the correlations of 2, 3, 13 and 14 letter words with Lexical density in the female data. 2 and 3 letter words appear to be in a statistically significant negative correlation (r_{2lw}= -0.354, r_{3lw} = -0.385) with the lexical density, meaning that increase in lexical density (i.e. more content words) relates inversely to the percentage of 2 and 3 letter words. On the other hand, 13 and 14 letter words have smaller but statistically significant positive correlation (r_{13lw} = 0.134, r_{14lw} = 0.144) with lexical density, meaning that as lexical density increases the percentage of longer words increases also but with smaller pace.

5.5 Part of Speech frequencies DFA analysis

The DFA with Part of Speech frequencies as independent variables and author's gender as dependent showed that only the usage of Adverbs (Wilk's $\lambda = 0.987$, $p = 0.003$) and the usage of Adjectives (Wilk's $\lambda = 0.995$, $p = 0.049$) present statistically significant differences between male and female authors. More specifically male authors use increased percentage of adverbs (M = 8.2, SD = 1.9) compared to female authors (M = 7.8, SD = 1.7) and female authors use increased percentage of adjectives (M = 8.2, SD = 2.2) compared to male authors (M = 7.9, SD = 1.9).

Adverb usage demonstrated strong relationship with the discriminant function with correlation coefficient -0.403 and standardized coefficient -0.435 whereas adjective usage exhibited weaker association with correlation coefficient 0.256 and standardized coefficient 0.464.

5.6 Lexical "richness" DFA analysis

The DFA with Lexical "richness" as independent variables and author's gender as dependent showed that only the Percentage of hapax legomena (Wilk's $\lambda = 0.993$, $p = 0.025$) present statistically significant differences between male and female authors. More specifically, male authors have higher percentage of hapax legomena (M = 41.2, SD = 7.6) compared to female authors (M = 39.9, SD = 7.2).

Percentage of hapax legomena demonstrated strong correlation with the discriminant function with correlation coefficient 0.542 and standardized coefficient 0.875.

A closer inspection of the association of the lexical "richness" variables with the author's gender reveals a complex and heterogeneous picture that is characteristic of the complexity of the relationship between author's gender and textual stylometric profile. Although Relative entropy and the percentage of words which do not belong to the most frequent 5000 words of the corpus theoretically measure the same abstract textual property, i.e. lexical "richness", appear to be inversely related to author's gender. Women write texts with rare vocabulary while men's texts present less lexical repetition and avoidance of standardized lexical patterns.

6 Conclusions

The present study investigated the role of the author's gender in the systematic differentiation observed in the stylometric profile of texts of men and women authors. Using a corpus compiled in a way to experimentally control text topic, genre and medium we studied a wide array of stylometric features and their usage distribution in men and women's texts. Multivariate statistical analysis (MANOVA followed by Discriminant Function Analysis) revealed that men and women use indeed differently most stylometric features, a fact that can be further exploited for the development of author's gender profiling systems.

References

Argamon, Shlomo; Dhawle, Sushant; Koppel, Moshe; Pennebaker, James (2005). Lexical predictors of personality type *Proceedings of the 2005 Joint Annual Meeting of the Interface and the Classification Society of North America: Theme: Clustering and Classification, 8-12 Jun 2005*. St. Louis, MO.

Argamon, Shlomo; Koppel, Moshe; Fine, Jonathan; Shimoni, Anat Rachel (2003). Gender, genre, and writing style in formal written texts. *Text, 23*(3), 321-346.

Argamon, Shlomo; Koppel, Moshe; Pennebaker, James W.; Schler, Jonathan (2007). Mining the blogosphere: Age, gender and the varieties of self–expression. *First Monday,12*(9). Retrieved from http://firstmonday.org/htbin/cgiwrap/bin/ojs/index. php/fm/article/view/2003/1878

Baillie, D. W. (1974). Authorship attribution in Jacobean dramatic texts. In J. L. Mitchell (Ed.), *Computers in the humanities*. Edinburgh: Edinburgh University Press.

Baron, Naomi S. (2004). See you Online. *Journal of Language and Social Psychology, 23*(4), 397-423. doi: 10.1177/0261927x04269585

Bray, James H.; Maxwell, Scott E. (1982). Analyzing and Interpreting Significant MANOVAs. *Review of Educational Research, 52*(3), 340-367. doi: 10.3102/00346543052003340

Clarke, Dave; Wheless, James; Chacon, Monica; Breier, Joshua; Koenig, Mary-Kay; McManis, Mark; Baumgartner, James (2007). Corpus callosotomy: A palliative therapeutic technique may help identify resectable epileptogenic foci. *Seizure, 16*(6), 545-553.

Corney, Malcolm Walter (2003). *Analysing e-mail text authorship for forensic purposes.* (Master), Queensland University of Technology, Queensland.

Flannery, Kathleen A.; Liederman, Jacqueline; Daly, Louise; Schultz, Jennifer K. (2000). Male prevalence for reading disability is found in a large sample of black and white children free from ascertainment bias. *Journal of the International Neuropsychological Society, 6*(4), 433-442.

Frederikse, Melissa E.; Lu, Angela; Aylward, Elizabeth; Barta, Patrick; Pearlson, Godfrey (1999). Sex Differences in the Inferior Parietal Lobule. *Cereb. Cortex, 9*(8), 896-901. doi: 10.1093/cercor/9.8.896

Gorman, Christine; Nash, Madeleine (1992, 20 January 1992). Sizing up the sexes. *TIME,* 36-43.

Hair Jr. Joseph F.; Anderson, Rolph E.; Tatham, Ronald L.; Black, William C. (1995). *Multivariate data analysis* (5th ed.). Upper Saddle River, NJ, USA: Prentice-Hall.

Harrington, Greg S.; Farias, Sarah Tomaszewski (2008). Sex differences in language processing: Functional MRI methodological considerations. *Journal of Magnetic Resonance Imaging, 27*, 1221-1228.

Holmes, Janet (1990). Hedges and boosters in women's and men's speech. *Language and Communication, 10*(3), 185-205.

Hoover, David (2003). Another perspective on vocabulary richness. *Computers and the Humanities, 37*, 151-178.

Hota, Sobhan R.; Argamon, Shlomo; Koppel, Moshe; Zigdon, Iris (2006). *Performing gender: Automatic stylistic analysis of Shakespeare's characters.* Paper presented at the Proceedings of Digital Humanities 2006, Paris.

Huberty, Carl J.; Morris, John D. (1989). Multivariate analysis versus multiple univariate analyses. *Psychological Bulletin, 105*(2), 302-308.

Kimura, Doreen (1993). *Neuromotor mechanisms in human communication*. Oxford: Oxford University Press.

Kimura, Doreen (2000). *Sex and cognition*. Cambridge, MA: MIT Press.

Kimura, Doreen; Hampson, Elizabeth (1994). Cognitive pattern in men and women is influenced by fluctuations in sex hormones. *Current Directions in Psychological Science, 3*(2), 57-61.

Koppel, Moshe; Argamon, Shlomo; Shimoni, Anat Rachel (2002). Automatically categorizing written texts by author gender. *Literary and Linguistic Computing, 17*(4), 401-412.

Kouklakis, George; Mikros, George K.; Markopoulos, George; Koutsis, Ilias (2007). *Corpus Manager: A tool for multilingual corpus analysis*. Retrieved from http://ucrel.lancs.ac.uk/publications/CL2007/paper/244_Paper.pdf.

Koutsis, Ilias; Kouklakis, George; Mikros, George K.; Markopoulos, George (2005). *MINOTAVROS: A tool for the semi-automated creation of large corpora from the Web*(Vol.1). Retrieved from http://www.corpus.bham.ac.uk/PCLC/minotavros.doc.

Ledger, Gerard; Merriam, Thomas (1994). Shakespeare, Fletcher, and the Two Noble Kinsmen. *Literary and Linguistic Computing, 9*, 235-248.

Levitin, Daniel (2006). *This is your brain on music: The science of a human obsession*. New York: Dutton Adult.

Linn, Marcia C.; Petersen, Anne C. (1985). Emergence and characterization of sex differences in spatial ability: a meta-analysis. *Child Development, 56*, 1479-1498.

Luyckx, Kim; Daelemans, Walter (2008a). Personae: A corpus for author and personality prediction from text. In Nicoletta Calzolari, Khalid Choukri, Bente Maegaard, Joseph Mariani, Jan Odjik, Stelios Piperidis & Daniel Tapias (Eds.), *Proceedings of the Sixth International Language Resources and Evaluation (LREC'08), 28-30 May 2008*. Marrakech, Morocco.

Luyckx, Kim; Daelemans, Walter (2008b). Using syntactic features to predict author personality from text *Proceedings of Digital Humanities 2008 (DH 2008)* (pp. 146-149).

McGlone, Jeanette (1980). Sex differences in human brain organization: a critical survey. *Behavioral Brain Science, 3*, 215-227.

Meyers, Lawrence S.; Gamst, Glenn; Guarino, A.J. (2006). *Applied multivariate research. Design and interpretation*. Thousand Oaks, CA: Sage.

Mikros, George K.; Argiri, Eleni K. (2007). Investigating topic influence in authorship attribution. In Benno Stein, Moshe Koppel & Efstathios Stamatatos (Eds.), *Proceedings of the SIGIR 2007 International Workshop on Plagiarism Analysis, Authorship Identification, and Near-Duplicate Detection* (Vol. 276, pp. 29-35). Amsterdam, Netherlands: CEUR.

Miranda, García Antonio; Calle, Martín Javier (2007). Function words in authorship attribution studies. *Literary and Linguistic Computing, 22*(1), 49-66.

Moir, Anne; Jessel, David (1992). *Brain sex: The real difference between men and women*. New York: Delta.

Mulac, Anthony; Bradac, James J.; Mann, Susan Karol (1985). Male/female language differences and attributional consequences in children's television. *Human Communication Research, 11*(4), 481-506.

Mulac, Anthony; Studley, Lisa B.; Blau, Sheridan (1990). The gender-linked language effect in primary and secondary students' impromptu essays. *Sex roles, 23*(9-10), 439-470.

Oakes, Michael P. (1998). *Statistics for corpus linguistics*. Edinburgh: Edinburgh University Press.

Petasis, Georgios; Karkaletsis, Vangelis; Paliouras, Georgios; Androutsopoulos, Ion; Spyropoulos, Constantine, D. (2002, 29-31 May 2002). *Ellogon: A new text engineering platform.* Paper presented at the Proceedings of the third International Conference on Language Resources and Evaluation (LREC 2002), Las Palmas de Gran Canaria, Spain.

Preisler, Bent (1986). *Linguistic sex roles in conversation: Social variation in the expression of tentativeness in English*. Berlin: Mouton de Gruyter.

Rayson, Paul; Leech, Geoffrey; Hodges, Mary (1997). Social differentiation in the use of English vocabulary: some analyses of the conversational component of the British National Corpus. *International Journal of Corpus Linguistics, 2*(1), 133-152.

Rudman, Joseph (1997). The state of authorship attribution studies: some problems and solutions. *Computers and the Humanities, 31*(4), 351-365.

Schler, Jonathan; Koppel, Moshe; Argamon, Shlomo; Pennebaker, James (2006). *Effects of age and gender on blogging.* Paper presented at the Proceedings of the 2006 AAAI Spring Symposium on Computational Approaches for Analyzing Weblogs.

Shaywitz, Bennet A.; Shaywitz, Sally E.; Pugh, Ken R.; Constable, Todd R.; Skudlarski, Pawel; Fulbright, Robert K.; Gore, John C. (1995). Sex differences in the functional organization of the brain for language. *Nature, 373*, 607-609. doi: 10.1038/373607a0

Sommer, Iris; Aleman, André; Bouma, Anke; Kahn, René (2004). Do women really have more bilateral language representation than men? A meta-analysis of functional imaging studies. *Brain, 127*(8), 1845-1852. doi: 10.1093/brain/awh207

Stevens, James P. (2002). *Applied multivariate statistics for the social sciences* (4rth ed.). Hillsdale, NJ: Erlbaum.

Tannen, Deborah (1991). *You just don't understand: Women and men in conversation*. London: Virago Press.

Tweedie, Fiona J., & Baayen, Harald R. (1998). How variable may a constant be? Measures of lexical richness in perspective. *Computers and the Humanities, 32*(5), 323-352.

Weinfurth, Kevin P. (1995). Multivariate Analysis of Variance. In Laurence G. Grim & Paul R. Yarnold (Eds.), *Reading and understanding multivariate statistics* (pp. 245-276). Washington, DC: American Psychological Association.

Word length: aspects and languages

Ioan-Iovitz Popescu, Bucharest
Sven Naumann, Trier
Emmerich Kelih, Vienna
Andrij Rovenchak, Lviv
Haruko Sanada, Tokyo
Anja Overbeck, Göttingen
Reginald Smith, Rochester
Radek Čech, Olomouc
Panchanan Mohanty, Hyderabad
Andrew Wilson, Lancaster
Gabriel Altmann, Lüdenscheid

Abstract. The article presents some evaluations of four aspects of word length in different languages and compares both models and data: distribution, smoothness, word length in sentence, word length in text. A general discussion of the theoretical background is offered. As is shown, even these four problems require great teams, in order to bring elementary concepts and decisions. The number of co-authors should not surprise: everybody did what (s)he could.

0. Introduction

The word has as many properties as we are able to establish conceptually. Some of these properties supported the rise of different disciplines like semantics, morphology, dialectology, historical linguistics, etc. Some laws of behaviour of these properties are known, but the number of surprises - possibly concealed behind the concept of boundary conditions - is still greater. Here nothing helps but incessant testing, modelling, different viewing of data, modification of hypotheses, collecting of data from new languages, etc. Every "new" language can falsify a beloved theory or force us to modify it. Here, it is not possible to take into account all known properties, we necessarily must restrict ourselves to some selected aspects.

1. Length distributions

If somebody processes the enormous discipline of word problems, his first sight falls on the word length. This property has been treated since the 19th century and belonged for a long time to the preferred research domain of K.-H. Best who maltreated his students misusing them for evaluating word length in about 50 languages using about 4000 texts. It has been shown that both the models themselves vary and even in one language the parameters of identical distributions are different in different texts. This is caused by the boundary conditions sticking to

every text sort, every author, every age, education, etc. Those who tested the law-like nature of word length distribution used a general model and a software that iteratively computed the parameters and automatically performed the chi-square test for goodness-of-fit. However, there were several critical cases in which one special class of words had to be treated separately; for example in Slavic languages there are words of zero syllabic length (some prepositions) which being synsemantics display an excessive frequency. However, if one considers them as proclitics, the problem of modification of the given distribution disappears - but new problems arise.

For evaluating word length one cannot use a different measurement unit but the number of syllables. It is the only way securing generality, because it can be used in all languages. Other candidates, e.g. number of letters, graphemes, phonemes, moras, morphemes, signs, have all their intrinsic problems: there are no letters in some languages; graphemes would be more appropriate but just as phonemes or letters they omit one level, i.e. they are not immediate constituents of words, hence the distributions may be distorted (rather fractal); morphemes do not measure length but complexity (cf. e.g. introflection). In Japanese one works also with moras but for every mora one can securely state in speech whether it is one or two syllables. Thus if we want to approach the mechanism of word formation from the general point of view, we must use syllables as counting units. Besidess, everything that is written belongs to the secondary language.

There is a long bibliographical list of model fittings (cf. Best 1997, 2001; Grzybek 2006; Schmidt 1996) but no language law will ever be fully corroborated, simply because nobody can evaluate all languages or all necessary texts even in one language. The other problem, viz. the finding of boundary conditions, will for ever incriminate the clear conscience of quantitative linguists.

The following distributions have been tested: binomial, modified binomial, extended positive binomial, geometric, Cohen-negative binomial, Cohen-Poisson, Consul-Jain-Poisson, Conway-Maxwell-Poisson, Dacey-Poisson, Fucks-Poisson, Fucks-Gačečiladze-Poisson, geometric, Hirata-Poisson, hyper-Pascal, hyper-Poisson, lognormal, Meyer-Thomas, mixed Poisson, modified negative binomial, modified Poisson, negative binomial, Palm-Poisson, Pandey-Poisson, Poisson, Poisson-uniform, Pólya, Singh-Poisson. Many of them can be derived from a general model, in turn modified, mixed (or compounded) or (Feller-)generalized but it is evident that one must find the "causes" of modifications or generalizations.

The languages from which samples have been taken up to now are as follows: Arabic, Belorussian, Bulgarian, Cheremis, Chinese, Croatian, Czech, Danish, Dutch, Early English, English, Estonian, Eskimo, Faroese, Finnish, French, Gaelic, German, Hindi, Middle German, Swiss, Early New High German, German dialects, Gothic, Greek, Hindi, Hungarian, Irish, Italian, Japanese, Kechua, Korean, Latin, Latvian, Lituanian, Low German, Lower Sorbian, Malayalam, Maori, Marathi, Mordvinian, New High German, New Icelandic, Nor-

wegian, Old Church Slavic, Old High German, Old Greek, Old Hebraic, Odia (earlier Oriya), Old Icelandic, Persian, Polish, Portuguese, Russian, Saami, Serbocroatian, Slovak, Slovenian, Spanish, Swedish, Tamil, Telugu, Turkish, Ukrainian, Usbek, Vogul, Welsh, Yiddish and yearly new languages are added. A comparative within-language and between-language treatment is still missing; this enormous field cannot be captured completely even in a team-work.

This short survey shows that Indo-European languages are better represented than all the other ones, hence one can be sure that the results are somewhat skewed. Many articles try to master the boundary conditions simply by introducing new distributions - as far as the software at our disposal (*Fitter*) contains them.

In the centre of modelling one finds the Poisson distribution which can be derived in many different ways. It is very simple, and if we believe in pure chance in length ordering, it is sufficient. But if the damned boundary conditions intervene, we must clutch at a straw and search for a remedy. Either the language itself brings about a word structure that deviates from our conjectures or the text sort has its own peculiarities (e.g. poetry) or the author wanted to give his text a special air. As a matter of fact, even if all texts of a language follow the same distribution, each of them must have different parameter values. But if the writers and text sorts strongly deviate, the model must be modified. Thus developing ever new models for word length distributions is a legal, desirable activity. There is no end of this activity.

For the sake of lucidity, we present a simple survey of these distributions. The authors of individual articles used the following procedure: first fit all available theoretical distributions to all data, then take that theoretical distribution which fits well to the majority (or all) data. This is an empirically secure way to obtain good results. But if we take into account that texts are not necessarily written spontaneously and what more, after being ready they are corrected, reduced, enlarged, etc. we must expect exceptions. The only text sort written spontaneously and not corrected any more is private letters (especially those written by hand).

Another problem is the maximal length of words. In some languages they are too short, and the fitting cannot be tested because of too few degrees of freedom. In this case one must take a distribution having only one parameter, e.g. Poisson or geometric. Many times the “longest” classes are not very frequent and the theoretical frequencies must be pooled in order to obtain at least $NP_x > 1$ for all x. The way to definitive decisions is very troublesome and one does not obtain always a satisfactory result.

The Poisson distribution, either in its usual form, or displaced to the right, or truncated above the zero point (positive Poisson d.) should be used at the very beginning of any investigation. If Poisson is adequate, we conjecture that the process of writing is performed randomly or spontaneously, without any binding. If boundary conditions play a role and some classes deviate, one can either

perform class modifications and obtain e.g. the Cohen-Poisson d., the Pandey-Poisson d. or the Singh-Poisson d. If all classes deviate, one generalizes the recurrence function by adding a new parameter and may obtain e.g. the Conway-Maxwell-Poisson d., the hyper-Poisson d. or the Palm-Poisson d.; or one takes a distribution which has Poisson as a limiting case, viz. the binomial d., the hyper-binomial d., the negative binomial d., the hyper-Pascal d. and the Pólya d. Even these can be punctually modified: one already found the Cohen-negative binomial d. and G. Djuraš (2012) found a number of other modifications. However, the situation can get even more complex and one performs the Feller-generalization, namely, one replaces the argument t in the probability generating function of the Poisson distribution, $G(t) = \exp(a(t-1))$ by another probability generating function. In this way one obtained up to now the Hirata-Poisson d., the Consul-Jain-Poisson d., the Poisson-uniform d., and the Meyer-Thomas d. Now, in texts, we have everywhere the problem of possible non-homogeneity evoking the impression that the text has several strata. In such cases, the data must be captured by mixing of distributions. In this way one obtained the simple mixed Poisson d., the Dacey-Poisson d. and the Fucks-Poisson d. Still another way is considering the parameter of the Poisson distribution a variable with its own distribution. These results can be seen in Table 1.1 All of these distributions may be displaced or truncated. Their interrelations are presented in Figure 1.1

Table 1.1
Word length distributions found and their relations to Poisson
(number of parameters)

Class modifications	(positive) Cohen-Poisson(2); Pandey-Poisson(2); Singh-Poisson(2)
Poisson as special case of	Conway-Maxwell-Poisson(2); hyper-Poisson(2); Palm-Poisson(2)
Poisson as limiting case of	binomial(2); negative binomial(2); hyper-Pascal (3); Pólya(3-4)
Feller-generalization of Poisson	Hirata-Poisson(2); Consul-Jain-Poisson(2); Poisson-uniform(2); Meyer-Thomas (3)
Mixing of Poisson	mixed Poisson(3); Dacey-Poisson(2); Fucks-Poisson($\geq$2)

Some of the above distributions have been modified, too, and in principle it can be done with every distribution (cf. Wimmer, Witkovský, Altmann 1999). But one strives for a theory in which both the models and the variants are linguistically substantiated. As can be seen, the majority of the models have two parameters. There are still some other models (geometric, lognormal, Merkyte, hyper-binomial, extended positive Poisson, etc.) but they could be corroborated only ad hoc.

Since some languages prefer some distributions, it can be conjectured that the boundary conditions can be found directly in the given language. That means, the present distributions could be used also for typology. But we are still far from hitting such a distant target.

A serious problem is always the text size. Whatever we compute, small sizes are not sufficient. There may be outliers, there may be too few degrees of freedom, etc. But if we take long texts, we must give up any homogeneity, we must not trust the chi-square test for goodness-of-fit because with increasing size the chi-square increases and yields "bad" results. The usual technique was to take a coefficient (Cramer, Pearson, Chuprov) which took sample size (or also the degrees of freedom) into account - but the greater the size, the smaller the coefficient, i.e. it is not reliable. Unfortunately, nobody can estimate the ideal text size. There is, of course, a remedy, namely to test the goodness-of-fit by means of the determination coefficient. But in that case we consider the probability mass function a usual continuous function. This is no overrunning the scientific moral, because there are many ways to truth and each of them is only an approximation. But whatever we do, the criterion of the goodness-of-fit is a subjective decision. Even if in statistics one sets $\alpha = 0.05$ for the chi-square or $R^2 > 0.90$ for a usual function, this has nothing to do with reality or with truth, it is a convention giving us a first look of the reliability of the acceptation or rejection of the hypothesis.

The interrelations of the above distributions are presented in Figure 1.1.

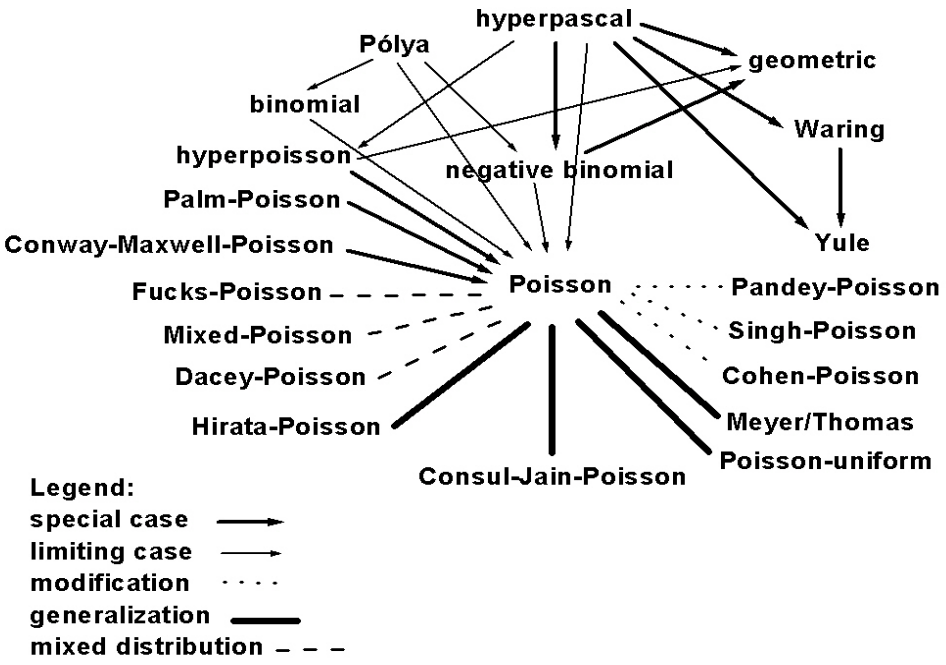


Figure 1.1. Relations of the Poisson distribution to some other ones used in word length research

In order to extend the investigation, we prepared 61 texts in 28 languages and different individual texts and try to show the general trend. The texts we analyzed are rather short hence a number of exceptions can occur. In Table 1.2 we show a survey of texts, the appropriate distributions and their parameters as well as the results of fitting. We shall try to draw some consequences. The testing by means of the chi-square test has been performed mostly with the smallest theoretical class-frequency > 1. But in several cases we pooled some classes in order to obtain better fitting. For the great majority of data we obtained several well fitting distributions. We chose either a common distribution for all texts of a certain language or we took only the "best" distribution from Table 1.1.

Table 1.2
Fitting some distributions to word length data in 61 texts in 28 languages

Language, Text alphabetically	**Distribution**	**Parameters**	X^2	*DF*	*P*	*I*	*S*
Akan Mma Nnsua Ade Bɔne	Positive Poisson	a = 1,0735	2,14	3	0,54	0,4495	1,2804
Akan Agya Yaw Ne Akutu Kwaa	Hirata-Poisson	a = 0,4352 b = 0,3817	6,75	2	0,03	0,4302	0,9519
Bamana Bamakɔ sigicogoya	Mixed Poisson	a = 1,1154 b = 0,0640 α = 0,5409	4,05	2	0,13	0,5469	1,5001
Bamana Masadennin	Mixed Poisson	a = 1,6608 b = 0,1995 α = 0,2662	3,47	2	0,18	0,6609	2,1723
Bamana Namakɔrɔba halakilen	Mixed Poisson	a = 1,8539 b = 0,2960 α = 0,1882	10,27	3	0,02	0,5814	1,9455
Bamana Sonsannin ani Surukuba	Mixed Poisson	a = 1,1755 b = 0,3686 α = 0,1479	0,63	1	0,43	0,3787	1,3181
Bulgarian* Ostrovskij, Kak se kaljavaše stomanata, Chapter 1	Cohen-Poisson	a = 1,2066 α = 0,1692	3,87	3	0,28	0,5737	0,8428
Czech Překvapení v justici	Singh-Poisson	a = 1,4105 α = 0,9131	3,67	3	0,30	0,5901	0,8296
Czech Marek Švehla: Voličův kalkul	Poisson	a = 1,2039	2,06	4	0,73	0,5872	1,4177
Czech Jan Macháček Slovenský (dobrý) příklad	Singh-Poisson	a = 1,3920 α = 0,9124	0,24	3	0,97	0,6158	1,0021
Czech Jan Čulík:O čem	Singh-	a = 1,4912	6,67	5	0,25	0,6652	1,2050

jsou dnešní Spojené státy?	Poisson	$\alpha = 0{,}9070$					
Czech Karel Hvížďala: O předem zpackané prezidentské volbě aneb Jak dlouho budeme bez prezidenta	Hyper-binomial	n = 7 m = 3,6230 q = 0,5693	4,61	3	0,20	0,5753	0,7364
French Dunkerque (Press)	Singh-Poisson	a = 1,1684 $\alpha = 0{,}6933$	0,60	2	0,74	0,6378	1,6376
German Assads Familiendiktatur (Press)	Hyper-Poisson	a = 5,1161 b = 8,1156	6,98	4	0,14	0,8533	2,3159
German ATT0012 (Press)	Hyper-Poisson	a = 6,9874 b = 11,6580	1,89	4	0,76	0,8754	2,1409
German Die Stadt des Schweigens (Press)	Hyper-Poisson	a = 7,0734 b = 12,5014	3,36	4	0,50	0,8756	2,4896
German Terror in Ost Timor (Press)	Hyper-Poisson	a = 2,9826 b = 4,7691	1,25	3	0,74	0,7352	1,9520
German Unter Hackern (Press)	Hyper-Poisson	a = 8,2168 b = 14,0931	10,71	5	0,06	0,8821	2,1106
Hindi Daily Hindi Milap, (31st May, 2012): After the sanction to love marriage, (page 4)	Hyper-poisson	a = 0,5963 b = 0,7097	1,10	2	0,48	0,3907	0,8140
Hindi Swatantra Varta, (31st July, 2012): The Anna Team on a cross-road (page 6)	Hyper-poisson	a = 0,4173 b = 0,5445	0,95	2	0,62	0,3356	0,7173
Hungarian A nominalizmus forradalma **(**Press)	Positive Singh-Poisson	a = 2,7021 $\alpha = 0{,}8788$	16,81	6	0,01	0,9216	1,3449
Hungarian Kunczekolbász **(**Press)	Hyper-Poisson	a = 3,5647 b = 3,5683	4,74	6	0,58	0,8883	1,4403
Indonesian Pengurus PSM terbelah **(**Press)	Conway-Maxwell-Poisson	a = 3,3373 b = 1,9779	6,98	3	0,07	0,3442	0,3323
Indonesian Sekolah ditutup **(**Press)	Cohen-binomial	n = 5 p = 0,3158 $\alpha = 0{,}0080$	4,94	1	0,03	0,4115	0,6787
Italian (Press, Online)	Singh-Poisson	a = 1,5762 $\alpha = 0{,}7773$	2,93	4	0,57	0,7588	1,4915
Japanese Miki, Jinseiron Note	Mixed Poisson	a = 2,6867 b = 0,6312 $\alpha = 0{,}2262$	3,32	4	0,51	0,8672	2,2278

Kikongo Bimpa: Ma Ngo ya Ma Nsiese	Binomial	n = 5 p = 0,1991	2,90	1	0,09	0,4205	0,7749
Kikongo Lumumba speech	Cohen-Poisson	a = 1,1487 α = 0,1540	3,94	3	0,27	0,5395	0,8396
Kikongo Nkongo ye Kisi Kongo	Hyperpa-scal	k = 0,0548 m = 0,0098 q = 0,2113	3,04	1	0,08	0,3392	1,0481
Latin Cicero, In Catilinam I	Extended positive binomial	n = 9 p = 0,1700 α = 0,7227	6,50	3	0,09	0,5666	0,6741
Latin Cicero, In Catilinam 2	Extended positive binomial	n = 9 p = 0,1801 α = 0,7166	8,05	2	0,02	0,6150	0,7940
Macedonian* Ostrovskij, Kako se kaleše čelkiot, Chapter 1	Singh-Poisson	a = 1,6427 α = 0,7696	1,09	3	0,78	0,7498	1,1283
Malayalam 1, Moralistic Hooligans	Positive Cohen-Poisson	a = 3,4255 α = 0,8565	5,01	4	0,54	0,6455	0,8924
Malayalam 2, No one should die	Positive Cohen-Poisson	a = 4,0324 α = 0,5349	4,34	6	0,63	0,7618	1,0833
Maninka Nko Doumbu Kende no. 2	Singh Poisson	a = 0,9601 α = 0,8058	1,95	3	0,58	0,5063	1,1282
Maninka Nko Doumbu Kende no. 7	Singh Poisson	a = 1,0551 α = 0,6788	0,41	3	0,94	0,5551	1,3906
Maninka Sìikán` (Constitution of Guinea, an excerpt)	Singh-Poisson	a = 1,3911 α = 0,6457	5,10	3	0,17	0,6755	1,3162
Maninka Teelen4	Singh-Poisson	a = 0,9500 α = 0,6731	1,73	3	0,63	0,5011	1,3116
Odia The Samaj, Bhubaneshwar (28 June 2012) Title: Who is great? (page 4)	Hyper-poisson	a = 0,9872 b = 0,0986	6,21	3	0,10	0,3580	0,9237
Odia The Dharitri, Balasore (12th February, 2012): Calculation for the District Council President (page 10)	Conway-Maxwell-Poisson	a = 3,7240 b = 1,7405	5,46	4	0,24	0,4056	0,4582
Romanian Paler, Aventuri solitare (excerpt)	1-d. Singh-Poisson	a = 1,5052 α = 0,7221	7,51	4	0,11	0,7235	1,3077
Romanian Popescu D.R.,	1-d.	a = 1,0786	2,02	3	0,56	0,5497	1,1578

Vânătoarea regală, Chapter 2	Singh-Poisson	α = 0,7540					
Romanian Steinhardt, Jurnalul fericirii, Trei soluții	Positive Singh-Poisson	a = 1,9659 α = 0,7901	7,70	4	0,10	0,7424	1,4402
Russian* Ostrovskij, Kak zakaljalas stal', Chapter 1	Singh-Poisson	a = 1,2427 α = 0,9374	1,57	4	0.81	0.5732	1,0362
Serbian* Ostrovskij, Kako se kalio čelik, Chapter 1	Singh-Poisson	a = 1,1926 α = 0,9171	0,28	2	0,87	0,5430	0,8474
Slovak E. Bachletová, Moja Dolná zem	Binomial	n = 11 p = 0,1141	1,01	3	0,80	0,4911	0,7038
Slovak E. Bachletová, Riadok v tlačive: nezamestnaný	Cohen-Poisson	a =1,4859 α = 0,1608	2,74	4	0,60	0,6211	0,7745
Slovenian* Ostrovskij, Kako se je kalilo jeklo, Chapter 1	Cohen-Poisson	a = 0.9856 α = 0.1580	2,51	3	0.47	0,5354	0,9902
Sundanese Agustusan (Salaka Online)	Hyper-poisson	a = 0,7204 b = 0,4222	9,66	3	0,02	0,4054	0,7763
Sundanese Aki Satimi (Salaka Online)	Hyper-poisson	a = 0,6441 b = 0,3345	0,11	2	0,95	0,3779	0,5943
Tagalog Hernandez, Limang Alas: Tatlong Santo	Hyper-poisson	a = 1,9456 b = 2,3391	6,33	5	0,28	0,6493	1,3059
Tagalog Hernandez, Magpisan	Mixed Poisson	a = 1,7016 b = 0,6537 α = 0,5287	8,80	4	0,07	0,6618	1,2883
Tagalog Rosales, Kristal Na Tubig	Mixed Poisson	a = 1,7416 b = 0,5924 α = 0,4877	7,69	4	0,10	0,6794	1,3693
Tamil (Press)	Positive Cohen-Poisson	a = 3,0521 α = 0,9115	5,73	6	0,45	0,6240	1,1262
Telugu Daily Andhra-bhoomi (4th August 2012) Train Journey without safety (page 4)	Hyper-poisson	a = 1,7924 b = 0,2301	4,59	4	0,33	0,5983	1,4391
Telugu Daily Andhra-bhoomi (4th August 2012): Trail-angaswamy: a biography, page10	Positive Cohen-Poisson	a = 3,2346 α = 0,7974	12,04	6	0,05	0,6526	1,0384
Vai Mu ja vaa lɔ	Positive	a = 0,5400	0,35	2	0,84	0,2883	0,4888

(T. Sherman)	Cohen-Poisson	α = 0,5049					
Vai Sabu Mua Ko	Positive Cohen-Poisson	a = 0,3355 α = 0,7572	2,30	1	0,13	0,2571	0,6697
Vai Vande bɛ Wu'u	Poisson	a = 0,4515	0,69	2	0,71	0,2939	0,8527
Welsh T1 Crynodeb Gweithredol	Cohen-binomial	n = 6 p = 0,1949 α = 0,3846	3,35	1	0,07	0,6010	1,0317
Welsh T2 Ffansi camu i esgidiau rhywun enwog?	1-d. Cohen-Poisson	a = 0,6713 α = 0,1270	5,80	1	0,06	0,3797	0,8398

For the analysis of the word length distribution in some Slavic languages marked with asterisk, the first chapter of the Russian novel "How the steel was tempered" (Chapter 1) by N. Ostrovskij and the translations into Slovene, Serbian, Bulgarian and Macedonian were used (for details cf. Kelih 2009). In all Slavic languages several distributions were adequate, we took only the "best" one.

Table 1.3
Some distributions used in word length research

Name	**Definition**
Poisson	$\frac{a^x e^{-a}}{x!},\ x = 0,1,2,...$
Positive Poisson	$\frac{a^x e^{-a}}{x!(1-e^{-a})},\ x = 1,2,3,...$
Pandey-Poisson	$\begin{cases} \frac{\alpha e^{-a} a^x}{x!}, & x = 0,1,2,...,c-1,c+1,c+2,. \\ 1-\alpha+\frac{\alpha e^{-a} a^c}{c!}, & x = c \end{cases}$
Positive Cohen-Poisson	$P_x = \begin{cases} \frac{(1-\alpha)a}{e^a-1-\alpha a}, & x = 1 \\ \frac{a^x}{x!(e^a-1-\alpha a)}, & x = 2,3,4,... \end{cases}$

Cohen-Poisson	$P_x = \begin{cases} e^{-a}(1+a\alpha), & x=0 \\ ae^{-a}(1-\alpha), & x=1 \\ \dfrac{a^x e^{-a}}{x!}, & x=2,3,\ldots \end{cases}$
Singh-Poisson	$P_x = \begin{cases} 1-\alpha+\alpha e^{-a}, & x=1 \\ \dfrac{\alpha a^x e^{-a}}{x!}, & x=2,3,\ldots \end{cases}$
Palm-Poisson	$\dfrac{R_{(x)}a^x}{{}_2F_0(-R,1;-a)}$, x = 0,1,…,$R$
Conway-Maxwell-Poisson	$\dfrac{a^x}{x!^b}P_0$, x = 0,1,2,…
Hyper-Poisson	$\dfrac{a^x}{b^{(x)}{}_1F_1(1;b;a)}$, x = 0,1,2,…
Negative binomial	$\binom{k+x-1}{x}p^k q^x$, x = 0,1,2,…
Hyperpascal	$\dfrac{\binom{k+x-1}{x}}{\binom{m+x-1}{x}}q^x P_0$, $x=0,1,2,\ldots$ $P_0 = ({}_2F_1(k,1;m;q))^{-1}$
Binomial	$\binom{n}{x}p^x q^{n-x}$, x = 0,1,…,n
Pólya (with 3 parameters)	$\dfrac{\binom{-p/s}{x}\binom{-q/s}{n-x}}{\binom{-1/s}{n}}$, $x=0,1,2,\ldots,n$
Hirata-Poisson	$\begin{cases} e^{-a}, & x=0 \\ \sum_{i=0}^{\left[\frac{x}{2}\right]}\binom{x-i}{i}\dfrac{e^{-a}a^{x-i}}{(x-i)!}b^i(1-b)^{x-2i}, & x=1,2,\ldots \end{cases}$

Consul-Jain-Poisson	$\begin{cases} e^{-a}, & x=0 \\ \dfrac{a(a+bx)^{x-1}e^{-(a+bx)}}{x!}, & x=1,2,... \end{cases}$
Poisson-uniform	$(b-a)^{-1}\left[e^{-a}\sum_{j=0}^{x}\frac{a^j}{j!}-e^{-b}\sum_{j=0}^{x}\frac{b^j}{j!}\right], \quad x=0,1,2,...$
Meyer-Thomas	$\frac{e^{-b}}{x!}\sum_{i=1}^{x}\binom{x}{i}ib^{i-1}(im)^{x-i}e^{-im}, \quad x=1,2,3,...$
Mixed Poisson	$\frac{\alpha a^x e^{-a}}{x!}+\frac{(1-\alpha)b^x e^{-b}}{x!}, \quad x=0,1,2,...$
Dacey-Poisson	$\frac{(1-\alpha)a^x e^{-a}}{x!}+\frac{\alpha x a^{x-1}e^{-a}}{x!}, \quad x=0,1,2,...$
Fucks-Poisson	$e^{-a}\sum_{k=0}^{\infty}(\varepsilon_k-\varepsilon_{k+1})\frac{a^{x-k}}{(x-k)!}, \quad x=0,1,2,...$

Discussion. All these distributions may be displaced (by defining the support as x = 1,2,3,… and replacing x by x-1 in the formula if the original support was x = 0,1,2,…); truncated (by setting x = 1,2,3,… and dividing the formula by 1-P_0); modified in different ways; generalized (e.g. by the Feller method), etc. In the literature on word length one finds a number of different adaptations. Unfortunately, they were made ad hoc and did not lead to a theoretical progress. Some of the distributions were ignored simply because they brought an innovation/ deviation. This study is an infinite process. It would be advisable to restrict the investigation to one family of distributions, e.g. that of Wimmer and Altmann (2005) in which one at least knows what kind of force or requirement the parameters may represent. The mixing of distributions is, from the theoretical point of view, no disadvantage, because the text may be stratified (= composed of different classes of elements, cf. Popescu, Altmann, Köhler 2010) but empirically it enlarges the number of parameters and the classical chi-square test does not bring a decisive result. In all cases one can alternatively consider the given distribution as a continuous function and test the deviations (or, decide the relevance of the deviations) by means of the determination coefficient. In most cases not only one of the given distributions is adequate, one always finds several well-fitting ones. The software tests automatically about 200 discrete distributions and it may happen that several of them are at the same time “adequate” for the same data. The curve-software tests about 8000 functions and one can add his own ones. There was only one case (Welsh) in which we were forced to use a more distant relative, the Cohen-binomial distribution, whose limiting cases are both Cohen-

Poisson and Poisson, and it is itself a modification of the binomial. It was not inserted in Figure 1.1.

The fact that one seldom meets the pure Poisson distribution in our data is a sign of special technique both in word formation and in syntax of the given languages. Or simply because languages have their own attractors for word length, and word length is always linked with some other properties.

The fate of this research area is characteristic for social sciences: the more we know, the more chaotically disintegrates the object of research and we cannot even imagine which forces were active.

In order to get a plastic picture of the results, we present the individual data using the criterion of J.K.Ord (1972), applied many times in quantitative studies. It can display groupings, trends, development, etc. One uses the first three moments of the distribution and defines

$$I = \frac{m_2}{m_1'}, \qquad S = \frac{m_3}{m_2}$$

where m'_1 is the mean and m_2, m_3 the second and third central moments of the empirical distribution (cf. Popescu et al. 2009: 154). The computed values are presented in Table 1.2. If we plot the values of I and S in a Carterian coordinate system we obtain the image presented in Figure 1.2. Even if we have a small number of texts form individual languages, one can see that they are fuzzily grouped. The Slavic languages are positioned around a straight line, Indonesian and Sundanese having strong mutual influence are very near to oneanother, Tagalog is more distant from them; Bamana and Vai have their own positions; Romanian occupies a straight line, etc. Thorough investigation of many texts of many text-sorts would furnish us - possibly - a view of text sorts and maybe also a look at the morphological typology of languages but this is not our aim. The straight line through these points shows merely the general trend strengthening our persuasion that there is some background law, a kind of attractor, which makes its way in every text taking into account the given boundary conditions.

If we compare the place of languages in Figure 1.2a, we can see that no language lies in the domain of the negative hypergeometric distribution characterized by $S < 2I - 1$. But this can be considered only the property of the languages studied. A negative m_3 is possible but not very probable. Surely, one can find languages in which $S \rightarrow 0$ (monosyllabic) but in order to be able to generalize, one must analyze a great number of languages.

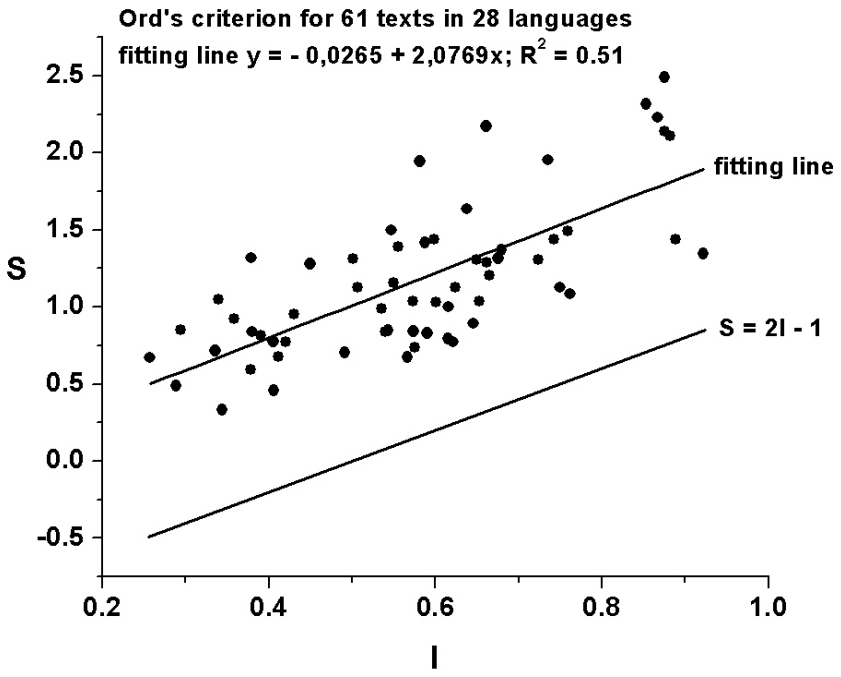


Figure 1.2a. Ord's criterion for 61 texts in 28 languages

Almost all points are placed in a kind of ellipse posited over the S = 2I-1 line. Preliminarily, it can only be said that the points occupy the domain of the hypergeometric and the beta-Pascal distributions. None of the points surpasses I = 1. The preliminary geometry of the ellipse is presented in Figure 1.2b. The ellipse center is defined by the coordinates I_C = mean I = 0,5815 and S_C = mean S = 1,1797 and the major axis by the fitting line y = - 0,0324 + 2,0846x. The distance between the most remote projections on the major axis approximates the major diameter 2*a* = 1,54 (corresponding to the pair Vai, *Sabu Mua Ko* – German, *Die Stadt des Schweigens*). Similarly, the distance between the most remote projections on the minor axis approximates the minor diameter 2*b* = 1,13 (corresponding to the pair Hungarian, *A nominalizmus forradalma* – Bamana, *Masadennin*). It results therefore a *flattening factor* $g = 1 - b/a = 0{,}27$. Though we considered grammatically and phonetically very different languages, the result is not definitive. The existence of outliers is always a reason for checking the complete computation, compare the definition of syllable in the given language with that in other ones, revise the whole theory but expecially to continue analyzing further data. It is also possible that all languages lie on the given straight line within a certain confidence interval.

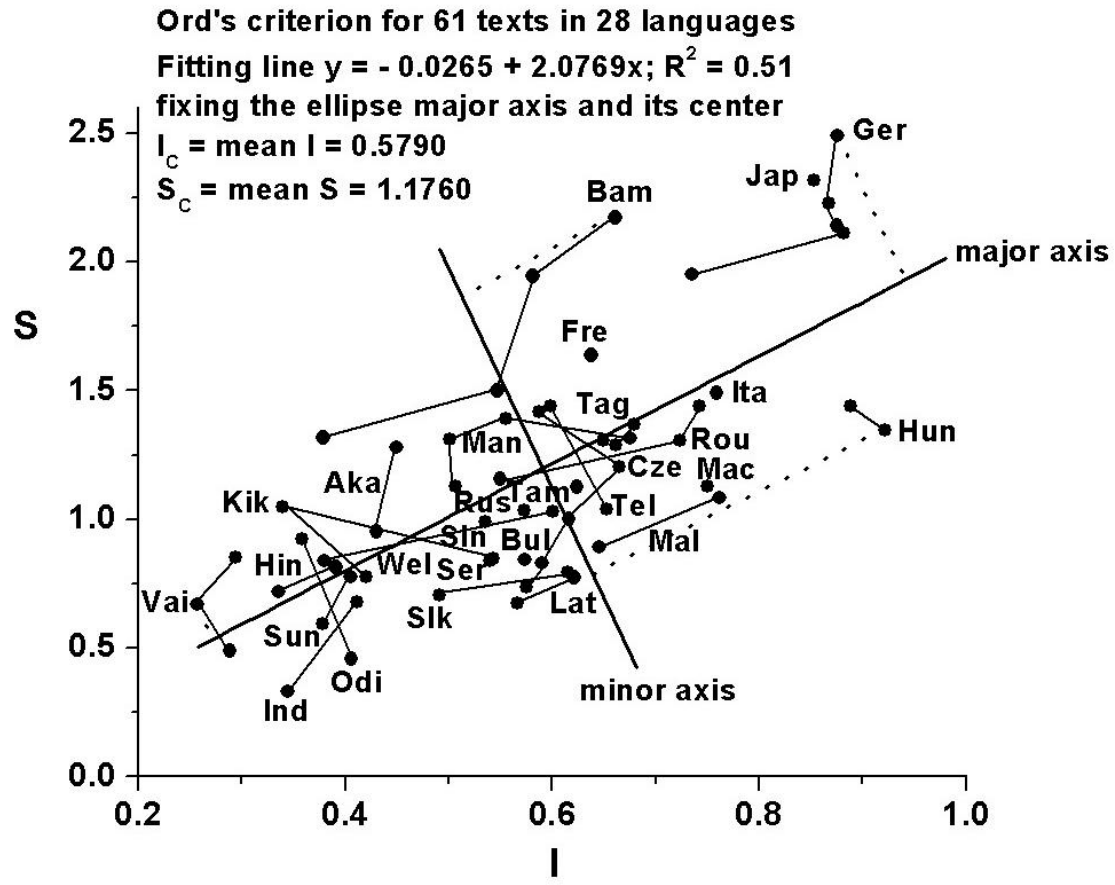


Figure 1.2b. Center and axes of the ellipse enclosing 61 texts in 28 languages (Aka = Akan, Bam = Bamako, Bul = Bulgarian, Cze = Czech, Fre = French, Ger = German, Hin = Hindi, Hun = Hungarian, Ind = Indonesian, Ita = Italian, Jap = Japanese, Kik = Kikongo, Lat = Latin, Mac = Macedonian, Mal = Malayalam, Man = Maninka, Odi = Odia, Rou = Romanian, Rus = Russian, Ser = Serbian, Slk = Slovak, Sln = Slovene, Sun = Sundanese, Tag = Tagalog, Tam = Tamil, Tel = Telugu, Vai = Vai, Wel = Welsh)

2. Smoothness

Length distributions represent an *a posteriori* ordered set of values, i.e., a pattern containing our(!) abstract view. But the sequence from which we obtain the numbers is not ordered. It follows some grammatical, thematic, psychological, text-sort and other rules or customs. These rules join autosemantics mostly by means of synsemantics. Synsemantics are more frequent and *eo ipso* they got historically shorter - unless all words were monosyllabic. In non-monosyllabic languages the sequence of lengths represents an oscillating structure which may be more or less smooth. The extent of oscillation and its (ir)regularity can be measured and expressed in many different ways. In time series analysis one uses mostly autocorrelations which enable us to see the degree of regularity of appear-

ing of the same length in some special distance (lag). In research on fractals one can compute the dimension of the resulting fractal, etc. On the other hand, we may ask about the distribution of distances between equal elements. There is a well developed theory of (random) distances (cf. Zörnig 1987) which may be compared with empirical data. In most cases they will deviate from randomness on different grounds: some grammars do not allow trespassing of special rules; for some entities the Skinner effect of reinforcement holds; some languages or text sorts do not use too long words and emphasize short distances between equal lengths, etc. These all are boundary conditions which should be investigated in individual cases.

For our purposes, we want to express the smoothness of a sequence by considering the local extremes. In the sequence 3,1,2,4,1,2,6,2 there are 6 extremes defined as numbers whose two neighbours are either both greater or both smaller. These are 1,4,1,6. The first and the last numbers are automatically extremes. Thus a sequence has *n* elements and *m* local extremes. The non-weighted non-smoothness can be expressed simply as

$$(1) \qquad NS = \frac{m-2}{n-2}.$$

But their simple proportion does not say anything about the strength of the oscillation. This can be computed by taking the arc length (*L*) between neighbouring elements using the Euclidian distance

$$(2) \qquad L = \sum_{i=1}^{n-1}[(x_i - x_{i+1})^2 + 1^2]^{1/2}.$$

Here x_i is the value of the *i*-th element in the sequence. For the above example we obtain

$$L = [(3 - 1)^2 + 1]^{1/2} + [(1 - 2)^2 + 1]^{1/2} + \ldots + [(6 - 2)^2 + 1]^{1/2}.$$

In long texts, this number is usually very great. Instead of taking the length directly we can divide each length by the maximal length present in the text, (in our example it is 6), that is, we can use x_i/x_{max} in the above formula and call it y_i. In this case, however, also the step must be reduced to $1/x_{max}$, otherwise the difference between neighbours almost disappears and the arc almost equals to $(n\text{-}1)^{1/2}$. However, we shall use directly (2). In order to set up a normalized indicator Popescu et al. (2010: 97) defined the roughness as $R = NS(L)/L_{max}$. For our purposes we take into account the omission of "words" of length 0 and obtain

$$(3) \qquad L_{\max} = (n-1)[(x_{\max} - 1)^2 + 1]^{1/2}$$

hence the roughness indicator has the form

$$(4) \qquad R = \frac{(m-2)L}{(n-2)L_{\max}}.$$

For the sake of illustration consider the roughness of the above sequence 3,1,2,4,1,2,6,2. Here

n = 8,
m = 6,
x_{max} = 6,
$L = [(3-1)^2+1]^{0.5} + [(1-2)^2+1]^{0.5} + [(2-4)^2+1]^{0.5} + [(4-1)^2+1]^{0.5} +$
$+ [(1-2)^2+1]^{0.5} + [(2-6)^2+1]^{0.5} + [(6-2)^2+1]^{0.5} = 18.7091$
$L_{max} = (8-1)[(6-1)^2+1]^{0.5} = 35.6931.$

Hence

R = (6-2)*18.7091/[(8-2)35.6931)] = 0.3494.

The roughness lies in the interval <0,1>. Computing this indicator for all our data, we obtain the results in Table 2.1, where *n* is the number of words of the considered text.

Table 2.1
Roughness of lengths in 61 texts of 28 languages

Language, Text alphabetically	**n**	**m**	**L**	**R**
Akan Mma Nnsua Ade Bɔne	143	60	218,6213	0,1536
Akan Agya Yaw Ne Akutu Kwaa	201	70	290,4720	0,1569
Bamana Bamakɔ sigicogoya	1138	382	1739,9743	0,1004
Bamana Masadennin	2615	758	4054,0607	0,0635
Bamana Namakɔrɔba halakilen	2392	718	3615,7292	0,0745
Bamana Sonsannin ani Surukuba	1406	356	1890,2265	0,0823
Bulgarian Ostrovskij, Kak se kaljavaše stomanata, Chapter 1	926	464	1644,8522	0,1744
Czech Překvapení v justici	289	116	500,9478	0,1355
Czech Marek Švehla: Voličův kalkul	288	127	482,1356	0,0911
Czech Jan Macháček Slovenský (dobrý) příklad	340	151	599,0170	0,1528
Czech Jan Čulík:O čem jsou dnešní Spojené státy?	2003	868	3633,3438	0,0867
Czech Karel Hvížďala :O předem zpackané prezidentské volbě aneb Jak dlouho budeme bez prezidenta	929	386	1615,0799	0,1185

French Dunkerque (Press)	1532	620	2558,3886	0,0955
German Assads Familiendiktatur (Press)	1415	573	2587,2710	0,0817
German ATT0012 (Press)	1146	478	2153,8356	0,0971
German Die Stadt des Schweigens (Press)	1567	648	2871,0648	0,0835
German Terror in Ost Timor (Press)	1398	591	2476,4021	0,0927
German Unter Hackern (Press)	1363	573	2558,3717	0,0979
Hindi Daily Hindi Milap, (31st May, 2012): After the sanction to love marriage, (page 4)	1106	463	1655,9122	0,1518
Hindi Swatantra Varta,(31st July, 2012): The Anna Team on a cross-road (page 6)	860	321	1212,2651	0,1272
Hungarian A nominalizmus forradalma **(**Press)	1314	666	2841,2580	0,0784
Hungarian Kunczekolbász **(**Press)	458	241	1016,7097	0,1446
Indonesian Pengurus PSM terbelah **(**Press)	345	130	537,8043	0,1144
Indonesian Sekolah ditutup **(**Press)	280	130	456,1638	0,1476
Italian (Press online)	2516	1304	4974,2034	0,1270
Japanese Miki, Jinseiron Note	2043	1099	3951,2210	0,1035
Kikongo Bimpa: Ma Ngo ya Ma Nsiese	956	500	1557,7629	0,1670
Kikongo Lumumba speech	824	375	1397,4363	0,1492
Kikongo Nkongo ye Kisi Kongo	768	360	1139,6341	0,1362
Latin Cicero, In Catilinam I	1064	456	1853,2258	0,1225
Latin Cicero, In Catilinam II	3095	1503	5632,4018	0,1249
Macedonian Ostrovskij, Kako se kaleše čelkiot, Chapter 1	1123	628	2251,3309	0,2197
Malayalam 1, Moralistic Hooligans	282	139	594,3144	0,1284
Malayalam 2, No one should die	288	144	668,4151	0,1434
Maninka Nko Doumbu Kende no. 2	1276	399	1913,3819	0,0917
Maninka Nko Doumbu Kende no. 7	1535	564	2394,5710	0,0941
Maninka Sìikán` (Constitution of Guinea, an excerpt)	1662	729	2950,3257	0,1526
Maninka Teelen4	1484	438	2182,7239	0,0849
Odia The Samaj, Bhubaneshwar (28 June 2012): Who is great? (page 4)	348	136	549,2713	0,1008
Odia The Dharitri, Balasore (12th February, 2012): Calculation for the District Council President (page 10)	630	313	1084,2770	0,1403
Romanian Paler, Aventuri solitare (excerpt)	891	456	1681,0150	0,1586

Romanian Popescu D.R., Vânătoarea regală, Chapter 2	1002	437	1658,3234	0,1413
Romanian Steinhardt, Jurnalul fericirii, Trei soluții	1511	708	2718,2766	0,1385
Russian Ostrovskij, Kak zakaljalas stal', Chapter 1	792	327	1316,6654	0,1128
Serbian Ostrovskij, Kako se kalio čelik, Chapter 1	1001	440	1703,1391	0,1811
Slovak E. Bachletová, Moja Dolná zem	872	382	1435,2595	0,1412
Slovak E. Bachletová, Riadok v tlačive: nezamestnaný	924	422	1655,5878	0,1343
Slovenian Ostrovskij, Kako se je kalilo jeklo, Chapter 1	977	376	1556,7049	0,1200
Sundanese Agustusan (Salaka Online)	416	181	664,2653	0,1357
Sundanese Aki Satimi (Salaka Online)	1283	584	2011,2286	0,1729
Tagalog Hernandez, Limang Alas: Tatlong Santo	1738	946	3238,0859	0,1434
Tagalog: Hernandez, Magpisan	1466	870	2838,6311	0,1625
Tagalog: Rosales, Kristal Na Tubig	1958	1201	3794,2703	0,1681
Tamil (Press)	384	167	771,8423	0,1080
Telugu Daily Andhrabhoo mi (4thAugust 2012): Train Journey without safety (p. 4)	665	304	1303,9431	0,0890
Telugu Daily Andhrabhoo mi (4th August 2012): Trailangaswamy:a biography (p. 10)	2295	144	616,9241	0,1261
Vai Mu ja vaa lɔ (T. Sherman)	3140	840	4079,9018	0,0842
Vai Sabu Mua Ko	495	136	631,3964	0,1099
Vai Vande bɛ Wu'u	426	159	571,2930	0,1574
Welsh T1 Crynodeb Gweithredol	985	502	1750,4984	0,1487
Welsh T2 Ffansi camu i esgidiau rhywun enwog?	1002	375	1441,3001	0,1303

As can be seen, the roughness is not very variable. This is based on the fact that x_{max} is used very seldom and R expresses merely the alternation of synsemantics and autosemantics. Hence it expresses formally a background of grammar. A similar indicator would be the fractal dimension of syllabic lengths.

The testing of the difference between two R is lengthy but possible (cf. Popescu, Mačutek, Altmann 2009: 49ff.). Here we shall merely order the languages/texts according to R and delay the problem of finding some variables responsible for the given magnitude of R. Since one finds very different values in the same language, one can conclude that text sort, style, education, etc. play an important role and hinder(!) setting up language typologies. Or, one would be

forced to propose different kinds of normalizations based on boundary conditions.

In any case we see that "great jumps" are not usual in the length sequences. The greatest roughness in our data has Macedonian followed by other Slavic languages. African languages are situated at the bottom of the realized values. This is perhaps the first hint at a possible boundary condition: since in Slavic languages we took the same translated text (Russian, Macedonian, Bulgarian, Serbian) - except for Slovenian - and obtained a great roughness. But Slovenian has a much lower roughness, and in Slovak texts it tends rather to the lower limit of the interval. Hence the above results are merely a first image of a property which can depend on all factors which are effective in a text.

3. Word length in sentence

The fact that the subject of sentence stands mostly at the beginning of sentence leads mechanically to its description, characterisation, circumstances, etc. in the rest of sentence. Here the specifications (attributes, predicates, modificators) of first order, then those of second order, etc. are placed and the longer the sentence, the longer get the specifications. This holds, of course, only in form of averages, not for individual sentences. The idea expressed by L. Uhlířová (1997) has been corroborated in some later investigations (cf. Fenk, Fenk-Oczlon 2006; Kelih 2010; Fan, Grzybek, Altmann 2010). Nevertheless, one can expect text-specific or language-specific exceptions, e.g. in Japanese where a long attributive clause can precede the subject, hence the given explication does not hold. In any case, the hypothesis can be tested using our data on a broad background.

The testing can be performed in two ways:

(1) For each text separately we form groups of sentences having the same number of words; for each group we compute the mean length at each position and investigate the course of length. This way seems to be simple but it can be investigated only using long texts in which every sentence length is repeated several times in order to obtain reliable results. One takes groups containing at least 5 sentences (but 10 is "better").

(2) One considers the text as a whole and computes the mean length of the first word, than that of the second, etc. in all sentences. We obtain the best estimation for the positions at the beginning, but the reliability of the mean decreases. If there are no more than 10 sentences having a certain length, one should cease computing. This approach has a certain disadvantage: short sentences may be constructed differently and destruct the smooth increase of length. Different additional hypotheses are possible: (i) If for different sentence lengths the course of mean word lengths is different, is this difference significant? That is, are the slopes different? (ii) Is it possible that in short sentences the long words stay at the beginning and in long ones (also) at the end? (iii) A quite different possibility

is the hypothesis that the more synthetic a language, the greater is the oscillation of word length in sentence (i.e. the more analytic the language, the smaller the oscillation?).

The views may be combined, different tests can be proposed and if a text does not abide by any of them, literary scientists may search for the causes of this phenomenon.

Evidently, the number of hypotheses can be increased and the way to a theory is still long, even if in some languages one can obtain clearer results. The methodological problem is: when can we say that a hypotheiss of this sort is sufficiently corroborated? Are all our trials membra disiecta? Was our analysis of words "correct" or is a negative result the consequence of a "false" analysis?

As an example of approach (1), we show the results using a Slovak text (E. Bachletová, *Moja Dolná zem*). Since only groups containing at least 5 sentences are relevant, we obtain the results as presented in Table 3.1.

Table 3.1
Mean words lengths in individual positions in sentence groups (Slovak text)
(Approach 1)

Sentence length	**Position in sentence**					
	1	**2**	**3**	**4**	**5**	**6**
2	2,6	2,4				
3	1,6	2	2,6			
5	2,2	2,2	2	2,1	2,6	
6	1,4	2,4	2	1,8	2	2,8

As can be seen, the course is not always monotonously increasing but in each group it is quite different. Nevertheless, the last value is mostly the greatest. It is, of course, possible that the groups of equally long sentences are not sufficiently represented. It is better if one performs this investigation with longer texts. Here, lengths 1, 4, and > 6 could not be taken into account because of too small samples.

Thorough scrutinizing of this hypothesis was performed in Fan, Grzybek, Altmann (2010) who nevertheless stated that there is an increase of mean length in sentence, but the longer the sentence the smaller is the slope *b*. The decrease of the slope is very regular. Unfortunately, for short texts this cannot be shown.

The same text scrutinized under approach (2) (considering all sentences) yields the results presented in Table 3.2 and Figure 3.1.

While the first approach yields not yet interpretable results, the second approach yields an approximately linear dependence presented in Figure 3.1. The strong oscillation is mostly ascribed to the existence of clauses within which the same tendency is repeated. In short texts this can cause a rejection of the hypothesis.

Table 3.2
Mean words lengths in individual positions in sentence groups (Slovak text)
(Approach 2)

Position	1	2	3	4	5	6	7	8	9	10
Mean	2,17	1,88	2,13	2,25	2,21	2,29	2,30	2,25	2,43	2,77

11	12	13	14	15	16	17	18	19	20
2,57	2,32	2,35	2,47	2,50	2,20	2,50	2,27	2,60	2,50

But even if we partition the sentence in clauses and make equal length groups, it is possible that e.g. the third clause in sentence behaves differently and in that case we create again inhomogeneous data.

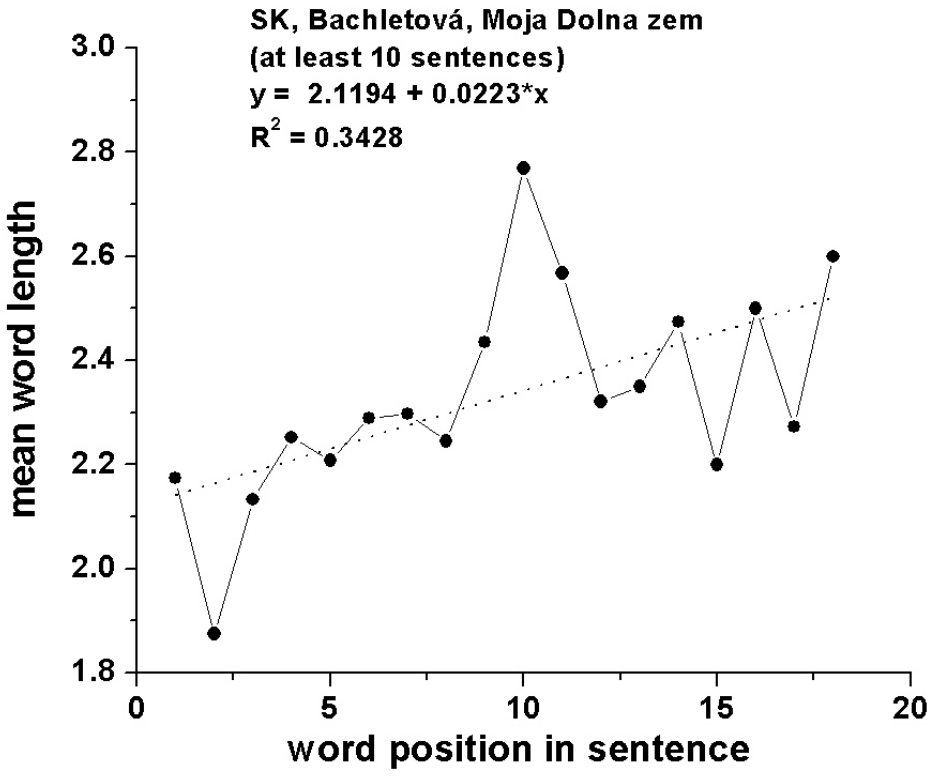


Figure 3.1. Mean word length in positions 1-18 in a Slovak text

For Bachletová's *Moja Dolná zem* we obtain *Mean length* = 2,1194 + 0,0223**Position*, with R^2 = 0,34 which is not sufficiently significant but the trend is visible inspite of great oscillation. The results for linear fitting from other texts are presented in Table 3.3. We notice, however, that power fitting generally

allows better results as, for instance, in this case, *Mean length* = 2,0148**Position*$^{0.0715}$, with R^2 = 0,40.

Table 3.3
The course of mean word lengths in 61 texts and 28 languages
(in contrast to Table 4, here *n* = number of sentences; the mean word length is taken over at least 10 sentences)

Language, Text alphabetically	**n**	**a**	**b**	**R^2**
Akan Agya Yaw Ne Akutu Kwaa	18	1,7453	-0,0327	0,176
Akan Mma Nnsua Ade	15	1,6650	-0,0087	0,007
Bamana Bamakɔ sigicogoya	84	1,6305	-0,0012	0,003
Bamana Masadennin	207	1,5829	0,0038	0,022
Bamana Namakɔrɔba halakilen	248	1,6318	-0,0048	0,053
Bamana Sonsannin ani Surukuba	168	1,5291	-0,0076	0,104
Bulgarian Ostrovskij, Kak se kaljavaše stomanata, Chapter 1	83	2,1471	-0,0008	0,0005
Czech Čulík, O čem jsou dnešní Spojené státy?	114	2,3508	0,0012	0,0023
Czech Hvížďala, O předem zpackané prezidentské volbě	37	2,3372	-0,0008	0,0009
Czech Macháček, Slovenský dobrý příklad	22	1,9034	0,0452	0,314
Czech Spurný, Prekvapení v justici	17	1,8970	0,0421	0,232
Czech Švehla, Editorial, Voličův kalkul	19	1,8672	0,0492	0,236
French Dunkerque (Press)	105	1,7267	0,0082	0,13
German Assads Familiendiktatur	119	2,0271	0,0064	0,028
German ATT0012 (Press)	83	1,9008	0,0210	0,233
German Die Stadt des Schweigens (Press)	120	1,9639	0,0034	0,015
German Terror in Ost Timor (Press)	98	1,9258	0,0054	0,027
German Unter Hackern (Press)	110	1,955	0,0122	0,128
Hindi After the sanction to love marriage	38	1,7734	-0,00003	0,000006
Hindi The Anna Team on a cross-road	42	1,6007	0,0045	0,03
Hungarian A nominalizmus forradalma **(**Press)	63	2,4654	0,0139	0,09
Hungarian Kunczekolbász **(**Press)	32	2,4196	0,0286	0,25

Indonesian Pengurus PSM terbelah **(**Press)	28	2,444	0,0222	0,13
Indonesian Sekolah ditutup **(**Press)	15	2,7691	-0,0170	0,077
Italian (Press, Online)	92	2.2088	0.0008	0.0014
Japanese Miki, Jinseiron Note, first 100 sentences	100	1,8826	0,0166	0,174
Kikongo Bimpa: Ma Ngo ya Ma Nsiese	79	1,9895	0,0149	0,086
Kikongo Lumumba speech	43	2,0584	-0,0057	0,088
Kikongo Nkongo ye Kisi Kongo	29	1,7176	0,0051	0,052
Latin Cicero, In Catilinam I	80	2,265	0,0095	0,13
Latin Cicero, In Catilinam II	180	2,3819	0,0015	0,004
Macedonian Ostrovskij, Kako se kaleše čelkiot, Chapter 1	89	2,2924	-0,0036	0,043
Malayalam 1, Moralistic Hooligans	32	3,5957	0,0527	0,336
Malayalam 2, No one should die	29	4,0148	0,0259	0,061
Maninka Nko Doumbu Kende no. 2	37	1,7707	-0,0006	0,003
Maninka Nko Doumbu Kende no. 7	34	1,7909	-0,0030	0,04
Maninka Sìikán` (Constitution of Guinea, an excerpt)	94	1,9545	-0,0056	0,076
Maninka Teelen4	29	1,6951	-0,0015	0,014
Odia Calculation for the District Council President	28	2,9565	-0,0018	0,0033
Odia Who is great?	36	2,9707	-0,0198	0,08
Romanian Paler, Aventuri solitare (excerpt)	17	1,93	0,0095	0,04
Romanian Popescu D.R., Vânătoarea regală, Chapter 2	61	1,7838	0,0002	0,00008
Romanian Steinhardt, Jurnalul fericirii, Trei soluții	85	1,975	0,0018	0,005
Russian Ostrovskij, Kak zakaljalas stal', Chapter 1	76	2,1320	0,0058	0,02
Serbian Ostrovskij, Kako se kalio čelik, Chapter 1	83	2,1685	-0,0123	0,168
Slovak E. Bachletová, Moja Dolná zem	92	2,1194	0,0223	0,34

Slovak E. Bachletová, Riadok v tlačive: nezamestnaný	78	2,2433	0,0245	0,17
Slovenian Ostrovskij, Kako se je kalilo jeklo, Chapter 1	84	1,9690	-0,0046	0,02
Sundanese Agustusan (Salaka Online)	53	2,158	0,0008	0,0003
Sundanese Aki Satimi (Salaka Online)	147	2,2089	-0,0069	0,09
Tagalog Hernandez, Limang Alas: Tatlong Santo	104	1,9915	0,0121	0,13
Tagalog Hernandez, Magpisan	111	1,9606	0,025	0,13
Tagalog Rosales, Kristal Na Tubig	139	2,1791	-0,0024	0,009
Tamil (Press)	32	3,4551	0,0059	0,0067
Telugu Trailangaswamy	51	3,4916	0,0479	0,131
Telugu Train Journey without safety	61	3,2949	0,0331	0,3049
Vai Mu ja vaa lɔ (T. Sherman)	193	1,4867	-0,0009	0,0060
Vai Sabu Mua Ko	39	1,5749	-0,0130	0,2536
Vai Vande bɛ Wu'u	35	1,4922	-0,0061	0,0270
Welsh T1 Crynodeb Gweithredol	40	2,0535	-0,0040	0,0241
Welsh T2 Ffansi camu i esgidiau rhywun enwog?	34	1,6492	-0,0021	0,0123

(i) As can be seen looking at parameter *b* there is a positive slope, but not with all texts. A Slovenian, Sundanese and Tagalog texts display a slightly decreasing tendency of length which can be considered simply zero. But even the positive slopes are very small, not significant.

(ii) Though the slope is very small, there is, nevertheless, an increasing tendency, but the determination coefficient is in all cases not high enough in order to speak of a real tendency. The oscillation is too great and destructs the linear regression. In all cases, the parameter *b* and the determination coefficient are too small.

Discussion. This tendency - if it exists at all - cannot be presented with persuasion. If we form groups of sentences with equal length, we mix up the text and in any case, we ignore the boundaries of clauses. Whatever we do, the oscillation remains very strong. The most conspicuous case is a stage play in which the speech of persons may be very different. Word length in one- or two-word sentences may strongly differ, e.g. "Yes!", "No!" against "Wonderful!", "Excellent!", G. "Absolut unwahrscheinlich!" That means, the relationship is not quite general and both the texts and the way of constructing data must be selected

according to pre-formulated criteria. Hence, here either there is no law or the boundary conditions necessary for the validity of a law are seldom fulfilled.

On the other hand, word is, perhaps, not the adequate unit for measuring the length rhythm in the sentence. Is it the clause, the phrase, the punctuation or something else?

Hence, the sequence of word lengths in sentence may be random or it may follow some trend which depends on boundary conditions like style, text sort, thematic concentration, text homogeneity and other circumstances which must be scrutinized in detail, in order to find the links of this phenomenon to other phenomena and subsume it under an extensive "word length theory".

4. Word length in increasing text

According to the above hypothesis, word length increases with increasing text. This is in principle impossible and can be realized only in short texts. Word length cannot increase continuously but there may be a limit to which it converges. The test can be performed by evaluating the texts "vertically", i.e. to compute the mean word length of say first 10 sentences, then of the next ten, etc. The grouping can be made differently, e.g. in short texts one takes only groups of 5 sentences, but in long ones one can take 100 or whole chapters.

A second method is to compute stepwise the mean of the first *x* words and observe the change of the mean. If the hypothesis of increasing length is correct, then the curve must be at least non-decreasing.

There is no "best" method and perhaps the testing must be made differently for every text.

Let us consider first the Slovak text *Moja Dolná zem* by Bachletová containing 92 sentences. We subdivide the text in sentence groups taking 10 sentences together, compute the mean word lengths in the first 10 sentences, then in the next 10 sentences, etc. and for the last group we take the rest. We obtain the result**s** plotted in Figures 4.1. The numerical results of several texts are presented in Table 4.1.

If we compute the linear regression in Bachletová's *Moja Dolná zem*, we even obtain a decreasing function ($R^2 = 0{,}12$)

$$\textit{Mean length} = 2{,}3278 - 0{,}0152\textit{Group},$$

which is not significant but at least shows that the trend is not that simple as conjectured in the hypothesis. Surely, the oscillation may disturb the trend but in any case we have here cases of rejection.

If the texts are short, one may form 5 groups which are enough for studying the tendency - if it is simple. Testing the hypothesis using our texts we obtained the results in Table 4.2. We took means of 10 subsequent sentences.

As can be seen, not only we do not have a significant increase of word length with deployment of text, just on the contrary. In the most cases the coefficient of linear regression *b* is negative and in some cases the regression is even significant (cf. Akan). The tendency may hold for some languages or texts (e.g. French) but it is at least not as general as supposed.

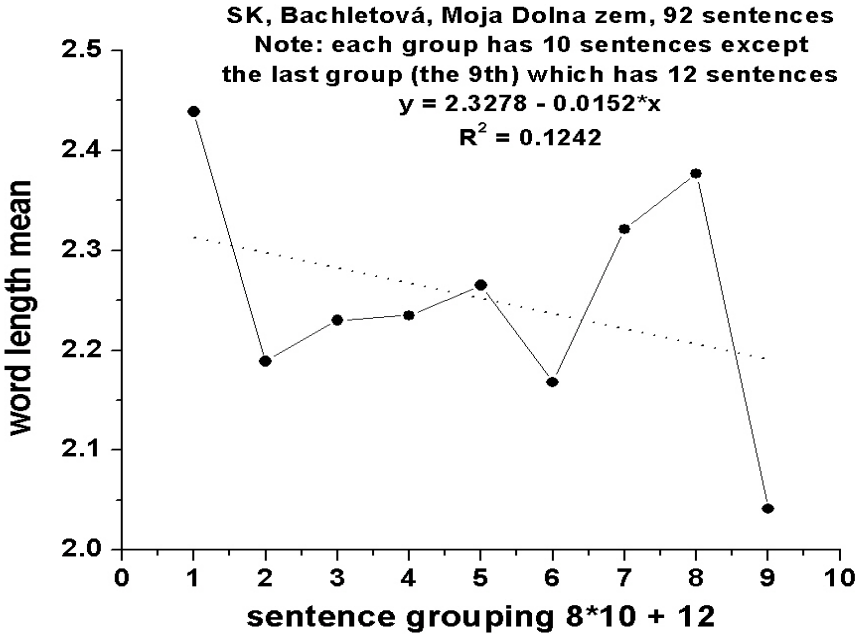


Figure 4.1a. Mean word length in groups of 10 subsequent sentences in *Moja Dolná zem* (Bachletová, Slovak)

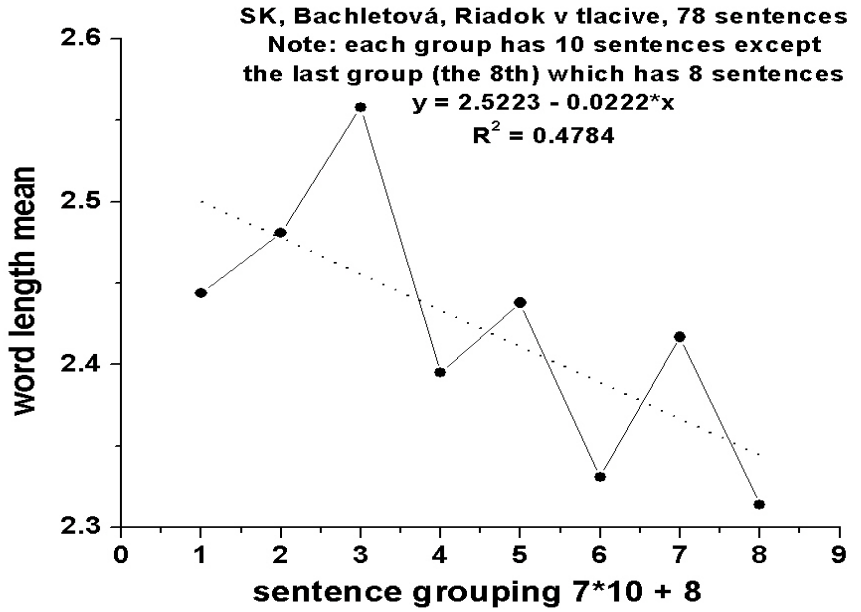


Figure 4.1b. Mean word length in groups of 10 subsequent sentences in *Riadok v tlačive* (Bachletová, Slovak)

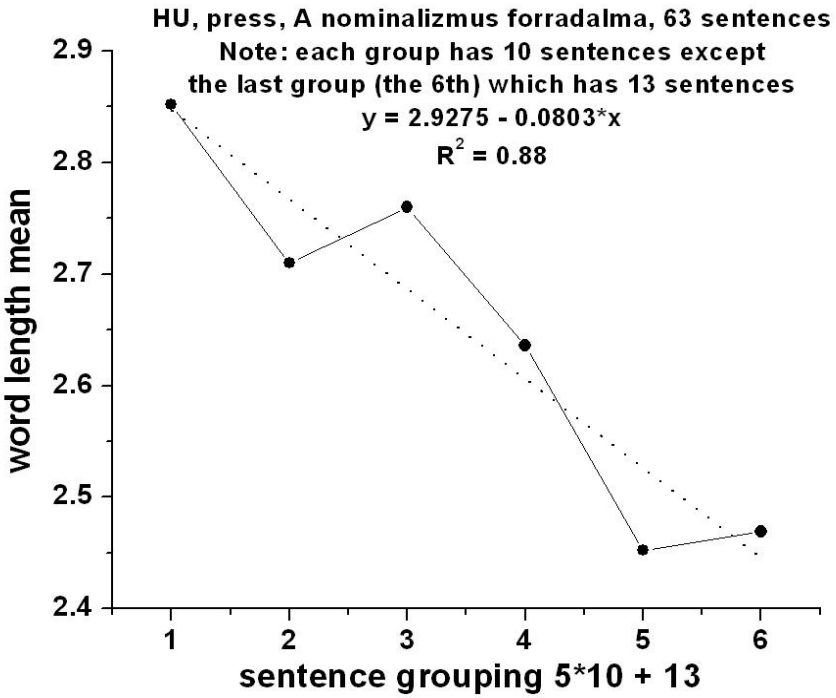


Figure 4.1c. Mean word length in groups of 10 subsequent sentences in *A nominalizmus forradalma* (press, Hungarian)

In spite of strong oscillations, the trend of the mean word length in deploying text can be considered also in detail, by subsequent single (not grouped) sentences, as revealed by comparing Figures 4.1a and 4.2 for Bachletová's *Moja Dolná zem.* In particular, in both plots the slope of the linear fitting is negative, indicating a slight decrease of the mean word length in deploying text. The data of all texts as processed sentence by sentence (that is not grouped as above in Table 4.2) are collected in Table 4.3 below. The only difference in slope *b* sign between single and grouped sentences was found for Latin and Romanian (D.R. Popescu). According to this table, 38 out of 61, that is about 62 % cases have negative *b*.

Both results show that increasing text does not necessarily mean increase of word length. This is perhaps caused by the fact that not only new words appear in the text - which may be longer - but the references to the preceding central words are shorter. As a matter of fact, none of the sequences display a significant increase of word length. Though there are positive results in the literature, we recommend further testing, selection of new texts and searching for boundary conditions.

Table 4.1
Mean word length in increasing texts (by subsequent grouped sentences)

Language, Text alphabetically	**1**	**2**	**3**	**4**	**5**	**6**	**7**	**8**	**9**	**10**
Akan Mma Nnsua Ade Bɔne	1,833	1,688	1,571	1,625	1,429	-	-	-	-	-
Bamana Bamakɔ sigicogoya	1,761	1,541	1,642	1,573	1,512	1,561	1,662	1,686	-	-
Bamana Masadennin	1,701	1,638	1,56	1,637	1,664	1,486	1,512	1,67	1,654	1,556
Bamana Namakɔrɔba halakilen	1,598	1,526	1,594	1,650	1,462	1,475	1,728	1,497	1,703	1,609
Bamana Sonsannin ani Surukuba	1,627	1,625	1,438	1,435	1,436	1,458	1,392	1,514	-	-
Bulgarian Ostrovskij, Kak se kalja-vaše stomanata, Chapter 1	2,385	2,118	2,017	2,146	2,147	1,979	2,057	2,272	-	-
Czech Čulík, O čem jsou dnešní Spojené státy?	2,351	2,434	2,199	2,626	2,356	2,266	2,195	2,244	2,385	2,427
Czech Hvížďala, O předem zpackané prezidentské volbě	2,179	2,436	2,33	2,596	2,306	2,319	2,374	-	-	-
Czech Macháček, Slovenský dobrý příklad	2,175	2,453	2,384	2,151	-	-	-	-	-	-
Czech Spurný, Prekvapení v justici	2,378	2,289	2,271	2,205	2,140	2,306	-	-	-	-
Czech Švehla, Editorial, Voličův kalkul										
French Dunkerque (Press)	1,772	1,758	1,807	1,843	1,873	-	-	-	-	-
German Assads Familiendiktatur (Press)	2,038	2,005	2,109	2,220	2,142	2,036	-	-	-	-
German ATT0012 (Press)	2,113	2,057	2,278	2,208	1,989	1,988	1,935	2,058	-	-
German Die Stadt des Schweigens (Press)	1,811	1,964	1,959	1,966	2,078	1,943	2,108	2,159	2,166	1,931
German Terror in Ost Timor (Press)	1,953	2,005	1,955	1,876	1,85	2,272	2,08	1,896	1,884	1,913

German Unter Hackern (Press)	2,106	2,189	2,275	2,101	2,112	1,722	1,851	2,135	2,048	1,97
Hindi After the sanction to love marriage	1,778	1,698	1,801	1,788	1,764	1,624	1,693	1,862	-	-
Hindi The Anna Team on a cross-road	1,591	1,624	1,566	1,652	1,748	1,727	1,677	1,573	-	-
Hungarian A nominalizmus forradalma **(**Press)	2,852	2,71	2,76	2,636	2,452	2,469	-	-	-	-
Hungarian Kunczekolbász **(**Press)	2,737	2,582	2,663	-	-	-	-	-	-	-
Indonesian Pengurus PSM terbelah **(**Press)	2,626	2,567	2,515	-	-	-	-	-	-	-
Indonesian Sekolah ditutup **(**Press)	2,644	2,540	2,562	-	-	-	-	-	-	-
Italian (Press, Online)	2,180	2,093	2,102	2,050	2,292	2,286	2,234	2,327	2,354	-
Japanese Miki, Jinseiron Note, first 100 sentences	2,075	2,291	2,071	2,052	1,897	1,995	2,116	2,101	2,094	2,321
Kikongo Bimpa: Ma Ngo ya Ma Nsiese	2,065	2,076	2,062	2,189	2,082	2,101	1,945	2,256	-	-
Kikongo Lumumba speech	1,947	2,032	1,991	2,173	2,000	1,956	1,923	2,095	2,016	-
Kikongo Nkongo ye Kisi Kongo	1,689	1,896	1,738	1,816	1,802	1,758	-	-	-	-
Latin Cicero, In Catilinam I	2,505	2,327	2,484	2,281	2,350	2,204	2,332	2,403	-	-
Latin Cicero, In Catilinam II	2,312	2,403	2,467	2,436	2,395	2,288	2,432	2,278	2,652	2,368
Macedonian Ostrovskij, Kako se kaleše čelkiot, Chapter 1	2,591	2,349	2,112	2,198	2,371	2,017	2,282	2,388	2,260	-
Malayalam 1, Moralistic Hooligans	4,200	4,026	3,760	3,660	3,875	3,472	-	-	-	-
Malayalam 2, No one should die	4,338	4,487	3,962	3,974	4,108	4,268	-	-	-	-
Maninka Nko Doumbu Kende no. 2	1,646	1,799	1,653	2,004	1,657	1,704	1,842	-	-	-
Maninka Nko Doumbu Kende no. 7	1,880	1,659	1,800	1,591	1,656	1,663	1,835	-	-	-
Maninka Sìikán` (Constitution of Guinea, an excerpt)	1,901	1,805	2,047	1,906	1,988	1,876	1,821	1,861	1,846	-
Maninka Teelen4	1,745	1,680	1,753	1,667	1,456	1,627	1,559	-	-	-
Odia Calculation for the District Council	3,000	3,180	3,115	2,712	2,750	3,135	-	-	-	-

President										
Odia Who is great?	3,020	2,553	2,772	3,068	2,656	2,606	3,264	-	-	-
Romanian Paler, Aventuri solitare (excerpt)	2,211	1,87	2,117	-	-	-	-	-	-	-
Romanian Popescu D.R., Vânătoarea regală, Chapter 2	1.922	1.667	1.674	1.798	1.839	1.795	-	-	-	-
Romanian Steinhardt, Jurnalul fericirii, Trei soluții	1,972	2,109	1,904	2,159	1,919	2,064	1,97	2,164	1,955	-
Russian Ostrovskij, Kak zakaljalas stal', Chapter 1	2,123	2,115	1,862	1,916	1,924	1,838	1,901	1,995		
Serbian Ostrovskij, Kako se kalio čelik, Chapter 1	2,248	2,297	1,971	1,991	2,090	1,933	2,138	2,042	-	-
Slovak E. Bachletová, Moja Dolná zem	2,439	2,189	2,230	2,235	2,265	2,168	2,321	2,377	2,041	-
Slovak E. Bachletová, Riadok v tlačive: nezamestnaný	2,444	2,481	2,558	2,395	2,438	2,331	2,417	2,314	-	-
Slovenian Ostrovskij, Kako se je kalilo jeklo, Chapter 1	2,000	2,088	1,912	1,706	2,000	2,500	2,059	2,147	-	-
Sundanese Agustusan (Salaka Online)	1,987	2,138	2,295	2,133	2,253	-	-	-	-	-
Sundanese Aki Satimi (Salaka Online)	2,114	2,159	2,137	2,065	2,223	2,085	2,244	-	-	-
Tagalog Hernandez, Limang Alas: Tatlong Santo	2,231	2,035	2,091	2,121	2,277	1,994	1,82	2,018	2,128	2,173
Tagalog Hernandez, Magpisan	2,224	2,264	2,207	2,248	2,159	2,216	2,316	2,154	2,055	2,132
Tagalog Rosales, Kristal Na Tubig	2,102	2,132	2,245	2,096	2,139	2,216	2,133	-	-	-
Tamil (Press)	3,509	3,553	3,436	3,397	3,559	3,734	-	-	-	-
Telugu Trailangaswamy	3,798	4,000	3,784	3,286	3,313	-	-	-	-	-
Telugu Train Journey without safety	3,625	3,495	3,374	3,699	3,681	3,244	-	-	-	-
Vai Mu ja vaa lɔ (T. Sherman)	1,447	1,433	1,466	1,493	1,486	1,478	1,481	1,484	1,500	1,505

Vai Sabu Mua Ko	1,617	1,393	1,486	1,477	-	-	-	-	-	-
Vai Vande bɛ Wu’u	1,467	1,455	1,410	1,500	-	-	-	-	-	-
Welsh T1 Crynodeb Gweithredol	2,144	1,986	1,925	2,014	2,028	1,945	1,943	2,164	-	-
Welsh T2 Ffansi camu i esgidiau rhywun enwog?	1,614	1,698	1,68	1,561	1,582	1,596	1,594	-	-	-

Table 4.2
Mean word length in deploying text by subsequent grouped sentences: all texts

Language, Text alphabetically	**n**	**sentence grouping**	**a**	**b**	**R^2**
Akan Mma Nnsua Ade Bɔne	15	5*3	1,8905	-0,0871	0,86
Bamana Bamakɔ sigicogoya	84	7*10 + 14	1,6293	-0,0027	0,006
Bamana Masadennin	207	9*20 + 27	1,6477	-0,0073	0,09
Bamana Namakɔrɔba halakilen	248	9*25 + 23	1,5475	0,0067	0,047
Bamana Sonsannin ani Surukuba	168	7*20 + 28	1,5921	-0,0226	0,376
Bulgarian Ostrovskij, Kak se kalja-vaše stomanata, Chapter 1	83	7*10 + 13	2,2049	-0,0144	0,069
Czech Čulík, O čem jsou dnešní Spojené státy?	114	9*11 + 15	2,3755	-0,005	0,013
Czech Hvížďala, O předem zpackané prezidentské volbě	37	6*5 + 7	2,3161	0,0117	0,038
Czech Macháček, Slovenský dobrý příklad	22	3*5 + 7	2,326	-0,0141	0,015
Czech Spurný, Prekvapení v justici	17	5*3 + 2	2,3521	-0,0249	0,318
Czech Švehla, Editorial, Voličův kalkul	19	3*5 + 4	2,3935	-0,072	0,509
French Dunkerque (Press)	105	4*20 + 25	1,7245	0,0287	0,894
German Assads Familiendiktatur (Press)	119	5*20 + 19	2,0405	0,0146	0,114
German ATT0012 (Press)	83	7*10 + 13	2,1899	-0,0248	0,271
German Die Stadt des Schweigens (Press)	120	12*10	1,8823	0,0229	0,369
German Terror in Ost Timor (Press)	98	9*10 + 8	1,984	-0,0028	0,005
German Unter Hackern (Press)	110	10*11	2,1859	-0,0246	0,207
Hindi After the sanction to love marriage	38	7*5 + 3	1,7506	0,0001	0,00001
Hindi The Anna Team on a cross-road	42	7*5 + 7	1,6063	0,0086	0,093
Hungarian A nominalizmus forradalma **(**Press)	63	5*10 + 13	2,9275	-0,0803	0,88

Hungarian Kunczekolbász **(**Press)	32	2*10 + 12	2,7347	-0,0370	0,23
Indonesian Pengurus PSM terbelah **(**Press)	28	2*10 + 8	2,6377	-0,0235	0,57
Indonesian Sekolah ditutup **(**Press)	15	3*5	2,664	-0,041	0,56
Italian (Press, Online)	92	8*10 + 12	2,0549	0,0316	0,605
Japanese Miki, Jinseiron Note, first 100 sentences	100	10*10	2,0588	0,0077	0,035
Kikongo Bimpa: Ma Ngo ya Ma Nsiese	79	7*10 + 9	2,0599	0,0082	0,048
Kikongo Lumumba speech	43	8*5 + 3	2,0057	0,0019	0,004
Kikongo Nkongo ye Kisi Kongo	29	5*5 +4	1,7691	0,0040	0,011
Latin Cicero, In Catilinam I	80	8*10	2,4390	-0,0174	0,18
Latin Cicero, In Catilinam II	180	10*18	2,3637	0,0072	0,04
Macedonian Ostrovskij, Kako se kaleše čelkiot, Chapter 1	89	8*10 + 9	2,3727	-0,0175	0,081
Malayalam 1, Moralistic Hooligans	32	5*5 + 7	4,2515	-0,1198	0,740
Malayalam 2, No one should die	29	5*5 + 4	4,337	-0,0421	0,1403
Maninka Nko Doumbu Kende no. 2	37	6*5 + 7	1,7004	0,0144	0,055
Maninka Nko Doumbu Kende no. 7	34	6*5 + 4	1,765	-0,0097	0,036
Maninka Sìikán` (Constitution of Guinea, an excerpt)	94	8*10 +14	1,9391	-0,0089	0,096
Maninka Teelen4	29	6*4 + 5	1,7783	-0,0343	0,49
Odia Calculation for the District Council President	28	5*5 + 3	3,0838	-0,0291	0,071
Odia Who is great?	36	6*5 + 6	2,7453	0,0258	0,042
Romanian Paler, Aventuri solitare (excerpt)	17	2*5 + 7	2,16	-0,047	0,07
Romanian Popescu D.R., Vânătoarea regală, Chapter 2	61	1 + 6*10	1,8656	-0,0164	0,16
Romanian Steinhardt, Jurnalul fericirii, Trei soluții	85	8*10 + 5	2,0128	0,0022	0,004
Russian Ostrovskij, Kak zakaljalas stal', Chapter 1	76	7*10 + 6	2,2137	-0,0074	0,014
Serbian Ostrovskij, Kako se kalio čelik, Chapter 1	83	7*10 + 13	2,2094	-0,0268	0,249
Slovak E. Bachletová, Moja Dolná zem	92	8*10 + 12	2,3278	-0,0152	0,12

Slovak E. Bachletová, Riadok v tlačive: nezamestnaný	78	7*10 + 8	2,5223	-0,0222	0,48
Slovenian Ostrovskij, Kako se je kalilo jeklo, Chapter 1	84	7*10 + 14	2,0790	-0,0273	0,41
Sundanese Agustusan (Salaka Online)	53	4*10 + 13	2,0031	0,0527	0,48
Sundanese Aki Satimi (Salaka Online)	147	6*20 + 27	2,0999	0,0117	0,142
Tagalog Hernandez, Limang Alas: Tatlong Santo	104	9*10 + 14	2,1362	-0,0086	0,04
Tagalog Hernandez, Magpisan	111	9*10 + 21	2,274	-0,0139	0,32
Tagalog Rosales, Kristal Na Tubig	139	6*20 + 19	2,1297	0,0055	0,044
Tamil (Press)	32	5*5 + 7	3,4209	0,0315	0,25
Telugu Trailangaswamy	51	4*10 + 11	4,1414	-0,1684	0,7
Telugu Train Journey without safety	61	5*10 + 11	3,6219	-0,0292	0,0893
Vai Mu ja vaa lɔ (T. Sherman)	193	9*20 + 13	1,4427	0,0063	0,7
Vai Sabu Mua Ko	39	3*10 + 9	1,575	-0,0327	0,208
Vai Vande bɛ Wu'u	35	3*10 + 5	1,4445	0,0054	0,035
Welsh T1 Crynodeb Gweithredol	40	8*5	2,0187	-0,00001	0,0000001
Welsh T2 Ffansi camu i esgidiau rhywun enwog?	34	6*5 + 4	1,6696	-0,0129	0,295

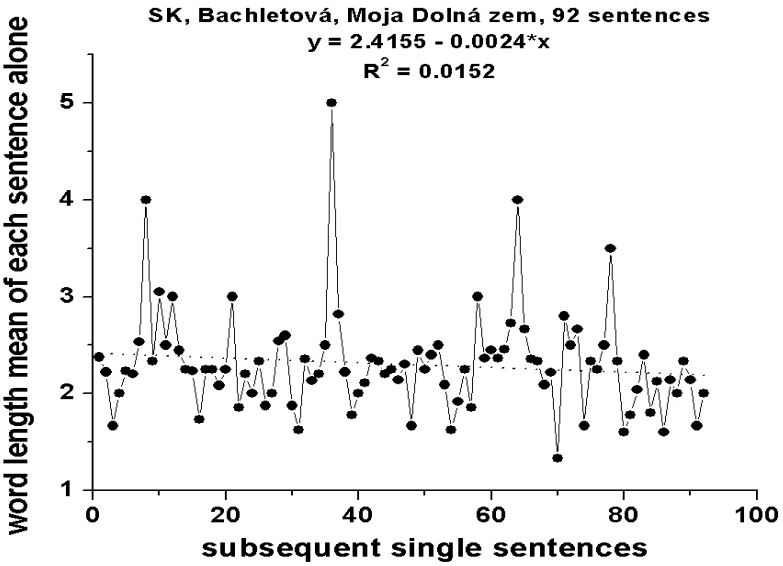


Figure 4.2, Mean word length of single subsequent sentences in *Moja Dolná zem* (Bachletová, Slovak)

Table 4.3
Mean word length in deploying text by subsequent single sentences: all texts

Language, Text alphabetically	**n**	**a**	**b**	**R^2**
Akan Mma Nnsua Ade Bɔnɛ	15	1,9105	-0,0308	0,269
Bamana Bamakɔ sigicogoya	84	1,6354	0,0009	0,0049
Bamana Masadennin	207	1,6034	0,00003	0,00003
Bamana Namakɔrɔba halakilen	248	1,6543	0,0002	0,0008
Bamana Sonsannin ani Surukuba	168	1,6494	-0,0005	0,0021
Bulgarian Ostrovskij, Kak se kaljavaše stomanata, Chapter 1	83	2,2393	-0,0017	0,0095
Czech Čulík, O čem jsou dnešní Spojené státy?	114	2,4454	-0,0016	0,017
Czech Hvížďala, O předem zpackané prezidentské volbě	37	2,3061	0,0022	0,0061
Czech Macháček, Slovenský dobrý příklad	22	2,3762	-0,0113	0,038
Czech Spurný, Prekvapení v justici	17	2,2814	-0,0043	0,0105
Czech Švehla, Editorial, Voličův kalkul	19	2,4003	-0,0189	0,129
French Dunkerque (Press)	105	1,7115	0,0009	0,008
German Assads Familiendiktatur (Press)	119	2,0519	0,0011	0,005
German ATT0012 (Press)	83	2,1897	-0,0033	0,04
German Die Stadt des Schweigens (Press)	120	2,0024	0,0006	0,001
German Terror in Ost Timor (Press)	98	2,0062	-0,0017	0,017
German Unter Hackern (Press)	110	2,2277	-0,0032	0,038
Hungarian A nominalizmus forradalma **(**Press)	63	2,8498	-0,008	0,16
Hungarian Kunczekolbász **(**Press)	32	2,8190	-0,0130	0,09
Indonesian Pengurus PSM terbelah **(**Press)	28	2,7030	-0,0058	0,03
Indonesian Sekolah ditutup **(**Press)	15	2,6414	-0,0010	0,0001
Italian (Press, Online)	92	2,0437	0,0038	0,1708
Japanese Miki, Jinseiron Note, first 100 sentences	100	2,0939	0,0006	0,003

Kikongo Bimpa: Ma Ngo ya Ma Nsiese	79	2,0596	-0,0008	0,0018
Kikongo Lumumba speech	43	2,1644	-0,0041	0,025
Kikongo Nkongo ye Kisi Kongo	29	1,8071	-0,0004	0,0008
Latin Cicero, In Catilinam I	80	2,3208	0,0003	0,0002
Latin Cicero, In Catilinam II	180	2,4055	0,00004	0,00003
Macedonian Ostrovskij, Kako se kaleše čelkiot, Chapter 1	89	2,4005	-0,0023	0,0145
Malayalam 1 (Press)	32	4,1761	-0,0187	0,07
Malayalam 2 (Press)	29	4,5570	-0,0202	0,0835
Maninka Nko Doumbu Kende no. 2	37	1,7203	0,0037	0,013
Maninka Nko Doumbu Kende no. 7	34	1,8257	-0,0039	0,022
Maninka Sìikán` (Constitution of Guinea, an excerpt)	94	1,9605	-0,0008	0,0065
Maninka Teelen4	29	1,7210	-0,0062	0,1526
Odia Calculation for the District Council President	28	3,1279	-0,0102	0,08
Odia Who is great?	36	2,7607	0,0078	0,038
Romanian Paler, Aventuri solitare (excerpt)	17	2,1140	-0,0031	0,005
Romanian Popescu D.R., Vânătoarea regală, Chapter 2	61	1,7864	0,0002	0,0001
Romanian Steinhardt, Jurnalul fericirii, Trei soluții	85	1,9558	0,0017	0,01
Russian Ostrovskij, Kak zakaljalas stal', Chapter 1	76	2,3033	-0,0017	0,004
Serbian Ostrovskij, Kako se kalio čelik, Chapter 1	83	2,2220	-0,0032	0,042
Slovak E. Bachletová, Moja Dolná zem	92	2,4155	-0,0024	0,015
Slovak E. Bachletová, Riadok v tlačive: nezamestnaný	78	2,5138	-0,0042	0,04
Slovenian Ostrovskij, Kako se je kalilo jeklo, Chapter 1	84	2,0996	-0,0033	0,05
Sundanese Agustusan (Salaka Online)	53	2,0697	0,0038	0,037
Sundanese Aki Satimi (Salaka Online)	147	2,1556	0,0007	0,005
Tagalog Hernandez, Limang Alas: Tatlong Santo	104	2,1218	-0,0002	0,0005
Tagalog Hernandez, Magpisan	111	2,2212	-0,0002	0,0004
Tagalog Rosales, Kristal Na Tubig	139	2,1674	0,0002	0,0006
Tamil (Press)	32	3,4603	0,0060	0,0147
Telugu Trailangaswamy	51	4,2411	-0,0184	0,1317

Telugu Train Journey without safety	61	3,6084	-0,002	0,005
Vai Mu ja vaa lɔ (T. Sherman)	193	1,4379	0,0004	0,0140
Vai Sabu Mua Ko	39	1,5679	-0,0045	0,0700
Vai Vande bɛ Wu'u	35	1,4627	-0,0007	0,0014
Welsh T1 Crynodeb Gweithredol	40	2,0516	-0,0004	0,0003
Welsh T2 Ffansi camu i esgidiau rhywun enwog?	34	1,6567	-0,0027	0,0427

For instance Kelih (2012) found a significant interrelation between text length and word length in Russian and Bulgarian texts: the longer the text, the longer the word length. This interrelation has been modelled by a (simple) power function. However, Kelih (2012) used a slightly different approach: firstly, single chapters of a novel have been cumulated stepwise, secondly word length has been measured stepwise in the cumulated chapters, thirdly, the strong and law-like interrelation has been obtained in word-form types only, and fourthly, the length of the used texts was over 100.000 tokens and over 18.000 types respectively. In the case of using word-form tokens, on the contrary, a decrase of word length with growing text length has be obtained (cf. Kelih 2012: 76f.), but this increase is less systematic than the one obtained on the types level. Thus, generally spoken, the analysis of Kelih (2012) is based on rather different "boundary conditions" than the analysis performed above.

The common feature of these sequences is a great oscillation at the beginning and a decrease at the end. Though in the Romanian text by Paler and in some Slavic languages the end slightly increases, the general trend is decreasing. Hence, whatever the way of measurement, the hypothesis that word length increases with increasing text is either falsified or we still neglect some boundary conditions which are not yet known. The sequences cannot be captured by a single function. The texts have their individualities which testify either to non-spontaneous creation or to changes made post-hoc.

As an example consider the course of mean word length in Akan**,** the beginning of the text *Mma Nnsua Ade Bɔne*, as presented in Table 4.4. Here it is sufficient to discard some of the first values and begin with a rather smooth part.

The course of values in individual texts, as shown in Figure 4.3, displays in some cases analogous behaviour (e.g. Vai texts) which does not bring about difficulties for a mathematician, but a linguist stays in front of the door to Dante's hell, reads the famous inscription and if he has courage, he mutters the famous rebellious words of Galilei and continues investigating this infinite domain.

Table 4.4
Cumulative mean word length in Akan

Position in text	Word length	Cumulative mean word length	
	2		
1	1	1.500	discarded
2	3	2.000	discarded
3	3	2.250	
4	1	2.000	
5	1	1.833	
6	1	1.714	
7	1	1.625	
8	4	1.889	
9	1	1.800	
10	3	1.909	
…	…	…	

Since the first word in text may be characteristic of text or language but it does not represent a mean, we propose to omit the first 9 words and begin to count taking the mean of the first 10 values as $x = 1$, then 1-11 as $x = 2$, etc. In this way the sequence changes and displays a more smooth course. If we use this method, i.e. if we smooth the beginning of the sequence, we obtain much more acceptable results. The course of means beginning with a concave course can be captured by the function

$$y = p_1 x^{p_2} (p_3 - \exp(-p_4 x)),$$

where p_i are the parameters. The first part of the function (the power part) represents the decrease of the mean word length, the exponential part represents the increasing beginning of the empirical sequence. The oscillation is not as strong as with the original measurement. However, sometimes even this function cannot capture the course, so that one must choose among three different courses expreesed by the curves

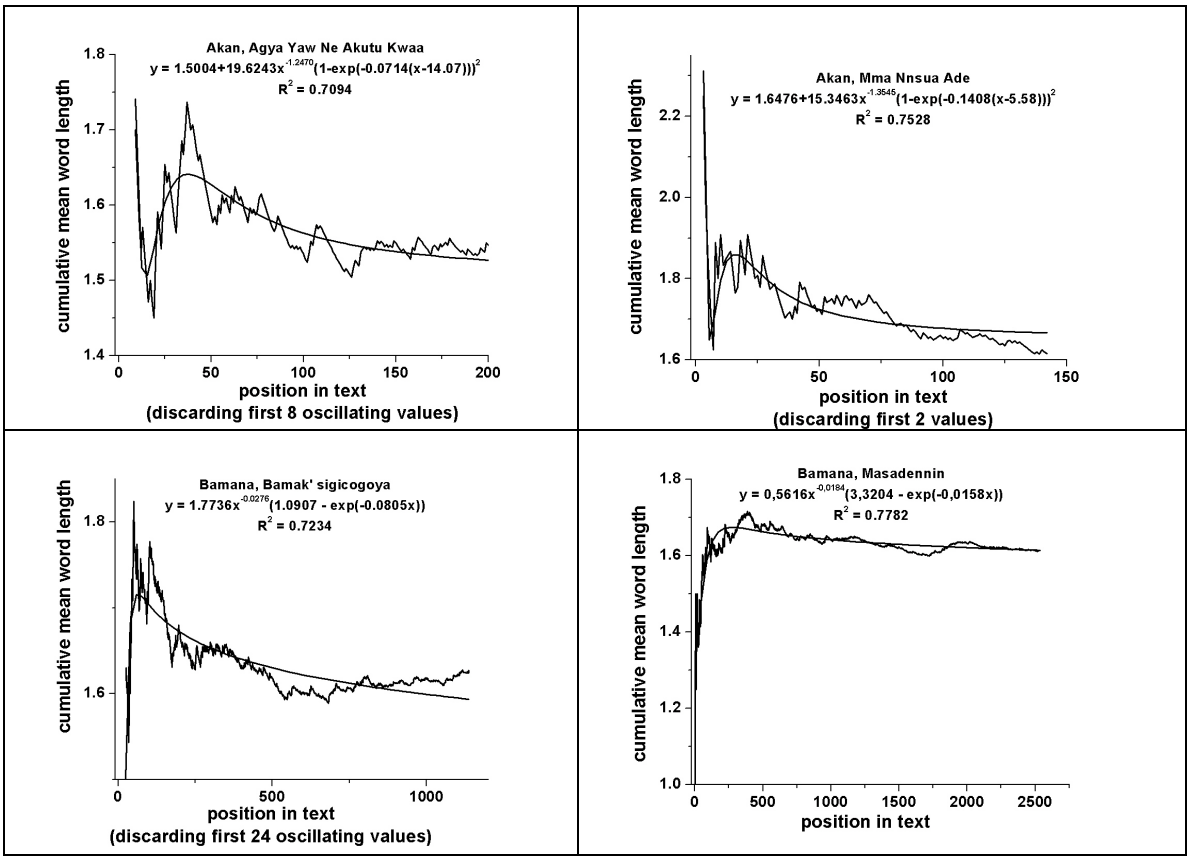
$$y = p_1 x^{-p_2}$$
$$y = p_1 + p_2 (1 - \exp(-p_3 (x - p_4)))^2 .$$

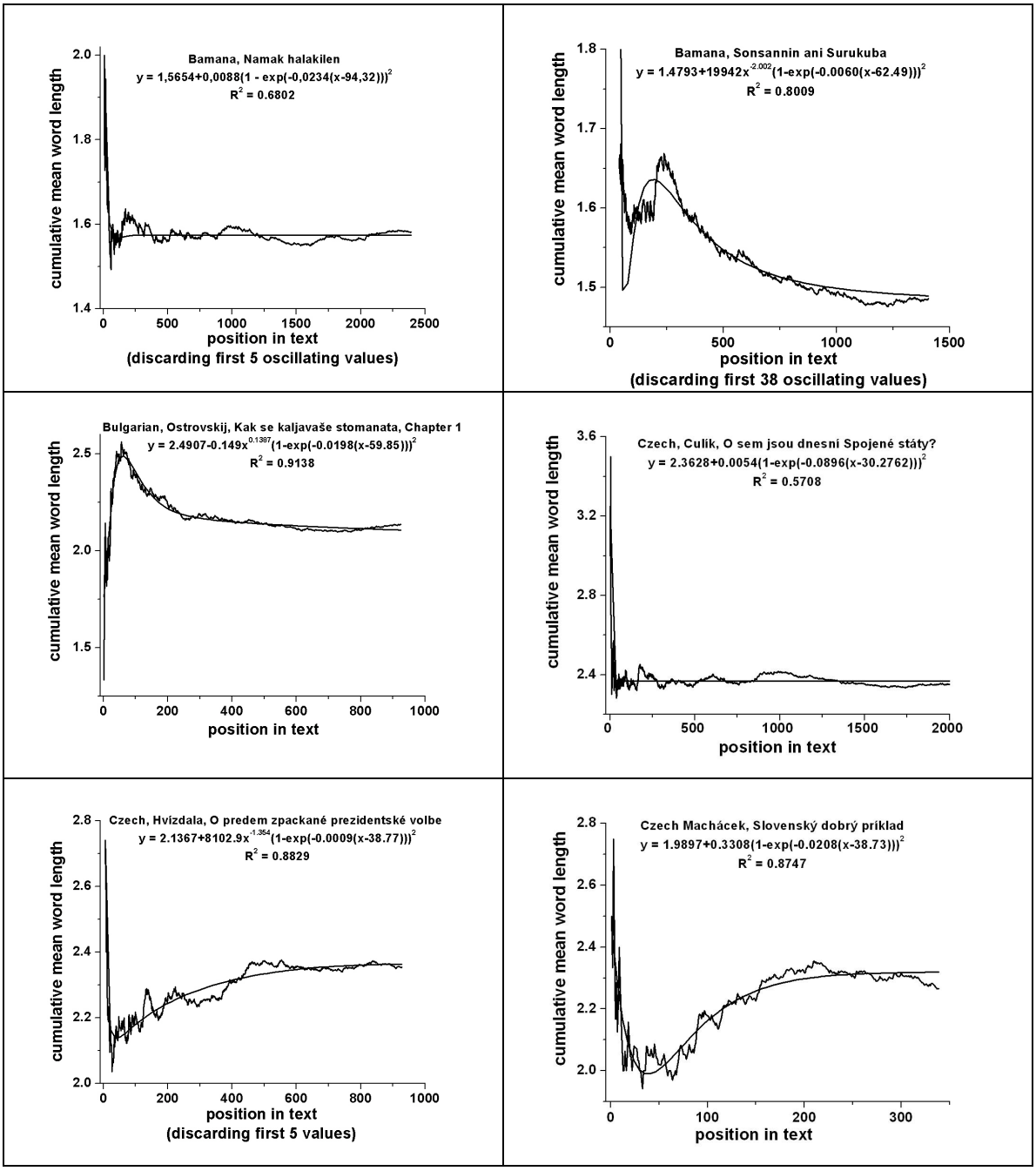
In some cases the first 10 (or more) values must simply be discarded - as has been made in Maninka (Doumbu Kende no 2), Slovak (Moja Dolná zem), or one must perform a further combination of functions, namely to modify the Morse function ad hoc by the power function to obtain

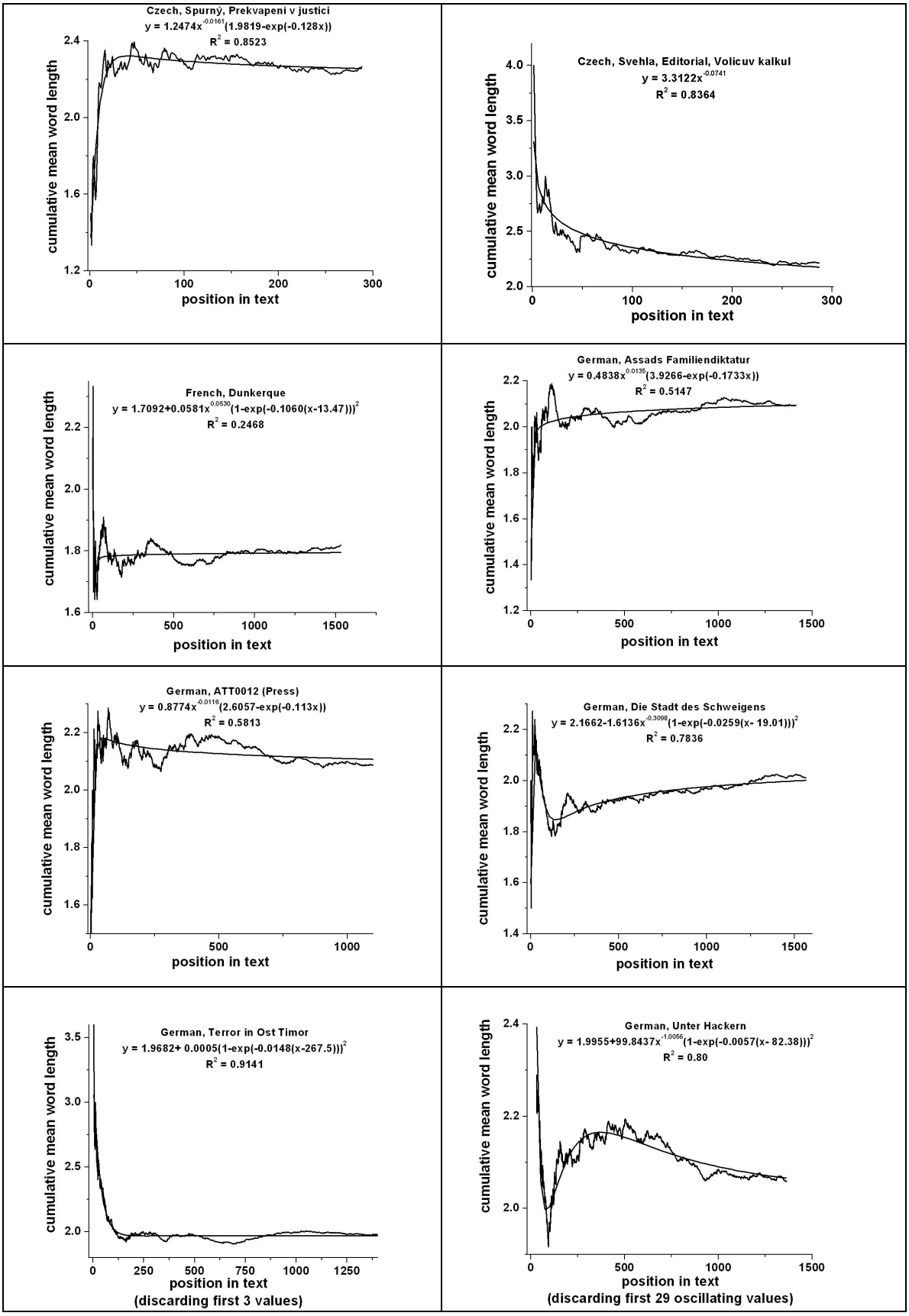
$$y = p_1 + p_2 x^{-p_3}(1 - \exp(-p_4(x - p_5)))^2 .$$

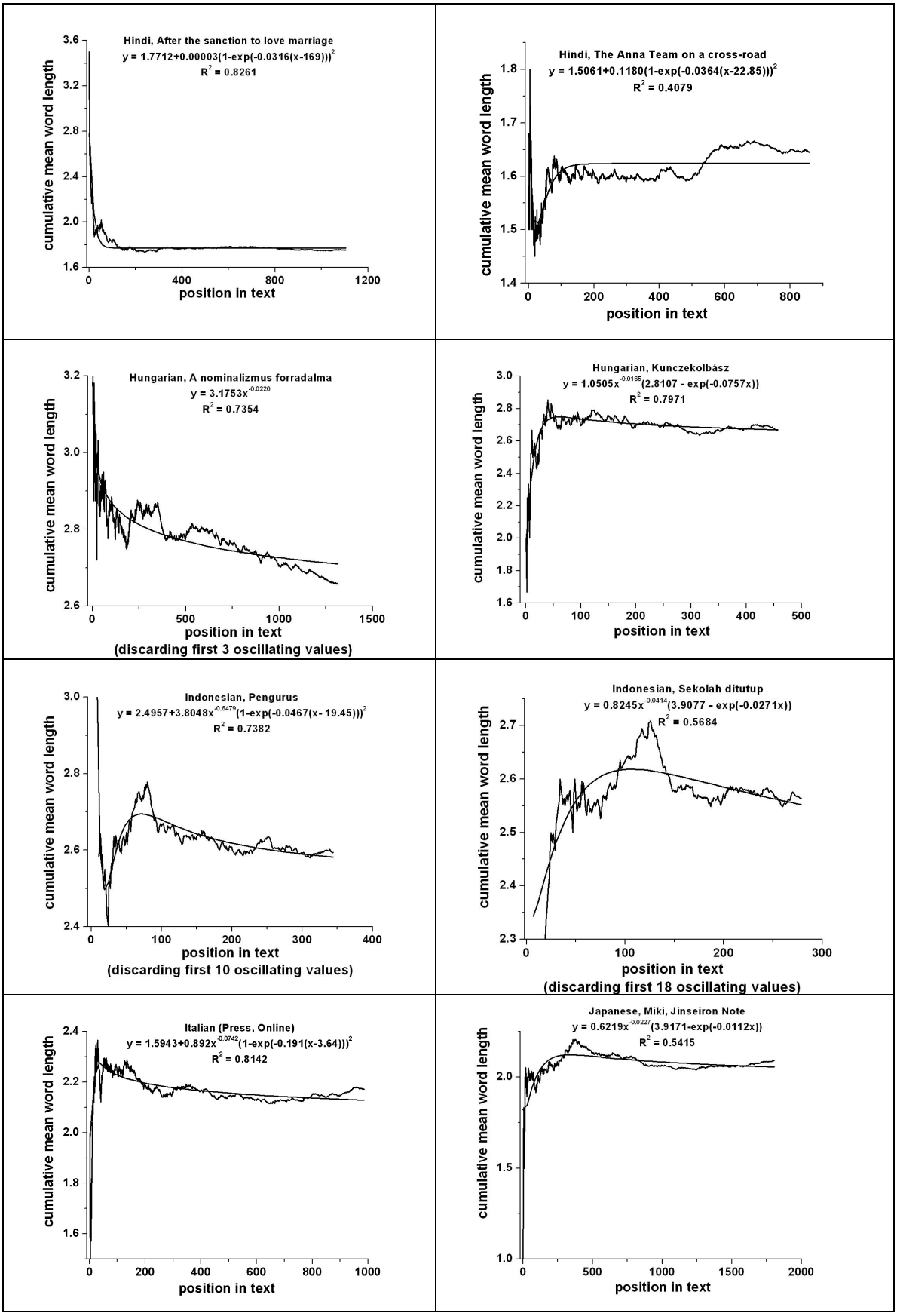
which can, however, roughly fit the general course of the stepwise mean word length of Bachletová's *Moja Dolná zem*. As indicated under the Ox axis title of the plots given in Figure 4.3, the non relevant initial oscillations were circumvented either by counting from the mean of the first 10 values (as in Vai, *Vande*; Maninka, *Sìikán`*; Maninka, *Teleen 4*) or simply discarding them (as in Maninka, *Nko Doumbu Kende no. 2*; Bachletová, *Moja Dolná zem*). Notice that not always these procedures are necessary.

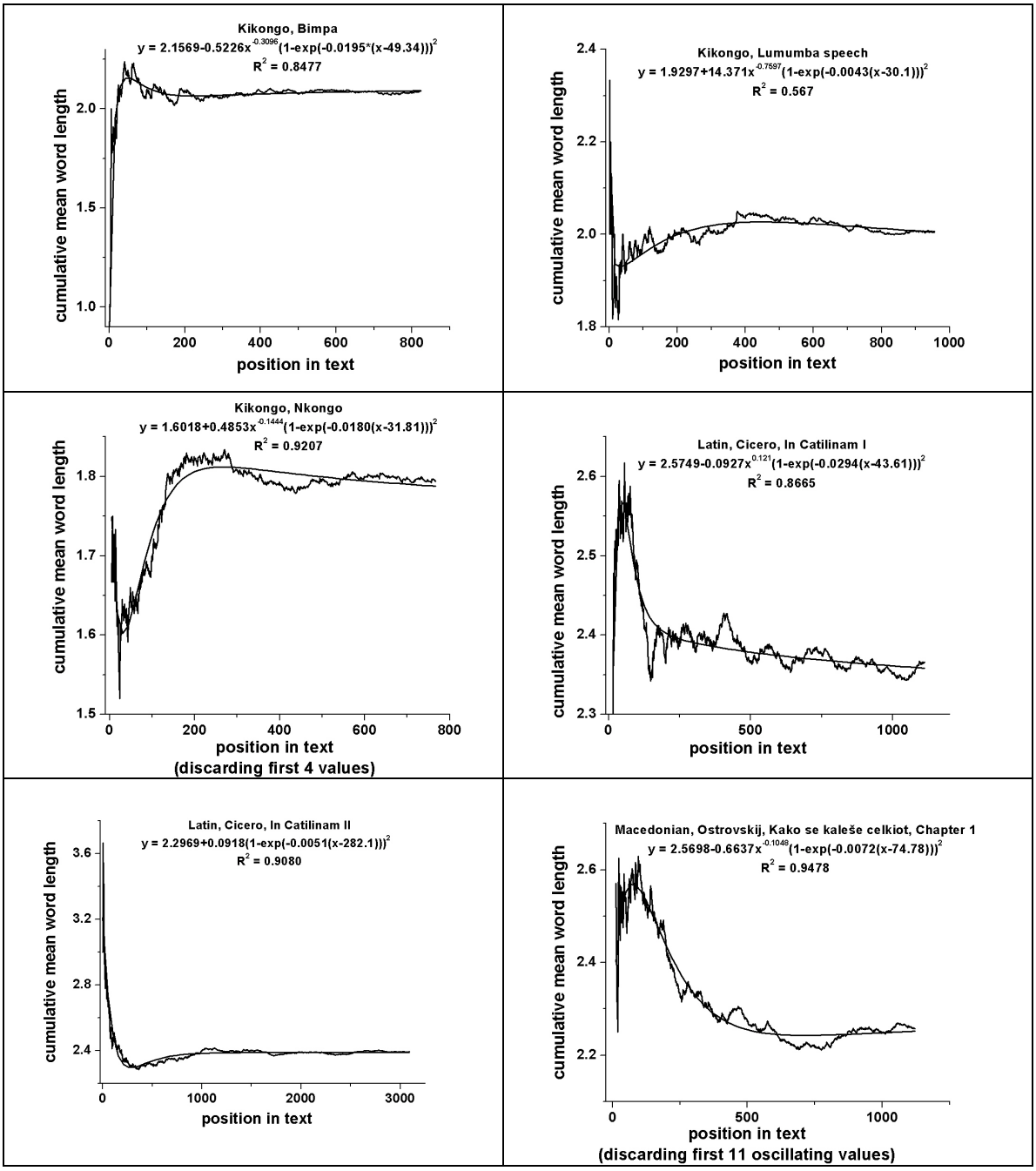
The individual functions are presented in Table 4.5 and Figure 4.4.

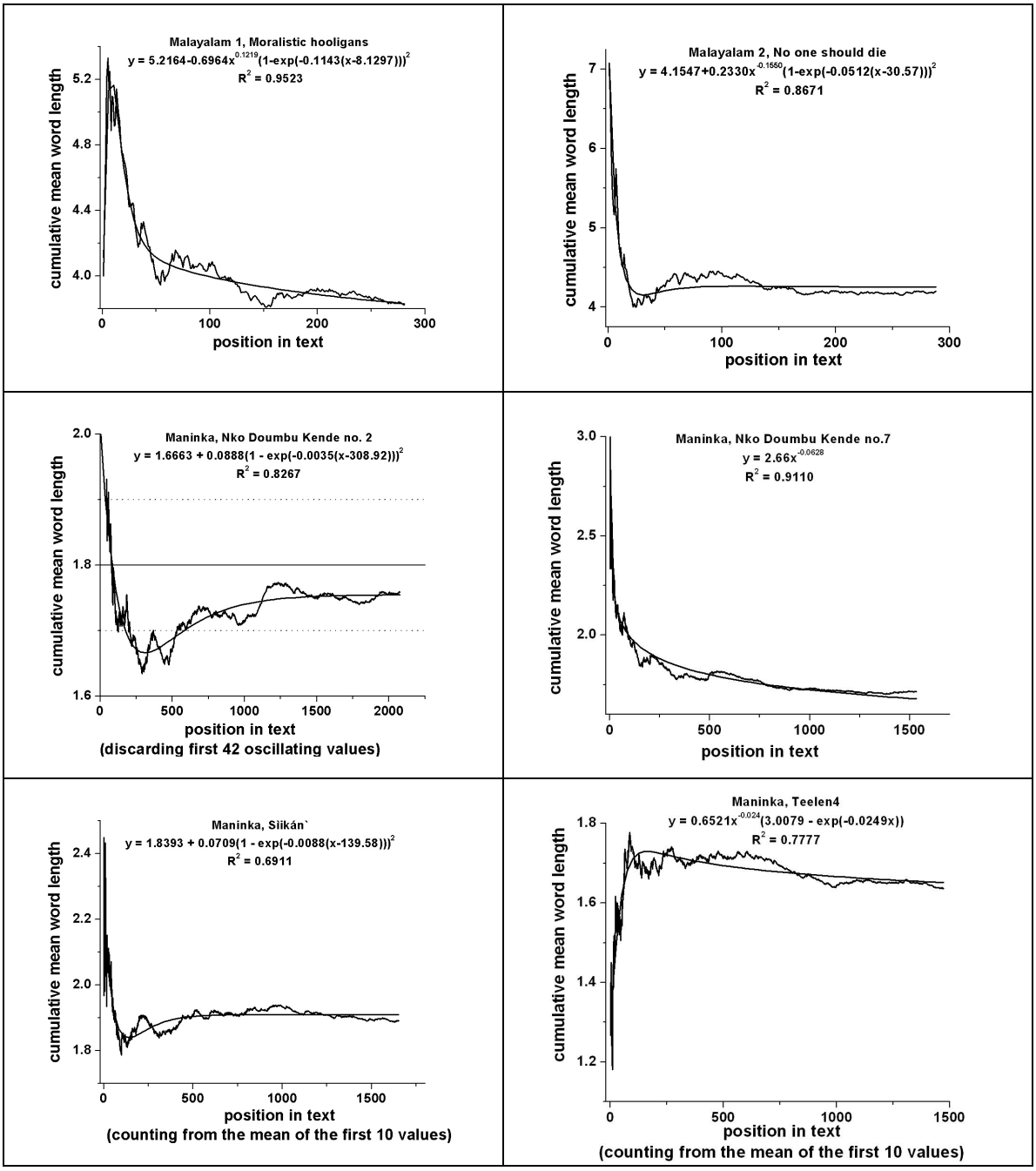


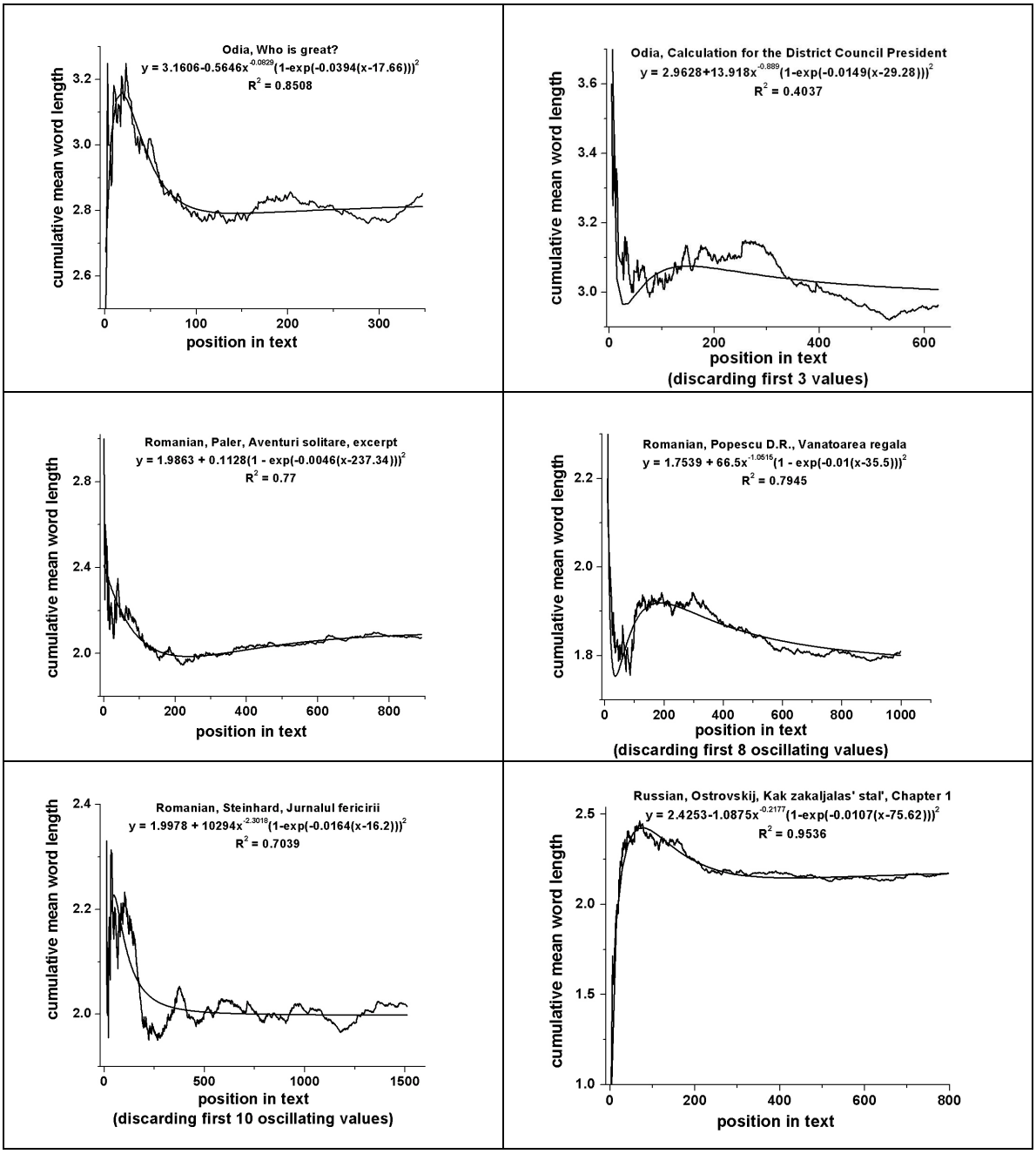

Bamana, Namak halakilen
y = 1,5654+0,0088(1 - exp(-0,0234(x-94,32)))^2
R^2 = 0.6802
cumulative mean word length
position in text
(discarding first 5 oscillating values)
Bamana, Sonsannin ani Surukuba
y = 1.4793+19942x^-2.002(1-exp(-0.0060(x-62.49)))^2
R^2 = 0.8009
cumulative mean word length
position in text
(discarding first 38 oscillating values)
Bulgarian, Ostrovskij, Kak se kaljavaše stomanata, Chapter 1
y = 2.4907-0.149x^0.1387(1-exp(-0.0198(x-59.85)))^2
R^2 = 0.9138
cumulative mean word length
position in text
Czech, Culik, O sem jsou dnesní Spojené státy?
y = 2.3628+0.0054(1-exp(-0.0896(x-30.2762)))^2
R^2 = 0.5708
cumulative mean word length
position in text
Czech, Hvízdala, O predem zpackané prezidentské volbe
y = 2.1367+8102.9x^-1.354(1-exp(-0.0009(x-38.77)))^2
R^2 = 0.8829
cumulative mean word length
position in text
(discarding first 5 values)
Czech Machácek, Slovenský dobrý príklad
y = 1.9897+0.3308(1-exp(-0.0208(x-38.73)))^2
R^2 = 0.8747
cumulative mean word length
position in text

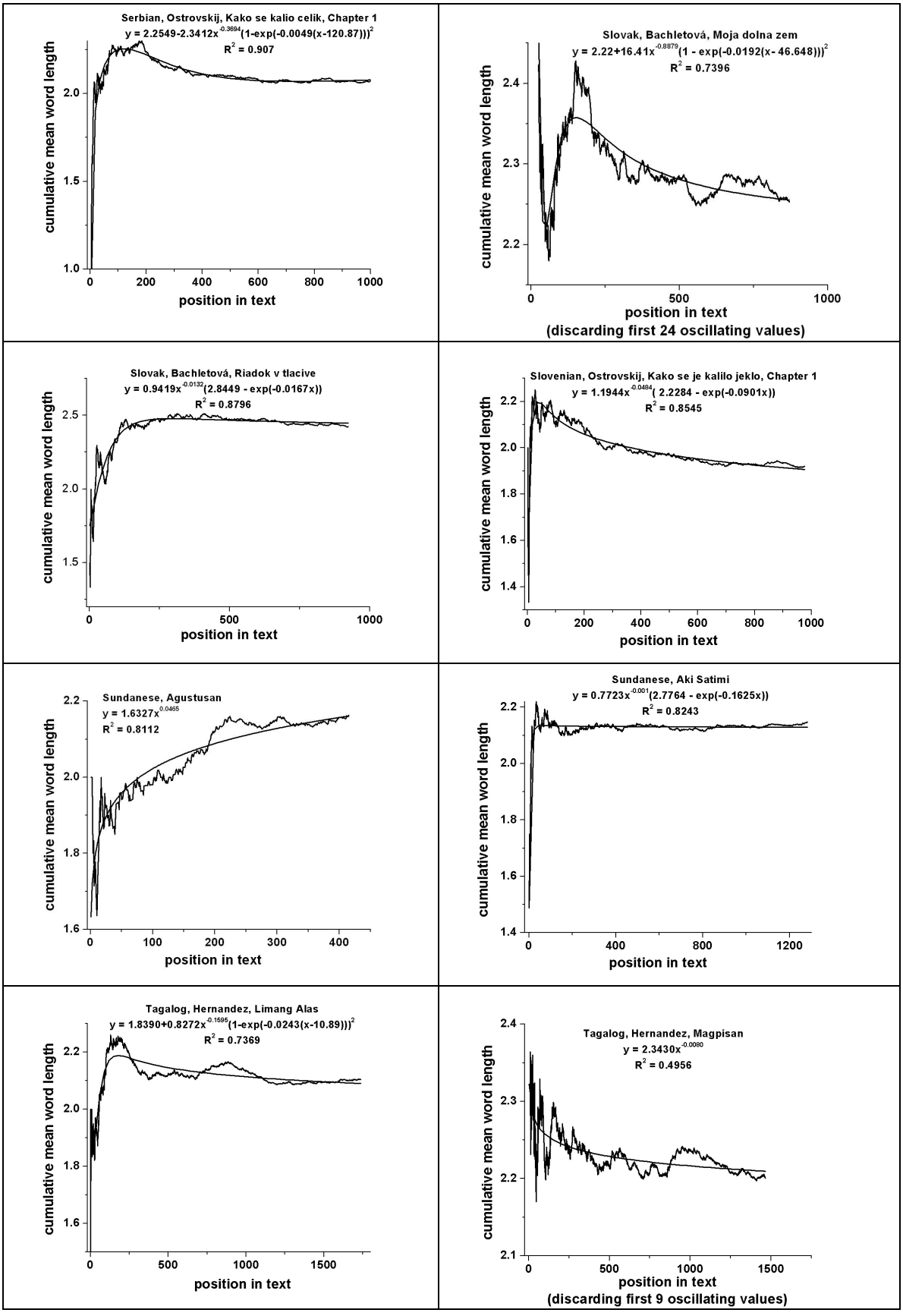

Czech, Spurný, Prekvapeni v justici
y = 1.2474x^-0.0161(1.9819-exp(-0.128x))
R^2 = 0.8523
cumulative mean word length
position in text
Czech, Svehla, Editorial, Volicuv kalkul
y = 3.3122x^-0.0741
R^2 = 0.8364
cumulative mean word length
position in text
French, Dunkerque
y = 1.7092+0.0581x^0.0530(1-exp(-0.1060(x-13.47)))^2
R^2 = 0.2468
cumulative mean word length
position in text
German, Assads Familiendiktatur
y = 0.4838x^0.0135(3.9266-exp(-0.1733x))
R^2 = 0.5147
cumulative mean word length
position in text
German, ATT0012 (Press)
y = 0.8774x^-0.0116(2.6057-exp(-0.113x))
R^2 = 0.5813
cumulative mean word length
position in text
German, Die Stadt des Schweigens
y = 2.1662-1.6136x^-0.3098(1-exp(-0.0259(x- 19.01)))^2
R^2 = 0.7836
cumulative mean word length
position in text
German, Terror in Ost Timor
y = 1.9682+ 0.0005(1-exp(-0.0148(x-267.5)))^2
R^2 = 0.9141
cumulative mean word length
position in text
(discarding first 3 values)
German, Unter Hackern
y = 1.9955+99.8437x^-1.0056(1-exp(-0.0057(x- 82.38)))^2
R^2 = 0.80
cumulative mean word length
position in text
(discarding first 29 oscillating values)

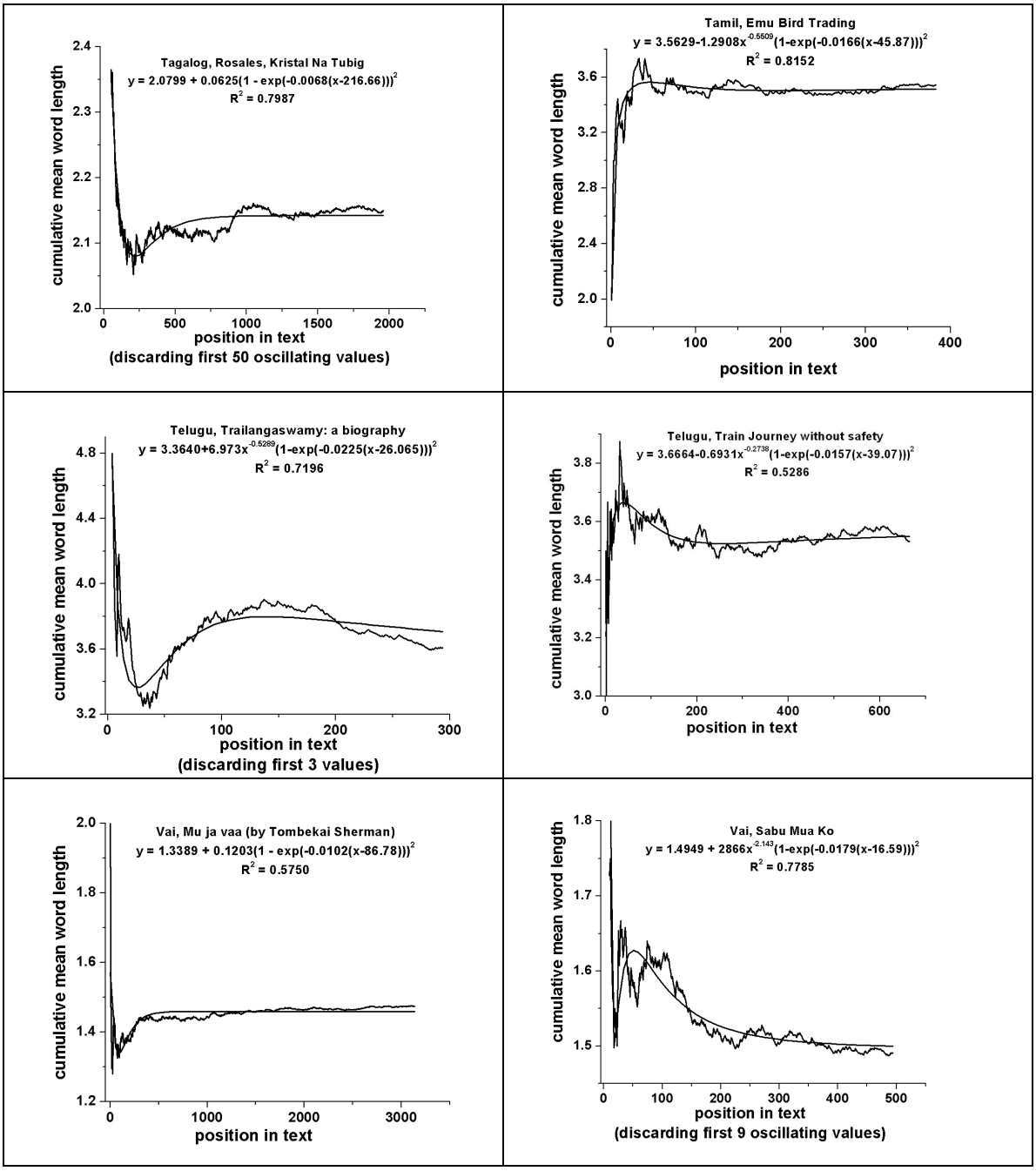

Hindi, After the sanction to love marriage
y = 1.7712+0.00003(1-exp(-0.0316(x-169)))^2
R^2 = 0.8261
cumulative mean word length
position in text
Hindi, The Anna Team on a cross-road
y = 1.5061+0.1180(1-exp(-0.0364(x-22.85)))^2
R^2 = 0.4079
cumulative mean word length
position in text
Hungarian, A nominalizmus forradalma
y = 3.1753x^-0.0220
R^2 = 0.7354
cumulative mean word length
position in text
(discarding first 3 oscillating values)
Hungarian, Kunczekolbász
y = 1.0505x^-0.0165(2.8107 - exp(-0.0757x))
R^2 = 0.7971
cumulative mean word length
position in text
Indonesian, Pengurus
y = 2.4957+3.8048x^-0.6479(1-exp(-0.0467(x- 19.45)))^2
R^2 = 0.7382
cumulative mean word length
position in text
(discarding first 10 oscillating values)
Indonesian, Sekolah ditutup
y = 0.8245x^-0.0414(3.9077 - exp(-0.0271x))
R^2 = 0.5684
cumulative mean word length
position in text
(discarding first 18 oscillating values)
Italian (Press, Online)
y = 1.5943+0.892x^-0.0742(1-exp(-0.191(x-3.64)))^2
R^2 = 0.8142
cumulative mean word length
position in text
Japanese, Miki, Jinseiron Note
y = 0.6219x^-0.0227(3.9171-exp(-0.0112x))
R^2 = 0.5415
cumulative mean word length
position in text


Kikongo, Bimpa
y = 2.1569-0.5226x^-0.3096(1-exp(-0.0195*(x-49.34)))^2
R^2 = 0.8477
cumulative mean word length
position in text
Kikongo, Lumumba speech
y = 1.9297+14.371x^-0.7597(1-exp(-0.0043(x-30.1)))^2
R^2 = 0.567
cumulative mean word length
position in text
Kikongo, Nkongo
y = 1.6018+0.4853x^-0.1444(1-exp(-0.0180(x-31.81)))^2
R^2 = 0.9207
cumulative mean word length
position in text
(discarding first 4 values)
Latin, Cicero, In Catilinam I
y = 2.5749-0.0927x^0.121(1-exp(-0.0294(x-43.61)))^2
R^2 = 0.8665
cumulative mean word length
position in text
Latin, Cicero, In Catilinam II
y = 2.2969+0.0918(1-exp(-0.0051(x-282.1)))^2
R^2 = 0.9080
cumulative mean word length
position in text
Macedonian, Ostrovskij, Kako se kaleše celkiot, Chapter 1
y = 2.5698-0.6637x^-0.1048(1-exp(-0.0072(x-74.78)))^2
R^2 = 0.9478
cumulative mean word length
position in text
(discarding first 11 oscillating values)


Malayalam 1, Moralistic hooligans
y = 5.2164-0.6964x^0.1219(1-exp(-0.1143(x-8.1297)))^2
R^2 = 0.9523
cumulative mean word length
position in text
Malayalam 2, No one should die
y = 4.1547+0.2330x^-0.1550(1-exp(-0.0512(x-30.57)))^2
R^2 = 0.8671
cumulative mean word length
position in text
Maninka, Nko Doumbu Kende no. 2
y = 1.6663 + 0.0888(1 - exp(-0.0035(x-308.92)))^2
R^2 = 0.8267
cumulative mean word length
position in text
(discarding first 42 oscillating values)
Maninka, Nko Doumbu Kende no.7
y = 2.66x^-0.0628
R^2 = 0.9110
cumulative mean word length
position in text
Maninka, Siikán`
y = 1.8393 + 0.0709(1 - exp(-0.0088(x-139.58)))^2
R^2 = 0.6911
cumulative mean word length
position in text
(counting from the mean of the first 10 values)
Maninka, Teelen4
y = 0.6521x^-0.024(3.0079 - exp(-0.0249x))
R^2 = 0.7777
cumulative mean word length
position in text
(counting from the mean of the first 10 values)


Odia, Who is great?
y = 3.1606-0.5646x^-0.0829(1-exp(-0.0394(x-17.66)))^2
R^2 = 0.8508
cumulative mean word length
position in text
Odia, Calculation for the District Council President
y = 2.9628+13.918x^-0.889(1-exp(-0.0149(x-29.28)))^2
R^2 = 0.4037
cumulative mean word length
position in text
(discarding first 3 values)
Romanian, Paler, Aventuri solitare, excerpt
y = 1.9863 + 0.1128(1 - exp(-0.0046(x-237.34)))^2
R^2 = 0.77
cumulative mean word length
position in text
Romanian, Popescu D.R., Vanatoarea regala
y = 1.7539 + 66.5x^-1.0515(1 - exp(-0.01(x-35.5)))^2
R^2 = 0.7945
cumulative mean word length
position in text
(discarding first 8 oscillating values)
Romanian, Steinhard, Jurnalul fericirii
y = 1.9978 + 10294x^-2.3018(1-exp(-0.0164(x-16.2)))^2
R^2 = 0.7039
cumulative mean word length
position in text
(discarding first 10 oscillating values)
Russian, Ostrovskij, Kak zakaljalas' stal', Chapter 1
y = 2.4253-1.0875x^-0.2177(1-exp(-0.0107(x-75.62)))^2
R^2 = 0.9536
cumulative mean word length
position in text


Serbian, Ostrovskij, Kako se kalio celik, Chapter 1
y = 2.2549-2.3412x^-0.3694(1-exp(-0.0049(x-120.87)))^2
R^2 = 0.907
cumulative mean word length
position in text
Slovak, Bachletová, Moja dolna zem
y = 2.22+16.41x^-0.8879(1 - exp(-0.0192(x- 46.648)))^2
R^2 = 0.7396
cumulative mean word length
position in text
(discarding first 24 oscillating values)
Slovak, Bachletová, Riadok v tlacive
y = 0.9419x^-0.0132(2.8449 - exp(-0.0167x))
R^2 = 0.8796
cumulative mean word length
position in text
Slovenian, Ostrovskij, Kako se je kalilo jeklo, Chapter 1
y = 1.1944x^-0.0484(2.2284 - exp(-0.0901x))
R^2 = 0.8545
cumulative mean word length
position in text
Sundanese, Agustusan
y = 1.6327x^0.0465
R^2 = 0.8112
cumulative mean word length
position in text
Sundanese, Aki Satimi
y = 0.7723x^-0.001(2.7764 - exp(-0.1625x))
R^2 = 0.8243
cumulative mean word length
position in text
Tagalog, Hernandez, Limang Alas
y = 1.8390+0.8272x^-0.1595(1-exp(-0.0243(x-10.89)))^2
R^2 = 0.7369
cumulative mean word length
position in text
Tagalog, Hernandez, Magpisan
y = 2.3430x^-0.0080
R^2 = 0.4956
cumulative mean word length
position in text
(discarding first 9 oscillating values)


Tagalog, Rosales, Kristal Na Tubig
y = 2.0799 + 0.0625(1 - exp(-0.0068(x-216.66)))^2
R^2 = 0.7987
cumulative mean word length
position in text
(discarding first 50 oscillating values)
Tamil, Emu Bird Trading
y = 3.5629-1.2908x^-0.5509(1-exp(-0.0166(x-45.87)))^2
R^2 = 0.8152
cumulative mean word length
position in text
Telugu, Trailangaswamy: a biography
y = 3.3640+6.973x^-0.5289(1-exp(-0.0225(x-26.065)))^2
R^2 = 0.7196
cumulative mean word length
position in text
(discarding first 3 values)
Telugu, Train Journey without safety
y = 3.6664-0.6931x^-0.2738(1-exp(-0.0157(x-39.07)))^2
R^2 = 0.5286
cumulative mean word length
position in text
Vai, Mu ja vaa (by Tombekai Sherman)
y = 1.3389 + 0.1203(1 - exp(-0.0102(x-86.78)))^2
R^2 = 0.5750
cumulative mean word length
position in text
Vai, Sabu Mua Ko
y = 1.4949 + 2866x^2.143(1-exp(-0.0179(x-16.59)))^2
R^2 = 0.7785
cumulative mean word length
position in text
(discarding first 9 oscillating values)

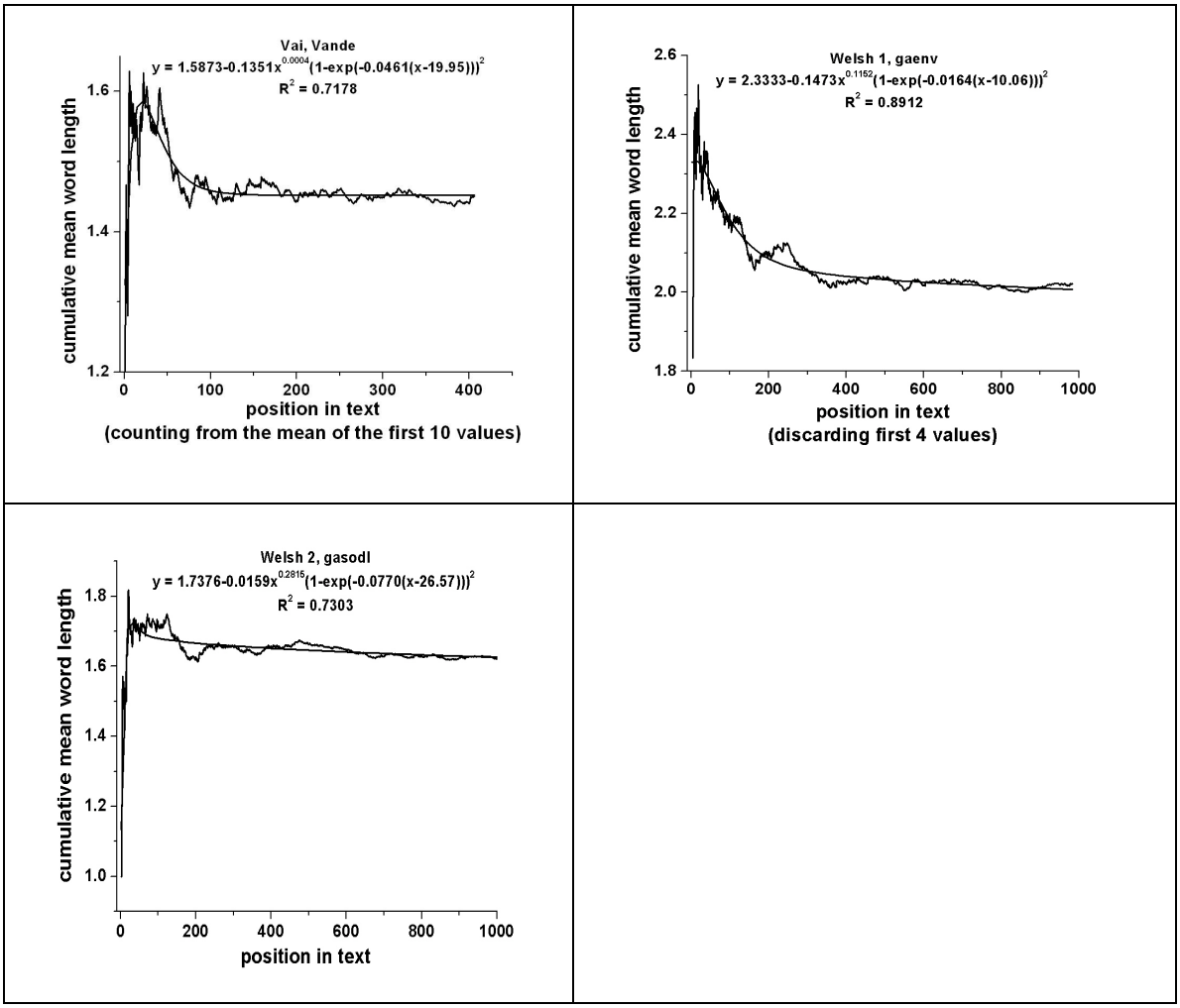


Figure 4.4. Course of mean word lengths in 61 texts of 28 languages fitted by function (5)

This is a purely explorative approach not having preliminarily a sound linguistic substantiation. Nevertheless, the function can be considered a family. The most general function is (J. Mačutek, private communication)

$$y = p_1 + p_2 x^{p_3}\left(p_4 - e^{p_5(x-p_6)}\right)^{p_7}, \tag{5}$$

resulting from the differential equation[1]

[1] Denoting the "shifted" mean word length by $Y = y - p_1$ the following differential equation (6) is of the type Y'/Y = difference between two functions of the position x in text. In other words, we get again the "fight" between two contrary actions along the text, one increasing the word length (autosemantics) and the other one decreasing the word length (synsemantics, auxiliaries).

$$(6) \qquad \frac{y'}{y-p_1} = \frac{p_3}{x} - \frac{p_5 p_7 e^{p_5 |x-p_6}}{p_4 - e^{p_5 (x-p_6)}}$$

where the parameter p_1 on the left hand side shows that the values of y must be greater than a certain constant; usually it is 1. Parameter p_3 is, as usually, the language constant; the functions in the last expression are somehat complex and represent non-linear relations between the force of hearer (numerator) and the force of community (selfregulation in denominator). This is a generalization of the Wimmer-Altmann (2005) model. From formula (5) we obtain all functions used in Table 11 by replacing some parameters by constant values. Hence

$$(7) \qquad y = p_1 + p_2 x^{p_3} \left(1 - e^{p_5 (x-p_6)}\right)^2$$

results if $p_4 = 1$, $p_7 = 2$;

$$(8) \qquad y = p_1 + p_2 \left(1 - e^{p_5 (x-p_6)}\right)^2$$

results if $p_3 = 0$, $p_4 = 1$, $p_7 = 2$;

$$(8) \qquad y = p_2 x^{p_3} \left(p_4 - e^{p_5 x}\right)$$

results if $p_1 = 0$, $p_6 = 0$, $p_7 = 1$;. and

$$(10) \qquad y = p_2 x^{p_3}$$

results if e.g. $p_1 = 0$ and $p_4 - e^{p_5 (x-p_6)} = 1$ or e.g., $p_4 = 2$, $p_5 = 0$, and many other parameter values. This case shows us that some phenomena behave quite differently than expected by the previous theory. Both its extension and inclusion in Köhler's control cycle would be very useful.

Table 4.6
The course of mean word lengths in texts in 61 texts of 28 languages fitted by function (5) (with smoothed or discarded beginning as indicated in the title of the Ox axis of plots given in Figure 5)

Language, Text alphabetically	**Function**	**R^2**
Akan Agya Yaw Ne Akutu Kwaa	$1,5004+19,6243x^{-1,2470}(1-\exp(-0,0714(x-14,07)))^2$	0,71
Akan Mma Nnsua Ade Bɔne	$1,6476+15,3463x^{-1,3545}(1-\exp(-0,1408(x-5,58)))^2$	0,75
Bamana Bamakɔ sigicogoya	$1,7736x^{-0,0276}(1,0907 - \exp(-0,0805x))$	0,72
Bamana Masadennin	$0,5616x^{-0,0184}(3,3204 - \exp(-0,0158x))$	0,78
Bamana Namakɔrɔba halakilen	$1,5654+0,0088(1 - \exp(-0,0234(x-94,32)))^2$	0,68
Bamana Sonsannin ani Surukuba	$1,4793+19942x^{-2,002}(1-\exp(-0,0060(x-62,49)))^2$	0,80
Bulgarian Ostrovskij, Kak se kaljavaše stomanata, Chapter 1	$2,4907-0,149x^{0,1387}(1-\exp(-0,0198(x-59,85)))^2$	0,91
Czech Čulík, O čem jsou dnešní Spojené státy?	$2,3628+0,0054(1-\exp(-0,0896(x-30,2762)))^2$	0,57
Czech Hvížďala, O předem zpackané prezidentské volbě	$2,1367+8102,9x^{-1,354}(1-\exp(-0,0009(x-38,77)))^2$	0,88
Czech Macháček, Slovenský dobrý příklad	$1,9897+0,3308(1-\exp(-0,0208(x-38,73)))^2$	0,87
Czech Spurný, Prekvapení v justici	$1,2474x^{-0,0161}(1,9819-\exp(-0,128x))$	0,85
Czech Švehla, Editorial, Voličův kalkul	$3,3122x^{-0,0741}$	0,84
French Dunkerque (Press)	$1,7092+0,0581x^{0,0530}(1-\exp(-0,1060(x-13,5)))^2$	0,25
German Assads Familiendiktatur (Press)	$0,4838x^{0,0135}(3,9266-\exp(-0,1733x))$	0,51
German ATT0012 (Press)	$0,8774x^{-0,0116}(2,6057-\exp(-0,113x))$	0,58
German Die Stadt des Schweigens (Press)	$2,1662-1,6136x^{-0,3098}(1-\exp(-0,0259(x- 19,0054)))^2$	0,78
German Terror in Ost Timor (Press)	$1,9682+ 0,0005(1-\exp(-0,0148(x-267,5)))^2$	0,91
German Unter Hackern (Press)	$1,9955+99,8437x^{-1,0056}(1-\exp(-0,0057(x- 82,38)))^2$	0,80
Hindi After the sanction to love marriage	$1,7712+0,00003(1-\exp(-0,0316(x-169)))^2$	0,83
Hindi The Anna Team on a cross-road	$1,5061+0,1180(1-\exp(-0,0364(x-22,85)))^2$	0,41

Hungarian A nominalizmus forradalma **(**Press)	$3{,}1753x^{-0{,}0220}$	0,74
Hungarian Kunczekolbász **(**Press)	$1{,}0505x^{-0{,}0165}(2{,}8107 - \exp(-0{,}0757x))$	0,80
Indonesian Pengurus PSM terbelah **(**Press)	$2{,}4957+3{,}8048x^{-0{,}6479}(1-\exp(-0{,}0467(x- 19{,}45)))^2$	0,74
Indonesian Sekolah ditutup **(**Press)	$0{,}8245x^{-0{,}0414}(3{,}9077 - \exp(-0{,}0271x))$	0,57
Italian (Press, Online)	$1{,}5943+0{,}892x^{-0{,}0742}(1-\exp(-0{,}191(x-3{,}64)))^2$	0,81
Japanese Miki, Jinseiron Note, first 100 sentences	$0{,}6219x^{-0{,}0227}(3{,}9171-\exp(-0{,}0112x))$	0,54
Kikongo Bimpa: Ma Ngo ya Ma Nsiese	$2{,}1569-0{,}5226x^{-0{,}3096}(1-\exp(-0{,}0195(x-49{,}34)))^2$	0,85
Kikongo Lumumba speech	$1{,}9297+14{,}371x^{-0{,}7597}(1-\exp(-0{,}0043(x-30{,}1)))2$	0,57
Kikongo Nkongo ye Kisi Kongo	$1{,}6018+0{,}4853x^{-0{,}1444}(1-\exp(-0{,}0180(x-31{,}81)))^2$	0,92
Latin Cicero, In Catilinam I	$2{,}5749-0{,}0927x^{0{,}121}(1-\exp(-0{,}0294(x-43{,}61)))^2$	0,87
Latin Cicero, In Catilinam II	$2{,}2969+0{,}0918(1-\exp(-0{,}0051(x-282{,}1)))^2$	0,91
Macedonian Ostrovskij, Kako se kaleše čelkiot, Chapter 1	$2{,}5698-0{,}6637x^{-0{,}1048}(1-\exp(-0{,}0072(x-74{,}78)))^2$	0,95
Malayalam 1, Moralistic Hooligans	$5{,}2164-0{,}6964x^{0{,}1219}(1-\exp(-0{,}1143(x-8{,}1297)))^2$	0,95
Malayalam 2, No one should die	$4{,}1547+0{,}2330x^{-0{,}1550}(1-\exp(-0{,}0512(x-30{,}57)))^2$	0,87
Maninka Nko Doumbu Kende no. 2	$1{,}6663 + 0{,}0888(1 - \exp(-0{,}0035(x-308{,}92)))^2$	0,83
Maninka Nko Doumbu Kende no. 7	$2{,}66x^{-0{,}0628}$	0,91
Maninka Siìkán` (Constitution of Guinea, an excerpt)	$1{,}8393 + 0{,}0709(1 - \exp(-0{,}0088(x-139{,}58)))^2$	0,69
Maninka Teelen4	$0{,}6521x^{-0{,}024}(3{,}0079 - \exp(-0{,}0249x))$	0,78
Odia Calculation for the District Council President	$2{,}9628+13{,}918x^{-0{,}889}(1-\exp(-0{,}0149(x-29{,}28)))^2$	0,40
Odia Who is great?	$3{,}1606-0{,}5646x^{-0{,}0829}(1-\exp(-0{,}0394(x-17{,}66)))^2$	0,85
Romanian Paler, Aventuri solitare (excerpt)	$1{,}9863 + 0{,}1128(1 - \exp(-0{,}0046(x-237{,}34)))^2$	0,77
Romanian Popescu D.R., Vânătoarea regală, Chapter 2	$1{,}7539 + 66{,}5x^{-1{,}0515}(1 - \exp(-0{,}01(x-35{,}5)))^2$	0,80
Romanian Steinhardt, Jurnalul fericirii, Trei soluții	$1{,}9978 + 10294x^{-2{,}3018}(1-\exp(-0{,}0164(x-16{,}2)))^2$	0,70
Russian Ostrovskij, Kak zakaljalas stal', Chapter 1	$2{,}4253-1{,}0875x^{-0{,}2177}(1-\exp(-0{,}0107(x-75{,}62)))2$	0,95
Serbian Ostrovskij, Kako se kalio čelik, Chapter 1	$2{,}2549-2{,}3412x^{-0{,}3694}(1-\exp(-0{,}0049(x-120{,}87)))^2$	0,91

Slovak Bachletová, Moja Dolná zem	$2,22+16,40x^{-0,8879}(1 - \exp(-0,0192(x- 46,648)))^2$	0,74
Slovak Bachletová, Riadok v tlačive: nezamestnaný	$0,9419x^{-0,0132}(2,8449 - \exp(-0,0167x))$	0,88
Slovenian Ostrovskij, Kako se je kalilo jeklo, Chapter 1	$1,1944x^{-0,0484}(2,2284 - \exp(-0,0901x))$	0,85
Sundanese Agustusan (Salaka Online)	$1,6327x^{0,0465}$	0,81
Sundanese Aki Satimi (Salaka Online)	$0,7723x^{-0,001}(2,7764 - \exp(-0,1625x))$	0,82
Tagalog Hernandez, Limang Alas: Tatlong Santo	$1,8390+0,8272x^{-0,1595}(1-\exp(-0,0243(x-10,89)))^2$	0,74
Tagalog Hernandez, Magpisan	$2.343x^{-0.008}$	0.50
Tagalog Rosales, Kristal Na Tubig	$2.0799 + 0.0625(1 - \exp(-0.0068(x-216.66)))^2$	0.80
Tamil (Press)	$3.5629-1.2908x^{-0.5509}(1-\exp(-0.0166(x-45.87)))^2$	0,82
Telugu Trailangaswamy	$3,3640+6,973x^{-0,5289}(1-\exp(-0,0225(x-26,065)))^2$	0,72
Telugu Train Journey without safety	$3,6664-0,6931x^{-0,2738}(1-\exp(-0,0157(x-39,07)))^2$	0,53
Vai Mu ja vaa lɔ (T. Sherman)	$1.3389 + 0.1203(1 - \exp(-0.0102(x-86.78)))^2$	0.58
Vai Sabu Mua Ko	$1.4949 + 2866x^{-2.143}(1-\exp(-0.0179(x-16.59)))^2$	0.78
Vai Vande bɛ Wu'u	$1,5873-0,1351x^{0,0004}(1-\exp(-0,0461(x-19,95)))^2$	0,72
Welsh T1 Crynodeb Gweithredol	$2,3333-0,1473x^{0,1152}(1-\exp(-0,0164(x-10,06)))^2$	0,89
Welsh T2 Ffansi camu i esgidiau rhywun enwog?	$1,7376-0,0159x^{0,2815}(1-\exp(-0,0770(x-26,57)))^2$	0,73

In any case, we can state that the development of word length in text may have very different courses. Just as above, it depends on boundary conditions, and the links of this phenomenon to other text properties are a future task for synergetic linguistics.

Resume

Word length is a stochastic phenomenon depending on a great number of factors disturbing the possibility of exact prediction of whatever kind. In long texts, even if their homogeneity cannot not presupposed, there may be a background mechanism arising e.g. from semantic, stylistic, educational, scientific, etc. grounds which get their way and display an observable tendency on some level. As a matter of fact, our results will for ever be marked by some uncertainty because we cannot take all influences into account. Every influencing factor should obtain

its own parameter in our formulas but in that case the number of observations or observation classes may turn out to be smaller than the number of necessary parameters and no test would be possible. The situation would not improve even if we tried to model word length using partial differential equations where the same problem with parameters would arise.

Hence our endeavours are trials and errors. Sometimes the same text sort in two different languages yields very similar results, while two texts of different text sorts in the same language may yield very different results. Since the number of languages is too great and the number of writers and aims multiply this number, we shall be able to set up formulas in explorative or deductive way but we shall never obtain a satisfactory explanation. Nevertheless, the formulas can be derived from a very general background giving space to all possible factors; but here, too, the shortness of words and the number of parameters in the distributions or functions will always collide. Hence, the only remedy is the acceptation of a variety of models and deriving them - in the best case - from a common background.

In order to bring order into this enormous field, individual investigations performed on (supposedly) homogeneous texts are not sufficient. With some groups of texts nice trends may appear and in turn another group displays an opposite trend. The field is far from being systematized.

Here we merely wanted to show the contradictory character of a simple language phenomenon in which linguists are interested since 150 years.

The main results obtained in this article are: (1) The distribution of word lengths in terms of syllable numbers abides by some models related to the family of Poisson distributions. (2) The roughness of the sequence of lengths moves in our data in the interval <0,06; 0,22> but preliminarily, it is not possible to show the "cause", i.e. another property of the text linked with length, which would be responsible for a concrete value. (3) Though in some texts word length increases monotonously from the beginning to the end of sentence, it is not necessarily so. (4) In the same way, the change of word length from the beginning of text to its end is not that smooth as supposed up to now.

Much individual research is necessary to find the control cycle of these four aspects.

References

Best, K.-H. (ed.) (1997). *Glottometrika 16*: *The Distribution of Word and Sentence Length.* Trier: WVT.

Best, K.-H. (ed.) (2001). *Häufigkeitsverteilungen in Texten.* Göttingen: Peust & Gutschmidt Verlag.

Djuraš, G. (2012). *Generalized Poisson Models for Word Length Frequencies in Texts of Slavic Languages.* Diss. University of Graz

Fan, F., Grzybek, P., Altmann, G. (2010). Word length in sentence. *Glottometrics 20, 70-109.*

Fenk, A., Fenk-Oczlon, G. (2006). Within-sentence distribution and retention of content words and function words. In: Grzybek, P. (ed.), *Contributions to the science of text and language. Word length studies and related issues: 157-170*. Dordrecht: Springer.

Grzybek, P. (ed.) (2006). *Word Length Studies and Related Issues*. Dordrecht: Springer.

Kelih, E. (2009): Preliminary analysis of a Slavic parallel corpus. In: Jana Levická und Radovan Garabík (eds.), *NLP, Corpus Linguistics, Corpus Based Grammar Research. Fifth International Conference Smolenice, Slovakia, 25-27 November 2009. Proceedings: 175-183.* Bratislava: Tribun.

Kelih, E. (2012): On the dependency of word length on text length. Empirical results from Russian and Bulgarian parallel texts. In: Naumann, S.; Grzybek, P.; Vulanović, R.; Altmann, G. (eds.), *Synergetic Linguistics. Text and Language as Dynamic Systems: 67-80*. Wien: Praesens.

Meyer, P. (1997). Word length distribution in Inuktitut narratives: empirical and theoretical findings. *Journal of Quantitative Linguistics 4(1-3), 143–155.*

Meyer, P. (1999). Relating word length to morphemic structure: a morphologically motivated class of discrete probability distributions. *Journal of Quantitative Linguistics 6(1), 66–69.*

Ord, J.K. (1972). *Families of frequency distributions*. London: Griffin.

Popescu, I.-I., Altmann, G., Grzybek, P., Jayaram, B.D., Köhler, R., Krupa, V., Mačutek J., Pustet, R., Uhlířová, L., Vidya, M.N. (2009). *Word Frequency Studies*. Berlin-New York: Mouton de Gruyter, XI + 278 pp.

Popescu, I.-I., Altmann, G., Köhler, R. (2010). Zipf's law - another view. *Quality and Quantity 44(4) 713-731.*

Popescu, I.-I., Kelih, E., Mačutek, J., Čech, R., Best, K.-H., Altmann, G. (2010). *Vectors and codes of texts.* Lüdenscheid: RAM.

Popescu, I.-I., Mačutek, J., Altmann, G. (2009). *Aspects of word frequencies*. Lündenscheid: RAM.

Schmidt, P. (ed.) (1996). *Glottometrika 15*: *Issues in General Linguistic Theory and the Theory of Word Length.* Trier: WVT.

Uhlířová, L. (1997). O vztahu mezi délkou slova a jeho polohou ve větě. *Slovo a slovesnost 57, 174-184.*

Wilson, A. (2012). Word lengths in Welsh: Further investigations on prose and verse. In: Naumann, S., Grzybek, P., Vulanović, R., Altmann, G. (eds.), *Synergetic linguistics. Text and language as dynamic systems: 257-265.* Wien: Praesens

Wimmer, G., **Altmann, G.** (2005). Unified derivation of some linguistc laws. In: Köhler, R., Altmann, G., Piotrowski, R.G. (eds.), *Quantitative Linguistics. An International Handbook: 791-807.* Berlin-New York: de Gruyter.

Wimmer, G., **Witkovský, V., Altmann, G.** (1999). Modification of probability distributions applied to word length research. *Journal of Quantitative Linguistics 6, 257-268*

Zörnig, P. (1987). A theory of distances between like elements in a sequence. *Glottometrika 8, 1-22.*

Texts

Akan: Agya Yaw Ne Akutu Kwaa, Mma Nnsua Ade Bɔne (short stories, supplied by kasahorow)

Bamana: Bamakɔ sigicogoya, Namakɔrɔba halakilen, Sonsannin ani Surukuba — from the Bamana text corpus (thanks to Valentin Vydrin);

Bamana: Masadennin — the Bamana translation by Bukari Jara of *The Little Prince (Le petit prince)*, a famous 1943 novella by Antoine de Saint-Exupéry (Bamako: Edition Jamana, 1989).

Bulgarian: Ostrovskij, Kak se kaljavaše stomanata, Chapter 1

Czech: Jan Čulík: O čem jsou dnešní Spojené státy? (2. 8. 2012). Britské listy. http://blisty.cz/art/64324.html

Czech: Karel Hvížďala: O předem zpackané prezidentské volbě aneb Jak dlouho budeme bez prezidenta 7. 8. 2012 Blog.aktualne.cz http://blog.aktualne.centrum.cz/blogy/karel-hvizdala.php?itemid=17155

Czech: Jan Macháček Slovenský (dobrý) příklad (4. 8. 2012). Respekt. http://respekt.ihned.cz/audit-jana-machacka/c1-56924030-slovensky-dobry-priklad

Czech: Jaroslav Spurný: Překvapení v justice (2.7. 2012) http://respekt.ihned.cz/komentar/c1-56386120-prekvapeni-v-justici

Czech: Marek Švehla: Voličův kalkul (4. 8.2012). http://respekt.ihned.cz/c1-56902580-editorial-volicuv-kalkul

French: Dunkerque – La route des dunes. Le Blog de François Béguin, fait partie de „Une année en France“, http://dunkerque.blog.lemonde.fr/ 07.05.2012

Hindi: Daily Hindi Milap, (31st May, 2012): After the sanction to love marriage, (page 4)

Hindi: Swatantra Varta,(31st July, 2012): The Anna Team on a cross-road (page 6)

Indonesian: Sekolah ditutup **(**Press online, 01.05.2012)

Indonesian: Pengurus PSM terbelah **(**Press online 01.05.2012)

Italian: [*Il bosone di Higgs scoperto dal Cern potrebbe essere un "impostore"*, http://www.meteoweb.eu/2012/07/il-bosone-di-higgs-scoperto-dal-cern-potrebbe-essere-un-impostore/143186/ http://www.meteoweb.eu/2012/07/fisica-scoperta-la-particella-di-dio-dati-molto-significativi-sul-bosone-di-higgs/142116/; http://www.meteoweb.eu/2012/07/scoperta-la-particella-di-dio-adesso-arrivo-tante-altre-sorprese-peter-higgs-verso-il-premio-nobel/142344/, 11.07.2012]

Japanese: Miki, K. (1941, 1995). Jinseiron Note (Essay on the life). Tokyo: Sôgensha. (CD-ROM edition included in Shinchô Bunko no 100 satsu); Shinchôsha (1995). Shinchô Bunko no 100 satsu (CD-ROM edition of 100 paperbacks extracted from Shinchô Bunko series). Tokyo: Shinchôsha

Kikongo: Independence Day Speech for the Democratic Republic of Congo. Patrice Lumumba, June 30, 1960

Kikongo: Bimpa: Ma Ngo ya Ma Nsiese (Tale: Mr. Leopard and Mr. Antelope) from ngunga.com

Kikongo: Nkongo ye Kisi *K*ongo (The People of Kongo - Customs and Traditions) by the honorable Mr. Ernesto Nzakundomba. Publisher: Imprensa Nacional - E.P. 1st Edition, Luanda, December. 2006.

Latin: Cicero, In Catilinam II, 180sentences, http://www.thelatinlibrary.com/cicero/cat.shtml,

Macedonian: Ostrovskij, Kako se kaleše čelkiot, Chapter 1

Malayalam: Newspaper. Malayala Manorama (21 June 2012). Place of Publication: Cochin. *Moralistic hooligans and their bad activities* (page 10)

Malayalam: Newspaper. Malayala Manorama (21 June 2012). Place of Publication: Cochin. *No one should die in hospital for not getting Oxygen* (page 10)

Maninka: Nko Doumbu Kende no. 2 (press);

Maninka: Nko Doumbu Kende no. 7 (press);

Maninka: Sìikán` (Constitution of Guinea, an excerpt).
Maninka texts (in Nko script) were supplied by Valentin Vydrin

Odia: The Samaj, Bhubaneswar (28 June 2012): Who is great? (page 4)

Odia: The Dharitri, Balasore (12th February, 2012): Calculation for the District Council President (page 10)

Romanian: O. Paler, Aventuri solitare (excerpt)

Romanian: D.R. Popescu, Vânătoarea regală, Chapter 2

Romanian: N. Steinhardt, Jurnalul fericirii, Trei soluții

Serbian: Ostrovskij, Kako se kalio čelik, Chapter 1

Slovenian: Ostrovskij, Kako se je kalilo jeklo, Chapter 1

Sundanese: Agustusan (Salaka Online)**;**

Sundanese: Aki Satimi (Salaka Online)

Tagalog: Hernandez, Limang Alas: Tatlong Santo; Hernandez, Magpisan; Rosales, Kristal Na Tubig

Tamil: Emu Bird Trading– Case filed against Sathiyaraj and Saratkumar http://tamil.oneindia.in/news/2012/08/08/tamilnadu-emu-scam-case-file-against-advertisement-modals-actor-159236.html

Telugu: Daily Andhrabhoo mi (4th August 2012): Train Journey without safety (p. 4)

Telugu: Daily Andhrabhoo mi (4th August 2012): Trailangaswamy: a biography (p. 10)

Vai: Mu ja vaa lɔ (by T. Sherman), supplied by Charles Riley and Tombekai Sherman

Vai: Sabu Mua Ko, Vande bɛ Wu'u — from a Vai book *Kכ' Tíɛ́ Banda Tɛiɛ Nú [Stories We Tell Diring Rice Harvest]*, 2nd edition (The Institute for Liberian Languages: Monrovia, Liberia, 1992; thanks to Valentin Vydrin)

Welsh (gaenv): Welsh-language executive summary from: *Preparing for climate change impacts on freshwater ecosystems (PRINCE).* Science Report SC030300/SR. Bristol: Environment Agency, 2007. http://publications.environment-agency.gov.uk/PDF/SCHO0507BMOJ-E-E.pdf

Welsh (gasodl): Ffansi camu i esgidiau rhywun enwog? *Y Cymro*, 8 July 2011. http://www.y-cymro.com/newyddion/c/44/i/449/desc/ffansi-camu-i-esgidiau-rhywun-enwog/

Alternative methods of goodness-of-fit evaluation applied to word length data

Ján Mačutek[1], *Gejza Wimmer*[2]

1. Introduction

The present paper can be seen as a continuation and an exemplification of ideas introduced by Mačutek and Wimmer (2013). Our goal is to demonstrate that using different methods for goodness-of-fit evaluation leads to different results. First, distributions which achieve a satisfying fit with respect to one method do not have to be among the best ones if another method is chosen. Second, quite different parameter estimations can be obtained.

The paper is organized as follows. In Section 2 we provide a list of ten distributions which were chosen for our analyses. Three methods of goodness-of-fit evaluation (the discrepancy coefficient based on the Pearson χ^2 statistic, the determination coefficient and the total variation distance) are used. In Section 3 we use the distributions to model word length data from three languages (English, German and Persian). The methods mentioned above are applied to evaluate their goodness-of-fit. Finally, Section 4 is dedicated to the discussion of achieved results.

The lists of the distributions and goodness-of-fit measures from this paper are by no means exhaustive, they serve only as examples.

2. Distributions and criteria used

Popescu et al. (2013) present many distributions which were used to model word length, all of them being related to the Poisson distribution. For our purposes, ten of them were chosen. They can be found in the discrete distribution dictionary by Wimmer and Altmann (1999). Their definitions are given in Table 1, together with acronyms used to denote the distributions further in the paper. The distributions are ordered alphabetically. As zero-syllable words are usually not considered in word length modelling (cf. Antić et al. 2006), some of them are shifted to the right by one (i.e., they are defined on the set 1,2,...).[3]

---

[1] Department of Applied Mathematics and Statistics, Comenius University, Mlynská dolina, 842 48 Bratislava, Slovakia, email: jmacutek@yahoo.com

[2] Mathematical Institute, Slovak Academy of Sciences, Štefánikova 49, 814 73 Bratislava, Slovakia and Faculty of Natural Sciences, Matej Bel University, Banská Bystrica

[3] For the sake of simplicity, we use the "original" names of the distributions, without emphasizing the shift (e.g., we use the Poisson distribution, not the 1-displaced Poisson distribution).

Table 1
Distributions used for the analyses

Name (acronym)	Definition
Cohen-Poisson (CP)	$P_1 = (1+\alpha a)e^{-a}$ $P_2 = (1-\alpha)ae^{-a}$ $P_x = \frac{e^{-a}a^{x-1}}{(x-1)!} \qquad x = 3,4,\ldots$
Consul-Jain-Poisson (CJP)	$P_1 = e^{-a}$ $P_x = \frac{a\left[a+b(x-1)^{x-2}\, e^{-\left[a+b(x-1)\right]}\right]}{(x-1)!} \qquad x = 2,3,\ldots$
Conway-Maxwell-Poisson (CMP)	$P_1 = \left(\sum_{i=0}^{\infty} \frac{a^i}{(i!)^b}\right)^{-1}$ $P_x = \frac{a^{x-1}}{\left[(x-1)!\right]^b} P_1 \qquad x = 2,3,\ldots$
Dacey-Poisson (DP)	$P_x = \frac{(1-\alpha)a^{x-1}e^{-a} + \alpha(x-1)a^{x-2}e^{-a}}{(x-1)!} \qquad x = 1,2,\ldots$
hyper-Poisson (HP)	$P_x = \frac{a^{x-1}}{{}_1F_1(1;b;a)b^{(x-1)}} \qquad x = 1,2,\ldots$
mixed Poisson (MP)	$P_x = \frac{\alpha a^{x-1}e^{-a} + (1-\alpha)b^{x-1}e^{-b}}{(x-1)!} \qquad x = 1,2,\ldots$
Poisson (P)	$P_x = \frac{e^{-a}a^{x-1}}{(x-1)!} \qquad x = 1,2,\ldots$
positive Cohen-Poisson (PCP)	$P_1 = \frac{(1-\alpha)a}{e^a - 1 - \alpha a}$ $P_x = \frac{a^x}{x!\left(e^a - 1 - \alpha a\right)} \qquad x = 2,3,\ldots$
positive Poisson (PP)	$P_x = \frac{e^{-a}a^x}{x!\left(1-e^{-a}\right)} \qquad x = 1,2,\ldots$
Singh-Poisson (SP)	$P_1 = 1-\alpha+\alpha e^{-a}$ $P_x = \frac{\alpha a^{x-1}e^{-a}}{(x-1)!} \qquad x = 2,3,\ldots$

The goodness-of fit of these distributions will be evaluated by the following criteria (f_i are the observed frequencies, N the sample size and P_i the probabilities from the theoretical model):

The discrepancy coefficient is the ratio of the Pearson χ^2 and the sample size, i.e.,

$$C = \frac{\chi^2}{N},$$

where

$$\chi^2 = \sum_{i=1}^{n} \frac{\left(f_i - NP_i\right)^2}{NP_i}.$$

The determination coefficient is defined as

$$R^2 = 1 - \frac{\sum_{i=1}^{n}\left(f_i - NP_i\right)^2}{\sum_{i=1}^{n}\left(f_i - \overline{f}\right)^2},$$

with $\overline{f}$ being the mean frequency, and the total variation distance as

$$D_{TV} = \frac{1}{2}\sum_{i=1}^{n}\left|P_i - \frac{f_i}{N}\right|.$$

3. Results

The distributions from Table 1 were used as models for word length in three texts (English, German and Persian). The word length data (cf. Table 2) were taken from Best (2009), Best (2010) and Best (2008), respectively.

Table 2
Word length in an English, German and Persian text

number of syllables	English	German	Persian
1	713	93	1043
2	226	60	2128
3	26	22	1147
4	8	9	308
5	1	2	88
6			14
7			1

Results of fitting and parameter values[4] which optimize the fit[5] (i.e., the best fit possible is achieved for the values) are presented in Tables 3a-5c. In each table, distributions are ordered according to the goodness-of fit measure, the distribution with the best fit being the first.

Table 3a
Evaluating goodness-of-fit with respect to the discrepancy coefficient *C*
for the English text (cf. Best 2009)

	C	*a*	*b*	α
MP	0.0046	1.5588	0.2936	0.0200
CJP	0.0062	0.3104	0.0297	
PCP	0.0064	0.5163		0.1659
CMP	0.0067	0.3021	0.7343	
HP	0.0071	0.4588	1.5130	
PP	0.0077	0.5843		
SP	0.0079	0.3604		0.8950
CP	0.0087	0.3328		0.0361
DP	0.0089	0.3248		0.0000
P	0.0089	0.3248		

Table 3b
Evaluating goodness-of-fit with respect to the determination coefficient R^2
for English text (cf. Best 2009).

	R^2	*a*	*b*	α
CMP	0.9999	0.3184	1.3063	
DP	0.9999	0.1777		0.1234
HP	0.9999	0.2138	0.6713	
PCP	0.9999	0.3813		0.4012
SP	0.9999	0.2547		1.1841
CP	0.9997	0.3123		0.0000
CJP	0.9997	0.3123	0.0000	
MP	0.9997	-	0.3123	0.0000
P	0.9997	0.3123		
PP	0.9988	0.6011		

[4] If one obtains $\alpha = 0$ in the mixed Poisson distribution (which is often the case for our data), the value of the parameter *a* becomes irrelevant.

[5] Programs for the optimizations were written in statistical software environment R by the authors. Optimized parameter values from Altmann-Fitter (1997) were used as initial values in the programs.

Table 3c
Evaluating goodness-of-fit with respect to the total variation distance D_{TV} for the English text (cf. Best 2009)

	D_{TV}	a	b	α
PCP	0.0056	0.4172		0.3419
SP	0.0060	0.2814		1.0926
CMP	0.0061	0.3170	1.1632	
DP	0.0061	0.2132		0.0940
HP	0.0061	0.2545	0.8028	
CP	0.0089	0.3119		0.0000
CJP	0.0089	0.3119	0.0000	
MP	0.0089	-	0.3119	0.0000
P	0.0089	0.3119		
PP	0.0164	0.5945		

Table 4a
Evaluating goodness-of-fit with respect to the discrepancy coefficient C for the German text (cf. Best 2010)

	C	a	b	α
MP	0.0028	1.0918	0.3895	0.5174
CMP	0.0031	0.6356	0.6465	
HP	0.0031	1.2289	1.9534	
PCP	0.0031	1.2368		0.0265
PP	0.0032	1.2513		
CJP	0.0034	0.6968	0.0769	
SP	0.0046	0.8782		0.8575
CP	0.0077	0.7979		0.1164
DP	0.0145	0.7589		0.0000
P	0.0145	0.7589		

Table 4b
Evaluating goodness-of-fit with respect to the determination coefficient R^2
for German text (cf. Best 2010)

	R^2	a	b	α
MP	0.9993	1.4399	0.5411	0.2367
CJP	0.9991	0.6914	0.0761	
CMP	0.9989	0.6397	0.6860	
PCP	0.9988	1.1864		0.0772
HP	0.9986	1.0777	1.6848	
PP	0.9983	1.2507		
SP	0.9980	0.8076		0.8967
CP	0.9973	0.7426		0.0778
DP	0.9940	0.6866		0.0001
P	0.9940	0.6867		

Table 4c
Evaluating goodness-of-fit with respect to the total variation distance D_{TV}
for German text (cf. Best 2010)

	D_{TV}	a	b	α
MP	0.0093	1.2947	0.5200	0.2949
CJP	0.0100	0.6913	0.0717	
CMP	0.0103	0.6452	0.7038	
PCP	0.0103	1.1909		0.0771
HP	0.0110	1.0589	1.6413	
SP	0.0141	0.8207		0.8931
PP	0.0155	1.2564		
CP	0.0171	0.7614		0.0928
DP	0.0258	0.6932		0.0000
P	0.0258	0.6932		

Table 5a
Evaluating goodness-of-fit with respect to the discrepancy coefficient C
for the Persian text (cf. Best 2008)

	C	a	b	α
HP	0.0016	0.7182	0.3525	
DP	0.0019	0.6435		0.5797
SP	0.0029	0.9801		1.2482
CMP	0.0055	1.8824	1.6748	
PCP	0.0068	1.3758		0.6760
CP	0.0562	1.2253		0.0000
CJP	0.0562	1.2253	0.0000	
MP	0.0562	-	1.2253	0.0000
P	0.0562	1.2253		
PP	0.1308	1.9023		

Table 5b
Evaluating goodness-of-fit with respect to the determination coefficient R^2
for the Persian text (cf. Best 2008)

	R^2	a	b	α
DP	0.9997	0.6297		0.5857
HP	0.9995	0.7119	0.3474	
CMP	0.9992	2.0061	1.8579	
SP	0.9980	1.0165		1.2251
PCP	0.9954	1.4671		0.6475
CP	0.9219	1.3148		0.0000
CJP	0.9219	1.3148	0.0000	
MP	0.9219	-	1.3148	0.0000
P	0.9219	1.3148		
PP	0.8278	2.1165		

Table 5c
Evaluating goodness-of-fit with respect to the total variation distance D_{TV} for the Persian text (cf. Best 2008)

	D_{TV}	a	b	α
DP	0.0065	0.6332		0.5846
HP	0.0076	0.7157	0.3508	
CMP	0.0105	2.0403	1.8655	
SP	0.0144	1.0138		1.2233
PCP	0.0235	1.4602		0.6422
CP	0.1056	1.4086		0.0000
CJP	0.1056	1.4086	0.0000	
MP	0.1056	-	1.4086	0.0000
P	0.1056	1.4086		
PP	0.1694	2.1270		

4. Conclusion

The results from Section 3 confirm that the notion of the distribution with the best fit is relative, depending on the goodness-of-fit measure chosen. For the English text, e.g., we obtain the first five distributions in the order MP – CJP – PCP – CMP - HP with respect to the discrepancy coefficient *C*, but PCP – SP – CMP – DP - HP on the first five places if we use the total variation distance (cf. Tables 3a and 3c, respectively).

A close cooperation among linguists and mathematicians is needed to interpret particular goodness-of-fit measures and to decide which of them should be used in which field of linguistics.

Obviously, different goodness-of-fit measures result also in different parameter estimations (cf. Tables 3a-5c). A researcher thus obtains more possibilities to choose interpretable parameter values and/or to reveal mutual influences of linguistic units and their properties.

Acknowledgement
Supported by the grants VEGA 2/0038/12 (J. Mačutek, G. Wimmer) and APVV-0096-10 (G. Wimmer).

Software

Altmann-Fitter (1997). Lüdenscheid: RAM-Verlag.

References

Antić, G., Kelih, E., Grzybek, P. (2006). Zero-syllable words in determining word length. In: Grzybek, P. (ed.), *Contributions to the Science of Text ad Language. Word Length Studies and Related Issues: 117-156.* Dordrecht: Springer.

Best, K.-H. (2008). Word length in Persian. *Glottometrics 16, 27-30.*

Best, K.-H. (2009). Wortlängen im Englischen. *Glottometrics 19, 1-10.*

Best, K.-H. (2010). Silben-, Wort- und Morphlängen bei Lichtenberg. *Glottometrics 21, 1-13.*

Mačutek, J., Wimmer, G. (2013). Evaluating goodness-of-fit of discrete distribution models in quantitative linguistics. *Journal of Quantitative Linguistics*, to appear.

Popescu, I.-I., Naumann, S., Kelih, E., Rovenchak, A., Sanada, H., Overbeck, A., Smith, R., Čech, R., Mohanty, P., Wilson, A., Altmann, G. (2013). Word length: aspects and languages. *This volume*.

Wimmer, G., Altmann, G. (1999). *Thesaurus of unvariate discrete probability distributions.* Essen: Stamm.

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

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

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

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

Für die Ermittlung der Lösungen lässt sich hier der sonst übliche Newton-Raphson-Algorithmus nicht anwenden, da ${}_1F_1[1;\lambda;\theta]$ nicht analytisch darstellbar ist. Gemäß Definition (2) ist offensichtlich, dass die Berechnung der partiellen Ableitungen von ${}_1F_1[1;\lambda;\theta]$ 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 $d = s^2/(\bar{x} - 1)$.

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
$\bar{\alpha}_{ML}$	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
$\bar{\alpha}_{ML}$	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
$\tilde{\alpha}_{ML}$	0.63	0.86	0.87	0.83
$\tilde{\theta}_{ML}$	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.

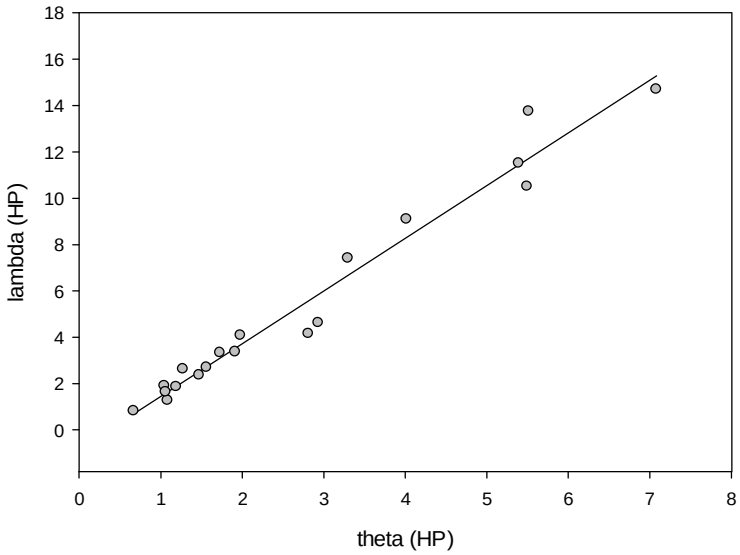


Abb. 1: Linearer Zusammenhang zwischen den Parametern θ und λ der Hyperpoisson-Verteilung

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.

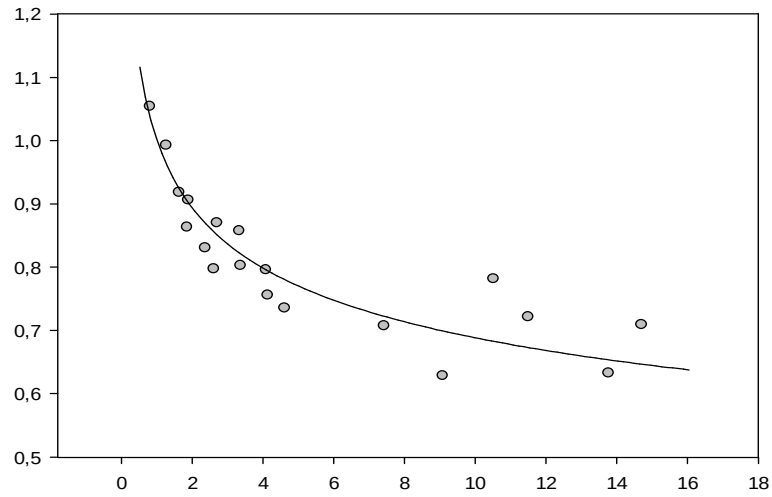


Abb. 2: θ (HP) – α (SP)

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.*

Corpus Independent Performance Metric for Language Models

Bahar Karaoğlan, Izmir
Bekir Taner Dinçer, Muğla
Tarık Kışla, Izmir
Senem Kumova Metin, Izmir

Abstract. The motivation behind this study is the need for a corpus independent performance metric for language models. Here, first a theoretical background for the proposed metric is established and then it is formulated for practical applications. The idea behind the proposed metric is very simple and relies on the fact that if a language model is evaluated by the amount of signal information over different corpora of same size, simply by the entropy, the performance of the model may vary due to the difference in amount of semantic information contained in the corpora. The contributions of the paper are threefold. First, we propose a method to measure the semantic information in a corpus by exploiting the correlation between the signal information and the semantic information within a written text. Second, we introduce a new notion: ideal corpus, which contains ideal amount of semantic information for a given number of words. Finally, we propose a performance metric in which semantic information contained in the corpus at hand relative to ideal semantic information that should be contained is considered in performance measurement. The parameters of the proposed metric for the amount of semantic information contained in a corpus of size N and in language L, $SEI(L,D_N)$, is obtained by exploiting Zipf and Heaps' power laws. The proposed metric is tested on different corpora. It is seen that as the size of the corpus increases, even the entropy changes, the value calculated for the proposed metric doesn't change which is an evidence of the independence of the metric from the corpus size.

1. Introduction

Language models for written corpora are usually based on the statistical data of frequencies of appearances of certain patterns. Considering these patterns as the signals, the general approach to measure the performance of a language model is through cross entropy and/or perplexity which are borrowed concepts from Shannon's information theory. In this respect a language model represents a language as much as it decreases the uncertainty hence the cross entropy. But, when we test the same language model over different corpora it is very likely that we get different entropies and we cannot be sure if the results are because of the language model or the corpus or how much of it comes from the corpus and how much comes from the model. Therefore, there is a need to alleviate the effect of the corpus in order to assess the true performance of the language model.

In this study, we propose a corpus independent performance metric for language models in which the relative amount of semantic information carried in the corpus at hand is normalized with respect to an ideal corpus with the same size and language and assumed to carry ideal amount of semantic information.

The idea behind the proposed method is very similar to frequency normalization where relative frequencies are considered, rather than the real frequencies to diminish the effect of the document length. Unfortunately, normalization of the corpus by semantic information is not that trivial as frequency normalization. Because, in order to measure the semantic information contained in the corpus, the smallest unit of semantic information, the meanings assigned to words, are to be counted. The frequencies of the word forms may give a clue on the number of meanings conveyed in the corpus but due to the effects of polysemy and synonymy, the same word form may have different meanings or different word forms may have the same meaning. Since the words are the only ingredients, we need to find a way to go from word form dimension to meaning dimension. Here, Zipf laws (Zipf, 1949) give us clear clues on how to measure the semantic information of a corpus and an ideal corpus of same size. Zipf has stated power relations between frequencies of words and rank of words; between frequencies of words and vocabulary at each frequency; and frequencies of words and average number of meanings at each frequency. By making use of these laws and the Heaps' law (Heaps, 1978) that states the expected vocabulary size of a corpus, we approximate the signal information derived from individual words to semantic information derived from the number of different meanings conveyed by these words. We formulize our metric by multiplying the proportion of the semantic information conveyed by the corpus at hand to semantic information that is assumed to be conveyed by an ideal corpus (the text of same length and same language) with the entropy generated by the language model as a factor of normalization.

We tested our metric on a Turkish corpus and two English corpora and showed that even the cross entropy changes, the proposed performance metric gives the same result for different corpora of different sizes. This, if not proves, strengthens our claim that the proposed metric is corpus size independent. The following two sections give brief overview about information and entropy and power laws that we consider in order to clarify the rationale behind the proposed metric. Section 4 explains the derivation of the proposed metric. Section 5 gives the results and section 6 is the conclusion.

2. Information and Entropy

"Information" is a philosophical term, thus open to discussion when looked at from different points of view. "*It is hardly to be expected that a single concept of infor-*

mation would satisfactorily account for the numerous possible applications of this general field" (Shannon 1993, p. 180). Following Shannon and Weaver (1949) three aspects related to information are: 1) quantification of information, 2) meaning of information and 3) impact of information. The first aspect is directly related to Shannon's theory of information which deals with objective, mathematical quantification of the amount of signal information transferred. The latter two deal with the quantification of semantic information which is a subjective issue. The meaning and impact of information differ according to person and context. Communication between people or machines is carried via messages coded in symbols (Figure. 1). The message is encoded into symbols as the "informer" understands or thinks best represents what (s)he wants to express and it is decoded by the informee as (s)he comprehends it. So both sides are subjective and change according to the understanding of the "informer" and the "informee". But the message, once coded in symbols is a physical phenomenon and can be quantified.

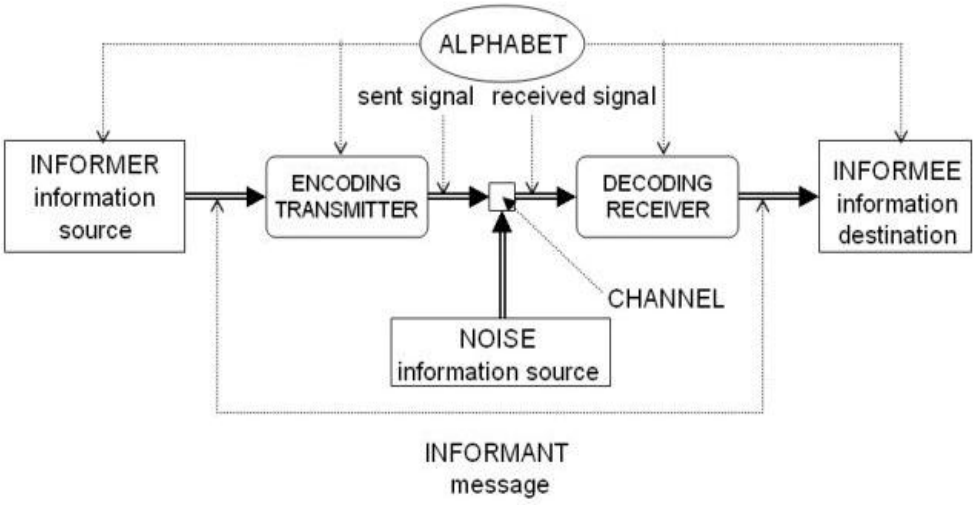


Figure 1. Communication model (Adapted from Shannon and Weaver, 1949)

The amount of information carried by a symbol is indirectly related to its probability of occurrence: rare events are more informative than usual events. If we know the time schedule of a very reliable railway transportation system, information that the train will be on time has no value. But, if we hear that some trips are cancelled due to some hazard, this information is valuable. The relation between the information carried by a symbol and its probability of occurrence is formulated as follows:

$$I(s) = \log_2 \frac{1}{P(s)} \quad \text{or } I(s) = -\log_2 P(s) \qquad (1)$$

Here $I(s)$ is the amount signal information carried by signal, s, and calculated by taking the logarithm in base 2 of inverse of $P(s)$, the probability of occurrence of the signal s. The numerical value obtained from this equation is the quantitative representation of the amount of signal information in terms of binary digits. Comprehension or realization of signal by the mind that is the semantics of the signal which is the main issue in cognitive psychology is not considered in signal information point of view.

Shannon's source coding theorem establishes that, on average, the number of *bits* needed to represent the result of an uncertain event is given by its entropy. In performance evaluation of language models, cross entropy (or perplexity) is accepted widely as a key measure of signal information (Rosenfeld, 2000).

$$H(m,L) = \sum_{1}^{N} \frac{1}{N} \log \frac{1}{m(x_{1n})} \qquad (2)$$

where N is the size of the training set (test corpus), $m(x1n)$ is the probability of event x estimated from the test corpus according to language model and L is language.

3. Power Laws in Language

Power laws in language are simply the empirical power laws which provide a frequency based mathematical model for natural languages in the field of information retrieval and natural language processing. In his book *Human Behavior and the Principle of Least Effort* (1949) Zipf claims that the principle of least effort in human actions is also fundamental in language and shows that theory and models coincide with the experimental results. Zipf's power laws are considered in the study to provide a formulization for vocabulary balance in language and the distribution of meanings. Additionally, Heaps' law (Heaps, 1978) of vocabulary is used to formulate the number of unique words (vocabulary size) in the corpus. Aforementioned power laws are summarized in the following paragraphs:

Distribution of Words

If the number of appearances of each word is counted up in a large corpus and the words are ranked in decreasing order of their frequency, the most frequent item hav-

ing the rank 1, the relation between the frequency and the rank of a word can be stated as:

$$f \times r = C \tag{3}$$

where *f* is the frequency, *r* is the rank and *C* is a constant. Mathematically, Zipf's law has the convenient property that, if we take the logarithm of both sides, we obtain a linear function

$$\log(f_r) = H_N - B_N \log(r) \tag{4}$$

where f_r is the frequency of the word with rank *r*, B_N and H_N are numerical constants that are inserted to the formulation.

Mandelbrot's derivation on Zipf's first law is accepted as one of the most challenging derivation of Zipf Law that gives better results on lowest and highest frequency fits. Mandelbrot's modified equation is given as:

$$\log(f_r) = H_N - B\log(r + W) \tag{5}$$

where $W>1$ is a parameter which saves the probability distribution when B_N in Zipf's equation is greater than one.

Vocabulary Balance:

This law defines a relation that enables the mathematical calculation of vocabulary size of a corpus. It states that if the numbers of different words having the same frequency of occurrence are ranked in descending order there is a linear but inverse relation between the rank and the frequency of words. The relation is formulated as

$$\log(i) = K_N - D_N \log(V(i, N)) \tag{6}$$

where $V(i,N)$ is the number of words seen i times in a corpus of N words, K_N and D_N are numerical constants dependent on N. The total vocabulary size can be calculated as the sum of number of words having a frequency of *i* (i ∈ 1, 2, … *n*). In other words $V = \sum_{i}^{n} V(i,N)$.

Actually this law of vocabulary has a strong relation with the first law and it is formulated by Kornai (2002) as given in equation 7. It depicts that if any distribution is consistent with the first law, it must also be consistent with the second.

$$D_N = B_N / (1 + B_N) \tag{7}$$

The law of vocabulary balance is stated with the following equation in order to predict the occurrence frequencies of different word forms in the study of Kornai (2002)

$$V(i,N) = mV(N) / i^{\rho+1} \tag{8}$$

where *m* is a constant and ρ is inversely proportional to the constant *B* in Zipf's first law (equation 4).

Vocabulary Size:

Heaps' law (1978) is an empirical law which gives the vocabulary size as a function of collection (corpus) size as in following formula:

$$V = kN^{\rho} \tag{9}$$

where *V* is the number of unique words in a corpus of *N* words. *k* and ρ are the numerical constants determined empirically. Kornai (2002) stated that the ρ given in equation 9 is a constant which is inversely proportional to the *B* in Zipf's first law (equation 4).

Distribution of Meanings:

Zipf's third law, often called as "Law of Meanings", states the relation between the frequencies of words and number of different meanings that these words may convey. In his book, Zipf states that

> *"…under the conflicting forces of Unification and Diversification the m number of different meanings to be verbalized will be distributed in such a way that no single word will have m different meanings and that on the other hand there will be fewer than m different words. As a consequence, we may expect that at least some words must have multiple meanings."*

The problem to be solved above is which words will have multiple meanings and how many meanings they will convey. To characterize the effect of conflicting forces Zipf has first turned attention to the most frequently used word. If the frequency of the most frequent word is F_1 and it has w_1 meanings, the equation below can be written

$$F_1 = w_1 \times f_1 \tag{10}$$

where f_1 represents the average frequency of the occurrence of the w_1 meanings. The forces of unification will tend to increase the number of w_1 in the direction of putting all meanings behind a single word. In contrast, the forces of diversification will tend to increase f_1 in the direction of reducing number of different meanings per word. The empirical study on frequency-rank curves shows that the effect of conflicting forces arises itself as a hyperbolic relationship between two variables. As a result w_1 and f_1 will also stand in a hyperbolic relationship with one another with a further result that w_1 will tend to equal f_1 generating the equations 11, 12.

$$F_1 = w_1 \times w_1 \tag{11}$$

$$\sqrt{F_1} = w_1 \tag{12}$$

The equation for the most frequent word can be generalized for all frequencies and worded as "the average number of meanings (w) grows proportionally to the frequency (f) and grows inversely proportional to the rank (r)". This is formulated as follows:

$$w \propto \sqrt{f} \qquad \text{or} \qquad \overline{w} \propto 1/\sqrt{r} \tag{13}$$

In this study, while measuring vocabulary balance and vocabulary size, we used the constants obtained and validated in the previous studies of Dinçer (2004) and Karaoğlan et.al. 2008). The studies utilize a Turkish corpus of 2,511,930 surface formed words which is a merged set of METU (Say, et. al., 2004) and Bilkent (Hakkani-Tür et. al, 2003) corpora; and an English corpus of 654,728 words which is a merged set of Time, Cranfield and Medlars corpora.

The curve fitting experiments on distribution of words in Turkish (Dinçer, 2004; Karaoğlan et. al., 2008) and English (Dinçer, 2004) using Zipf's and Mandelbrot's equations with constants *W = 10, W = 100, W = 1000, W = 10000* showed that Mandelbrot's equation outperforms Zipf's equation in modeling the distribution of word frequencies. The best curve fits are obtained with the constants *W=1000* and *B = 1.2786* where coefficient of determination[1] is $R^2 = 0.998$ for Turkish ($F = 1.400E+08$, $\alpha = 0.000$). The constant values *W = 1000* and *B = 1.4316* gave the best fitting result with $R^2 = 0.994$ ($F = 5.506E+06$, $\alpha = 0.000$) in English corpus.

[1] Coefficient of determination is the ratio of explained variation compared to total variation and the value of determination range from 0 to 1.

While measuring the vocabulary size (equation (9)), we accepted the constant $\rho = (1/1.2786)$ for Turkish and $\rho = (1/1.4316)$ for English. In equation (8) which defines the vocabulary balance in language, the constant *m* may be accepted as $6/\pi^2$ (which is obtained in cases where *B=1)*, keeping the distribution consistent for *B>1*, in order to support the relation $\sum_{i=1}^{\infty} V(i,N) = V(N)/i^{\rho+1}$. (Actually the constant is calculated by the equation $m = 1/\zeta(\rho+1)$ in which ζ is the Reimann function) . In this study we used the constants $m = 6/\pi^2$ and $\rho = (1/1.2786)$ for Turkish and $\rho = (1/1.4316)$ for English in experiments involving vocabulary balance relations.

4. Proposed Metric

The proposed metric is solely based on our argument that there should be a strong link between the signal information and the semantic information. The rationale behind this is very simple: the semantic information that the sender wants to transmit determines the set of symbols that compose the message.

A language model is assessed by the degree to which it represents the language. Hence, its performance can be formulated as the ratio between the amount of information calculated by the language model on a corpus of size *N* (signal information), and the ideal amount of information that should be carried in a text with the same size as the corpus. More formally we state that:

"Performance of a language model *m* on corpus *D* of *N* words, written in language *L*, $B(m,L,D_N)$, can be measured in terms of amount of semantic information represented with respect to signal information conveyed and thus, can be formulated as the ratio between the signal information (*SII*) and semantic information (*SEI*)."

$$B(m,L,D_N) = \frac{SII(m,L,D_N)}{SEI(L,D_N)} \qquad (14)$$

As can be seen from the formula the semantic information ($SEI(L,D_N)$) employed in corpus *D* is independent of the language model. Here, the semantic information represents the expected amount of information that should be carried in a corpus of size *N* and is calculated by the use of Zipf parameters driven as a result of experimental studies. If we consider the signal information $SII(m,L,D_N)$ in a corpus as being equal to the cross entropy $H(m,L)$ calculated for the language model, then the only dependence on the size of the corpus of the proposed performance metric will be sourced from the calculation of the semantic information.

$$B(m,L,D_N) = \frac{H(m,L)}{SEI(L,D_N)} \qquad (15)$$

An important point to note here is that, if two different language models are compared on the same corpus the only factor that effects their performance is left as the signal information which is actually the cross entropy, $H(m,L)$, created by the language model, since the semantic information will be the same in both cases.

A corpus independent, performance metric, $B(m,L)$, of a language model m on a corpus of size N, written in language L can be defined as follows:

$$B(m) = \frac{SEI(L_N)}{SEI(D_N)} \times H(m,L) \qquad (16)$$

$SEI(L_N)$ denotes the semantic information that should be carried in any corpus of size N, written in language L. In other words, it denotes the semantic information of an ideal corpus which contains ideal amount of semantic information for a given number of words. And $SEI(D_N)$ is the semantic information carried by the available corpus D of size N. Therefore, *RSEI,* the ratio of the semantic information carried in a particular corpus of size N in language L to the semantic information that should be carried in any corpus in language L and of size N is $SEI(L_N) / SEI(D_N)$.

If $SEI(L_N) / SEI(D_N) = RSEI > 1$, we conclude that the corpus D carries less semantic information than is anticipated for the size of text written in the same language.

If $SEI(L_N) / SEI(D_N) = RSEI < 1$, we conclude that the corpus on hand conveys more information (greater cross entropy value) than the anticipated average value for the language.

If $SEI(L_N) / SEI(D_N) = RSEI = 1$, we conclude that the corpus on hand carries equal amount of information to the expected amount of information carried by the language. Therefore, we can say that the actual performance metric of the language model can now be given by the cross entropy value.

In the proposed performance metric power laws are taken as the basis to quantify the semantic information, $SEI(L_N)$, in a corpus of size N, written in language L. Two major factors in the metric of semantic information that affect the semantic variety of a natural language in written texts are:

1. The size of the vocabulary of the corpus – vocabulary effect – $V(N)$
2. Average number of meanings that can be attained by distinct words – polysemic effect – ϖ_r

In fact, synonymy, more than one word having the same meaning, is another major factor that contributes to semantic information. But, having no Zipf law that

can be related with this property left it outside of the scope of this study. Taking Zipf power laws as the basis, the semantic information ratio contained in a corpus *D* of length *N* with respect to the natural language it is written in, (*RSEI*), can be formulated as follows:

$$RSEI(D_N) = \frac{SEI(L_N)}{SEI(D_N)} = \frac{N^{\rho} \sum_{f=1}^{N} V_L(f,N) \cdot \varpi_L(f)}{V_D \sum_{f=1}^{N} V_D(f,N) \cdot \varpi_L(f)} \quad (17)$$

In the above equation, the numerator represents the number of meanings that should be conveyed by the number of words that should be seen in a corpus of size *N* written in language *L*. Here, N^{ρ}, by Heaps' law, is the expected number of distinct words, vocabulary, in any corpus of *N* words, written in language *L*. The denominator represents the number of meanings that is conveyed by the number of words that is seen in the corpus on hand. V_D is the actual size of the corpus *D* which is actively used in the performance metric. $V_L(f,N)$ is the number of distinct words expected to appear with frequency *f* in any corpus of size *N* written in language *L*; $V_D(f,N)$ is the actual number of distinct words that appear with frequency *f* in corpus *D*. $\varpi_L(f)$ denotes the expected average number of meanings of words having frequency of *f* and is calculated by the formula 13.

5. Test Results of the Proposed Performance Metric

To show that our proposed language model performance metric is independent from size of the corpus, we applied it to 20 different corpora of different sizes in 2 languages. For this purpose a Turkish corpus: METU (Say, et. al., 2004) and English corpora: Medlars and Time which are summarized in table 1 are used as the master corpora and as a benchmark cross entropy is assumed. Sizes of the corpora vary from 5,000 words to 100,000 words.

Table 1

The general statistics of the corpora used in the study (METU, Medlars, Time)

	METU	**Medlars**	**Time**
Corpus Size (*N*) in words	~2 M	161,605	249,567
Vocabulary Size (*V*)	200,000	12,609	20,856
Number of Documents	978	1,034	425

Table 2
Test results of proposed performance metric $B(m)$ and cross entropy $H(m,L)$

	METU Corpus		**Medlars Corpus**		**Time Corpus**	
N	***H*(*m*,*L*)**	***B*(*m*)**	***H*(*m*,*L*)**	***B*(*m*)**	***H*(*m*,*L*)**	***B*(*m*)**
5,000	3.2147	0.4788	2.4008	0.3649	2.4811	0.2550
10,000	3.3801	0.4841	2.5129	0.3442	2.5947	0.2375
15,000	3.4844	0.4811	2.5651	0.3494	2.6452	0.2350
20,000	3.5508	0.4716	2.5969	0.3513	2.6756	0.2337
25,000	3.6034	0.4660	2.6229	0.3547	2.6988	0.2352
30,000	3.6242	0.4839	2.6397	0.3637	2.7223	0.2352
35,000	3.6678	0.4620	2.6555	0.3652	2.7385	0.2374
40,000	3.6844	0.4806	2.6687	0.3667	2.7468	0.2414
45,000	3.7176	0.4736	2.6803	0.3691	2.7599	0.2417
50,000	3.7237	0.4816	2.6881	0.3746	2.7686	0.2435
55,000	3.7450	0.4835	2.6963	0.3767	2.7770	0.2469
60,000	3.7631	0.4944	2.7013	0.3845	2.7845	0.2480
65,000	3.7801	0.4996	2.7091	0.3834	2.7904	0.2494
70,000	3.7810	0.4985	2.7149	0.3862	2.7961	0.2520
75,000	3.7933	0.4976	2.7218	0.3865	2.8006	0.2550
80,000	3.8041	0.5001	2.7256	0.3914	2.8061	0.2552
85,000	3.8163	0.5037	2.7293	0.3967	2.8087	0.2596
90,000	3.8134	0.5115	2.7349	0.3960	2.8127	0.2608
95,000	3.8277	0.5094	2.7380	0.4003	2.8154	0.2634
100,000	3.8373	0.5154	2.7419	0.4036	2.8205	0.2633

In Table 2, the first column labeled as N, lists the corpora sizes in terms of word counts. For each considered size, 30 sample corpora are obtained randomly from every master corpus. Cross entropy for 1-gram language model ($H(m,L)$) according to formula 2 is calculated for every corpus of the same size and the averages are listed under column ($H(m,L)$) as a representative of the actual cross entropy of the corpus of the relevant size. The same is repeated for the proposed metric and the averages are listed under the column ($B(m)$). As it is seen in the table, the averaged cross entropy values increase in parallel with the increase in corpus size where as the proposed language model performance metric stays around for 0.50 for METU; 0.38 for Medlars and 0.25 for Time corpus.

The arguments given above can also be proven statistically. To determine if the average cross entropies calculated over corpora of different sizes are statistically different we look at the confidence intervals of the population means. If the sample size is large enough; $n \geq 30$; to calculate the confidence interval for the mean average cross entropy of a corpus of any size we do not need to check if the cross entropies show Gaussian distribution or not. That is why 30 is chosen as the number of sam-

ples for this work. For a corpus of a given size, the confidence interval for population mean (μ) of cross entropies is

$$\bar{x} - \frac{s}{\sqrt{n}} z_{0.025} \leq \mu \leq \bar{x} + \frac{s}{\sqrt{n}} z_{0.025}$$

Here, $\bar{x}$ is the mean of the samples (the mean values given in Table 3) and *s* is the mean sum of squared deviations of empirical values of 30 cross entropy values from the mean, and henceforth it is simply called as the standard deviation. The standard error is obtained by dividing the standard deviation by the square root of the sample size (n=30). Based on the central limit theorem, for large ($n \geq 30$) sample sizes, the mean of the samples has Gaussian distribution around the population mean, μ. At this point, it has to be underlined that the distribution mentioned above belongs to the mean of 30 cross entropy values, not the individual cross entropy values. As a result, for a confidence interval of 5%, the mean of the population where the samples are drawn must be in the interval $[-z_{0.025}, z_{0.025}]$ ([-1.96, 1.96]) of standard normal distribution. The confidence interval obtained by the interval mentioned above is known as the "95% confidence interval of population mean". For corpora of different sizes, if at least one of the mean cross entropy is within the confidence interval of the other, it is stated that there is no significant difference between the cross entropies. In Table 3, the confidence intervals for all corpora of different sizes are listed.

Table 3

95% confidence intervals of population means with different corpus sizes for cross entropy and proposed performance metric (95% CI).

		METU Corpus		**Medlars Corpus**		**Time Corpus**	
N		***H(m,L)***	***B(m)***	***H(m,L)***	***B(m)***	***H(m,L)***	***B(m)***
5,000	**mean**	**3.2147**	**0.4788**	**2.4008**	**0.3649**	**2.4811**	**0.2550**
	95%	3.1046	0.3924	2.3451	0.3056	2.4285	0.2257
	CI	3.3248	0.5753	2.4565	0.4243	2.5338	0.2843
10,000	**mean**	**3.3801**	**0.4841**	**2.5129**	**0.3442**	**2.5947**	**0.2375**
	95%	3.2779	0.3534	2.4828	0.3104	2.5494	0.2169
	CI	3.4822	0.6309	2.5430	0.3781	2.6400	0.2581
15,000	**mean**	**3.4844**	**0.4811**	**2.5651**	**0.3494**	**2.6452**	**0.2350**
	95%	3.3951	0.4183	2.5400	0.3210	2.6043	0.2180
	CI	3.5738	0.5483	2.5903	0.3779	2.6860	0.2519
20,000	**mean**	**3.5508**	**0.4716**	**2.5969**	**0.3513**	**2.6756**	**0.2337**
	95%	3.43120	0.3282	2.5752	0.3240	2.6395	0.2195

	CI	3.6704	0.6354	2.6186	0.3787	2.7117	0.2479
25,000	**mean**	**3.6034**	**0.4660**	**2.6229**	**0.3547**	**2.6988**	**0.2352**
	95%	3.5304	0.3822	2.6059	0.3319	2.6685	0.2235
	CI	3.6764	0.5589	2.6400	0.3777	2.7290	0.2468
30,000	**mean**	**3.6242**	**0.4839**	**2.6397**	**0.3637**	**2.7223**	**0.2352**
	95%	3.5433	0.3619	2.6260	0.3378	2.6949	0.2249
	CI	3.7052	0.6226	2.6535	0.3898	2.7497	0.2455
35,000	**mean**	**3.6678**	**0.4620**	**2.6555**	**0.3652**	**2.7385**	**0.2374**
	95%	3.6123	0.3892	2.6484	0.3389	2.7216	0.2276
	CI	3.7234	0.5411	2.6627	0.3916	2.7555	0.2473
40,000	**mean**	**3.6844**	**0.4806**	**2.6687**	**0.3667**	**2.7468**	**0.2414**
	95%	3.6024	0.3864	2.6587	0.3428	2.7284	0.2302
	CI	3.7664	0.5851	2.6787	0.3908	2.7652	0.2526
45,000	**mean**	**3.7176**	**0.4736**	**2.6803**	**0.3691**	**2.7599**	**0.2417**
	95%	3.6568	0.3952	2.6692	0.3399	2.7432	0.2322
	CI	3.7785	0.5594	2.6915	0.3984	2.7766	0.2512
50,000	**mean**	**3.7237**	**0.4816**	**2.6881**	**0.3746**	**2.7686**	**0.2435**
	95%	3.6500	0.3736	2.6759	0.3453	2.7521	0.2315
	CI	3.7975	0.6027	2.7003	0.4039	2.7851	0.2556
55,000	**mean**	**3.7450**	**0.4835**	**2.6963**	**0.3767**	**2.7770**	**0.2469**
	95%	3.6873	0.3683	2.6840	0.3510	2.7661	0.2362
	CI	3.8026	0.6138	2.7086	0.4026	2.7879	0.2576
60,000	**mean**	**3.7631**	**0.4944**	**2.7013**	**0.3845**	**2.7845**	**0.2480**
	95%	3.6919	0.3840	2.6884	0.3554	2.7750	0.2385
	CI	3.8343	0.6191	2.7143	0.4138	2.7940	0.2576
65,000	**mean**	**3.7801**	**0.4996**	**2.7091**	**0.3834**	**2.7904**	**0.2494**
	95%	3.7170	0.3891	2.6968	0.3551	2.7811	0.2389
	CI	3.8431	0.6248	2.7215	0.4118	2.7998	0.2599
70,000	**mean**	**3.7810**	**0.4985**	**2.7149**	**0.3862**	**2.7961**	**0.2520**
	95%	3.7212	0.3830	2.7016	0.3580	2.7864	0.2421
	CI	3.8408	0.6294	2.7283	0.4145	2.8059	0.2620
75,000	**mean**	**3.7933**	**0.4976**	**2.7218**	**0.3865**	**2.8006**	**0.2550**
	95%	3.7369	0.3950	2.7083	0.3587	2.7908	0.2443
	CI	3.8497	0.6117	2.7354	0.4144	2.8103	0.2657
80,000	**mean**	**3.8041**	**0.5001**	**2.7256**	**0.3914**	**2.8061**	**0.2552**
	95%	3.7480	0.3768	2.7086	0.3592	2.7959	0.2438
	CI	3.8602	0.6402	2.7426	0.4237	2.8163	0.2666
85,000	**mean**	**3.8163**	**0.5037**	**2.7293**	**0.3967**	**2.8087**	**0.2596**
	95%	3.7677	0.3818	2.7156	0.3670	2.7990	0.2472
	CI	3.8649	0.6421	2.7430	0.4263	2.8185	0.2719
90,000	**mean**	**3.8134**	**0.5115**	**2.7349**	**0.3960**	**2.8127**	**0.2608**
	95%	3.7724	0.4035	2.7222	0.3665	2.8051	0.2515

	CI	3.8543	0.6317	2.7478	0.4255	2.8202	0.2702
95,000	**mean**	**3.8277**	**0.5094**	**2.7380**	**0.4003**	**2.8154**	**0.2634**
	95%	3.7765	0.3887	2.7249	0.3687	2.8035	0.2505
	CI	3.8789	0.6556	2.7513	0.4319	2.8273	0.2762
100,000	**mean**	**3.8373**	**0.5154**	**2.7419**	**0.4036**	**2.8205**	**0.2633**
	95%	3.7899	0.3967	2.7304	0.3762	2.8136	0.2545
	CI	3.8847	0.6496	2.7535	0.4311	2.8274	0.2720

In Table 3 for each corpus size, cross entropy ($H(m,L)$ and performance value of the proposed metric ($B(m)$ produced over *1-gram* language model averaged over 30 samples and their 95% confidence intervals are listed. For example, in the case of METU corpus of size N = 5,000, the calculated average value over 30 samples for cross entropy is 3.2147; average performance value is 0.4788. For a corpus of this size, the lower limit for the mean cross entropy in 95% confidence interval is 3.1046 and the upper limit is 3.3248 (the interval is [3.1046, 3.3248]). To state that mean cross entropy value calculated over METU corpus of size N = 5,000 is significantly different from that of another corpus of different size it has to be greater than the upper limit of the calculated 95% confidence interval i.e. 3.3248. As it is seen in Table 3 for N = 10,000, mean cross entropy value is 3.3801 which is greater than the upper limit of the confidence interval for N = 5,000 which is 3.3248. Thus we can confidently say that for significance level $\alpha = 0.05$ mean cross entropy values obtained over corpora of sizes N = 5,000 and N = 10,000 are significantly different. We see similar results for mean cross entropies in case of Medlars and Time corpora in Table 3.

On the other hand, 95% confidence interval calculated for N = 5,000 encapsulates all other mean performance values in case of our proposed metric in all corpora involved. In other words, the performance of the proposed metric generated over *1-gram* language model does not show statistically significant difference when applied on corpora of different sizes. In summary, we have reached experimental evidences that show cross entropies calculated over different sized corpora are significantly different. But, there is no evidence that shows the performance values obtained using the proposed metric are significantly different.

The cross entropy and "proposed performance metric" results obtained from different corpus sizes are shown graphically on METU, Medlars and Time Corpora in Figure 2, 3 and 4 respectively.

As seen in Figure 2, 3 and 4 clearly, when corpus size is increasing, cross entropy values (upper curve) are increasing, but the values of proposed performance metric (lower curve) are oscillating around a constant value. The estimated linear trend of aforesaid values is also shown in the figures. When the equations of the estimated linear trend of cross entropy and proposed performance metric are com-

pared, it is seen that the proposed performance metric is more stable and consistent than cross entropy in spite of the change in corpus size.

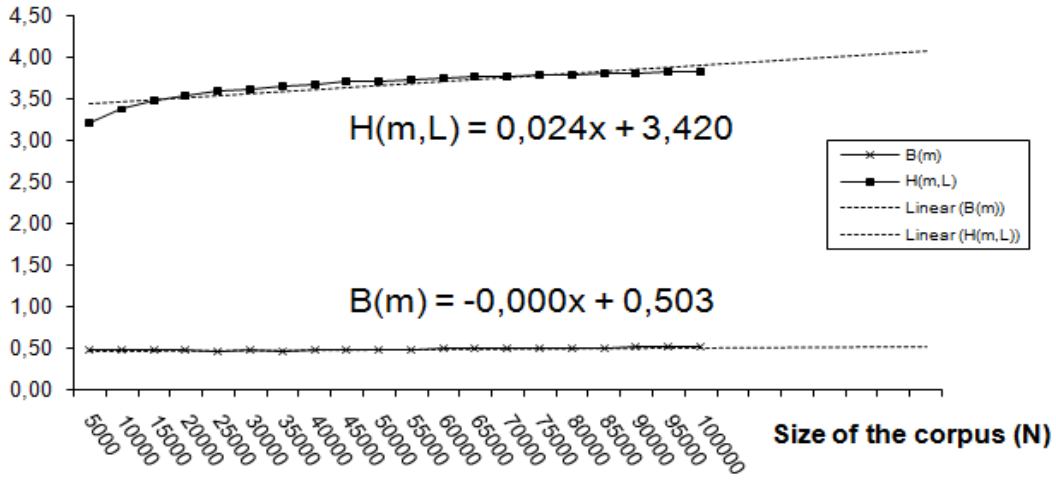


Figure 2: The cross entropy (*H(m,L)*) and "proposed performance metric" (*B(m)*) results obtained from different corpus size for METU corpus.

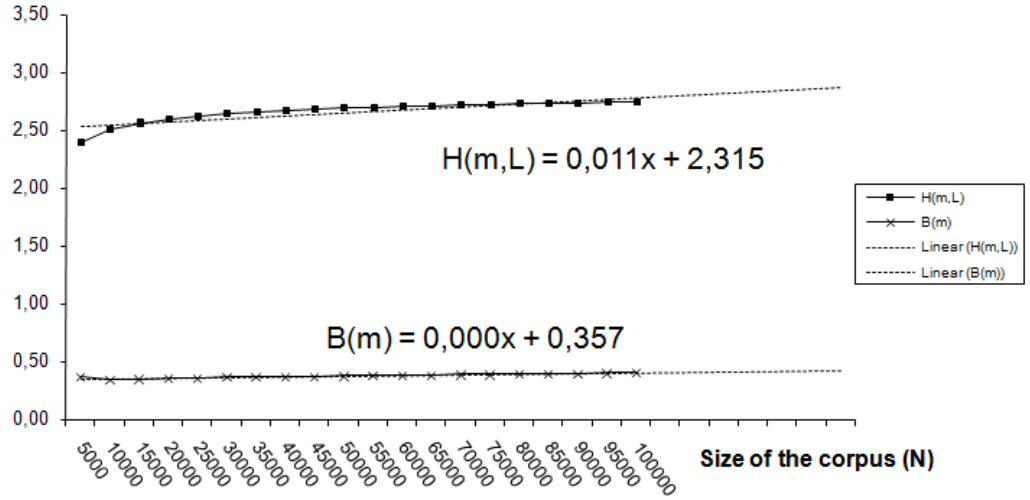


Figure 3: The cross entropy (*H(m,L)*) and "proposed performance metric" (*B(m)*) results obtained from different corpus size for Medlars corpus.

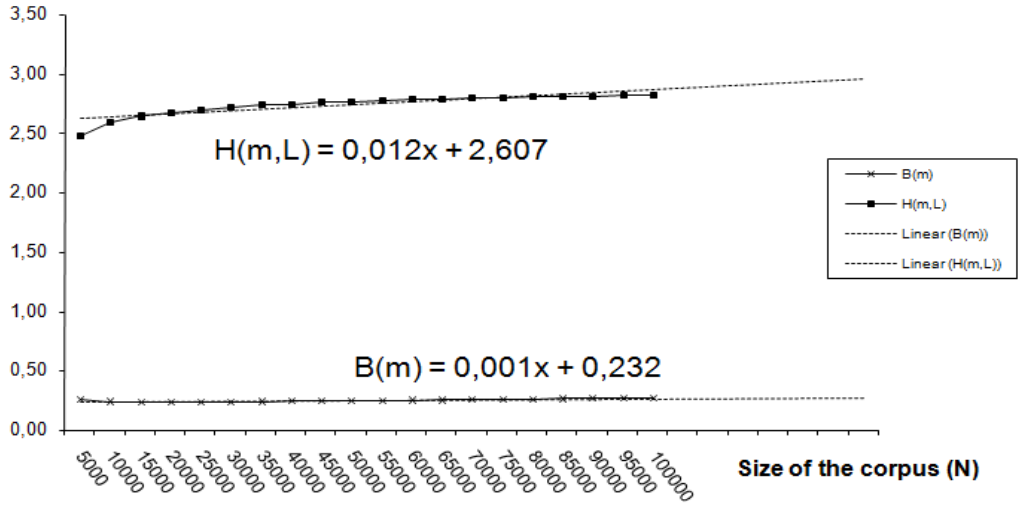


Figure 4: The cross entropy (*H(m,L)*) and "proposed performance metric" (*B(m)*) results obtained from different corpus size for Time corpus..

6. Conclusion

In this study, a language model performance metric which is independent of corpus size is formulated exploiting Zipf's and Heaps' Power Laws. The rational behind relies on our intuitive argument that there should be a strong link between the signal information and the semantic information since the semantic information that the sender wants to transmit determine the set of symbols that compose the message. The empirical results obtained from the experimental studies provide enough evidence to reject the common view that the semantic information and the signal information are weakly correlated.

The proposed metric is evaluated on test corpora of 20 different sizes in two different languages: English and Turkish. The test results showed that as the corpus size increases, the amount of signal information increases as expected, on the other hand the metric, considering the relative amount of semantic information contained in the corpus, gives stable values strengthening the claim of corpus size independency. Further studies should be carried out with corpora in different languages to investigate the language independency of the metric.

Acknowledgement
This study was funded by The Scientific and Technical Research Council of Turkey (EEEAG- 104E120) and Ege University, Science and Technology Application and Research Center (2006/BIL/013).

References

Heaps, H.S. (1978). *Information Retrieval: Computational and Theoretical Aspects*. Academic Press, *Heaps' law is proposed in Section 7.5 (pages 206–208).*

Dinçer, B.T. (2009). *Türkçe İçin İstatistiksel Bir Bilgi Getirim Sistemi*. Phd. Thesis. International Computer Institue, Ege University

Ilgen B., Karaoğlan B. (2007). Investigation of Zipf's 'Law-of-Meaning' on Turkish Corpora. *22nd International Symposium on Computer and Information Sciences, Middle East Technical University, Ankara, Turkey, IEEE*

Karaoğlan B., Dinçer, T., Kışla, T., Kumova Metin, S., Kocabaş, İ. (2008). *Zipf Kanunları Esasında Güncel Yazılı Türkçe'nin Nicel Dilbilim Ölçütleri*. TUBITAK Project: EEEAG 104E120.

Karaoğlan B., Kumova Metin S., Dinçer B.T. (2011). Investigating Zipf's Laws on Turkish, *2nd Int. Symposium on Computing in Science & Engineering, Kuşadası, İzmir, Turkey.*

Kocabaş, I., Karaoğlan, B., Kışla T. (2007). Zipf's Law of Burstiness in Turkish: The Length of Intervals Between Repetitions. *22nd ISCIS, Ankara, Turkey, IEEE Computer Society 7-9 November*

Kornai, A. (2002), How many words are there? *Glottometrics 2002(4), 61-86.*

Kumova, S., Karaoglan, B., Dinçer B.T. (2006). Kelime Sayısı - Kelime Dağarcığı İlişkisinin Belirlenmesi. *Turkish Symposia on Artificial Intelligence and Neural Networks, pp. 269-276*

Rosenfeld, R. (2000). Two decades of Statistical Language Modeling: Where Do We Go From Here? *Proceedings of the IEEE, 88(8).*

Say B., Zeyrek D., Oflazer K., Özge U. (2002). Development of a Corpus and a Treebank for Present day Written Turkish. *Proceedings of the Eleventh International Conference of Turkish Linguistics.*

Shannon, C.E., Weaver, W. (1949). *The Mathematical Theory of Communication*, Urbana: University of Illinois Press.

Shannon, C. E. (1993). *Collected Papers*, edited by N. J. A. Sloane and A. D. Wyner, New York: IEEE Press.

Tür,G., Hakkani-Tür, D., Oflazer K.(2002). A Statistical Information Extraction System for Turkish. *Natural Language Engineering ,9(2)*, 181-210.

Zipf, G.K. (1949). *Human Behavior and the Principle of Least-Effort*. Cambridge, MA: Addison-Wesley

Wie kommuniziert eine Gesellschaft aus linguistischer Sicht?[1]

Sigurd Wichter, Göttingen

1. Einführung

Wie kommuniziert eine Gesellschaft aus linguistischer Sicht? Diese Frage möchte ich behandeln und am Ende eine Antwort unternehmen auf der Grundlage meiner Theorie der sprachlichen Kommunikation, der Reihentheorie.

Der Weg zur Antwort ist der: Sozialität und Zeitlichkeit des Kommunizierens genauer zu beachten. Zu Beginn gilt es wahrzunehmen, dass situationsabgeschlossene Kommunikationseinheiten wie das Gespräch oder die Kommunikation eines Textes (wir nennen diese Kommunikationseinheiten „**Kommunikate**“) in der Kommunikation nicht isoliert und kommunikativ beziehungslos sind. Ein Kommunikat steht vielmehr in einer Folge von Kommunikaten, die untereinander sozial kohärent sind. Beispiele für solche Folgen sind etwa Folgen von Briefen in einem Briefwechsel, Folgen von Emails in einer geschäftlichen Verbindung, Folgen von Gesprächen in einem Gremium oder Folgen von Gesprächen im privaten Kreis.

Diese Folgen von Kommunikaten sind sozial kohärent über die Sozialität ihrer Gruppe, und sie sind geprägt durch die zeitlichen Erstreckungen des Kommunikationsablaufs mit den resultierenden Einbettungsverhältnissen.

Ein Textkommunikat beispielsweise besteht aus **Sprechakten**. Diese werden repräsentiert durch Sätze. Diese bestehen ihrerseits aus Satzteilen etc. Das Textkommunikat bettet hier also ein.

[1] Udo Kipper habe ich für ausführliche Gespräche und Kritik zu danken.

Beim Briefwechsel dagegen hat dieses Textkommunikat eine andere Position. Hier ist es selbst eingebettet in die Folge, die über die reine Abfolge von Textkommunikaten hinaus durch soziale Kohärenz einschließlich thematischer Bindungen als Einheit bestimmt ist und die wir als „**Reihe**“ terminologisieren werden.

Aus diesen Beobachtungen heraus samt weiteren Analysen gehen wir davon aus, dass alle sprachlichen Kommunikationseinheiten alternativ **drei Ebenenbereichen** zugeordnet werden können. Diese sind:

1. der Ebenenbereich der **Sprechakte** und Sprechaktsequenzen

2. der Ebenenbereich der **Kommunikate** als der situationsabgeschlossenen Kommunikationseinheiten (Gespräche, Textkommunikate u.a.)

3. der Ebenenbereich der **Reihen** (als der sozial kohärenten Folgen von Kommunikaten)

Die Einheiten dieser Ebenenbereiche sind nun genauer zu bestimmen. Unser Hauptaugenmerk gilt weniger dem Ebenenbereich der kleineren Einheiten, die überdies per Sprechakttheorie- und Grammatikliteratur gut, wenngleich manchmal etwas bunt dargestellt sind. Unser Hauptaugenmerk gilt vor allem den **Reihen**. Denn mit dem auf ihrer Erfassung basierenden Konzept, der **Reihentheorie**, können wir die **Kommunikation einer Gesellschaft**, die **Kommunikation der Gesellschaften untereinander** und die **globale Kommunikation** aus unserer Sicht bestimmen.

Die Reihentheorie habe ich 2011 in der Monographie „Kommunikationsreihen aus Gesprächen und Textkommunikaten“[2] vorgestellt. Ich stütze mich im Folgenden auf diese Arbeit und entwickle sie gleichzeitig weiter.

Um die zentrale Größe der Reihe gut bestimmen zu können, müssen zuvor deren Elemente, die Kommunikate vor allem, genauer bestimmt werden. Auf dieses Feld nun begeben wir uns jetzt. Am Anfang einer Kommunikatbesprechung stehen jeweils Beispiele, dann folgt die Bestimmung.

Wir beginnen mit dem Textkommunikat.

2. Beispiele für Textkommunikate

Wir unterbreiten einen kurzen, zweischrittigen Schriftwechsel zwischen dem Geschäftsführer eines Instituts und einem Mitglied. Der erste Schritt ist ein Brief, der zweite Schritt eine Email. Wir glauben dabei nicht, dass wir dem geneigten Leser neue Einblicke in die Theorie der Geschäftsführung vermitteln. Es geht uns

[2] Sigurd Wichter (2011): Kommunikationsreihen aus Gesprächen und Textkommunikaten. Zur Kommunikation in und zwischen Gesellschaften. Berlin/Boston: de Gruyter (= Reihe Germanistische Linguistik 294) (zitiert als „‚Kommunikationsreihen’ (2011)“).

lediglich und vielmehr um die Exemplifizierung der Spezifik von Textkommunikaten und deren Folgebeziehung.

Die Wiedergabe der Texte erfolgt anonymisiert, auch in den Zeitangaben, die nunmehr geändert sind. Die Sprechakte oder kurz Akte[3] sind durchgezählt und stehen nacheinander auf je eigener Halbzeile, die die Produktionshälfte der Zeile bildet und die sich in der Spalte des Produzenten befindet. Die Rezeptionshälfte der Zeile bleibt leer, solange normales Verstehen unterstellt werden kann. Andernfalls trägt sie entsprechende Vermerke zu Verstehensdifferenzen zwischen Produktion und Rezeption.

Einladung

Geschäftsführer NN1	**Mitglied NN2**
1] Institut für […] an der […] Universität NN0	
2] NN1	
3] Geschäftsführer des Instituts für […]	
4] […] [Straße und Hausnummer]	
5] […] [Postleitzahl und Ort]	
6]NN0, den 4.2.2008	
7] Herrn Prof. Dr. NN2	
8] […] [Straße und Hausnummer]	
9] […] [Postleitzahl und Ort]	
10] **Institut für […], NN0**	
11] **Mitgliederversammlung am 28.3. 2008**	
12] Lieber Herr Prof. NN2,	
13] ich möchte Sie als Mitglied des Instituts für […] zu unserer jährlichen Mitgliederversammlung am Freitag, dem 28.3.2008, von 10 bis 12 Uhr, in den Konferenzraum […] im […] einladen.	
14] Die Tagesordnung folgt noch. 15] Im Jahr 2007 fand leider aufgrund zeitlicher Probleme vieler Mitglieder keine Mitgliederversammlung statt.	
16] Bitte geben Sie mir bis zum 4.3.2008 kurz Bescheid, ob Sie zu unserer Mitgliederversammlung am 28.3. kommen können –	
17] vielen Dank im Voraus.	
18] Gibt es übrigens eine Emailadresse, über die ich Sie gut erreichen kann?	

[3] Akte werden realisiert durch Formulierungstypen (Satz, Wortgruppe, Wort, tonales Zeichen, ggf. Geste und mimischer Ausdruck und weitere Zeichen), vgl. „Kommunikationsreihen“ (2011:40 u. 2.2.2.C3).

19] Mit besten Grüßen aus NN0	
20] [Unterschrift]	
21] NN1	

Die Antwort lautet wie folgt:

Niederlegung der Mitgliedschaft

Geschäftsführer NN1	**Mitglied NN2**
	1][Emailabsender NN2]
	2][Emailempfänger NN1]
	3][Datum]
	4]Lieber Herr NN1,
	5] haben Sie Dank für Ihre Einladung zur jährlichen Mitgliederversammlung des Instituts.
	6] Da ich aber, wie Sie vielleicht wissen, nicht mehr in NN0 wohne, erscheint es mir in dieser Situation doch sinnvoll, dem Institut nicht mehr als Mitglied anzugehören.
	7] Ich lege also hiermit meine Mitgliedschaft nieder.
	8] In eins damit möchte ich sagen, dass ich sehr gerne mitgemacht habe.
	9] Dem Institut und Ihnen als Geschäftsführer wünsche ich für die Zukunft alles Gute.
	10] Mit herzlichen Grüßen verbleibe ich
	11] Ihr NN2

Mit diesem Schriftwechsel haben wir eine Reihe vor uns, die aus zwei Textkommunikaten besteht, aus dem Brief und der Email. Die Textkommunikate bilden deshalb eine Reihe, weil sie sozial kohärent sind als Kommunikationen zwischen zwei Mitgliedern einer Gruppe. Sie sind überdies auf explizite Weise auch inhaltlich kohärent, da im Einzelnen enge Bindungen[4] bestehen, unter anderem etwa die Erwartungen, die der Brief formuliert (Kenntnisnahme, Bescheid, Emailadresse ggf. zukommen lassen) und die in der Antwort erfüllt werden. Die Textkommunikate ihrerseits bestehen aus Akten.

Die betrachtete Reihe stellen wir als Folge der Textkommunikate dar:

1. Einladung (Geschäftsführer an Mitglied)
2. Niederlegung der Mitgliedschaft (Mitglied an Geschäftsführer)

[4] Zur Intertextualität und zu deren Einordnung in eine übergreifende Intersegmentalität vgl. „Kommunikationsreihen“ (2011:202-214).

Wir werden im Folgenden zunächst das Textkommunikat bestimmen, und damit den Typ des Briefes und der Email in einen allgemeineren typologischen Zusammenhang stellen, wobei wir auf die Akte nicht mehr prinzipiell eingehen werden, da wir sie in der Informiertheit des Lesers sicher geborgen fühlen. Danach folgt die Exemplifizierung und Bestimmung weiterer Kommunikate.

3. Bestimmung des Textkommunikats

Was ist ein Textkommunikat genauer? Zunächst: Ein Textkommunikat[5] ist eine bestimmte Einheit der Textkommunikation. Die **Textkommunikation** ist dabei eine **monologische** sprachliche Kommunikation, in der ein Produzent einen Inhalt für einen Rezipienten formuliert und dieser die Formulierung rezipiert. Die Produktion der Formulierung erfolgt dabei insgesamt unbeeinflusst und wird nicht unterbrochen.

Textkommunikate sind nun (ad 1) solche Einheiten der Textkommunikation, die inhaltlich und äußerlich abgeschlossen sind. Der Produzent hat für den Augenblick, für die jetzige Kommunikationssituation, alles gesagt, was zu sagen ist. Textkommunikate sind also **lokal abgeschlossen**. (Das gilt im Übrigen nicht für Akte).

Textkommunikate können sich (ad 2) zum einen inhaltlich auf die Vergangenheit und auf die Zukunft beziehen. Und sie stehen zum anderen auch jenseits dieser Explizitheit des Vor- und Nachbezugs in der sozialen Vor- und Nachgeschichte. Denn der Produzent ist sich in der Regel bewusst, dass es vor seinem Schreiben Textkommunikate (und übrige Kommunikate) gab, die mit dem jetzigen geplanten Textkommunikat in sozial relevantem Zusammenhang stehen, und dass sein jetziges Schreiben nicht nur im respice finem des nächsten Kommunikats, sondern überhaupt in respice finem weiterer zukünftiger Textkommunikate und weiterer zukünftiger Kommunikate überhaupt steht. Der Schreiber des Textes oder der Vortragende hat im Moment der Kommunikation oder im Moment der inneren Vorwegnahme der Kommunikation in aller Regel wohl eine Agenda, eine Agenda, wie verschwommen auch immer, wie bedacht auch immer, für das, was nach seinem Kommunikat passieren könnte, für die in Zukunft möglichen Kommunikate und weitere Entwicklungen. Kurz: Textkommunikate sind **global offen.**

Textkommunikate sind, wie die Textkommunikation überhaupt, (ad 3) **kommunizierte Texte**. Die Kommunikation durch eine tatsächliche Rezeption muss vollzogen sein. Die Stellungnahme, die ungelesen im Papierkorb landet und ungelesen bleibt, mag zwar ein Text, vielleicht auch ein exzellenter Text sein. Sie ist kein Textkommunikat.

Für Textkommunikate gelten (ad 4) einige Besonderheiten in folgenden Kategorien:

[5] Ausführlich in „Kommunikationsreihen“ (2011:77-114).

Anzahl der Kommunikationspartner,
Rechte der Partner,
mediale Gestaltung,
Akt als Basiseinheit und
Umfang der Einheit.

Stichwort Hermeneutik. Zu betonen ist: Wenn ein Produzent ein und denselben Text an zwei Adressaten schickt und diese jeder für sich den Text rezipieren, dann liegen **zwei** Textkommunikate vor. Denn es sind zwei selbständige Kommunikationen. Für viele Bereiche kann man sich behelfen, indem man bei Textgleichheit die verschiedenen Textkommunikate zu einem einzigen zusammenfasst und dann in der Analyse fortfährt auf der Basis eines einzigen Textkommunikats (etwa in der Massenkommunikation), diese Vereinfachung gleichwohl im Gedächtnis behält. Denn wenn man als Vorbedingung für die Fortsetzung der Analyse die genaue Ausdifferenzierung der jeweiligen Rezeption fordert, ist man bei sehr vielen Rezipienten in Schwierigkeiten (etwa bei einer Million; aber auch hunderttausend Rezeptionen sind schon schwierig). Zu unterscheiden sind natürlich einfache Texte „Wir treffen uns morgen nachmittag um 4 im Café Grotemeyer“ von etwas komplexeren, etwa dem Neuen Testament, manchen Gesetzestexten, manchen literarischen Texten, überhaupt von all jenen Texten, die wir antreffen, wenn wir den Spuren der allgemeinen Hermeneutik und denen der jeweiligen Haushermeneutik folgen.

Wenn man die Komplexität arbeitstechnisch zunächst ausspart, muss man natürlich eine Vorstellung davon haben, was man anrichtet und ob man das Angerichtete wieder einfangen muss, und wenn ja, wie man das macht. Angesiedelt und potentiell hilfreich in diesem Bereich sind mehrere Disziplinen, die erwähnte **Hermeneutik** etwa und ihre Dependancen in den einzelnen textbezogenen Fächern. Wünschenswert wäre ein Ausbau einer linguistischen Hermeneutik, die sich auch um den Bereich der Massenmedien kümmern könnte.

Zur typologischen Reichweite. Als Textkommunikate kommen nun nicht nur die Typen Briefe und Emails in den Bereich der Erfassung, sondern sehr viele weitere Typen. Den Textkommunikaten zugrunde liegende Texte stammen aus sehr vielen Bereichen, etwa aus dem politischen, dem wirtschaftlichen, dem religiösen und dem kulturellen und nicht zuletzt aus dem privaten Bereich, um nur die Ausdifferenzierung nach einer, nach der Kategorie des Sektors, zu nennen. Auch zur Texttypologie gibt es eine umfangreiche Literatur.

Das Textkommunikat ist wie bemerkt ein monologisches Kommunikat. Wenn wir nun zum Gespräch übergehen, haben wir im Gegensatz zu diesem ein dialogisches Kommunikat vor uns, oder genauer: ein in sich dialogisches Kommunikat, das wir dann als „zeitdirekt dialogisches Kommunikat“ terminologisieren werden.

4. Beispiele für Gespräche

Auch Gespräche können eine Reihe bilden. Betrachten wir einige der Gespräche zwischen einem Blumenhändler und seinen Kundinnen (als Auswahl aus der gesamten Reihe der Verkaufsgespräche im Geschäft des Blumenhändlers). Wir zitieren partiell vier Gespräche. Diese entnehmen wir dem Corpus von Kipper 1968.[6]

Aus dem Corpus von Kipper (1968) ist zwar nicht zu ersehen, in welcher Weise diese vier Gespräche aufeinander folgen. Aber sie liegen alle in einem bestimmten Zeitraum von einigen Monaten und sind tatsächlich Schritte einer Reihe, Schritte einer Reihe deshalb, weil die Gespräche inhaltlich auf einen bestimmten, dem Blumenhändler und den Kundinnen gemeinsamen Gegenstandsbereich gerichtet sind, weil in jedem Gespräch potentiell Partner- und Sachwissen aus früheren Gesprächen bereit besteht und relevant ist und weil die Konstellation aus Blumenhändler und jeweiliger Kundin eine Gruppe bildet mit bestimmten sozialen Bindungen.

Für die Transkription verwenden wir nicht das für die Gesprächstranskription bzw. die Transkription gesprochener Sprache übliche Zeilenmodell. Wir erweitern vielmehr die oben für die Modellierung der Textkommunikate eingeführte Spaltentranskription.

Das Verfahren bleibt im Prinzip dasselbe. Es sei noch einmal kurz rekapituliert. Die Akte als Ganze werden also von oben nach unten auf je eine Halbzeile geschrieben innerhalb der Spalte des jeweiligen Sprechers. Die andere Halbzeile, die die Halbzeile der Rezeption ist, bleibt leer, sofern normales Verstehen angenommen werden kann. Sie symbolisiert lediglich auf diese Weise den Rezeptionsprozess im Fall anzunehmender normaler Rezeption. Anders im Fall anzunehmender Verstehensdifferenzen zwischen der Produktionsformulierung und der Rezeption: Hier ist die Rezeptionshalbzeile der Ort entsprechender Vermerke.

Da die Gesprächspartner sich (im Unterschied zum Textkommunikat) wechselseitig äußern, gibt es natürlich keine leere Spalte. Vielmehr sind beide Spalten mit Akten besetzt.

Simultanes Sprechen wird *zusätzlich* gekennzeichnet. Die simultanen Akte bzw. Segmente aus Akten selbst werden entgegen ihrer Simultanität zeilenmäßig getrennt aufgeschrieben nach dem genannten Prinzip: ein Akt auf eine Halbzeile, um die Rezeptionszeilenhälfte leer oder aber für entsprechende Einträge frei zu halten. Das Prinzip bedingt, dass die genaue Nachstellung des Zeitverlaufs für

[6] Die Gespräche mit ihren Transkriptionen stammen aus: Udo Kipper (1968): Untersuchungen zur deutschen Umgangssprache der Gegenwart (unter besonderer Berücksichtigung des Verkaufs-Gesprächs). Hausarbeit der Fachprüfung für das Lehramt an Gymnasien. Bochum, Bd. 2, S. 6 (Gespräch „Erdpflege“), S. 7-9 (Gespräch „Gummibaum“), S. 10 u. 12-13 (Gespräch „Friedhof“) und S. 14 (Gespräch „Hochzeit“).

diese Stelle nur über die Zusatzkennzeichnung (Doppelzeile, Fett- und Kursivdruck, s.u.) erfolgt.

Erdpflege

Blumenhändler	**Kundin**
[…][7]	
	[…]
	1] und dann wollte ich etwas für Erdpflege
	2] ham Sie da auch so irgend was
	3] was kostet so was
4] zwo Mark	
5] **reißen Se sich das** hier auf ***n das is aber auch***	
	6] **n das is aber auch**
	7] ja
8] oh da kommen Se aber weit mit	
	9] ja
10] für Läuse für allet	
11] und denn können Se pudern	
	12] n
13] kommen Se weit mit	
	14] was kost das
15] eine Mark	
16] och da reichen Se den ganzen Sommer mit	
	17] schönen Dank auch

Gummibaum[8]

Blumenhändler	**Kundin**
	1] Morgen
2] Guten Morgen Frau B	
	3] Ich möchte gern ein paar Nelken haben
4] ja aber gern	
[…]	

[7] Die Transkription setzt erst im Verlauf des Gesprächs ein.
[8] Die ausgelassenen Akte werden nicht mitgezählt.

	[…]
	5] und dann hab ich noch eine Frage Herr XX
6] aber gern	
	7] mein Gummibaum woll der wächst hoch
8] n	
	9] nun hat man mir gesagt den könnt man abschneiden
10] auf Ihre Verantwortung	
	11] ja ich mach das nich selber
12] also Frau B **???** ***denn***	
	13] **denn** gestern warn se da zum Schnibbeln
	14] da vonne Dings die Bäume da ham **se da** ***n***
15] **n**	
	16] da war en Gärtner bei
	17] der sagte ja sicher können Se das machen
	[…]
[…]	
18] ja also passen Se auf	
19] ich will Ihnen das erklären	
[…]	
	[…]
	20] ja ich will das ja garnich machen
	21] Sie wenn’s geht, sollen Sie das ja machen
22] ja das tue ich auch Frau B	
	[…]
[…]	
	23] wenn Se ma vorbeikommen
24] ja das tue ich auch Frau B	
[…]	
	[…]
	24] hat noch so’n Stück bis zur Decke
25] dann können we noch bis morgen warten	
	26] woll also

27] jaja	
	28] also auf Wiedersehen
29] Schönen Dank	
30] Wiedersehen	

Dieses Gespräch besonders zeigt die Vertrautheit zwischen dem Blumenhändler und einer Kundin. Der Blumenhändler ist außerordentlich hilfsbereit, und die Kundin bringt ihm volles Vertrauen entgegen. Der Umgang schließt Hausbesuche des Händlers mit ein.

Friedhof

Blumenhändler	**Kundin**
1] Guten Morgen	
	2] morgen
3] bitte schön Frau C	
	4] ich wollte gern en paar Blümchen haben
5] aber gern	
6] fürn Friedhof	
	7] ja
[…]	
	[…]
8] Frau C	
9] Sie ham unbeschränkt Kredit bei mir	
10] wenn se alle so wärn wie Sie	
	11] oh ja YY ich sag schon Herr NN wir ham doch viel für unsere Omma getan woll
12] jo	
	13] jede Woche jede Woche jede Woche en Strauß
14] ja	
	15] jede Woche bald für zehn Mark Blumen auf'n Friedhof woll
16] recht schönen Dank	
	17] jo kann ich heute Nachmittag ma mit Ruhe darauf gehen
18] jo sicher	
	19] so
20] is ja auch schönes Wetter	

	21] ja also auf Wiedersehen
22] Wiedersehen Frau C	

Auch dieses Gespräch zeigt eine enge Verbindung (die Kundin lässt im Verlauf des Gesprächs einen Teilbetrag anschreiben) und bezeugt eine längere Vorgeschichte vertrauten Umgangs.

Hochzeit

Blumenhändler	**Kundin**
1] Morgen Frau D	
2] bitte schön	
	3] bin ich schon dran
4] jawohl Frau D	
	5] ich hätt gern en Blümchen eine Topfblume ja
6] zum Verschenken	
	7] zur Hochzeit sollte es sein
8] zur Hochzeit	
9] ja dann nehm' Se doch so ne Y **oder** Z ***ja***	
	10] **ja**
	11] ja wo se mehr von haben
	12] nich wo se länger was von haben
[…]	
	[…]

Die Gespräche zeigen insgesamt die soziale Kohärenz zwischen dem Blumenhändler und seinen Kundinnen, und zwar nicht nur in der Minimalanforderung an soziale Kohärenz zwischen Verkäufer und Käufer überhaupt, sondern in der besonderen Ausprägung des vertrauten und vertrauensvollen Umgangs.

Als Elemente der Reihe, die wir „Blumenladenreihe" nennen wollen, ergeben sich:

1. Erdpflege (Gespräch zwischen Blumenhändler und Kundin)
2. Gummibaum (Gespräch zwischen Blumenhändler und Kundin)
3. Friedhof (Gespräch zwischen Blumenhändler und Kundin)
4. Hochzeit (Gespräch zwischen Blumenhändler und Kundin)

5. Bestimmung des Gesprächs als Kommunikat

Was ist ein Gespräch genauer? Ein **Gespräch**[9] ist zunächst eine **zeitdirekt dialogische** Kommunikation. Die Kommunikation erfolgt zwischen Akteuren als prinzipiell gleichberechtigten Kommunikationspartnern in der Form von wechselweise vorgebrachten Darlegungen zu einem Inhalt. „zeitdirekt" bedeutet dabei: Die wechselweise vorgebrachten Einzelbeiträge („Gesprächsschritte") folgen zeitlich unmittelbar aufeinander.

Ein **Gespräch** ist dann des weiteren (ad 1) eine Kommunikation, die inhaltlich und äußerlich abgeschlossen ist. Wenn man sich trennt, also äußerlich abschließt, haben die Gesprächspartner in der Regel zum Zeitpunkt der Trennung alles gesagt, was zu sagen ist.

Wenn einem späterhin einfällt, dass man doch das eine oder andere hätte sagen sollen, dann liegt diese Bewusstwerdung nicht mehr vor oder auf dem Zeitpunkt der Trennung. Sie kommt zu spät.

Wenn man immerhin noch während des Gesprächs denkt, dieses oder jenes noch sagen zu müssen, sich aber dagegen entscheidet oder es vergisst, dann hat man zwar rechtzeitig gedacht, dies aber nicht rechtzeitig umgesetzt, so dass man resultativ wiederum nichts anderes gemacht als zum Zeitpunkt der Trennung dem Partner gezeigt zu haben, dass man alles Nötige gesagt hat.[10] Denn das Nötige ist hier das gesprächsöffentlich Nötige, das gezeigte Nötige. Auf das verschwiegene einseitig Nötige kann man sich nicht berufen.

Das Gespräch ist also **lokal geschlossen**.

Das Gespräch besitzt gleichwohl (ad 2) im Bewusstsein der Gesprächspartner aber auch seine Position in der Folge früherer Gespräche und kommender Gespräche. Wenn nicht explizit inhaltlich („ich hab Dir doch letztens erzählt, dass …"), dann doch dadurch, dass ein gemeinsamer inhaltlicher und sozialer Horizont aufgebaut wurde. Und was die Zukunft angeht, so gilt auch hier, auch jenseits des Expliziten („Den Rest erzähl ich Dir beim nächsten Mal"), im gegenwärtigen Kommunizieren der Bedacht der Zukunft: respice finem. Das Gespräch ist also **global offen**.

Ein Gespräch ist (ad 3) per se eine **vollzogene Kommunikation**.

Anders als bei den Textkommunikaten bieten sich der Rezeption keine vorfestgelegten, vielmehr freie und spontane Formulierungen an.

Für Gespräche gelten (ad 4) einige Besonderheiten in folgenden Kategorien:

---

[9] Ausführlicher in „Kommunikationsreihen" (2011:7-76).

[10] Dass es Gesprächsfragmente gibt, in denen die Partner die Gespräche abrupt abbrechen und auseinander laufen, ist davon unbenommen.

Anzahl der Kommunikationspartner,
Rechte der Kommunikationspartner,
mediale Gestaltung,
Gesprächsschritt und Akt als Basiseinheiten und
Umfang der Einheit.

Zur typologischen Reichweite. Die Vielfalt der Gesprächstypen zeigt sich in der Vielfalt der leitenden Zwecksetzungen (Verkaufsgespräche, Beratungsgespräche, Unterrichtsgespräche, Dienstgespräche etc.), um nur eine der Klassifikationskategorien zu nennen. Auch hier gibt es umfangreiche Literatur.

Wir wenden uns nun dem Programmdialog zu, einem Kommunikat, das sowohl mit dem Gespräch als auch mit dem Textkommunikat verwandt ist. Die Verwandtschaft mit dem Gespräch besteht darin, dass das Programmkommunikat zeitdirekt dialogisch ist. Die Verwandtschaft mit dem Textkommunikat ergibt sich daraus, dass Teile des Programmdialogs vorfestgelegte Formulierungen aufweisen. Wie das geht, sollen die beiden nächsten Abschnitte zeigen.

6. Beispiele für Programmdialoge

Um zunächst nur ungefähr zeigen zu können, worum es geht, seien Beispiele für die von mir so genannten Programmdialoge gegeben. Der Lebensbereich der Beispiele ist der automatische Telefonservice eines großen Kinohauses.

Es gibt zwei Beispiele. Was die Authentizität anbelangt, so entstammen diese zwei Beispiele nicht genau zwei Telefonaten bzw. Durchläufen. Denn die Durchläufe sind zu schnell, um sie in einem Zug mitschreiben zu können. Die Durchläufe insgesamt (am 14.11.2012) ergaben aber ein Mitschriftmaterial, das die Konstruktion der beiden Beispiele ermöglichte, derart, dass diese Dialoge tatsächlich jeweils in einem Zug hätten erfolgen können.

Im ersten Beispiel geht es für den Nutzer um zwei Dinge. Zum einen möchte er sich die Filme des Tages seines Anrufs ansagen lassen, zum zweiten möchte er einen Platz für den Kinobesuch reservieren. Der Telefonservice, also die Programmseite in unserem Dialog, erfüllt diesen zweiten Wunsch nach Reservierung nicht, denn für die Wahlmöglichkeit „heute'“ teilt die Stimme mit: „Es gibt keinen weiteren Vorstellungstermin.“

Der Nutzer möchte dann aber für einen anderen Film zu einem anderen Tag reservieren. Darum der zweite Programmdialog. Auch hier wird der Reservierungswunsch nicht erfüllt. Hier zeigt sich aber, dass das Reservierungsprogramm (mindestens für den Zeitraum meiner Anrufe) offensichtlich defekt ist. Es kommt über die Wiederholungen der Akte „heute“ und „Es gibt keine weiteren Vorstellungstermine“ nicht hinaus. Mein `wilder´ Eingriff als Nutzer mit der Wahl „2“ führt dann zwar zur nächsten Abfrageserie, nämlich der nach der gewünschten Uhrzeit des geplanten Filmbesuchs. Da die Abarbeitung dieser Fragen

aber sinnlos ist, weil ja nicht bestimmt ist, um welchen Tag es sich handeln soll, habe ich das Telefonat abgebrochen.

Wunsch nach Programmansage und Reservierung

Telefonprogrammansage und Telefonreservierung der Firma Cineplex Münster	**Nutzer, Interessent, möglicher Kinobesucher**
Telefonservice bereit	NULL
	1] Anruf unter 98 71 23 45
2] Willkommen bei der Telefonreservierung der Cineplex Münster.	
3] Drücken Sie die 1, um eine Vorstellung zu reservieren.	
4] Drücken Sie die 2 für die Programmansage.	
	5] 2
6] Möchten Sie die heutigen Filme angesagt bekommen, so drücken Sie die 2, ansonsten die 3, um einen anderen Tag zu wählen.	
	7] 2
8] Drücken Sie die 1, um einen Film zurückzugehen, die 2, um einen Film zu reservieren, oder die 3, um zum nächsten Film zu gelangen.	
9] Wir spielen heute „Cloud Atlas“ um 20 Uhr 15	
10] „Dredd“ (3D) um 20 Uhr 15	
11] „Argo“ um 17 Uhr 30, um 20 Uhr und um 22 Uhr 45	
12] „Das Schwergewicht“ um 15 Uhr, um 20 Uhr 15 und um 22 Uhr 45	
13] „Harodim – Nichts als die Wahrheit“ um 18 Uhr 30 und um 23 Uhr	
	14] 3
15] „Skyfall“ um 14 Uhr 30, um 16 Uhr 30, um 17 Uhr 15, um 17 Uhr 45, um 19 Uhr 45, um 21 Uhr, um 22 Uhr 45 und um 23 Uhr	
16] Vielleicht lieber morgen um 16 Uhr 20 und um 18 Uhr 30?	

17] „Die Hochzeit unserer dicksten Freundin“ um 23 Uhr	
18] „Die Vermessung der Welt“ in 3D um 17 Uhr 30 und um 20 Uhr 15	
	19] 3
	20] 3
21] „Asterix und Obelix – Im Auftrag Ihrer Majestät“ in 3D um 15 Uhr	
[…]	
	[…]
22] „Gregs Tagebuch – Ich war’s nicht!“ um 14 Uhr 30	
23] Heute laufen keine weiteren Filme.	
24] Die Programmansage ist beendet.	
25] Drücken Sie die 1, um eine Vorstellung zu reservieren.	
26] Drücken Sie die 2 für die Programmansage.	
27] Drücken Sie die 1, um eine Vorstellung zu reservieren.	
28] Drücken Sie die 2 für die Programmansage.	
	29] 1
30] Welchen Film möchten Sie besuchen?	
31] Bestätigen Sie den gewünschten Film mit der 2, ansonsten 1, um einen Film zurückzugehen, oder 3 zum nächsten Film.	
32] „Cloud Atlas“	
	33] 2
34] An welchem Tag möchten Sie den Film sehen?	
35] Bestätigen Sie den gewünschten Tag mit der 2, ansonsten 1, um einen Tag zurückzugehen, oder 3 zum nächsten Tag.	
36] heute	
	37] 2
38] Es gibt keinen weiteren Vorstellungstermin.	
	39] Beendigung des Telefonats
Telefonservice bereit	NULL

Wunsch nach Reservierung

Telefonprogrammansage und Telefonreservierung der Firma Cineplex Münster	**Nutzer, Interessent, möglicher Kinobesucher**
Telefonservice bereit	NULL
	1] 2. Anruf unter 98 71 23 45
2] Willkommen bei der Telefonreservierung der Cineplex Münster.	
3] Drücken Sie die 1, um eine Vorstellung zu reservieren.	
4] Drücken Sie die 2 für die Programmansage.	
	5] 1
6] Welchen Film möchten Sie besuchen?	
7] Bestätigen Sie den gewünschten Film mit der 2, ansonsten 1, um einen Film zurückzugehen, oder 3 zum nächsten Film.	
8] „Cloud Atlas“	
9] „Das Geheimnis der Feenflügel“	
10] „Dredd“ 3D	
11] „Argo“	
	12] 2
13] An welchem Tag möchten Sie den Film sehen?	
14] Bestätigen Sie den gewünschten Tag mit der 2, ansonsten 1, um einen Tag zurückzugehen, oder 3 zum nächsten Tag.	
15] heute	
16] Es gibt keinen weiteren Vorstellungstermin.	
17] heute	
18] Es gibt keinen weiteren Vorstellungstermin.	
	19] 3
20] heute	
21] Es gibt keinen weiteren Vorstellungstermin.	
	22] 2
23] Wann möchten Sie den Film besuchen?	
24] Bestätigen Sie die richtige Uhrzeit mit der 2, ansonsten 1, um zurückzugehen, oder 3 zur nächsten Uhrzeit.	
	25] Beendigung des Telefonats
Telefonservice bereit	NULL

In der Untersuchung von Programmdialogen ist es nicht ohne Interesse, den Erwartungshorizont soweit zu öffnen, dass auch die Programmseite fehlerhaft sein kann. Im vorliegenden Fall sind die Defekte spätestens beim einfachen Nachhaken ersichtlich. Nicht auszuschließen sind aber auch weniger plakative Fehler bei der sogenannten Benutzerführung.

Insofern ist die oben in den Beispielen dokumentierte Begegnung mit einem partiell fehlerhaften Programm lehrreich. Ich möchte aber betonen, dass es mir jetzt nicht um fehlerhafte Programmdialoge geht und mir die fehlerhaften Beispiele, wenn ich so sagen darf, empirisch schlicht zugelaufen sind. Aber sie bieten selbst in der Fehlerhaftigkeit der Programmseite Anschauung genug, um zum Kern der Sache kommen zu können: zu den Programmdialogen, wie sie ihre Schöpfer, die Unternehmen bzw. Organisationen und die umsetzenden Mitarbeiter, selbst anbieten und durchführen, und wie sie mittlerweile von sehr vielen Kommunikationspartnern bzw. Akteuren genutzt werden.

Die beiden Programmdialoge bilden eine Reihe:[11]

1. Wunsch nach Programmansage und Reservierung
2. Wunsch nach Reservierung

7. Bestimmung des Programmdialogs als Kommunikat

Bevor wir den Programmdialog einordnen und näher bestimmen, sind drei Fragen grundsätzlicherer Natur zu stellen. Die weitestreichende Frage zielt auf unsere bisher naiv vorgetragene Annahme, dass es sich beim Programmdialog um Kommunikation handelt. Die beiden anderen Fragen setzen eine hierauf positive Antwort voraus und betreffen die sortenbezogene Einordnung des Programmdialogs.

Bei der Einordnung der skizzierten Kommunikationsformen bzw. Kommunikationseinheiten stellen sich also folgende Fragen.

1. Liegt tatsächlich Kommunikation vor?
2. Wenn Kommunikation vorliegt, handelt es sich dann um ein Gespräch?
3. Wenn Kommunikation vorliegt, es sich aber nicht um ein Gespräch handelt, liegen dann Textkommunikate vor?

Zur Frage der Kommunikation. In der Richtung Unternehmen an Kunden liegt Kommunikation vor, da ein sprachlich formulierter Sachverhalt mitgeteilt und rezipiert wird. Dass es sich um ein Unternehmen handelt und um

[11] Der zweite Programmdialog auf Beispielebene war wie angedeutet gar nicht geplant und ergab sich auf Grund fehlender Information beim ersten Anruf. Das Fehlen der Information erwies sich dann später als Folge eines Programmdefekts. De facto ergab sich auf der Ebene des empirischen Zugriffs eine Arbeit nicht ohne Sysiphuscharakter für den der Programmseite hingegeben lauschenden Analysator. Mithin: Es ergab sich de facto eine umfangreichere Reihe mit vielen Programmdialogfragmenten.

einen anonymen Formulierer, schränkt das nicht ein, da wir auch beim Textkommunikat die Möglichkeit der anonymen bzw. kollektiven Autorschaft haben, wir also auch etwa die nachrichtenagenturgeborenen Nachrichten auf der 1. Zeitungsseite als Kommunikation akzeptieren.

Anders liegt der Fall in der Frage der Nutzerantwort. Hier könnte es Nutzer geben, deren Einstellung so beschaffen ist, dass sie, wenn sie beispielsweise auf Geheiß der Vorgabe die Art der Auswahl treffen, etwa die Zahl 2 auswählen und dann drücken, dass sie dies also nicht im Bewusstsein tun, mit einem Partner in kommunikativer Verbindung zu stehen, also nicht im Kommunikationsbewusstsein, sondern im Bedienungsbewusstsein, also im Bewusstsein, ein Gerät zu bedienen, wie etwa den heimischen Staubsauger.

Und was tatsächlich den Staubsauger angeht, so liegt bei seiner Bedienung gewiss keine Kommunikation vor. Denn welches sollte die Person sein, die das Schalten und Einstellen am Staubsauger empfängt, oder welches sollte das Personenkollektiv sein, das als Partner die Gerätesteuerungen des Staubsaugernutzers rezipiert. In der Tat: Es wird keine andere Person oder Personengruppe erreicht, alle Schaltungen und Einstellungen und auch das Saugen selbst bleiben im oder am Gerät, einem Gegenstand, dessen ausschließliche Häuslichkeit zu behaupten keine Schwierigkeiten bereitet, jedenfalls bis zur unmittelbaren Gegenwart.

Im Fall des Telefonats, bei dem sich ein Anrufer in das Programm eines Unternehmens begibt, dürfte jedoch das Kommunikationsbewusstsein des Anrufers beim Reagieren (Drücken von Zahlen etc.) der Regelfall sein: Seine auf Rahmen- und Zeitpunkt-Geheiß des Programms erfolgenden Reaktionen dürfte er kommunikativ meinen und als Produktionen verstehen, wobei das Unternehmen diese als solche auch akzeptiert und mithin rezipiert:

> da es eine gesprächsimitierende Stimme ist, die führt;
>
> da die Zahlen, die man drückt, wohl auf die eine oder andere Weise beim Unternehmen aufgenommen werden mindestens zur Beobachtung des Marktes, was auch dem Nutzer nicht fremd sein dürfte;
>
> da man beim Reservieren mit dem Unternehmen nicht nur unverbindlich in Kontakt kommt, sondern eine Reservierungsvereinbarung eingeht;
>
> da die andere Seite des Kontakts nicht mehr im eigenen, überschaubaren Bereich liegt (wie etwa der Staubsauger vor mir), sondern mir entzogen ist und
>
> da durchaus damit zu rechnen ist, dass der automatische Service aus welchen Gründen auch immer unterbrochen wird und sich ein Mitarbeiter einschaltet.

Unter den Bedingungen also, unter denen auch auf der Nutzerseite Kommunikation mit dem Unternehmen als Adressat unterstellt werden kann, und genau auch eben für diesen Nutzertyp spreche ich für beide zusammen von einer „Programm**kommunikation**“. (Für den anderen Nutzertyp, den Typ mit Bedienungsbewusstsein, ergibt sich, dass das Telefonat kommunikativ gesehen nur eine Textkommunikation darstellt, mit dem Anrufer lediglich als Rezipient).

Zur Frage eines möglichen Gesprächsstatus. Die Programmkommunikation ist keine Gesprächskommunikation. Mit der Gesprächskommunikation teilt sie zwar die zeitdirekte Dialogizität. Aber die Programmseite (die Unternehmensseite) ist vorformuliert ohne Spielräume dafür, wann und welche Formulierungen abzuspielen sind; und die Nutzerseite kann zwar, wenn das Programm es vorsieht, ihre Wünsche anbringen, aber nur so, dass der vorgegebene Formulierungs- und Zeitpunktrahmen beachtet wird.

Zur Frage eines möglichen Textkommunikatstatus. Die Programmkommunikation ist keine Textkommunikation. Zwar ist die Programmseite (die Unternehmensseite) vollständig vorformuliert ohne die genannten Spielräume. Darin gleicht die Programmkommunikation der Textkommunikation. Die Nutzerseite aber hat dagegen wie bemerkt eine gewisse Wunschfreiheit. Zentral gilt aber der Unterschied in der Kommunikationsrichtung ‚dialogisch – monologisch’: Die Programmkommunikation ist in sich dialogisch, genauer: zeitdirekt dialogisch, im Gegensatz zur unbedingten Monologizität, die der Textkommunikation eignet.

Bestimmung des Programmdialogs: der Programmdialog als Kommunikatform der Programmkommunikation. Die **Programmkommunikation** ist eine **asymmetrische, zeitdirekt dialogische sprachliche Kommunikation**. Es gibt zwei Partnerseiten: die Programmseite und die Nutzerseite. Die **Programmseite** ist in der Regel ein Kollektiv (Unternehmen oder Organisation einschließlich des oder der Mitarbeiter, die das Kommunikationsdesign, die `Kommunikationsoberfläche´, herstellen). Die **Nutzerseite** ist in der Regel eine einzelne Person.

Die Programmseite äußert sich in **vorfestgelegten** Formulierungen, diese dabei ohne Spielräume dafür, wann und welche Formulierungen zu emittieren sind. Die Nutzerseite äußert sich jeweils mit der aktuellen **Selbstfestlegung** gemäß eigenem Wunsch, muss diese Selbstfestlegung aber ausdrücken als eine Position innerhalb des jeweils angebotenen Auswahlrahmens.

Die Programmkommunikation ist per se immer schon **vollzogene Kommunikation**. Der Text, den der Kommunikationsdesigner für den Telefonservice einsetzt, ist als solcher noch keine Kommunikation. Erst, wenn dieser Text als Programmstimme abläuft und von einem Anrufer gehört und ggf. beantwortet wird, wird er zum Teil der Kommunikation. Wenn also ein Nutzer anruft und sich einlässt, wird Kommunikation vollzogen. Da wir Programmkommunikation bestimmt haben als Gemeinschaftsleistung von Programmseite und Nutzerseite, ist sie, wie gesagt, jedes Mal vollzogene Kommunikation.

Es gibt **lokal geschlossene**, also mit Anfang und Ende versehene Programmkommunikationen, die für sich selbst genug sind, für die Situation des Kommunizierens. Den zeitlichen Anfang einer solchen Programmkommunikation bestimmt der Nutzer. Die Frage der Beendigung ist nicht undifferenziert zu beantworten. Wir nehmen als Beendigung den Zeitpunkt, den der Nutzer wählt, ob er nun ein Programm aus Sicht der Programmseite abgearbeitet hat (etwa einen Kaufvertrag abgeschlossen hat) oder nicht. Wird der Nutzer vom Programm `rausgeworfen´, nehmen wir ein Fragment an.

Lokal geschlossene Programmkommunikationen sind, wenn wir über den Tellerrand der lokalen Geschlossenheit in das zeitliche Davor und Danach schauen, in der Regel gleichzeitig auch **global offen**. Der dienende Charakter der Programmkommunikationen führt sie zu einer Leistung hin (etwa zu einem Kauf oder einer Reservierung oder einem anschließenden Gespräch), oder sie birgt eine Leistung in sich, etwa ein Textkommunikat wie bei Wikipedia oder die Möglichkeit, auf einer Stadtwebseite durch eine auf Dauer gestellte Kamera sehen zu können, die eine Straßenansicht zeigt, wie sie sich im Moment ergibt, vulgo: live und online. Mindestens aber stehen sie sowohl programmseitig als auch nutzerseitig in der Folge vorangegangener und noch kommender Kommunikationen.

Mithin gelten als Kriterien:

ad 1: vollzogene Kommunikation
ad 2: lokal geschlossene Kommunikation
ad 3: global offene Kommunikation

Da diese Kriterien vorliegen, werten wir eine solche Programmkommunikationseinheit als Kommunikat und sprechen vom „Programmkommunikat“ oder, in der Kommunikationsrichtung spezifischer, vom **„Programmdialog“**.

Programmdialoge sind also **lokal geschlossene** und **global offene Programmkommunikationen**.

Hinzu kommen, teils wie schon angedeutet, Besonderheiten in folgenden Kategorien:

Anzahl der Kommunikationspartner,
Rechte der Kommunikationspartner,
mediale Gestaltung,
Dialogschritt und Akt als die Basiseinheiten und
Umfang der Einheit,

wobei diese Besonderheiten andernorts noch zu beschreiben wären.

Was den Einsatz der Programmkommunikation durch die Unternehmen bzw. Organisationen angeht, so ist festzuhalten, dass das, was der Nutzer von der Programmseite aktuell in seinen Programmdialogen sieht oder hört, im Hinblick

auf den Gesamtzusammenhang, in dem die Programmseite vom Unternehmen geschaffen wurde, nur eine **Oberfläche** darstellt.

Was die eigentlichen Überlegungen und Motivationen in den Unternehmen anbelangt, die zu der Programmseite führen, die dann tatsächlich nach außen hin in Erscheinung tritt, so bleiben diese dem Nutzer am Telefon oder vor dem Monitor verborgen. Ob also ein Unternehmen etwa Daten sammelt, diese auswertet und die Auswertung zu zielgenaueren Platzierungen nutzt oder ob ein Unternehmen Objektivität vorgibt, de facto aber bestimmte Produkte unter Bruch der Objektivität verdeckt bevorzugt, ist an der Oberfläche der Programmseite nicht zu erkennen, dieser ja auch ggf. gewiss mit Fleiß entzogen.

Allerdings gilt diese Separierung von Oberfläche und Hintergrund sehr viel genereller, weit über die telefonische Programmkommunikation hinaus. Auch in vielen Bereichen, in denen `echte´ Gespräche geführt und Emails geschickt werden, etwa im Bankensektor, ist der Hintergrund für den einfachen Kunden doch recht tief und intransparent.

Zur typologischen Reichweite. Programmdialoge werden nicht nur übers Telefon geführt. Sehr häufig bieten sich Gelegenheiten im Internet. So ist die Benutzung von Google, Amazon, Wikipedia, von Webseiten von Unternehmen und Organisationen, von Städten und Regionen, von Spiegel online und anderen publizistischen Unternehmen etc., aber bereits auch schon die Benutzung des eigenen Emailservice und nicht zuletzt die des Bankomaten, dessen kommunikative Anbindung an die Bank sicherlich sehr sorgfältig gehandhabt wird, jeweils ein Programmdialog, wobei Programmdialoge auch andere Programmdialoge sowie Textkommunikate und Gespräche einbetten können.

Als Beispiel für einen Internetprogrammdialog sei das Aufsuchen von Wikipedia-Artikeln unterbreitet:

Wikipedia Programmdialog Teil I (1-3),
Wikipedia Programmdialog Teil II (4-8) und gleichzeitig Artikel „Karl-Heinz Best“ (4, 6, 8),
Wikipedia Programmdialog Teil III (9),
Wikipedia Programmdialog Teil IV (10-13) und gleichzeitig Artikel „Quantitative Linguistik“ (10, 12, 14)

Adresszeile bereit	NULL
	1]Aufruf der Webseite „Wikipedia“ durch http://wikipedia.de
2] Webseite „Wikipedia“ erscheint mit Suchfläche	
	3] Sucheintrag „Karl-Heinz Best“ in die Suchfläche

4] Anfang des Artikels „Karl-Heinz Best“ erscheint einschließlich Suchfläche	Rezeption des sichtbaren Textteils, genaue Lektüre
	5] Scrollen bis zum 2. Bildschirmabschnitt
6] 2. Bildschirmabschnitt erscheint	Genaue Lektüre
	7] Scrollen bis zum 3. Bildschirmabschnitt
8] 3. Bildschirmabschnitt erscheint	Genaue Lektüre bis zum Artikelende
	9] Sucheintrag „Quantitative Linguistik“ in die Suchfläche
10] Anfang des Artikels „Quantitative Linguistik“ erscheint einschließlich Suchfläche	Rezeption des sichtbaren Textteils, genaue Lektüre
	11] Scrollen bis zum 2. Bildschirmabschnitt
12] 2. Bildschirmabschnitt erscheint	Genaue Lektüre
	[…] jeweils Scrollen[12]
[…] jeweiliger Bildschirmabschnitt	jeweils genaue Lektüre
	13] Scrollen bis zum 9. Bildschirmabschnitt
14] 9. Bildschirmabschnitt erscheint	Genaue Lektüre bis zum Artikelende
	15] Beenden durch Webseitenwechsel
Neue Webseite erscheint	NULL

Die breitere Umrandungslinie markiert jeweils die in den Programmdialog eingebetteten Textkommunikate.

Wir gestatten uns abschnittsabschließend eine eher übergreifende Bemerkung:

Der **Ökonomisierungsdruck** hat im Verein mit und unter Ausnutzung von Digitalisierungstechnologien die Programmkommunikation hervorgebracht. Im Wirtschaftssektor, in der staatlichen Administration, aber auch in der Kultur erlaubt das Verfahren der Programmkommunikation die u.a. personalsparende, überhaupt aber eben die ökonomische und vielleicht auch die machteffektivere Steuerung von bis zu Millionen zählenden Publika. Man kann vielleicht sagen, dass der Nutzer, der Kunde, der Leser, der Bürger etc. schon einen bedeutenden kommunikativen `Domestizierungsprozess´ hinter sich hat einschließlich des Erwerbs vieler Bedienungsbefähigungen. Gewiss ist damit das Serviceangebot an ihn gewaltig gewachsen. Der Saldo allerdings aus kommunikativer Selbstunterwerfung und Serviceentgelt (dieser Saldo im Übrigen auch eine Komponente der

[12] Die nicht aufgeführten Zeilen zähle ich hier der Einfachheit halber nicht mit.

Lebensqualität) wird nicht so schnell offenbar werden wie beim Trojanischen Pferd. Dieser Saldo offenbarte sich ja schon über Nacht.

Verwandte mit der Programmkommunikation gibt es wohl seit langem, à part bemerkt vielleicht, seit es die spielerische Vorgabe gibt: „die rechte oder die linke Hand?“, oder die nicht ganz so spielerische: „Ja oder Nein!“. Sicherlich aber gehört die Abarbeitung von Listen an Personen dazu qua dialogischer Vorgabekommunikation, eben in der ökonomischen Kommunikationsform „1: viele“ („einer, der mit der Liste, erfasst viele“) (Zugehörigkeiten, Steuern, Besitz, Einstellung etc.). Der Unterschied ist dann, wer zu wem kommt und warum, und wie das Machtprofil aussieht. Hierher gehören wohl auch düstere Kommunikationseinbettungen.

Auch das **Prüfungsgespräch** gehört zu den Verwandten der Programmkommunikation. Wir haben hier eine Vorgabeseite und eine Seite, die den Vorgaben folgen muss, gewiss dabei mit großen Spielräumen beiderseits, deren Ausnutzung im Übrigen manchmal zu wunderbaren Gesprächen führen kann. Die Vorgabeseite hat eine Bewertungspflicht, was die Asymmetrie erhöht.

Weiterhin kann man auf Grund der Erhebungstechnik das **explorative Interview** zu den Verwandten des Programmdialogs zählen, da diese Form von Interview nach einem Frageleitfaden auf der Seite des Interviewers vorstrukturiert ist. In gewisser Weise analog dazu ist dann auch das **journalistische Interview**, wenn sich der Interviewer einen Satz von Fragen zur Abarbeitung vorgenommen hat.

Auch eine Art von Programmkommunikation haben wir in der Satire „**Buchbinder Wanninger**“ von Karl Valentin.[13] Der Buchbinder trifft am Telefon nicht gleich auf den zuständigen Mitarbeiter der Firma, sondern wird jeweils abgewimmelt und durchgereicht und, nun ja, verliert nicht nur die syntaktische Souveränität. Die Kette derjenigen, die durchgereicht haben, ist die Vorgabeseite. Der Buchbinder ist die Seite, die folgen muss. Bis der Buchbinder endlich das große Los zieht und den zuständigen Mitarbeiter am Telefon hat – nun ja, zum kurzen Glück. Der Mitarbeiter macht Feierabend.

Der Programmdialog liegt überkreuz mit beiden, überkreuz mit dem Gespräch und überkreuz mit dem Textkommunikat. Es gibt nun ein weiteres Kommunikat, das ebenfalls mit beiden überkreuz liegt. Allerdings spezifisch anders, wie der folgende Abschnitt zeigen möchte.

8. Zum Extemporemonolog als Kommunikat

Wenn jemand, etwa bei einer Feier, aufsteht und spontan eine Rede hält, dann ist diese nicht ganz ungewöhnliche Kommunikation, obzwar monologisches Sprechen, **keine Textkommunikation**. Denn die Formulierungen sind nicht vorfest-

[13] Eine mündliche Anregung von Udo Kipper; vgl. auch „Kommunikationsreihen“ (2011:386-390).

gelegt. Sie sind, wie gesagt, spontan. Der Redner spricht aus dem Stegreif, wie die überkommene Redewendung sagt, das heißt eben, aus dem Steigbügel, und steigt zum Sprechen erst gar nicht vom Pferd. Er ist der Redner ohne Vorbereitung, aus dem aktuellen Moment heraus formulierend, aus dem Augenblick, „ex tempore".

Die spontane Rede ist aber auch **keine Gesprächskommunikation**. Sie ist zwar spontan, also frei formuliert. Aber sie ist eben eine Rede, und kein Dialog.

Kurz: Die spontane Rede ist **monologisch** und **frei formuliert**.

Eine solche Rede, die per normalem Verständnis des Wortes „Rede" **lokal abgeschlossen** ist mit Anfang und Ende, die **global offen** ist mit ihren möglichen inhaltlichen Bezügen in die Vergangenheit und die Zukunft und überdies eine Position hat in einer Folge von Kommunikationen in der Bezugsgruppe, und die notwendig eine **vollzogene Kommunikation** darstellt, eine solche Rede also terminologisieren wir als „Extemporekommunikat" oder, spezifischer in Sachen Kommunikationsrichtung, als „**Extemporemonolog**".

Wir haben leider keine authentischen Beispiele für Extemporemonologe, derart, dass wir sie per Transkription vorlegen könnten. Aber es sei aus zwei Berichten aus dem amerikanischen Wahlkampf 2012 zitiert, die nicht nur zeigen oder mindestens nahelegen, dass es Extemporemonologe gibt, sondern sie auch näher charakterisieren.

Zunächst aus dem Artikel „Dirty Joe" (sc. Joe Biden, amtierender amerikanischer Vizepräsident) der Süddeutschen Zeitung, platziert als Nachricht auf der ersten Seite:

> […] Redner zur besten Sendezeit sind er selbst, sein Vize Joe Biden und der frühere Präsident Bill Clinton. Beide sind beliebt, auch deswegen, weil sie gern drauflosreden. ‚Loose cannon' sagt man im Politjargon, wenn jemand der verbale Colt locker sitzt. Immerhin sorgt das für etwas Spannung im Wahlkampf – so wie der Auftritt der republikanischen Wunderwaffe Clint Eastwood, dessen Einlage beim Parteitag der Republikaner vergangene Woche die Wahlkampfstrategen verwirrte. […]
>
> Risikofaktor Nummer zwei [für den amtierenden Präsidenten] ist Vize-Präsident Joe Biden. Als Obama ihn vor vier Jahren auswählte, rätselten seine Berater, wie sie Biden bis zur Wahl so eindämmen könnten, dass keine Katastrophe passiere. Kürzlich in Virginia erklärte er, Romney werde die Börse von der Leine lassen und ‚euch wieder in Ketten legen'. Er sprach vor Hunderten Schwarzen. Obamas Wahlkampfmanager waren tagelang beschäftigt, den schiefen Vergleich zu erklären. Beim Parteitag soll Biden vom Teleprompter ablesen. (Süddeutsche Zeitung, 3. Sept. 2012, S. 1)

Dann gibt es eine Art verallgemeinernde Begleitbetrachtung, 3 Tage später, auf der zuständigen Seite 2 der Süddeutschen Zeitung:

> Joe Biden ist keiner, der seine Worte auf die Goldwaage legt. Er redet von der Leber weg, ohne Teleprompter, meist frei nun ohne Skript. Das geht manchmal schief. Vor einem Jahr etwa verglich der Demokrat rechte Tea-Party-Sympathisanten im Kongress mit ‚Terroristen', und erst vor zwei Wochen warnte er, der republikanische Herausforderer wolle das Volk ‚wieder in Ketten legen'. Das hat dem 69-jährigen Vize-Präsidenten den Ruf eingebracht, gleichsam eine verbale Schrotflinte zu sein. (Süddeutsche Zeitung, 6. Sept. 2012, S. 2)

Mithin: der Extemporemonolog ist keine Angelegenheit allein der spontanen stilistischen sprachlichen Gestaltung. Im vorliegenden Fall ist er vielmehr vor allem eine Angelegenheit der Inhalte. Eine allzu sprechsprachliche Syntax und eine teils hastige Lexemwahl, vorausgesetzt, sie bleibt innerhalb der roten Thema- und Beurteilungslinien, wären vielleicht nur ein Hauch von einem Nebenproblem. Aber es wäre kein Hauptproblem, keine „loose cannon", keine über Deck rollende Kanone, die den Gegner zum Daumen drückenden Zuschauer machte.

Der Extemporemonolog als Rede ist vor zeitdirekten dialogischen Einsprechungen geschützt, wenn er, wie in den vorzitierten Bereichen, in formalisierter Kommunikationsweise abläuft. Das heißt: Der Redner hat die alleinige Gewalt über das Mikrophon, und das Publikum kann sich nur durch kollektive, sprachlich nicht im Einzelnen pointierte Regungen melden.

Auf Übergangsmöglichkeiten ist hinzuweisen. Der Extemporemonolog kann Formulierungspartien enthalten, die vorfestgelegt sind, sei es, dass diese aus des Redners Gedächtnis stammen, wenn sie ob häufiger Wiederholungen gedächtnispräsent sind, sei es, dass sie Notizen entnommen sind, die der Redner auf Papier oder sonstwie präsentiert vor sich hat.

Eine weitere Übergangsmöglichkeit: Bei einer eher zahlenmäßig überschaubaren und im Prinzip informell verlaufenden Feier kann der Ausflug in den Extemporemonolog leicht in ein Gespräch münden, so dass wir dann insgesamt eher ein Gespräch denn einen Monolog hätten, oder eine zeithybride Einheit mit einem Erst- und einem Zweitabschnitt.

Der Absicherung halber hinzuzufügen ist, dass es nach unserer Bestimmung keinen Extemporemonolog ohne Hörerschaft gibt. Der Extemporemonolog ist vollzogene Kommunikation. Das Reden vor sich selbst und mit sich selbst ist keine Kommunikation.

Im Übrigen: Der klassische Bühnenmonolog, wenn ohne heimlich erwünschten und tatsächlich geneigten Lauscher auf der Bühne, dieser Monolog ist auf und innerhalb der Bühne keine Kommunikation im Sinne einer genuinen Kommunikation, denn er ist keine Kommunikation mit den dramatis personis, mit denen die Monologfigur zu tun hat.

Aber der Bühnenmonolog spaltet quasi märchenhaft oder eben poetisch (und routiniert) von sich ab einen gleichlautenden, gleichzeitigen, praktisch ununterscheidbaren Monolog für eine Zauberkommunikation ins Publikum, mit dem Publikum, das ob seiner Neugierde, auch seiner Lauschneugierde, nichts weniger als sehr geneigt ist.

Auch schriftliche Formen des Extemporemonologs sind vorfindlich. Wenn etwa jemand seine flüchtigen, spontanen, unvorbereiteten, kurz: seine wilden Skizzen, ohne sie nochmals korrigierend durchzusehen, aus der Hand gibt an jemanden anderen zur Lektüre, dann ist das eben ein solcher Extemporemonolog, allerdings nur, wenn der spontane Schreiber es selbst ist, der seine Skizzen aus der Hand gibt zur erfolgenden Lektüre.

9. Kommunikate

Wir haben in unseren bisherigen Ausführungen vier Kommunikate, also situationsabgeschlossene Kommunikationseinheiten, vorgestellt. Es sind dies:

das Gespräch
das Textkommunikat
der Programmdialog
der Extemporemonolog

Diese vier Kommunikate lassen sich nach zwei Kategorien mit jeweiliger zweifacher Ausprägung ordnen. Diese Kategorien sind:

die Kommunikationsrichtung
- zeitdirekt dialogisch
- monologisch

der Freiheitsgrad der Kommunikation
- frei (spontan)
- teils oder vollständig vorformuliert

Damit ergibt sich für diese vier Kommunikate folgende Systematik:

Kommunikationsrichtung / Freiheitsgrad	zeitdirekt dialogisch	monologisch
frei (spontan)	**Gespräch**	**Extemporemonolog**
teils oder vollständig vorformuliert	**Programmdialog**	**Textkommunikat**

Gespräch und **Textkommunikat** habe ich bereits, wie oben angemerkt, in „Kommunikationsreihen“ (2011) vorgestellt. Dabei konnte ich mich dankbar auf

eine Fülle an Arbeiten in den ihnen zugeordneten Spezialdisziplinen der Gesprächsanalyse und der Textwissenschaft stützen.

Ich habe dann in „Kommunikationsreihen“ (2011) keine weiteren Kommunikate angenommen, bin also damals innerhalb der Zweiteilung der Forschungslandschaft in Gesprächsanalyse und Textwissenschaft geblieben.

Die Aufnahme des **Programmdialogs** und des **Extemporemonologs** nun ist gegenüber „Kommunikationsreihen“ (2011) neu.

Anstoß gab die nähere Betrachtung der Servicekommunikation per Telefon und per Internet. Sie ließ die Eigenständigkeit des **Programmdialogs** als Kommunikat erkennen. Denn dieser Dialog war wie ausgeführt weder Gespräch noch Textkommunikat.

Was den **Extemporemonolog** angeht, so zeigte er sich zunächst der Neugier deduktiver systematischer Betrachtungen des Tabellenplatzes rechts oben.

Verändert hat sich damit die Auffassung vom Verhältnis zwischen der Kategorie der Kommunikationsrichtung und der Kategorie des Freiheitsgrades. Das Verhältnis ist also nicht eine Äquivalenz zwischen den Ausprägungen (Es gilt also nicht: ‚frei ~ zeitdirekt dialogisch’ und ‚vorformuliert ~ monologisch’). Vielmehr sind die beiden Kategorien, wie in der obigen Tabelle festgehalten, **kreuzklassifiziert**.

Was nun die Aktualität der Ansätze angeht: Der systematische, kommunikationstheoretische Ansatz des Programmdialogs und des Extemporemonologs ist, soweit ich sehe, in dieser Pointierung in der Literatur neu.

In einem weiteren Schritt wollen wir die Ausprägung „teils oder vollständig vorformuliert“ in die Ausprägungen „**teils vorformuliert**“ und „**vollständig vorformuliert**“ ausdifferenzieren und nach weiteren Kommunikaten fragen.

Die folgende Tabelle gibt diese Ausdifferenzierungen wieder und führt im Kommunikatfeld Ergänzungen auf:

Kommunikationsrichtung Freiheitsgrad	zeitdirekt dialogisch	monologisch
frei (spontan)	**Gespräch**	**Extemporemonolog**
teils vorformuliert	Typ A: Programmseite: vorformuliert Nutzerseite: in Grenzen frei **Programmdialog**	

	Typ B: Programmseite: in Grenzen frei Nutzerseite: vorformuliert **Nachsprechdialog (?)**	
vollständig vorformuliert	**Zeremonieller Dialog**	**Textkommunikat**

Die Dialogrichtung im Einzelnen:

Gespräch: wie besprochen
Programmdialog (zeitdirekt dialogisch, teils vorformuliert, Typ A): wie besprochen
Nachsprechdialog (?) (zeitdirekt dialogisch, teils vorformuliert, Typ B): s.u.
Zeremonieller Dialog (zeitdirekt dialogisch, vollständig vorformuliert): s.u.

Die Monologrichtung im Einzelnen:

Extemporemonolog: wie besprochen
Textkommunikat: wie besprochen

Zum provisorischen Ansatz des von uns so genannten Nachsprechdialogs. Dieser Ansatz steht unter einem prinzipiell weitläufiger zu besprechenden möglichen Vorbehalt, einem Vorbehalt nämlich, dass hier keine Inhalte verbreitet werden, sondern nur Fertigkeiten eingeübt werden. Beispiele, an denen entlang man diskutieren kann, sind die folgenden: Ein Sprach- oder auch ein Rhetoriklehrer gibt von ihm frei zu wählende Formulierungen vor, die der Lernende jeweils nachsprechen muss. Auch der zeitdirekte Dialog, der zwischen Mutter und Kleinkind stattfindet, mit der Mutter, die vorspricht, und dem Kleinkind, das nachspricht, wäre hier einzuordnen. Immerhin kann man argumentieren, dass der Vorsprechende auch Bedeutungen vermitteln kann und dass das Nachsprechen nicht nur ein rein phonetisches Imitieren sein muss, sondern auch tatsächliche Rezeption sein kann.

Zum zeremoniellen Dialog. Um ein Beispiel aus dem Bereich religiöser Kommunikation zu nennen: Der Vorbeter gibt vor, und die Gemeinde antwortet, und jeder der beiden Seiten trägt vor nach jeweils vorfestgelegter Formulierung. Was den Verbreitungsvorbehalt, das heißt den Vorbehalt, dass de facto keine in der Sache neuen Inhalte verbreitet werden, angeht, so gilt doch: Die per Wiederholung `mittransportierten´ Erneuerungen der Selbstverpflichtung, der Dank

barkeit gegenüber Gott und der Segenswünsche an den Kommunikationspartner sind eine sprachliche Kommunikation.

10. Bestimmung der Reihe als Folge von Kommunikaten

Die Kommunikate sind, wie beschrieben und beobachtet, situationsabgeschlossene Kommunikationseinheiten. Aber es gibt, so sicher wie trivial, im Zeitverlauf eben nicht nur eine einzelne Rede- oder Schreibsituation, mithin nicht nur das eine einzige und alleinstehende Kommunikat, ohne Vorgängerkommunikat und ohne Nachfolgerkommunikat, sondern deren viele auf dem partnergemeinsamen Weg zweier Akteure oder vieler Akteure in den verschiedenen Sektoren einer Gesellschaft oder einer Gesellschaftengruppe.

Wir haben oben schon andeutend solche zeit- und sozialgebundenen Ensembles von Kommunikaten als Reihen angesprochen und eine Reihe flüchtig umrissen als sozial kohärente Folge von Kommunikaten. Wir stellen nun eine genauere Bestimmung der Reihe hierher. Die Bestimmung lautet:[14]

(1) Eine **Reihe 1. Stufe** (oder auch: **„Basisreihe**“) ist dadurch bestimmt,
dass sie eine nach zeitlicher Sukzession und zeitlicher Parallelität geordnete Menge von Elementen, kurz: eine Folge[15] von Elementen ist derart,

(1.1) dass ein einzelnes dieser Elemente entweder ein Kommunikat oder ein Kommunikatsegment oder ein Kommunikatsegmentkomplex ist,

(a) wobei ein Kommunikatsegment einen Akt oder eine Sequenz von Akten innerhalb eines Kommunikats umfasst und

(b) ein Kommunikatsegmentkomplex aus mehreren Segmenten ein und desselben Kommunikats gebildet wird,

(1.2) dass sie folgenden Zwecksetzungen entspricht:

(a) den Zwecksetzungen der sie durchführenden Akteur- bzw. Kommunikationspartnergruppe A,

(b) ggf. auch den Zwecksetzungen einer institutionell übergeordneten Akteur- bzw. Kommunikationspartnergruppe B (B umfasst A) und/oder den Zwecksetzungen einer institutionell nachgeordneten Akteurgruppe C (C wird umfasst von A),
und dementsprechende Funktionen und dementsprechende Propositionenstrukturen enthält und
(1.3) dass sie auch Beziehungen zu parallel stattfindenden kommunikationsbezo-

[14] Teils in Verbesserung von „Kommunikationsreihen“ (2011:146f.).
[15] Unter Verwendung von „Folge“ in umgangssprachlicher Bedeutung (z.B.: „Das zog eine Folge von Ereignissen nach sich“.

genen Separata[16] und kommunikativ-physischen Handlungshybriden[17] aufweisen kann.

(1.4) Wir nennen die Elemente einer Reihe auch „Segmente der Reihe“ oder „Schritte der Reihe“.

(2) Eine **Reihe n-ter Stufe** (n > 1) enthält in ihrer Elementfolge eine oder mehrere Reihen der Stufen (n-1), ggf. auch Reihen niedrigerer Stufen (jeweils nach der intern höchsten Stufe eingeordnet)[18] und ggf. einzelne Kommunikate, Kommunikatsegmente und Kommunikatsegmentkomplexe. Diese sind unmittelbare Reihensegmente von R. Die Eigenschaften (1.2) und (1.3) der Reihe 1. Stufe gelten analog sowie die Benennungsweise (1.4).

(3) Eine **Reihe** ist eine Reihe 1. oder n-ter Stufe (n > 1).

---

[16] Separata sind für sich vollzogene Reflexionen, deren Struktur zum Teil kommunikat- und reihenähnliche Strukturen aufweisen (Selbstgespräche etc.).

[17] Ein kommunikativ-physisches Handlungshybrid sei am Beispiel verdeutlicht: Der Schiedsrichter auf dem Fußballfeld zeigt auf den Elfmeterpunkt. Das ist eine reine Kommunikation, weil ein jeder zeigen kann: Die Physis des Zeigens, die als solche ja durchaus gegeben ist, verschwindet völlig hinter der kommunikativen Funktion, wird als Begleitmoment in der Regel gar nicht erst bewusst wahrgenommen, wie das entsprechend auch und überhaupt für das Sprechen, Hören, Lesen, Schreiben gilt.

Was vom Spieler, an den die Ausführung des Elfmeters dann delegiert wurde, verlangt wird, ist nicht nur zu vermitteln, dass er auf den (nichtsprachlichen) direktiven Akt des Schiedsrichters gehorchend reagiert (die kommunikative Komponente). Von ihm wird gleichzeitig verlangt, dass er den Schuss professionell ausführt, professionell aus der Perspektive des Schiedsrichters und professionell aus der Perspektive seiner Mannschaft. Es wird gleichzeitig also eine besondere physische Leistung verlangt.

Der Schuss ist damit mehr als Kommunikation, das heißt, mehr als die Übermittlung eines Inhalts (wobei die Übermittlung des Inhalts als solche ja auch, gleichzeitig, stattfindet). Er ist mehr, weil das physische Mehr samt Inszenierung geistig-körperlicher Kunstfertigkeit konventionell bewusst wahrgenommen und konventionell als eigenständige Komponente der Handlung gewertet wird. Die Zuschauer selbst sind am Mehr interessiert und nicht am für sie Selbstverständlichen, nämlich, dass der Spieler dem Schiedsrichter und dem Publikum erfolgreich Gehorsam vermitteln konnte.

(Die Dinge lägen natürlich anders in einer Kommunikationsgemeinschaft, in der etwa die Antwort „ja“ konventionell durch einen Elfmeterschuss zu geben und jeder dazu auch in der Lage wäre. Bei Licht besehen wäre das allerdings etwas umständlich, da man immer einen Ball und ein Tor bei sich haben müsste und bei puristischer Regelauslegung auch einen Torhüter.)

[18] Ein Beispiel: Eine Reihe R der Stufe n-3 ist nur dann *unmittelbares* Segment der Reihe der Stufe n, wenn die Reihe R nicht in einer Reihe von höherer Stufe als sie selbst eingebettet ist (also in eine Reihe etwa der Stufe n-2 oder der Stufe n-1).

Auf dem Hintergrund dieser Bestimmung sei eine Reihe etwas holzschnittartiger auch wie folgt charakterisiert:

> Eine **Reihe** ist eine von ihrer Gruppe getragene und den Gruppenzwecksetzungen entsprechende Folge von Gesprächen, Textkommunikaten und weiteren Kommunikaten. Sie kann ein- oder mehrstufig sein.

Es gibt im Bereich der Zwecksetzungsbestimmung in (1.2) zwei Arten von Klammern. Die eine Klammer ist die der gemeinsamen Durchführung einer Reihe (1.2a). Nennen wir sie „**Durchführungsklammer**“. Alle die, die mit der Reihe als Durchführende zu tun haben, sind als unmittelbare Kommunikationspartner und Akteure beteiligt. Die andere Klammer ist die der institutionellen Beteiligung. Nennen wir sie „**Institutionsklammer**“.

Betrachten wir die Kommunikation einer etwas größeren Einrichtung, etwa die einer Fakultät. Basisreihen innerhalb dieser Institution sind zum Beispiel Vorlesungen und Seminare mit ihren wöchentlichen Kommunikaten. Basisreihen ergeben sich auch aus den Sitzungen der einzelnen Gremien. Diese Basisreihen stehen aber nicht nur unter der Zwecksetzung der jeweils an ihnen unmittelbar Beteiligten (Durchführung einer Veranstaltung als Dienstaufgabe, Teilnahme an einer Veranstaltung als Studienpensum etc.).

Die Basisreihen stehen vielmehr auch unter der Zwecksetzung der übergeordneten Institution, die die Aufgabe hat, in einem Fächerbereich Forschung und Lehre koordiniert durchzuführen. Insofern sind die Basisreihen auch Reihen der Fakultät, und diese Basisreihen ergeben (wenn wir Zwischenzusammenfassungen hier einfachheitshalber übergehen (Fachgruppenreihen, Institutsreihen etc.)) zusammengenommen die Reihe der Fakultät als Reihe aus den Basisreihen, so dass zu den Basisreihen auch mittelbar Beteiligte hinzutreten. Die Folge nun der Basisreihen der Fakultätsmitglieder, ob unmittelbar oder mittelbar beteiligt, entspricht den Zwecksetzungen der Fakultät und ist mithin eine Reihe, eben die der Fakultät.

Bestimmung (1.2) und deren analoge Erweiterung über (2) und (3) bestimmen die Reihe also als potentiell rekursiv, als **rekursionsfähig**.

Es ist kein Zweifel, dass sich für die institutionelle Klammerung und die institutionell begründete Rekursivität eine Fülle von typologischen Differenzierungen ergibt. Denn die **Stärke einer institutionellen Bindung** kann sehr variieren. So ergibt sich ein vieldimensionales Feld zwischen den äußeren Bereichen einer starken und einer schwachen Bindung zusammen mit einem breiten Übergangsbereich.

Bei möglichen Positionen auf diesem Feld der Bindungsstärke denke man an so verschiedene Institutionen wie etwa an diktatorische Regime, an militärische Zwänge, an Institutionen mit demokratisch geregelten Dienst- und Ar-

beitsverhältnissen, an den demokratischen Staat als Institution, an freiwillige Gemeinschaften, an Statusgruppen oder an direkte und indirekte Nachbarschaften in einem Viertel.

Noch ein Wort zu (1.1), zur Ausdifferenzierung von Kommunikaten, Kommunikatsegmenten und Kommunikatsegmentkomplexen als den Elementen einer Reihe. Man kann sich fragen, ob die Abspaltung von **Kommunikatsegmenten** und **Kommunikatsegmentkomplexen** nicht des Guten zuviel ist an Ausdifferenzierung.

Das ist nun in keiner Weise der Fall. Denn vor allem thematisch gebundene Reihen wie etwa Diskurse (s. auch weiter unten) haben ihren Ort, ihren Verhandlungsplatz, oft nicht in Kommunikaten in ihrer Ganzheit, sondern nur in einzelnen Kommunikatsegmenten und Kommunikatsegmentkomplexen.

Oder anders gesagt: In vielen Kommunikaten werden mehrere Themen behandelt, so oft in Alltagsgesprächen mit ihren Themen Wetter, Kultur, Politik, Lokales, Häusliches etc. Die Verfolgung nun **ein und desselben Themas** führt also **auch über Segmente und Segmentkomplexe von Kommunikaten** und nicht nur über Kommunikate in ihrer jeweiligen Ganzheit.

Entscheidend für die Reihenanalyse ist es nun, zu sehen, dass Reihen **Entwicklungsschichten** haben. So habe ich in „Kommunikationsreihen“ (2011: 149-151) an Schichten unterschieden:

1. die **Schicht der Zweckentwicklung**

> Der Zweck einer Reihe, der als solcher ja für sie konstitutiv ist, kann im Lauf der Entwicklung reihenerhaltend verändert werden. Es liegt dann eine Modifizierung der Reihe vor. Beispiele seien die folgend genannten. Ein Verein kann sich eintragen lassen. Eine politische Partei kann bürgerlich werden wollen. Man kann heiraten.

2. die **Schicht der Kettenentwicklung**

> Die Kette der Reihe ist nur die Folge der Zeitpunkte der Kommunikate oder sonstigen Reihenelemente ohne Betracht der Inhalte oder sonstiger Besonderheiten. Der Rhythmus der Kette kann mit Beginn der Reihe schon feststehen oder sich ändern. Sitzungen können sich häufen, Sitzungen können seltener werden.

3. die **Schicht der Wissensentwicklung**

Inhalte bzw. Wissenskomplexe [19] werden über Reihen verbreitet oder auch zurückgedrängt: Die Verbreitungsprofile werden modifiziert. Die Änderungen

[19] Inhalte sind an bestimmte Formulierungen gebunden. Wir sprechen hier vom „Inhalt“ eines Textes. Wissenskomplexe haben ihren Ort in der Kompetenz einer Person oder einer Personengruppe. Wir sprechen hier vom „Wissen“ einer Person oder vom „Wissen“ eines Faches. Der Inhalt ist gelagertes Wissen. Das Wissen einer Person prä-

sind konstitutiv für Reihen, denn die Elemente der Basisreihen, die Kommunikate, sind ja nichts anderes als Kommunikationen, und diese wiederum nichts anderes als die Vermittlung von Inhalten bzw. Wissenskomplexen. Das Wissen von Personen und Personengruppen wird so durch die Reihen modifiziert.

4. die **Schicht der Sozialentwicklung**

> Eine Modifizierung des Wissens beeinflusst auch die Sozialentwicklung einer Person oder Personengruppe, da mit der Modifizierung des Wissens auch eine Modifizierung von Bewertungen einschließlich der Partnerbewertungen und Partnerpositionierungen einhergeht.

5. die **Schicht der Umgebung**

> Die Umgebung einer Reihe (also die Räumlichkeiten und sonstigen physischen Gegebenheiten und Veränderungen sowie die benachbarten sozialen Gegebenheiten und Aktivitäten, wenn sie nicht unmittelbar einbezogen sind) gehört nicht zur Reihe, beeinflusst diese aber und wird ggf. auch ihrerseits beeinflusst. Darum sei diese Schicht, die ebenfalls auf ihre Weise dem Wandel unterworfen ist, hier mit hinzugezogen.

11. Reihen: Exemplifizierung

Zu einer kurzen Exemplifizierung der Reihe sei zunächst auf Beispielmaterial verwiesen, das wir weiter oben unterbreitet haben.

Es ergaben sich oben folgende Beispiele für Reihen, Reihensegmente und Reihenensembles:[20]

1. |||**Geschäftsführungsreihe mit M**||| =
   ||Textkommunikat Einladung|| _
   ||Textkommunikat Niederlegung der Mitgliedschaft||

2. |||**Blumenladenreihe**||| =
   ||Gespräch Erdpflege||,
   ||Gespräch Gummibaum||,
   ||Gespräch Friedhof||,
   ||Gespräch Hochzeit||

---

sentes Wissen. Beide Formen der Bindung von Wissen sind individuell und kulturell komplex aufeinander bezogen.

[20] Mehrere, nicht nach direkter oder indirekter Sukzession und auch sonst nicht geordnete Elemente einer Reihe nenne ich ein „Reihenensemble“. – Zur Notation: Reihen, Reihensegmente und Reihenensembles erscheinen in „|||“-Klammern, Kommunikate in „||“-Klammern und Akte in „|“-Klammern. (Vgl. überhaupt „Kommunikationsreihen“ (2011: Abschnitt 5.2.2)).

3. |||**Filmreservierungsreihe**||| =
   ||Programmdialog Wunsch nach Programmansage u. nach Reservierung|| _
   ||Programmdialog Wunsch nach Reservierung||

4. |||**Wikipediareihe**||| =
   ||Programmdialog Wikipedia Teil I|| _
   ||Programmdialog Wikipedia Teil II|| und
      versetzt parallel dazu [21] ||Textkommunikat Artikel „Karl-Heinz Best“|| _
   ||Programmdialog Wikipedia Teil III|| _
   ||Programmdialog Wikipedia Teil IV|| und
      versetzt parallel dazu ||Textkommunikat Artikel „Quantitative Linguistik“||

5. |||**Vizepräsidentenreihe**||| =
   ||Extemporemonolog Vergleich Tea-Party-Sympathisanten mit Terroristen||,
   ||Extemporemonolog „[…] Romney werde […] ‚euch wieder in Ketten legen’“||

Reihen können genauer **dokumentiert** werden wiederum mit der **Spaltenpartitur**. Ein einfaches, konstruiertes Beispiel sei das folgende. Es handelt sich um eine Reihe aus den Kommunikaten 1,2,3 und 4 zwischen den Personen A und B, wobei die Nummer des einzelnen Kommunikats unter jedem der Teilnehmer steht:

A	**B**
1	1
2+	2-
3-	3+
4	4

Die Kommunikate 1 und 4 sind Gespräche, die Kommunikate 2 und 3 sind Textkommunikate. „+“ kennzeichnet den Produzenten, „-“ den Rezipienten.

Eine Reihe zwischen A, B und C kann etwa wie folgt verlaufen:

A	**B**	**C**
1	1	1
2+	2-	
3+		3-

[21] „versetzt parallel“ meint: Die Kommunikate verlaufen als Ganze zeitlich parallel, aber abschnittsweise zeitlich versetzt (vgl. das zugehörige Transkript weiter oben im Abschnitt 7).

4-	4+	
5-		5+
6+	6-	6-
7	7	7

Eine Reihe kann kommunikatweise (und ggf. darüber hinaus) kommentiert werden. Als Beispiel sei die authentische Nachbarschaftsreihe aus „Kommunikationsreihen“ (2011: 362-365) herangezogen mit den Nachbarn A und B.

A	**B**	**Kommentar**
1	1	Die deutliche Bitte von A an B, überhängende Zweige zu schneiden
2	2	Die deutliche Gegenbitte von B an A, dass A ebenfalls seine zu B überhängenden Zweige schneiden solle
3	3	Mahnung von A an B, nun endlich zu schneiden, und vorsorgliche Abwehr durch B an A
	[Separatum von B]	B stellt fest, dass A noch nicht gekürzt hat.
4	4	A und B streiten sich, regeln aber dann das Problem friedlich

Die oben schon erwähnte Reihe „Buchbinder Wanninger“ kann man wie folgt aufzeichnen mit dem Buchbinder Wanninger als „Buchbindermeister“ und den übrigen Personen als seinen Gesprächspartnern:

Buchbindermeister	Portier	Sekretariat	Direktion	Verwaltung	Nebenstelle 33	Ingenieur Plaschek	Architekt Klotz	Direktor	Abteilung III	Buchhaltung
1	1									
2		2								
3			3							
4				4						
5					5					
6						6				
7							7			
8								8		
9									9	
10										10

Die Filmreservierungsreihe kann man wie folgt festhalten, wobei „P+“ den Programmanbieter kennzeichnet und „P-“ den Programmnutzer:

Programmanbieter (Unternehmen)	**Programmnutzer (Anrufer)**	**Kommentar**
1P+	1P-	Wunsch nach Programmansage und Reservierung
2P+	2P-	Wunsch nach Reservierung

Die Wikipediareihe sei wie folgt notiert:

Wikipedia[22]	Artikel „Karl-Heinz Best“	Artikel „Quantitative Linguistik“	Nutzer
1 P+ (Teil I)			1 P- (Teil I)
1 P+ (Teil II)	2+		1 P- (Teil II) u. vers. parallel 2-
1 P+ (Teil III)			1P- (Teil III)
1 P+ (Teil IV)		3+	1 P- (Teil IV) u. vers. parallel 3-

Ausführlich dokumentierte Reihen als Beispiele finden sich im Anhang von „Kommunikationsreihen“ (2011), so etwa die kommerzielle Reihe |||**Schadenregulierung 2005**||| mit 30 Kommunikaten und mehreren Subreihen, die eben notierte Nachbarschaftsreihe |||**Kirschbaum 2006**||| mit 4 Gesprächen als Reihenschritten oder die Regierungserklärungsreihe |||**Agenda 2010; 2003**|||, bestehend aus Vorkommunikaten und dem Hauptkommunikat.

Die **Wikipediareihe** zeichnet sich durch **zeitliche Überlappung** zwischen dem Programmdialog und den beiden Textkommunikaten aus. Insofern besteht hier eine besonders enge Bindung.

Besonders enge Bindungen gibt es im Übrigen auch in weiteren Fällen.

So ist das **Vorlesen** eine Reihe gleichzeitiger, aber verschiedener Kommunikate. Die Reihe besteht (1) aus dem Textkommunikat, in dem der Vorleser den zu lesenden Text rezipiert (Text auf Papier als Produzentenseite – Vorleser als Rezipient). Die Reihe besteht (2) aus dem Textkommunikat, in dem der Hörer den Text, der vorgelesen wird, rezipiert (Produzentenseite: der Vorleser – Rezipient: der Hörer). Zu denken ist auch an das **Diktat**: (1) Die Vorgabe für den Diktierenden ist ein Text, den er rezipieren muss. Und (2): Die Vorgabe für den Schreibenden ist der Text, wie er ihn vom Diktierenden hört.

Ein weiterer Fall ergibt sich aus der Art der kommunikativen Teilnahme in einem Seminar, in dem, so die Planung des Referenten, der Rezipient sowohl einer **Tischvorlage** folgt als auch gleichzeitig dem **Referat**.

[22] Die Kommunikate werden hier durchgezählt.

Weiterhin können besonders enge Bindungen zwischen zeitlich direkt aufeinanderfolgenden Kommunikaten bestehen, wenn sie in einem **übergreifenden Konstrukt** zusammengehalten werden. Ein Beispiel wäre etwa ein **Gottesdienst** mit den einzelnen zeremoniellen Dialogen und Textkommunikaten.

Wir wollen die Reihen, deren Kommunikate durch die genannten besonders engen Bindungen zusammengehalten werden, **„Konstruktreihen“** nennen.

Wenn man Beispiele für Reihen sucht, hat man eine fast grenzenlose Auswahl. Denn Reihen sind überall, und ein einzelnes Kommunikat, das aus allen Reihen auszuschließen wäre, gibt es nicht. Allerdings: Wenn man etwas mehr Platz hätte, wäre etwa der Briefwechsel zwischen Friedrich dem Großen und Voltaire hierher zu ziehen, zum einen, weil gerade Briefwechsel (und neuerdings eben Emailwechsel) sehr viel über Reihen erzählen können, und zum andern, weil der genannte Briefwechsel ein großes Beispiel für einen Briefwechsel darstellt, nicht allein und kaum schwergewichtig seines Umfangs wegen, vielmehr auf Grund seiner Intensität im Benehmen der beiden ihr Zeitalter prägenden und von ihrem Zeitalter geprägten Personen, eine Reihe, in der sichtbar und hintergründig viele Reihen der Zeit mitlaufen.

12. Reihen: Typologie

In Anbetracht, dass Reihen gesellschaftlich sozusagen flächendeckend sind, verwundert die Schwierigkeit nicht, die Fülle der Reihen in eine Typologie zu bringen. In „Kommunikationsreihen“ (2011: Kap. 6 u. 7) wurde dazu ein Versuch unternommen. Diese Skizze sei hier kurz rekapituliert.

Wir unterscheiden an Kategorien:

(1) Reihentypen nach Umfang

Zeitumfang: kurze versus lange Reihen
Personalumfang: kleine versus große Reihen
Linienzahl: einlinige und mehrlinige Reihen[23]
Phasenzahl: einphasige und mehrphasige Reihen[24]
Besetzungsdichte: dünn und dicht mit Elementen besetzte Reihen

Die |||**Geschäftsführungsreihe mit M**||| als Reihe speziell zwischen den genannten Personen, die in Sachen Geschäftsführung ansonsten noch nie etwas

[23] Es gibt den Pol einer einzigen Reihe und den Pol parallel verlaufender mehrerer Reihen.

[24] Es gibt den Pol einer einzigen, nicht in Subreihen gegliederten Reihe und den Pol mehrerer zeitlich nacheinander verlaufender Subreihen einer Reihe, also mehrerer Phasen der einen Reihe.

miteinander zu tun hatten, ist eine Reihe aus zwei Elementen, also eine **kurze**, genau besehen sogar die kürzestmögliche Reihe. Lang dagegen ist beispielsweise die Reihe zur friedlichen Nutzung der Atomenergie in Deutschland. Lang hiergegen wiederum ist die Reihe zwischen Katholiken und Protestanten in Europa, wenn man die Parteien einfach so zusammenfassen und nur auf die Kommunikation abstellen darf. Und dies überbietend **lang**, vielleicht menschheitslang ist der Diskurs über die Migration.

Das Minimum des Personalumfangs sind zwei Personen. Für eine **kleine**, sogar wiederum für die kleinstmögliche Reihe kann abermals die |||**Geschäftsführungsreihe mit M**||| stehen. **Große** Reihen mit der kommunikationspartnerschaftlichen Beteiligung von sechs-, sieben- oder achtstelligen Publika bieten etwa Zeitungen und Fernsehsender.

Einlinig und **einphasig** ist trivialerweise wiederum die |||**Geschäftsführerreihe mit M**|||. Die Geschäftsführung mit allen Mitgliedern erbringt M**ehrlinigkeit**, d.h. parallele Subreihen. Sollte sich der Vereinszweck verschieben oder sich sonstwie Änderungen ergeben, hätten wir eine neue Phase und mithin eine **mehrphasige Reihe**.

Ein Glückwunsch im Jahr: eine **dünn** besetzte Reihe. Jeden Tag telefonieren: eher eine **dichte** Besetzung.

(2) Reihentypen nach Ablaufstrukturen

geplante und ungeplante Reihen
Hauptreihen und Nebenreihen
nichtperiodische und periodische Reihen
knappe und ausführliche Reihen
funktionsparallele und Konzentrationsreihen
gleich bleibende und sich ändernde Reihen
konstruktfreie und Konstruktreihen[25]

Geplant sind etwa die Reihen in den Programmen der Schulen und Hochschulen oder auch die Nachrichtensendungen in den Medien oder das Erscheinen von Tages- und Wochenzeitungen. Der private Umgang dagegen kennt bisweilen auch **ungeplante Reihen**.

Die **Hauptreihe** etwa in einer Institution ist die Reihe der Sitzungen, auf der die entscheidenden Beschlüsse gefasst werden. Zuarbeitende Kommissionen führen von daher **Nebenreihen**.

Periodische Reihen liefern die sonntäglichen Gottesdienste einer Gemeinde. **Nichtperiodische Reihen** ergeben sich beispielsweise, wenn sich Briefpartner hin und wieder schreiben, zu Zufallszeitpunkten.

Knappe Reihen liefern die Kurznachrichten oder die Teaser auf der Startseite von publizistischen Online-Auftritten, **ausführliche Reihen** die mit dem

[25] In „Kommunikationsreihen“ (2011) noch nicht berücksichtigt.

Teaser verlinkten Volltexte oder die zentralen Nachrichtensendungen. Motivation für die Opposition der genannten Beispiele ist die Anpassung der Detaillierungsstufe an die jeweilige Aufnahmebereitschaft des Rezipienten. Eine andere Motivation, eine, die speziell in die Richtung der Ausführlichkeit drängt, ist das Bedürfnis nach Heraushebung, nach Zelebrieren.

Mehrzügige Schulen führen den Unterricht in Parallelklassen durch, mithin in **funktionsparallelen** Jahrgangsreihen. Die gegenläufige Richtung ist die Zusammenlegung von Parallelklassen, die zu **Konzentrationsreihen** führt.

Eine gleich bleibende Reihe ist eine Reihe, in deren Verlauf die Parameter gleich bleiben: ein Briefwechsel etwa immer nur zwischen denselben beiden Partnern, in immer gleichem Rhythmus, in Kommunikaten von immer gleichem Duktus. Anders der Diskurs zur Rechtschreibreform: Er nahm an Heftigkeit und Größe zu, und ebbte relativ rasch ab nach dem Hauptkommunikat in Gestalt des Urteils des Bundesverfassungsgerichts.

Für die **Konstruktreihen** gilt das oben Ausgeführte.

(3) Reihentypen nach den Verhältnissen zwischen den Akteuren

weniger und stärker konventionalisierte Reihen
kooperative und kompetitive Reihen
nichtöffentliche und mehr oder weniger öffentliche Reihen (einschließlich Massenreihen)
interne und externe Reihen
statusgeprägte Reihen (Laien-, Laien-Experten- und Expertenreihen; dominante und dominierte Reihen sowie Milieu-, Alters- und Regionalprägungen)
Reihentypen nach Wirklichkeitsbezug (pragmatische, literarische, religiöse Reihen)

Die akademischen Unterrichtsreihen vor dem Bolognaprozess waren **konventionalisiert**. Mit der Einführung der Bolognaregelungen waren viele Reihen im Anfangsstadium zum Teil improvisiert, eben **weniger konventionalisiert** als ihre Vergleichsstücke zuvor.

Koalitionsgründungen versprechen Kooperation und **kooperative Reihen**, die Positionen Regierung und Opposition Kompetition und **kompetitive Reihen**. Die unterschiedlichen Zielsetzungen der Kooperation und Kompetition und deren Verfolgung in entsprechenden Reihen finden sich wohl in allen Gruppen, seien dies Staaten untereinander, Gesellschaftsschichten untereinander (`Klassen´), Marktteilnehmer untereinander, Gruppen im privaten Bereich untereinander und, auch überall, Paare von Einzelpersonen.

Nichtöffentliche Reihen sind private Brief- oder Emailwechsel. **Öffentliche Reihen** sind die Reihen zwischen Produzenten und Rezipienten von Massenkommunikation. Differenzierungen ergeben sich aus der Bezugnahme auf

bestimmte Publika (etwa die sog. Universitätsöffentlichkeit). Nicht unbedeutende Tendenzen zu Grenzverschiebungen zwischen privat und öffentlich ergeben sich aus der Nutzung von sozialen Netzwerken wie „facebook“ bis hin zum Angebot auf den Verzicht von Privatheit überhaupt.

Interne Reihen zum Beispiel sind die Reihen eines Universitätsinstituts innerhalb des Vorstands, der Kommissionen, des Forschungs- und Lehrbereichs. Die Genehmigung des Lehrprogramms, institutionelle Umstrukturierungen, Berufungsverfahren etc. erfordern dann **externe Reihen** mit der Fakultät, dem Senat und der Universitätsleitung.

Der **Status** prägt Reihen. Reparaturannahmen in einer KFZ-Niederlassung etwa bieten eine Reihe aus Experten-Laien-Kommunikaten. Laien unter sich dort und Experten untereinander vollziehen dann jeweils andere Reihen.

Sich selbst als solche bekennende Phantasiewelten vor Augen zu führen ist eine weitesthin anerkannte Teilhabe an der gesellschaftlichen Kommunikation. Wir haben hier entsprechende, im weitesten Qualitätssinn als **literarisch verstandene Reihen**. Anders sind **religiöse Reihen** zu sehen, da deren Realitätsgehalt von ihren Vertretern anders beurteilt wird als von nichtreligiösen Gesprächspartnern. In gewisser Weise Bezugsgröße sein dürften in beiden Fällen dabei die **pragmatischen (oder faktualen) Reihen**.

(4) Reihentypen nach Medien

Gegenübermedium
traditionelle Medien
Neue Medien

Das sog. face-to-face-Gespräch, allerdings eben nicht über Videoverbindung, sondern genau nur im eigentlichen, im räumlichen Gegenüber auf Armeslänge, ist das von uns so genannte **Gegenübermedium**.

Die **traditionellen Medien** sind das Schreiben, der Postversand, das Telefon, die Massenmedien über Papier, Radio und Fernsehen.

Die **Neuen Medien** sind die des Internet mit ihren Anwendungen (Email, Chat, Handel, Enzyklopädie, soziale Netzwerke etc.).

(5) Domänenspezifische Reihentypen[26]

private Reihen
nichtprivate Reihen (Staat, Organisation, Öffentlichkeit)

Wir erfassen die **Privatheit** als den einen großen Pol der gesellschaftlichen sektoralen Strukturiertheit. Überdies wäre ohne diesen Lebensbereich (dies ist allerdings nur unsere Überzeugung, die wir nicht zwingend begründen können) eine

[26] Vgl. insbesondere „Kommunikationsreihen“ (2011:286-298).

humane Gesellschaft nicht denkbar, was sich in Gegensatz stellt zu den Begleitentwicklungen der Neuen Medien in Gestalt nicht nur von radikaler Infragestellung von Privatheit, sondern auch von ihrer tatsächlichen Bedrohtheit.

Demgegenüber stehen wie erwähnt die **nichtprivaten** Sektoren, in denen die Privatheit von vornherein eingeschränkt oder aufgehoben ist.

(6) Der Diskurs als gesellschaftlich übergreifende Reihe[27]

Wenn die Behandlung eines Themas nicht nur von den entsprechenden Fachleuten getragen wird, sondern vor allem auch der private Sektor mehr oder weniger umfangreich eingreift. sprechen wir von einem „Diskurs". Der Diskurs ist dabei nichts anderes als eine hochstufige Reihe.

(7) Reihen auf gesellschaftlicher Ebene

die Gesellschaftsreihe
die Gesellschaftengruppenreihe
die globale Reihe

Soweit mein Vorschlag (in Kürze) für eine einfache Typologie. Die letztgenannte Kategorie „Reihen auf gesellschaftlicher Ebene" ist dabei für mein besonderes Interesse an eben dieser Ebene von Bedeutung. Die Bestimmung dieser drei Reihentypen ist meine Antwort auf die Frage danach, wie eine Gesellschaft aus linguistischer Sicht kommuniziert.

Diesen Bestimmungen sei der nächste Abschnitt gewidmet.

13. Wie kommuniziert eine Gesellschaft aus linguistischer Sicht? Eine Antwort

Die Kommunikation einer Gesellschaft ist die Gesellschaftsreihe. Sie wird wie folgt bestimmt:

Die Kommunikation einer Gesellschaft ist die aus einer gesellschaftsinternen und einer gesellschaftsexternen Subreihe bestehende **Gesellschaftsreihe**, die bedingend und bedingt eingebunden ist in die geschichtliche Entwicklung der Gesellschaft und geprägt ist durch bestimmte Bedingungen der Gesellschaft, insbesondere durch

die interne Verfasstheit,
die Sprachensituation,
die domänen- und sektorbezogene Differenziertheit,
die internationale Positionierung und die Struktur des externen Raums,

[27] Vgl. insbesondere „Kommunikationsreihen" (2011:298-314).

die aktuellen Zweckprofile der jeweiligen Akteure der Gesellschaft einschließlich der herausragenden Themen,
die geschichtliche Ausgangssituation und deren Fortentwicklung,

wobei sie diese Bedingungen ihrerseits beeinflusst.

Der externe Raum der Gesellschaftsreihe sei etwas näher bestimmt durch die Bestimmung einer näheren Umgebung, der jeweiligen Gesellschaftengruppenreihe, und durch die Bestimmung des Gesamtrahmens, der globalen Reihe.

Die Gesellschaftengruppenreihe wird wie folgt bestimmt:

Die Kommunikation einer Gesellschaftengruppe ist die aus einer gruppeninternen und einer gruppenexternen Subreihe bestehende **Gesellschaftengruppenreihe**, die bedingend und bedingt eingebunden ist in die geschichtliche Entwicklung der Gesellschaftengruppe und geprägt ist durch bestimmte Bedingungen der Gesellschaftengruppe, insbesondere durch

die besonderen Geprägtheiten der einzelnen Mitgliedsgesellschaften und die Verfasstheit der Gruppe,
die Art der inneren Kohärenz,
die Sprachensituation,
die domänen- und sektorbezogene Differenziertheit,
die internationale Positionierung und die Struktur des externen Raums,
die aktuellen Zweckprofile der einzelnen Mitgliedsgesellschaften einschließlich der herausragenden Themen,
die geschichtliche Ausgangssituation und deren Fortentwicklung,

wobei sie diese Bedingungen ihrerseits beeinflusst.

Die globale Reihe schließlich wird so bestimmt:

Die globale Kommunikation als die Kommunikation der Weltbevölkerung ist die **globale Reihe**, das heißt die Reihe der Reihen aller Gesellschaften und aller Gesellschaftengruppen.

Mit Bedacht sei nun auf die Schlussbemerkung verwiesen.

14. Schlussbemerkung

Es sei erlaubt, mit einem Zitat zu schließen. Der Cantzler Johann Peter von Ludewig der Universität Halle bemerkt in seiner „Vorrede über das Universal=

Lexicon“[28] des Verlegers Johann Heinrich Zedler im ersten Band (S. 13) unter anderem:

> […] Und unser Wissen bleibet wohl /auch in weltlichen Dingen/ ein blosses Stückwerk. Ein Tag lehret den andern und der letztere wird öfters zum Meister des erstern.
> […]
> Halle, den 30 Sept. 1731. Johann Peter von Ludewig
> Cantzler der Universität Halle

Wir hatten nun nicht die Gelegenheit, uns kennenzulernen. Aber im Hinblick auf das Zitierte fühle ich mich dem Cantzler sehr verbunden.

[28] In „Grosses vollständiges Universal Lexicon Aller Wissenschaften und Künste […] Erster Band. A. – Am. Halle und Leipzig, Verlegts Johann Heinrich Zedler, Anno 1732“, zitiert nach der Neuauflage durch „Akademische Druck- und Verlagsanstalt“, Graz 1961, S. 1-16. Zitat: S. 13, Unterzeichnung S. 16.

Diversification of English Valency Patterns

Petra Steiner, University of Rostock

1. Introduction

Within quantitative linguistics, diversification has become a widely investigated and well-explained phenomenon on different linguistic levels (see Altmann 1985, Altmann & Best 1996, Altmann et al. 1996, Best 1996, Best 2005a: 256ff; Best 2005b: 262ff, Best & Brynjólfson 1997, Rothe 1991b, Wimmer & Altmann 1996: 112). However, concerning syntactic units and properties, the research on diversification was restricted for some time. Rothe (1991a) observes diversification of the German genitive case, Steiner (2009) and Steiner & Prün (2007) show the effects of diversification for Icelandic and German noun inflection.

The notion of valency, with its focus on the properties of verbs, nouns and adjectives to open syntactic and semantic argument slots, was even less in the focus of linguistics. Since the 1960s, valency grammar and valency dictionaries seem to be mainly located in German linguistics (Helbig & Schenkel 1968, Engel & Schumacher 1976, Sommerfeldt & Schreiber 1996, Schumacher et al. 2004).[1] It is certainly no coincidence that the latest valency dictionary for English, *A Valency Dictionary of English* (VDE) (Herbst et al. 2004), has been (co-)edited by German linguists. Within the linguistic community, the dominant perspective of research in syntax was and is one of phrase structures, however, Fillmore's case grammar (Fillmore 1968, 1970, 1971) and the later development to Frame Semantics (Fillmore 1977, 1982, 1985) and its application in the FrameNet project (Fillmore/Johnson/Petruck 2003, Fillmore 2007) can be considered as an important alternative to standard models of syntactic description, yielding a complex and precise description of the semantic-syntactic properties of lexical units and their concomitant arguments.

Within Quantitative Linguistics, Köhler (2012: 92ff) investigates valency patterns for German, Russian, Czech, Chinese, Hungarian and Finnish language data. Among other models, the number of verb patterns for single verbs is considered analogous to typical distributions of word sense (see Altmann 1985), showing the typical derivations of statistical hypotheses. The distribution of the number of sentence structures for Helbig and Schenkel (1991) can be fitted by the Zipf-Mandelbrot dis-

---

[1] See Somers (1987: 4ff) for an overview of the time before 1985 and Herbst & Schüller (2008: 109) for further references.

tribution, another function which is typical for distributional phenomena.

The following section provides first a definition of valency and then gives information on the English verb valency patterns as they are classified in the Erlangen valency patternbank. Then two linguistic hypotheses on diversification are derived and formulated with regard to Köhler's (2012: 92ff) investigations of verb valency. After the description of the data, the statistical hypotheses are tested and finally discussed.

2. Verb valency in the Erlangen Valency Patternbank

Valency is either defined as (a) the number of complements a governing element possesses or more generally as (b) the ability of binding and government. There are many definitions of valency[2], but for the purpose of this paper the following might be sufficient: Valency is the ability of verbs, nouns, adjectives, adverbs or particles to open up arguments slots to be filled by complements (see Herbst & Schüller 2008: 209). The differentiation of complements and adjuncts will not be discussed in this context and the classification of the complements for the VDE (*A Valency Dictionary of English*, Herbst et al. (2004)) is taken as justified and consistent. Optionality is not specifically marked. Instead all possible valency patterns are included.

The corpus-based data of the VDE was made available to the scientific community in the Erlangen Valency Patternbank (Herbst & Uhrig 2009). Besides adjective and noun patterns, it comprises 511 descriptions of verbs (Herbst 2009: 1). Verbs which show merely mono- or divalent patterns are not included. This means that verbs such as *eat* cannot be found inside the database. Frequency counts are provided (i) for the number of lexemes (lemmata) for which a pattern exists in the VDE and (ii) for the number of sense entries for which a pattern exists in the VDE (Herbst 2009: 2f). Herbst (2009:2) refers to these as *lexical units*, Köhler (2012: 92) as *verb variants*. Furthermore, the counts can be sorted by the number of the complements.

3. Linguistic and statistical hypotheses

Human capacity of memory and time for retrieval is restricted, therefore it is necessary to reduce the inventory of linguistic items. At the hypothetical extreme point of such a process, this would lead to one single form, for instance a certain grammatical function, with many (grammatical) meanings. In Quantitative Linguistics, this is

[2] For an overview of different definitions see Steiner (forthcoming).

called *semantic diversification*. However, such language would be hard to decipher and will not develop in natural contexts. To ease the so-called *need of the minimization of the decoding effort* (see Köhler 2005: 766ff), there exists an antagonistic process, which is referred to as *unification*. On the level of meaning, semantic unification is the tendency to minimize the effort of disambiguation. At the hypothetical extreme point of such a process, this would lead to forms with just one (grammatical) meaning, for instance pattern elements or valency patterns, which are unique. As described above, formal unification - the reduction of forms - comes along with semantic diversification. As both tendencies are operating on both the level of form and of meaning, a state of equilibrium is maintained and characteristic distributions of frequencies will occur: Only a few forms are used often, while the rest of the inventory is used with a low frequency (Rothe 1991b). Altmann (1991) derives some of these typical distributions, of which the negative binomial distribution and the Zipf-Alekseev distribution were among the most common.

For the current investigations, two hypotheses are postulated.

a. As demonstrated by Köhler (2012: 111), dependency types and semantic (case) roles usually show effects of diversification. This should also be the case for the frequencies of pattern elements of the VDE.

b. In analogy to Köhler's (2012: 94) investigation on the distribution of the numbers of sentence structures, it can be expected that the frequencies of verb valency patterns can be fitted by a Zipfian distribution. This should hold for lexemes as well as for lexical units.

4. Data and Tests of the Hypotheses

All the frequency counts of the following investigations were retrieved from the Erlangen Valency Patternbank (Herbst & Uhrig 2009). The fitting was performed with the Altmann-Fitter (Altmann 2000).

Hypothesis (a). The pattern elements with their observed and expected frequencies are shown in Table 1. This inventory comprises pattern elements for verbs, nouns and adjectives. The most frequent complements are noun phrases with 2381. The second class consists of 734 by_phrases, accounted for by the lists of passive constructions. The patternbank differentiates prepositional phrases and phrasal verbs by their particles, which leads to classes such as *to_NP* or *with_NP*. Other complements such as *a lot/much* are restricted to lexical units as *it says a lot for … that …*. This list is certainly not exhaustive. The frequency distribution can be modeled using the right truncated negative binomial distribution

$$(1)\qquad P_x = \frac{\binom{k+x-1}{x} p^k q^x}{F(R)}, \quad x = 0,1,2,...R;$$

$$k > 0; 0 < p < 1; q = 1 - p; R \in \mathrm{N}; F(R) = \sum_{i=0}^{R} \binom{k+i-1}{i} p^k q^i$$

with a very good fit.

Table 1
Observed and expected frequencies of the pattern elements
of the Erlangen Valency Patternbank

X	***Pattern element***	f_x	NP_x
1	NP	2381	2391.31
2	by_phrase	734	689.68
3	to_NP	460	441.11
4	V-ing	235	333.77
5	with_NP	225	272.08
6	to_INF	215	231.35
7	that_CL	210	202.14
8	adjective	174	179.99
9	it	173	162.53
10	from_NP	167	148.35
11	for_NP	147	136.56
12	out	127	126.58
13	up	111	118.00
14	ADV	108	110.54
15	wh_CL	106	103.97
16	on_NP	101	98.15
17	on	99	92.93
18	n/a	98	88.24
19	NP_V-ing	86	83.98
20	as_NP	85	80.10
21	NP_and_NP	82	76.55
22	about_NP	75	73.28
23	of_NP	73	70.27
24	CL	62	67.47
25	AdjP	61	64.88
26	for_NP_to_INF	59	62.46
27	$NP_{pl/group}$	59	60.19
28	to	59	58.07
29	by_NP	55	56.08
30	off	55	54.20
31	[it]	52	52.43
32	about_V-ing	52	50.76
33	in_NP	52	49.18
34	against_NP	51	47.68
35	in	51	46.26
36	down	49	44.90
37	SENTENCE	48	43.61
38	back	46	42.38

X	***Pattern element***	f_x	NP_x
39	over	46	41.21
40	at_NP	43	40.08
41	wh_to_INF	43	39.01
42	QUOTE	40	37.98
43	about	39	36.99
44	as_AdjP	39	36.05
45	on_V-ing	39	35.14
46	about_wh_CL	38	34.27
47	NP_{pl}	37	33.43
48	NP_1	36	32.62
49	NP_2	36	31.85
50	with	35	31.10
51	into_NP	34	30.37
52	to_V-ing	34	29.67
53	away	31	29.00
54	for_V-ing	31	28.35
55	together	31	27.72
56	on_wh_CL	29	27.11
57	ReflPron	29	26.52
58	in_V-ing	28	25.95
59	about_wh_to_INF	27	25.39
60	as_V-ing	27	24.86
61	over_NP	27	24.34
62	on_wh_to_INF	26	23.83
63	from_V-ing	25	23.34
64	NP:QUANT	24	22.87
65	upon_NP	23	22.40
66	between_ NP_and_NP	22	21.96
67	between_NP_{pl}	22	21.52
68	around	21	21.09
69	like_NP	21	20.68
70	round	21	20.28
71	of_V-ing	20	19.89
72	of	19	19.51
73	against_V-ing	18	19.14
74	by_V-ing	18	18.78
75	through	18	18.43
76	about_NP_V-ing	17	18.09
77	after_NP	17	17.75
78	like_V-ing	17	17.43
79	of_wh_CL	17	17.11
80	for	16	16.80
81	forward	16	16.50
82	across	15	16.20
83	ahead	15	15.91
84	in_favour_of_NP	15	15.63
85	upon	15	15.36
86	against	14	15.09
87	apart	14	14.83
88	at	14	14.57
89	from	14	14.32
90	something/a_lot/etc.	14	14.07
91	upon_V-ing	14	13.83
92	with_V-ing	14	13.60
93	of_NP:QUANT	13	13.37
94	off_NP	13	13.15
95	home	12	12.93
96	of_NP_V-ing	12	12.71
97	something/ little/ what/etc.	12	12.50
98	AdjP_pattern	11	12.30

X	*Pattern element*	f_x	NP_x
99	along	11	12.09
100	by_NP:QUANT	11	11.90
101	much/nothing/etc.	11	11.70
102	something/much/etc.	11	11.51
103	NUM	10	11.33
104	there	10	11.15
105	INF	9	10.97
106	into_V-ing	9	10.79
107	NP_pattern	9	10.62
108	out_of_NP	9	10.46
109	SCORE	9	10.29
110	to_ReflPron	9	10.13
111	toward_NP	9	9.97
112	towards_NP	9	9.82
113	as_if_CL	8	9.66
114	as_to_NP	8	9.52
115	by	8	9.37
116	for_NP_V-ing	8	9.23
117	like_CL	8	9.08
118	of_wh_to_INF	8	8.95
119	over_V-ing	8	8.81
120	over_wh_CL	8	8.68
121	through_NP	8	8.55
122	what	8	8.42
123	you	8	8.29
124	among_$NP_{pl/group}$	7	8.17
125	amongst_$NP_{pl/group}$	7	8.05
126	as_NP_pattern	7	7.93
127	at_NP:QUANT	7	7.81
128	at_V-ing	7	7.69
129	for_wh_CL	7	7.58

X	*Pattern element*	f_x	NP_x
130	NP_to_INF	7	7.47
131	to_NP_V-ing	7	7.36
132	toward_V-ing	7	7.25
133	towards_V-ing	7	7.15
134	as_AdjP_pattern	6	7.04
135	as_though_CL	6	6.94
136	as_to_wh_CL	6	6.84
137	aside	6	6.74
138	behind	6	6.65
139	by_SCORE	6	6.55
140	for_ReflPron	6	6.46
141	go	6	6.37
142	if_CL	6	6.28
143	NP_from_NP	6	6.19
144	to_NP:QUANT	6	6.10
145	whether_CL	6	6.01
146	with_NP_V-ing	6	5.93
147	by_NUM	5	5.85
148	DESCRIPTION	5	5.76
149	for_NP:QUANT	5	5.68
150	forth	5	5.60
151	I	5	5.53
152	into_NUM	5	5.45
153	into_wh_CL	5	5.37
154	NUM_and_NUM	5	5.30
155	off_V-ing	5	5.23
156	on_NP_V-ing	5	5.16
157	out_of_V-ing	5	5.08
158	wh	5	5.01
159	ADV:QUALITY	4	4.95
160	around_NP	4	4.88

X	***Pattern element***	f_x	NP_x
161	as	4	4.81
162	between	4	4.75
163	by_NP:QUANT/ SCORE	4	4.68
164	for_NP's_V-ing	4	4.62
165	in/out	4	4.56
166	in_favour_of	4	4.49
167	in_favour_of_V-ing	4	4.43
168	into	4	4.37
169	onto_NP	4	4.32
170	Pron	4	4.25
171	to_wh_CL	4	4.20
172	under_NP	4	4.15
173	upon_NP_V-ing	4	4.09
174	upon_wh_CL	4	4.04
175	what/much/etc.	4	3.98
176	why_CL	4	3.93
177	about_DESCRIPTI ON	3	3.88
178	beyond_NP	3	3.83
179	by_wh_CL	3	3.78
180	dumb	3	3.73
181	how_CL	3	3.68
182	in_favour	3	3.63
183	in_NUM	3	3.58
184	in_wh_CL	3	3.54
185	into_NP_and_NP	3	3.49
186	into_NP_{pl}	3	3.44
187	NP's_V-ing	3	3.40
188	of_NP_{pl}	3	3.36
189	ORDINAL	3	3.31

X	***Pattern element***	f_x	NP_x
190	otherwise	3	3.27
191	out_of_NUM	3	3.23
192	PART	3	3.19
193	past_NP	3	3.14
194	right/wrong	3	3.10
195	short	3	3.06
196	so/not	3	3.02
197	than_NP	3	2.99
198	than_V-ing	3	2.95
199	to(_INF)	3	2.91
200	to_NP_QUOTE	3	2.88
201	to_NP_SENTENCE	3	2.84
202	to_NUM	3	2.80
203	upon_ReflPron	3	2.76
204	way	3	2.73
205	wh_CL_often_negat ive	3	2.69
206	with_Pron	3	2.66
207	without	3	2.63
208	after	2	2.59
209	against_NP_V-ing	2	2.56
210	as_CL	2	2.53
211	as_DESCRIPTION	2	2.20
212	as_to_wh_to_INF	2	2.47
213	as_V-ed	2	2.44
214	at_NP_V-ing	2	2.40
215	at_wh_CL	2	2.37
216	before_NP	2	2.34
217	behind_NP	2	2.32
218	by_NP_V-ing	2	2.29
219	from_AdjP	2	2.26

X	*Pattern element*	f_x	NP_x
220	from_wh_CL	2	2.23
221	goodbye/good_night	2	2.20
222	in_NP_V-ing	2	2.17
223	in_wh_to_INF	2	2.15
224	into_wh_to_INF	2	2.12
225	it/one	2	2.10
226	it_+_pattern_of_II	2	2.07
227	it_for_NP_to_INF	2	2.04
228	NP_AdjP	2	2.02
229	NP_ADV	2	1.99
230	NP_and_V-ing	2	1.97
231	NP_INF	2	1.95
232	NP_{group}	2	1.92
233	NP_QUANT	2	1.90
234	NP_to	2	1.87
235	NP_to_INF_often_passive	2	1.85
236	NP_with_NP	2	1.83
237	of_it	2	1.81
238	of_ReflPron	2	1.78
239	on_how_to_INF	2	1.76
240	on_NP_to_INF	2	1.74
241	on_ReflPron	2	1.72
242	on_SCORE	2	1.70
243	on_to_NP	2	1.68
244	out_NP	2	1.66
245	out_QUOTE	2	1.64
246	out_SENTENCE	2	1.62
247	over_wh_to_INF	2	1.60
248	so	2	1.58
249	so/not/otherwise	2	1.56

X	*Pattern element*	f_x	NP_x
250	that	2	1.54
251	to_AdjP	2	1.52
252	to_be_V-ed	2	1.50
253	to_wh_to_INF	2	1.49
254	toward	2	1.47
255	towards	2	1.45
256	upon_it	2	1.43
257	upon_wh_to_INF	2	1.42
258	V-ing_and_NP	2	1.40
259	V-ing_and_V-ing	2	1.38
260	when_CL	2	1.37
261	with_wh_CL	2	1.35
262	within_NP	2	1.33
263	[there]	1	1.32
264	a_lot/much	1	1.30
265	about_how_to_INF	1	1.29
266	above_NP	1	1.27
267	across_NP	1	1.26
268	against_wh_CL	1	1.24
269	also_after_noun	1	1.23
270	aso_after_premodifier	1	1.21
271	aso_with_premodifier	1	1.20
272	and_INF	1	1.18
273	as_if_to_INF	1	1.17
274	as_QUOTE	1	1.16
275	as_to_V-ing	1	1.14
276	as_to_whether_CL	1	1.13
277	at_it	1	1.12
278	at_wh_to_INF	1	1.10
279	at_what	1	1.09
280	attributive	1	1.08

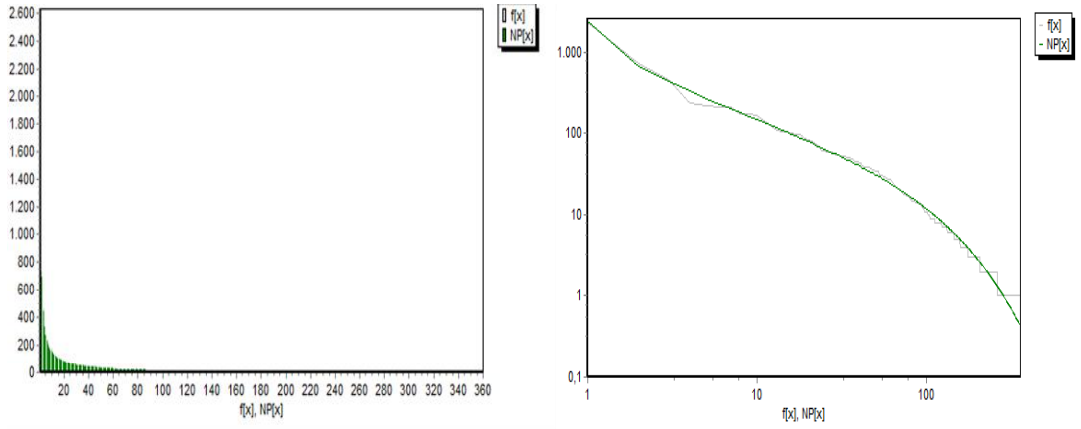
X	***Pattern element***	f_x	NP_x
281	away/in/out	1	1.06
282	behind_NP_V-ing	1	1.05
283	behind_why_CL	1	1.04
284	below_NP	1	1.03
285	beneath_NP	1	1.02
286	between_ NP:QUANT_and_NP:QUANT	1	1.00
287	between_ NP_and_V-ing	1	0.99
288	between_V-ing_ and_NP	1	0.98
289	between_V-ing_and_V-ing	1	0.97
290	be-tween_wh_CL_and_wh_CL	1	0.96
291	betweenV-ing_and_V-ing	1	0.95
292	but_INF	1	0.94
293	counter_to_NP	1	0.92
294	for_NP_1	1	0.91
295	for_NP_2	1	0.90
296	for_to_INF	1	0.89
297	for_wh_to_INF	1	0.88
298	for_what_CL	1	0.87
299	for_when_CL	1	0.86
300	for_where_CL	1	0.85
301	from_NP:QUANT	1	0.84
302	from_NP:QUANT_to_NP:QUANT	1	0.83
303	from_NP_V-ing	1	0.82
304	from_NUM	1	0.81

X	***Pattern element***	f_x	NP_x
305	goodbye	1	0.80
306	in_favour_of_NP_V-ing	1	0.79
307	in_NP/_V-ing	1	0.79
308	in_QUOTE	1	0.78
309	in_SENTENCE	1	0.77
310	in_support_of_NP	1	0.76
311	in_support_of_V-ing	1	0.75
312	in_that_CL	1	0.74
313	it_if_CL	1	0.73
314	it_when_CL	1	0.72
315	lest_CL	1	0.72
316	LETTERS	1	0.71
317	NP_/_NP_and_NP_{pl}	1	0.70
318	NP_by_NP	1	0.69
319	NP_p	1	0.68
320	NP_V-ed	1	0.68
321	NP}	1	0.67
322	of_NP_and_NP	1	0.66
323	of_NP_V-ed	1	0.65
324	of_NP_Ving	1	0.65
325	of_NUM	1	0.64
326	on_DESCRIPTION	1	0.63
327	on_how_CL	1	0.62
328	one/something	1	0.62
329	out_of	1	0.61
330	over_ReflPron	1	0.60
331	predicative	1	0.60
332	round_NP	1	0.59
333	so/otherwise	1	0.58

X	*Pattern element*	f_x	NP_x
334	than_INF	1	0.58
335	than_to_INF	1	0.57
336	than_wh_CL	1	0.56
337	there_+_NP	1	0.56
338	there_be	1	0.55
339	through_V-ing	1	0.54
340	to_be	1	0.54
341	to_it	1	0.53
342	to_judge/judging	1	0.53
343	to_NP_INF	1	0.52
344	to_where_CL	1	0.51
345	under	1	0.51
346	unlike_NP	1	0.50
347	up_to_NUM	1	0.50
348	V-ed	1	0.49
349	well	1	0.49
350	wh_to	1	0.48
351	what/something/etc.	1	0.47
352	what_CL	1	0.47
353	where_CL	1	0.46
354	who_CL	1	0.46
355	with_AdjP	1	0.45
356	with_NP/V__ing	1	0.45
357	with_ReflPron	1	0.44
358	without_NP	1	0.44
359	without_V-ing	1	0.43
360	yes/no	1	0.43

$k = 0.2911$, $p = 0.0092$, $R = 360$
$\chi^2 = 93.07$, $DF = 316$, $P(\chi^2) \approx 1.00$

Figure 1a and Figure 1b present the results in graphic form. As the distribution is very steep, the bi-logarithmic transformation gives a better impression of the similarity of the distributions.



(a) No transformation (b) bi-logarithmic transformation

Figure 1: Fitting the truncated negative binomial distribution to the data in Table 1

Hypothesis (b). For the sake of brevity, the hypothesis on the diversification of va-

lency patterns is tested solely on the frequencies of the active patterns of the verbal lexemes in the Erlangen Valency Patternbank, the first of which are presented in Table 2. *SCU* (subject complement unit) is a coarse-grained category which comprises noun phrases or clauses functioning as subjects.

Table 2
Head of the counts from the Erlangen Valency Patternbank
for active verb patterns, sorted by the frequency of lexical units

patterns	**lexemes**	**lexical units**
SCU VHC_{act}.....NP	467 (incl. 1 spec.)	1173 (incl. 1 spec.)
SCU VHC_{act}.....	358 (incl. 4 spec.)	577 (incl. 5 spec.)
SCU VHC_{act}..... NP.....ADV	137 (incl. 4 spec.)	267 (incl. 5 spec.)
SCU VHC_{act}.....ADV	117	256

The data under investigation contains 1324 different valency patterns. The most frequent one comprises 469 different lexemes and 1173 different sense entries. It is the typical ditransitive construction. The second one is the intransitive pattern with occurs with 358 lexemes, followed by two constructions with adverbials.

The mixed frequencies of all valency classes do not lead to any sufficiently good fit of a characteristic distribution. However, if the valency patterns are sorted by the number of valency elements, the typical Zipfian order appears. This is demonstrated for the valency classes of 1, 2, and 3 pattern elements. While valency classes 1 and 2 can be modelled by the right truncated modified Zipf-Alekseev distribution

$$(2)\, P_x = \begin{cases} \alpha, & x = 1 \\ \dfrac{(1-\alpha)x^{-(a+b\ln x)}}{\sum_{i=j}^{n} j^{-(a+b\ln j)}}, & x = 2,3,... \end{cases} \qquad a,b \in R, 0 < \alpha < 1$$

(see Table 3 and 4 and Figure 2 and 3), for valency class 3, only the Zipf-Mandelbrot distribution yields an acceptable result (see Table 5 and Figure 4). Due to their combinatorial potential, those classes are very different and it becomes obvious why their union does not lead to any satisfactory modeling.

Table 3
Observed and expected frequencies of active verb valency patterns
for valency 1 from the Erlangen Valency Patternbank

X	**f_x**	**NP_x**
1	358	358.00
2	7	7.37
3	5	5.43
4	4	4.30
5	4	3.56
6	4	3.04
7	3	2.64
8	3	2.33

X	**f_x**	**NP_x**
9	3	2.09
10	2	1.89
11	1	1.72
12	1	1.58
13	1	1.45
14	1	1.35
15	1	1.26

$a = 0.6164$, $b = 0.0769$, $n = 15$, $\alpha = 0.8995$
$\chi^2 = 1.8750$, $DF = 10$, $P(\chi^2) = 0.9972$

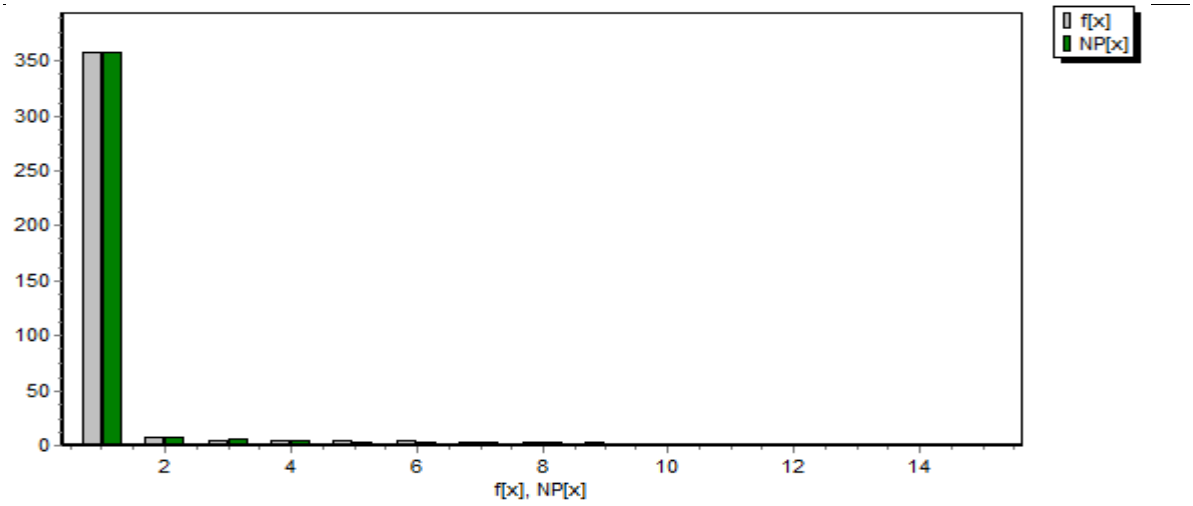


Figure 2: Observed and expected frequencies of active verb valency patterns
for valency 1 from the Erlangen Valency Patternbank

Table 4
Observed and expected frequencies of active verb valency patterns
for valency 2 from the Erlangen Valency Patternbank

X	**f_x**	**NP_x**
1	467	467.00
2	358	272.71
3	139	219.46
4	117	182.19
5	117	154.83
6	108	133.96
7	104	117.54
8	97	104.31
9	93	93.43
10	93	84.35
11	91	76.66
12	84	70.08
13	84	64.38
14	82	59.42
15	69	55.05
16	58	51.18
17	54	47.74
18	53	44.66
19	50	41.89
20	45	39.38
21	44	37.11
22	43	35.04
23	39	33.15
24	38	31.42
25	36	29.83
26	34	28.36
27	32	27.00
28	30	25.74
29	29	24.57
30	23	23.48
31	21	22.47
32	20	21.52
33	19	20.64
34	18	19.81
35	18	19.03
36	17	18.30
37	17	17.61

X	**f_x**	**NP_x**
38	16	16.95
39	16	16.34
40	15	15.76
41	15	15.21
42	15	14.69
43	14	14.20
44	14	13.73
45	14	13.28
46	13	12.86
47	13	12.46
48	12	12.07
49	12	11.71
50	11	11.36
51	11	11.02
52	11	10.70
53	10	10.40
54	10	10.11
55	10	9.83
56	10	9.56
57	10	9.30
58	10	9.05
59	10	8.81
60	9	8.59
61	8	8.37
62	8	8.16
63	8	7.95
64	8	7.76
65	7	7.57
66	7	7.39
67	7	7.21
68	7	7.04
69	7	6.88
70	7	6.72
71	7	6.57
72	6	6.42
73	6	6.28
74	6	6.14

X	**f_x**	**NP_x**
75	6	6.01
76	6	5.88
77	6	5.76
78	6	5.64
79	6	5.52
80	6	5.40
81	5	5.29
82	5	5.19
83	5	5.08
84	5	4.98
85	5	4.89
86	5	4.79
87	4	4.70
88	4	4.61
89	4	4.52
90	4	4.44
91	4	4.36
92	4	4.28
93	4	4.20
94	4	4.12
95	4	4.05
96	4	3.98
97	4	3.91
98	4	3.84
99	3	3.77
100	3	3.71
101	3	3.64
102	3	3.58
103	3	3.52
104	3	3.46
105	3	3.41
106	3	3.35
107	3	3.29
108	3	3.24
109	3	3.19
110	3	3.14
111	3	3.09

X	f_x	NP_x
112	3	3.04
113	3	2.99
114	3	2.95
115	3	2.90
116	3	2.86
117	3	2.81
118	3	2.77
119	3	2.73
120	2	2.69
121	2	2.65
122	2	2.61
123	2	2.57
124	2	2.54
125	2	2.50
126	2	2.47
127	2	2.43
128	2	2.40
129	2	2.36
130	2	2.33
131	2	2.30
132	2	2.27
133	2	2.24
134	2	2.21
135	2	2.18
136	2	2.15
137	2	2.12
138	2	2.09
139	2	2.06
140	2	2.04
141	2	2.01
142	2	1.98
143	2	1.96
144	2	1.93
145	2	1.91
146	2	1.88
147	2	1.86
148	2	1.84
149	2	1.81
150	1	1.79
151	1	1.77
152	1	1.75
153	1	1.73

X	f_x	NP_x
154	1	1.71
155	1	1.69
156	1	1.67
157	1	1.65
158	1	1.63
159	1	1.61
160	1	1.59
161	1	1.57
162	1	1.55
163	1	1.53
164	1	1.52
165	1	1.50
166	1	1.48
167	1	1.47
168	1	1.45
169	1	1.43
170	1	1.42
171	1	1.40
172	1	1.39
173	1	1.37
174	1	1.36
175	1	1.34
176	1	1.33
177	1	1.31
178	1	1.30
179	1	1.28
180	1	1.27
181	1	1.26
182	1	1.24
183	1	1.23
184	1	1.22
185	1	1.21
186	1	1.19
187	1	1.18
188	1	1.17
189	1	1.16
190	1	1.14
191	1	1.13
192	1	1.12
193	1	1.11
194	1	1.10
195	1	1.09

X	f_x	NP_x
196	1	1.08
197	1	1.07
198	1	1.06
199	1	1.05
200	1	1.04
201	1	1.03
202	1	1.02
203	1	1.01
204	1	1.00
205	1	0.99
206	1	0.98
207	1	0.97
208	1	0.96
209	1	0.95
210	1	0.94
211	1	0.93
212	1	0.92
213	1	0.92
214	1	0.91
215	1	0.90
216	1	0.89
217	1	0.88
218	1	0.88
219	1	0.87
220	1	0.86
221	1	0.85
222	1	0.84
223	1	0.84
224	1	0.83
225	1	0.82
226	1	0.81
227	1	0.81
228	1	0.80
229	1	0.79
230	1	0.79
231	1	0.78
232	1	0.77
233	1	0.77

$a = 0.2481$, $b = 0.1606$
$n = 233$, $\alpha = 0.1318$
$\chi^2 = 147.18$, $DF = 21$
$P(\chi^2) \approx 0.9998$

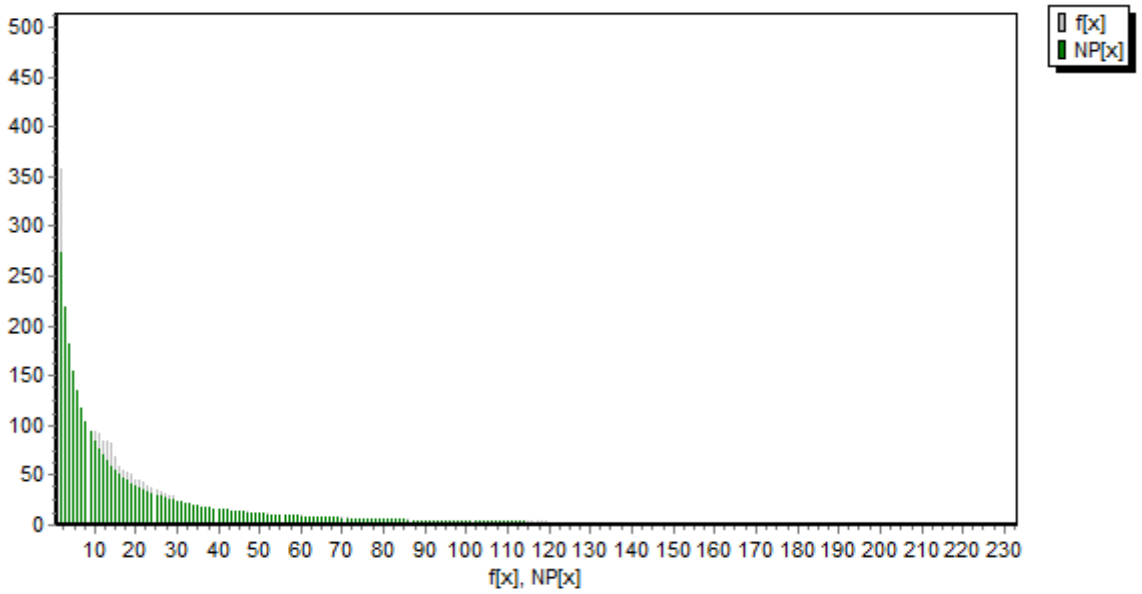


Figure 4: Observed and expected frequencies of active verb valency patterns for valency 2 from the Erlangen Valency Patternbank

Table 5
Observed and expected frequencies of active verb valency patterns for valency 3 from the Erlangen Valency Patternbank

x	**f_x**	**NP_x**	**x**	**f_x**	**NP_x**	**x**	**f_x**	**NP_x**
1	137	148.42	18	27	32.90	35	13	15.95
2	134	128.03	19	27	31.16	36	13	15.43
3	124	111.96	20	27	29.58	37	12	14.94
4	116	99.02	21	26	28.12	38	12	14.48
5	104	88.42	22	21	26.79	39	12	14.03
6	103	79.60	23	21	25.55	40	11	13.61
7	92	72.17	24	20	24.41	41	11	13.21
8	87	65.84	25	19	23.36	42	10	12.83
9	84	60.40	26	18	22.38	43	10	12.47
10	66	55.67	27	16	21.46	44	9	12.13
11	65	51.54	28	16	20.61	45	9	11.80
12	60	47.90	29	16	19.81	46	9	11.49
13	55	44.67	30	16	19.07	47	9	11.18
14	43	41.79	31	15	18.37	48	9	10.90
15	33	39.22	32	15	17.71	49	8	10.62
16	30	36.90	33	15	17.09	50	8	10.36
17	29	34.80	34	15	16.50	51	8	10.11

x	f_x	NP_x
52	8	9.86
53	8	9.63
54	8	9.41
55	8	9.19
56	7	8.99
57	7	8.79
58	7	8.59
59	7	8.41
60	7	8.23
61	7	8.06
62	7	7.89
63	7	7.73
64	7	7.58
65	7	7.43
66	7	7.28
67	6	7.14
68	6	7.00
69	6	6.88
70	6	6.75
71	5	6.63
72	5	6.51
73	5	6.39
74	5	6.28
75	5	6.17
76	5	6.06
77	5	5.96
78	5	5.86
79	5	5.76
80	5	5.67
81	4	5.58
82	4	5.49
83	4	5.40
84	4	5.32
85	4	5.23
86	4	5.15
87	4	5.07
88	4	5.00
89	4	4.92
90	4	4.85

x	f_x	NP_x
91	4	4.78
92	4	4.71
93	3	4.64
94	3	4.58
95	3	4.51
96	3	4.45
97	3	4.39
98	3	4.33
99	3	4.27
100	3	4.21
101	3	4.16
102	3	4.10
103	3	4.05
104	3	4.00
105	3	3.95
106	3	3.90
107	3	3.85
108	3	3.80
109	3	3.75
110	3	3.71
111	3	3.66
112	3	3.62
113	3	3.57
114	3	3.53
115	3	3.49
116	3	3.45
117	3	3.41
118	3	3.37
119	3	3.33
120	3	3.29
121	3	3.26
122	3	3.22
123	3	3.19
124	3	3.15
125	2	3.12
126	2	3.08
127	2	3.05
128	2	3.02
1299	2	2.99

x	f_x	NP_x
130	2	2.95
131	2	2.92
132	2	2.89
133	2	2.86
134	2	2.83
135	2	2.81
136	2	2.78
137	2	2.75
138	2	2.72
139	2	2.70
140	2	2.67
141	2	2.64
142	2	2.62
143	2	2.59
144	2	2.57
145	2	2.54
146	2	2.52
147	2	2.50
148	2	2.47
149	2	2.45
150	2	2.43
151	2	2.41
152	2	2.38
153	2	2.36
154	2	2.34
155	2	2.32
156	2	2.30
157	2	2.28
158	2	2.26
159	2	2.24
160	2	2.22
161	2	2.20
162	2	2.18
163	2	2.16
164	2	2.15
165	2	2.13
166	2	2.11
167	2	2.09
168	2	2.08

x	**f_x**	**NP_x**
169	2	2.06
170	2	2.04
171	2	2.03
172	2	2.01
173	2	1.99
174	2	1.98
175	2	1.96
176	2	1.95
177	2	1.93
178	2	1.92
179	2	1.90
180	2	1.89
181	2	1.87
182	2	1.86
183	2	1.84
184	2	1.83
185	2	1.82
186	2	1.80
187	2	1.79
188	2	1.78
189	2	1.76
190	2	1.75
191	2	1.74
192	2	1.72
193	2	1.71
194	1	1.70
195	1	1.69
196	1	1.68
197	1	1.66
198	1	1.65
199	1	1.64
200	1	1.63
201	1	1.62
202	1	1.61
203	1	1.60
204	1	1.58
205	1	1.57
206	1	1.56
207	1	1.55

x	**f_x**	**NP_x**
208	1	1.54
209	1	1.53
210	1	1.52
211	1	1.51
212	1	1.50
213	1	1.49
214	1	1.48
215	1	1.47
216	1	1.46
217	1	1.45
218	1	1.44
219	1	1.44
220	1	1.43
221	1	1.42
222	1	1.41
223	1	1.40
224	1	1.39
225	1	1.38
226	1	1.37
227	1	1.36
228	1	1.36
229	1	1.35
230	1	1.34
231	1	1.33
232	1	1.32
233	1	1.32
234	1	1.31
235	1	1.30
236	1	1.29
237	1	1.28
238	1	1.28
239	1	1.27
240	1	1.26
241	1	1.25
242	1	1.25
243	1	1.24
244	1	1.23
245	1	1.23
246	1	1.22

x	**f_x**	**NP_x**
247	1	1.21
248	1	1.21
249	1	1.20
250	1	1.19
251	1	1.19
252	1	1.18
253	1	1.17
254	1	1.17
255	1	1.16
256	1	1.15
257	1	1.15
258	1	1.14
259	1	1.13
260	1	1.13
261	1	1.12
262	1	1.12
263	1	1.11
264	1	1.10
265	1	1.10
266	1	1.09
267	1	1.09
268	1	1.08
269	1	1.08
270	1	1.07
271	1	1.06
272	1	1.06
273	1	1.05
274	1	1.05
275	1	1.04
276	1	1.04
277	1	1.03
278	1	1.03
279	1	1.02
280	1	1.02
281	1	1.01
282	1	1.01
283	1	1.00
284	1	1.00
285	1	0.99

x	f_x	NP_x
286	1	0.99
287	1	0.98
288	1	0.98
289	1	0.97
290	1	0.97
291	1	0.96
292	1	0.96
293	1	0.95
294	1	0.95
295	1	0.94
296	1	0.94
297	1	0.93
298	1	0.93
299	1	0.93
300	1	0.92
301	1	0.92
302	1	0.91
303	1	0.91
304	1	0.90
305	1	0.90
306	1	0.90
307	1	0.89
308	1	0.89
309	1	0.88
310	1	0.88
311	1	0.88
312	1	0.87
313	1	0.87
314	1	0.86
315	1	0.86
316	1	0.86
317	1	0.85
318	1	0.85
319	1	0.84
320	1	0.84
321	1	0.84
322	1	0.84
323	1	0.83
324	1	0.83

x	f_x	NP_x
325	1	0.82
326	1	0.82
327	1	0.82
328	1	0.81
329	1	0.81
330	1	0.81
331	1	0.80
332	1	0.80
333	1	0.80
334	1	0.79
335	1	0.79
336	1	0.79
337	1	0.78
338	1	0.78
339	1	0.78
340	1	0.77
341	1	0.77
342	1	0.77
343	1	0.76
344	1	0.76
345	1	0.76
346	1	0.75
347	1	0.75
348	1	0.74
349	1	0.74
350	1	0.74
351	1	0.74
352	1	0.73
353	1	0.73
354	1	0.73
355	1	0.73
356	1	0.72
357	1	0.72
358	1	0.72
359	1	0.71
360	1	0.71
361	1	0.71
362	1	0.71
363	1	0.70

x	f_x	NP_x
364	1	0.70
365	1	0.70
366	1	0.70
367	1	0.69
368	1	0.69
369	1	0.69
370	1	0.68
371	1	0.68
372	1	0.68
373	1	0.68
374	1	0.67
375	1	0.67
376	1	0.67
377	1	0.67
378	1	0.66
379	1	0.66
380	1	0.66
381	1	0.66
382	1	0.65
383	1	0.65
384	1	0.65
385	1	0.65
386	1	0.64
387	1	0.64
388	1	0.64
389	1	0.64
390	1	0.64
391	1	0.63
392	1	0.63
393	1	0.63
394	1	0.63
395	1	0.62
396	1	0.62
397	1	0.63
398	1	0.62
399	1	0.61
400	1	0.61
401	1	0.61
402	1	0.61

x	**f_x**	**NP_x**
403	1	0.61
404	1	0.60
405	1	0.60
406	1	0.60
407	1	0.60
408	1	0.60
409	1	0.59
410	1	0.59
411	1	0.59
412	1	0.59
413	1	0.59
414	1	0.58
415	1	0.58
416	1	0.58
417	1	0.58
418	1	0.58
419	1	0.58
420	1	0.58
421	1	0.57
422	1	0.57
423	1	0.57
424	1	0.56
425	1	0.56
426	1	0.56
427	1	0.56
428	1	0.56
429	1	0.55
430	1	0.55
431	1	0.55
432	1	0.55
433	1	0.55
434	1	0.55
435	1	0.54
436	1	0.54
437	1	0.54
438	1	0.54
439	1	0.54
440	1	0.53
441	1	0.53
442	1	0.53
443	1	0.53
444	1	0.53
445	1	0.53
446	1	0.52
447	1	0.52
448	1	0.52
449	1	0.52
450	1	0.52
451	1	0.52
452	1	0.51
453	1	0.51
454	1	0.51
455	1	0.51
456	1	0.51
457	1	0.51
458	1	0.51
459	1	0.50
460	1	0.50
461	1	0.50
462	1	0.50
463	1	0.50
464	1	0.50
465	1	0.49
466	1	0.49
467	1	0.49
468	1	0.49
469	1	0.49
470	1	0.49
471	1	0.49
472	1	0.48
473	1	0.48
474	1	0.48
475	1	0.48
476	1	0.48
477	1	0.48
478	1	0.48
479	1	0.47
480	1	0.47
481	1	0.47
482	1	0.47
483	1	0.47
484	1	0.47
485	1	0.47
486	1	0.46
487	1	0.46
488	1	0.46
489	1	0.46
490	1	0.46
491	1	0.46
492	1	0.46
493	1	0.45
494	1	0.45
495	1	0.45
496	1	0.45
497	1	0.45

$a = 1.4531$, $b = 8.3390$, $n = 497$,
$\chi^2 = 154.3648$, $DF = 380$, $P(\chi^2) \approx 1.00$

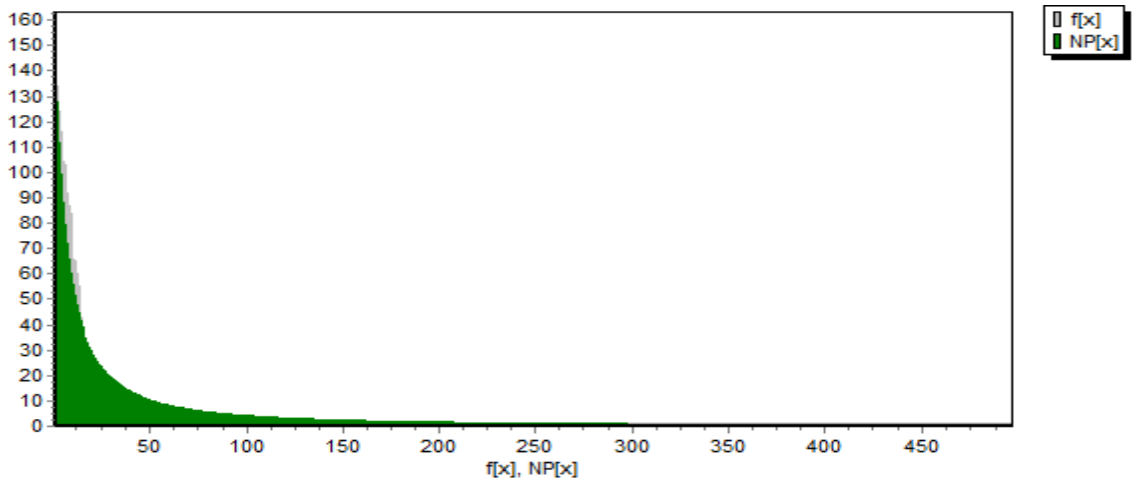


Figure 5: Observed and expected frequencies of active verb valency patterns for valency 2 from the Erlangen Valency Patternbank

5. Discussion

Both hypotheses on English valency patterns can be corroborated on the basis of the statistical fitting. For hypothesis (a), the truncated form of the distribution calls for an explanation. Certainly the list of the patterns is not complete as there might be some unrecognized or unaccounted forms which should augment the set of different pattern elements.

Hypothesis (b) leads also to two truncated distributions. This can be interpreted as a second indicator that the sample does not show the complete picture. Obviously, the tails of the distributions for valency 1 and 2 should be longer. The fact, that verbs without diverse valency patterns beside mono- and ditransitive patterns are not included in the sample (see section 2) cannot account for this: The head of the distributions would simply get steeper, as these patterns are the most frequent ones: the pattern *SCU* + VHC_{act} with a frequency of 358 lexemes for valency 1 and *SCU* + VHC_{act} + *NP* for valency 2, which has a count of 467 lexemes. The "missing" data is rather to be looked for in the classes of prepositional complements and phrasal verbs.

In general, the field of valency patterns could be ploughed much deeper, especially with due regard for the semantic interface or to the other side of the coin of valency: case marking and inflection. However, when the honored reader comes to these lines, he will be certainly already suffused with new ideas for his projects.

References

Altmann, Gabriel (1985). Semantische Diversifikation. *Folia Linguistica 19. 177-200.*

Altmann, Gabriel (1991). Modelling diversification phenomena in language. In: Rothe, Ursula (ed.), *Diversification Processes in Language: Grammar*: *33-46.* Hagen: Rottmann.

Altmann, Gabriel (2000). *Altmann-Fitter 2.1 for Windows 95*. Lüdenscheid: RAM-Verlag.

Altmann, Gabriel & Best, Karl-Heinz (1996). Zur Länge der Wörter in deutschen Texten. In: Schmidt, Peter (ed.), *Glottometrika 15*: *Issues in General Linguistic theory and the theory of word length: 166-180.* Trier: WVT Wissenschaftlicher Verlag Trier.

Altmann, Gabriel; Erat, Erkan & Hřebíček, Luděk (1996). Word length distribution in Turkish texts. In: Schmidt, Peter (ed.), *Glottometrika 15*: *Issues in General Linguistic theory and the theory of word length: 185-204.* Trier: WVT Wissenschaftlicher Verlag Trier.

Best, Karl-Heinz (1996). Zur Wortlängenhäufigkeit in schwedischen Pressetexten. In: Schmidt, Peter (ed.), *Glottometrika 15*: *Issues in General Linguistic theory and the theory of word length: 147-157.* Trier: WVT Wissenschaftlicher Verlag Trier.

Best, Karl-Heinz (2005a). Morphlänge. In: Köhler, Reinhard, Altmann, Gabriel & Rajmund G. Piotrowski (eds.), *Quantitative Linguistik/Quantitative Linguistics: Ein internationales Handbuch/An international handbook: 255-260.* Berlin/New York: de Gruyter. (Handbücher zur Sprach- und Kommunikationswissenschaft /Handbooks of Linguistics and Communicative Science 27).

Best, Karl-Heinz (2005b). Wortlänge. In: Köhler, Reinhard, Altmann, Gabriel & Rajmund G. Piotrowski (eds.), *Quantitative Linguistik/Quantitative Linguistics: Ein internationales Handbuch/An international handbook: 260-273.* Berlin/New York: de Gruyter. (Handbücher zur Sprach- und Kommunikationswissenschaft /Handbooks of Linguistics and Communicative Science 27).

Best, Karl-Heinz & Brynjólfsson, Einar (1997). Wortlängen in isländischen Briefen und Pressetexten. *Skandinavistik: Zeitschrift für Sprache, Literatur und Kultur der nordischen Länder 72(1). 24–40.*

Engel, Ulrich & Helmut Schumacher (1976). *Kleines Valenzlexikon deutscher Verben*. Tübingen: Narr.

Fillmore, Charles J. (1968). The Case for Case. In: Emmon Bach & Robert T. Harms (eds.), *Universals in linguistic theory: 1-88.* New York: Holt, Rinehart and Winston.

Fillmore, Charles J. (1970). The Grammar of 'Hitting' and 'Breaking'. In: Jacobs, Roderick A. & Peter S. Rosenbaum (eds.), *Readings in Transformational Gram-*

mar: 120-134. Waltham, Mass.: Ginn.

Fillmore, Charles J. (1971). Some Problems for Case Grammar. In: O'Brien, Richard, ed. *Report of the Twenty-Second Annual Round Table Meeting on Linguistics and Language Studies: 33-56.* Washington, D.C. University Press.

Fillmore, Charles J. (1977). Scenes-and-Frames Semantics. In: Zampolli, Antonio, (ed.), *Linguistic Structures Processing: 55-81*. Amsterdam/New York/Oxford: North Holland.

Fillmore, Charles J. (1982). Frame Semantics. In: Linguistic Society of Korea (ed.), *Linguistics in the Morning Calm: 111-137*. Seoul: Hanshin Publishing Co.

Fillmore, Charles J. (1985). Frames and the semantics of understanding. *Quaderni di Semantica 6(2), 222-254.*

Fillmore, Charles J., Christopher R. Johnson & Miriam R. L. Petruck (2003). Background to FrameNet. *International Journal of Lexicography 16(3) 235-250.*

Fillmore, Charles J. (2007). Valency issues in FrameNet. In: Thomas Herbst & Katrin Götz-Votteler (eds.), *Valency: Theoretical, Descriptive and Cognitive Issues: 129-160*. Berlin/ New York: de Gruyter. (Trends in Linguistics. Studies and Monographs 187).

Helbig, Gerhard & Wolfgang Schenkel (1968). *Wörterbuch zur Valenz und Distribution deutscher Verben*. Leipzig: VEB Verlag Enzyklopädie.

Helbig, Gerhard & Wolfgang Schenkel (1991). *Wörterbuch zur Valenz und Distribution deutscher Verben*. 8th ed. Leipzig: VEB Verlag Enzyklopädie.

Herbst, Thomas (2004). *A Valency Dictionary of English: Corpus-based Analysis of the Complementation Patterns of English Verbs, Nouns and Adjectives.* In cooperation with David Heath, Ian Roe, and Dieter Götz. Berlin: de Gruyter. (Topics in English Linguistics 40).

Herbst, Thomas & Peter Uhrig (2009). *Erlangen Valency Patternbank*: *a corpus-based research tool for work on valency and argument structure constructions*. [http://www.patternbank.uni-erlangen.de/cgi-bin/patternbank.cgi, retrieved on 29/12/2012].

Herbst, Thomas & Schüller, Susen (2008). *Introduction to Syntactic Analysis: A Valency Approach*. Tübingen: Narr.

Herbst, Thomas (2009). Introduction. Herbst, Thomas & Peter Uhrig. *Erlangen Valency Patternbank*: *a corpus-based research tool for work on valency and argument structure constructions*. [http://www.patternbank.uni-erlangen.de/cgi-bin/patternbank.cgi?do=introtxt; retrieved on 29/12/2012].

Köhler, Reinhard (2005). Synergetic linguistics. In: Köhler, Reinhard, Gabriel Altmann & Rajmund G. Piotrowski (eds.), *Quantitative Linguistik/Quantitative Linguistics: Ein internationales Handbuch/An international handbook: 760-774.* Berlin/New York: de Gruyter. (Handbücher zur Sprach- und Kommunikationswissenschaft /Handbooks of Linguistics and Communicative Science 27).

Köhler, Reinhard (2012). *Quantitative Syntax Analysis*. Berlin/Boston: de Gruyter.

(Quantitative linguistics 65).

Rothe, Ursula (1991a). Diversification of the case in German: genitive. In: Rothe, Ursula (ed.), *Diversification Processes in Language: Grammar*: *140-156.* Hagen: Rottmann.

Rothe, Ursula (ed.) (1991b). *Diversification Processes in Language: Grammar*. Hagen: Rottmann.

Schumacher, Helmut, Jacqueline Kubczak, Renate Schmidt & Vera de Ruiter (2004). *VALBU - Valenzwörterbuch deutscher Verben*. Tübingen: Narr. (Studien zur deutschen Sprache 31).

Somers, Harald (1987). *Valency and case in computational linguistics*. Edinburgh: Edinburgh University Press. (Edinburgh Information Technology Series 3).

Sommerfeldt, Karl-Ernst & Herbert Schreiber (1996). *Wörterbuch der Valenz etymologisch verwandter Wörter*: *Verben, Adjektive, Substantive*. Tübingen: Niemeyer.

Steiner, Petra (2009). Diversification in Icelandic inflectional paradigms. In: Köhler, Reinhard (ed.), *Issues in Quantitative Linguistics: 126-154.* Lüdenscheid: RAM-Verlag. (Studies in Quantitative Linguistics 5).

Steiner, Petra (forthcoming). *Quantitative Laws of Valency and Case.*

Steiner, Petra & Claudia Prün (2007). The effects of diversification and unification on the inflectional paradigms of German nouns. In: Grzybek, Peter &. Reinhard Köhler (eds.), *Exact methods in the study of language and text: dedicated to Professor Gabriel Altmann on the occasion of his 75th birthday: 623-631*. Berlin: de Gruyter.

Wimmer, Gejza & Altmann, Gabriel (1996). The theory of word length: some results and generalizations. In: Schmidt, Peter (ed.), *Glottometrika 15*: *Issues in general linguistic theory and the theory of word length: 112-133.* Trier: WVT Wissenschaftlicher Verlag Trier.

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

Quantitative Untersuchungen zur Valenz deutscher Substantive

Tim Duwaerts, Gereon Ullmann, TRIER

1. Einleitung

Die Valenz eines Substantivs beschreibt die Anzahl an Argumentstellen, welche von ihm geöffnet werden. Vorangegangene Untersuchungen haben gezeigt, dass im Bereich der Syntax verschiedene Phänomene zu finden sind, die bestimmten sprachlichen Gesetzmäßigkeiten zu folgen scheinen. Ergebnisse aus (Köhler, 2005), einer Untersuchung zur Valenz deutscher Verben, untermauern ein weiteres Mal die Existenz von Sprachgesetzen im Bereich der Syntax. Parallel dazu soll nun durch eine quantitative Untersuchung zur Valenz der Substantive geprüft werden, ob auch dieser Bereich den entsprechenden Gesetzmäßigkeiten unterworfen ist. Als Datengrundlage dient das *Wörterbuch zur Valenz und Distribution der Substantive* (Sommerfeld/Schreiber, 1980).

Dieses Wörterbuch umfasst insgesamt 750 Einträge und enthält, neben Informationen zur Polysemie und der Anzahl möglicher Satzbaupläne, Angaben zur Valenz und zur semantischen Form der Aktanten des entsprechenden Substantivs. Als Beispiel für den Aufbau der Daten soll hier der Eintrag zum Substantiv „*Bau*“ dienen:

Bau
V 1 = 'Prozeß des Errichtens'
1.1 → (2)
1.2 → Sg, pS (durch)
1.3 → fest: Sg + pS (durch)
(der Bau der Brücke durch den Großbetrieb)
V 2 = 'Struktur'
1.1 → (1)
1.2 → Sg
1.3 → (der Bau des Dramas)

Im Zuge der Untersuchung wurden die Substantive manuell aus diesem Valenzwörterbuch extrahiert. Die Stichprobe wies insgesamt 30 Substantive auf, die aufgrund ihrer zweifelhaften Valenzangaben aus der Stichprobe herausgenommen wurden, was den Umfang auf 720 Substantive verringerte. Die durch Polysemie bedingten Varianten mitgezählt, betrug der Stichprobenumfang insgesamt 1.192 Substantive.

Ausgehend von den Erkenntnissen in (Köhler, 2005) wurde auch hier angenommen, dass die Verteilung der Satzbaupläne durch eine Zipf-Mandelbrot-

Verteilung angenähert werden kann. Dieser Punkt bildete den ersten Untersuchungsgegenstand dieser Studie. Im Zusammenhang mit der Valenz wurden zwei Aspekte untersucht: Das Verhalten der mittleren Valenz bei unterschiedlichem Grad an Polysemie sowie das Verhalten der Valenz bei unterschiedlichen Wortlängen. Darüber hinaus wurde geprüft, ob die vorliegenden Daten den Zusammenhang zwischen Wortlänge und Polysemie bestätigen und ob der Grad an Polysemie sich auf die Verteilung der Satzbaupläne auswirkt.

Neben dem bereits gut belegten Zusammenhang zwischen Wortlänge und Polysemie (Genzor, 1999) scheinen zu den anderen Untersuchungen noch keine Studien vorzuliegen.

2. Verteilung der Satzbaupläne

Verglichen mit der Datenlage in (Köhler, 2005) fällt auf, dass im Bereich der Valenz deutscher Substantive deutlich weniger unterschiedliche Satzbaupläne existieren als im Bereich der Verben. Waren es bei diesen 205, werden bei den Substantiven nur 47 verschiedene Satzbauplan-Arten unterschieden. Trotz dieses

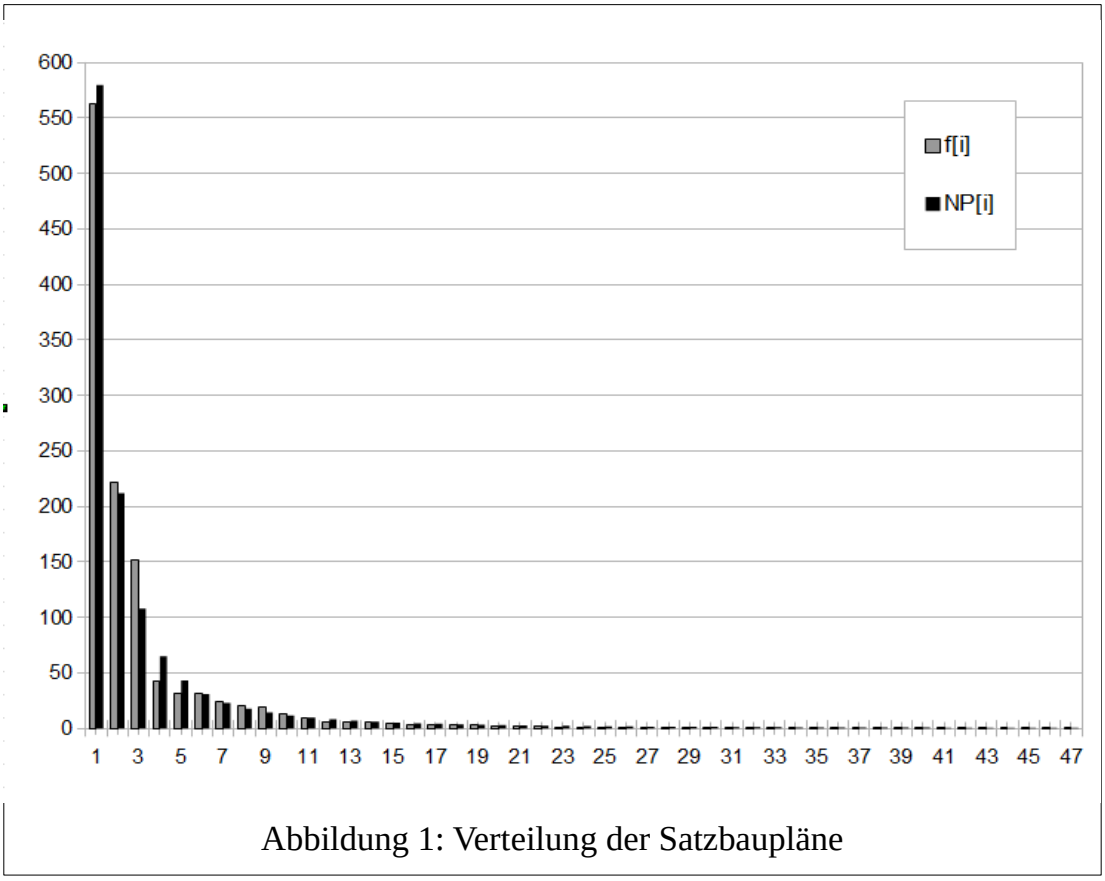


Abbildung 1: Verteilung der Satzbaupläne

relativ großen Unterschieds und der geringen Anzahl verbleibender unterschiedlicher Satzbauplanarten können die Häufigkeitsverhältnisse auch in diesem Falle

noch sehr gut durch die Zipf-Mandelbrot-Verteilung beschrieben werden (Abbildung 1 und Tabelle 2). Der errechnete r^2-Wert von 0,9918 bestätigt dies.

Mit einer Häufigkeit von 563 ist „Sg, pS“, also ein Substantiv im Genitiv mit präpositionalem Substantiv als zweiten Aktanten (vgl. Tabelle 1: Abkürzungen), der am häufigsten auftretende Satzbauplan in der Stichprobe. Die zweithäufigste Kombinationsmöglichkeit bildet mit 222 gezählten Vorkommen das Substantiv im Genitiv „Sg“. Mit 25 unikal auftretenden Satzbauplänen folgt die Verteilung beinahe exakt den Formulierungen des Zipf'schen Gesetzes.

3. Polysemie und Valenz

Das verwendete Wörterbuch weist jedem Substantiv einen Satzbauplan zu. Existieren bedingt durch Polysemie mehrere Varianten eines Substantivs, so können sich die Satzbaupläne dieser Varianten voneinander unterscheiden.

Bei steigendem Grad an Polysemie konnte in den vorliegenden Daten eine Abnahme der mittleren Valenz beobachtet werden (Abbildung 2: Zusammenhang zwischen der Polysemie und der mittleren Valenz). Der Verlauf der Abnahme kann durch die Funktion $y = 2.4817\exp(-0.1041x)$ angenähert werden. Die Güte der Annäherung beträgt $r^2 = 0.80$ und liegt damit knapp unter der konventionellen Grenze von 85 %.

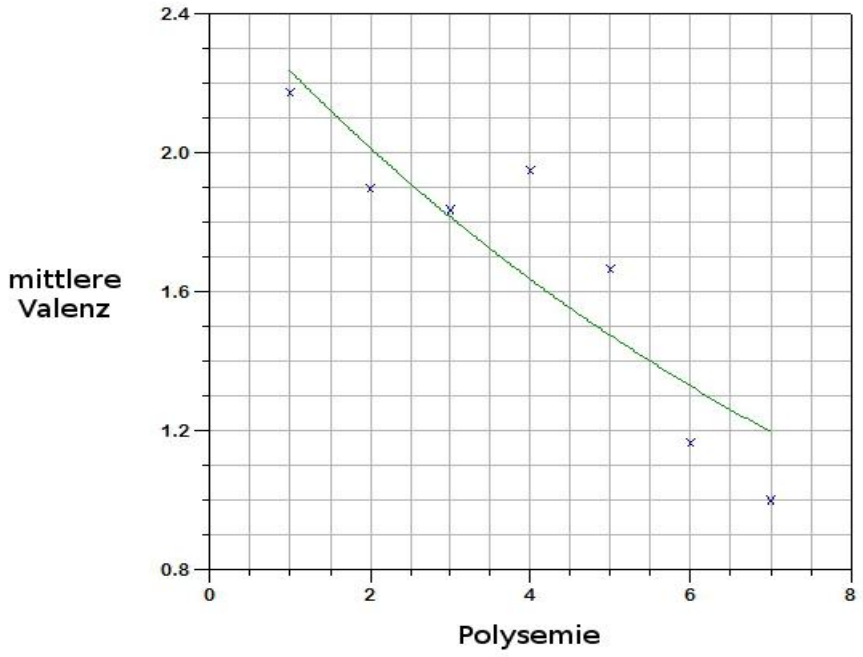


Abbildung 2: Zusammenhang zwischen der Polysemie und der mittleren Valenz

4. Wortlänge und Valenz

Auch in anderen quantitativen Untersuchungen, wie beispielsweise in Studien zur Wortfrequenz oder Polysemie, hat sich die Wortlänge als relevante Größe erwiesen. Aus diesem Grund schien es sinnvoll der Frage nachzugehen, ob sich dieser Faktor auch im Zusammenhang mit der Valenz als Indikator für Ab- oder Zunahme der mittleren Valenz eignet. Da sowohl die Wortlänge als auch die

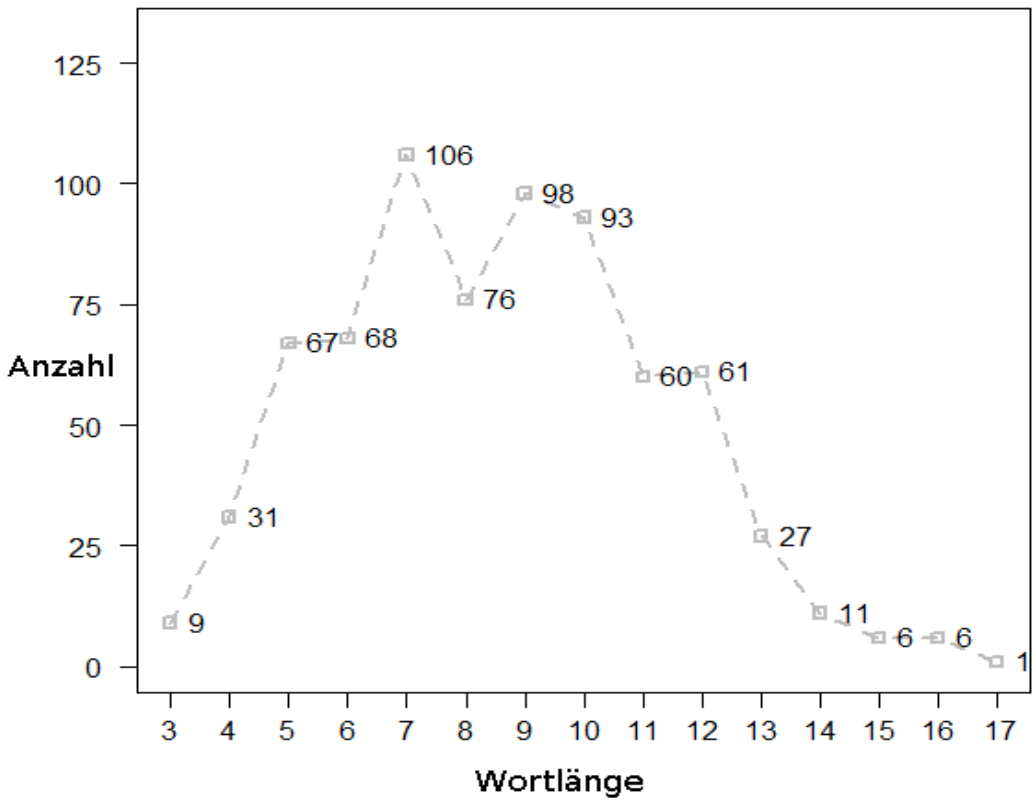


Abbildung 3: Verteilung der Wortlängen

Valenz bei steigender Polysemie abnehmen (siehe Abschnitt 5), wurde untersucht, ob hier ein näherungsweise linearer Zusammenhang zwischen den beiden Größen festgestellt werden kann. Aufgrund der Tatsache, dass für Wörter mit einer Länge von weniger als vier sowie für Wörter mit einer Länge von mehr als zwölf Buchstaben keine ausreichende Datenmenge zur Verfügung stand, wurden die Parameter nur über den Abschnitt der Wörter aus den repräsentativen Gruppierungsbereichen (Anzahl > 30) von vier bis zwölf untersucht. Tatsächlich lieferte die daraus resultierende Formel $y = 1.4166x^{0.0754}\exp(-0.0242x)$ ($r^2 = 0.93$) eine sehr gute Annäherung an die real gemessenen Werte für die Gruppierung der Wortlängen vier bis zwölf (s. Tabelle 6).

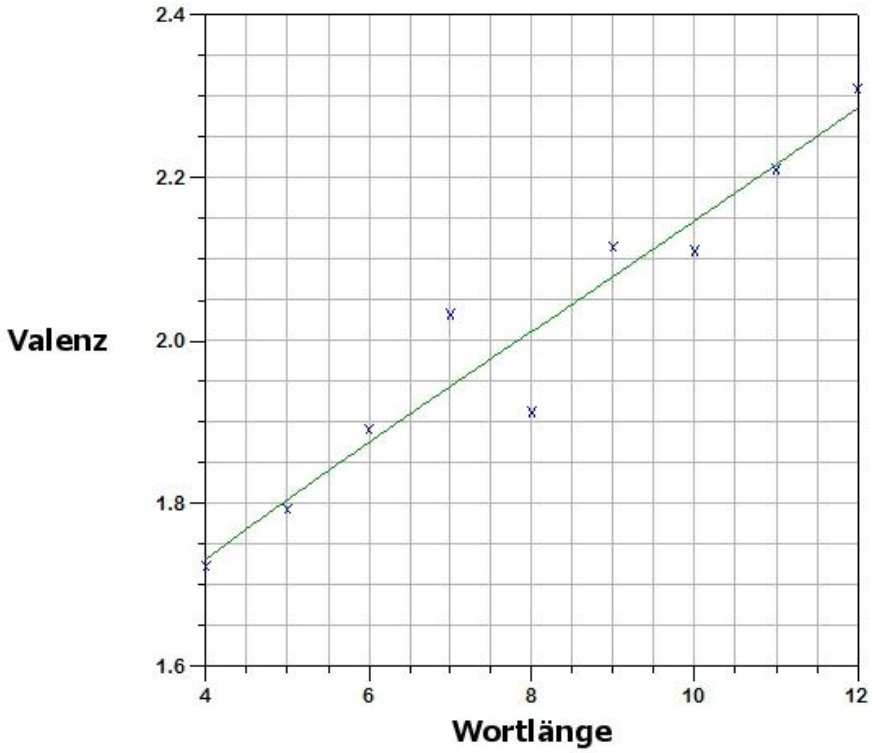


Abbildung 4: Zusammenhang zwischen der Wortlänge und der mittleren Valenz

5. Wortlänge und Polysemie

Im Zuge der Untersuchung ergab sich die Möglichkeit, die Verteilung der Wortlängen im Zusammenhang mit der Polysemie zu untersuchen. Wie bereits angesprochen handelt es sich hierbei um einen sehr gut bestätigten (vgl. Köhler 1986) Zusammenhang der besagt, dass die Polysemie mit steigender Wortlänge abnimmt (Abbildung 5: Zusammenhang zwischen der Wortlänge und der Polysemie). Zu erklären ist dies mit der Tatsache, dass die Bedeutung längerer Wörter meist von recht spezifischer Natur ist, wohingegen kürzere Wörterin der Regel eine allgemeinere Bedeutung aufweisen. Je allgemeiner die Bedeutung eines Wortes ist, desto unabhängiger ist sie vom kontextuellen semantischen Umfeld und kann entsprechend häufiger verwendet werden.

Wie erwartet, wurde der Zusammenhang auch durch unsere Daten bestätigt (Tabelle 5). Aufgrund der in Abschnitt 4 angesprochenen Häufigkeitsausprägung der Wortlängen wurden auch in diesem Punkt jene Wortlängen ignoriert, deren Häufigkeit einen Wert von 30 unterschritten. Der Verlauf lässt sich approximieren durch

$$y = 3.0717x^{-0.2949},\ r^2 = 0.89.$$

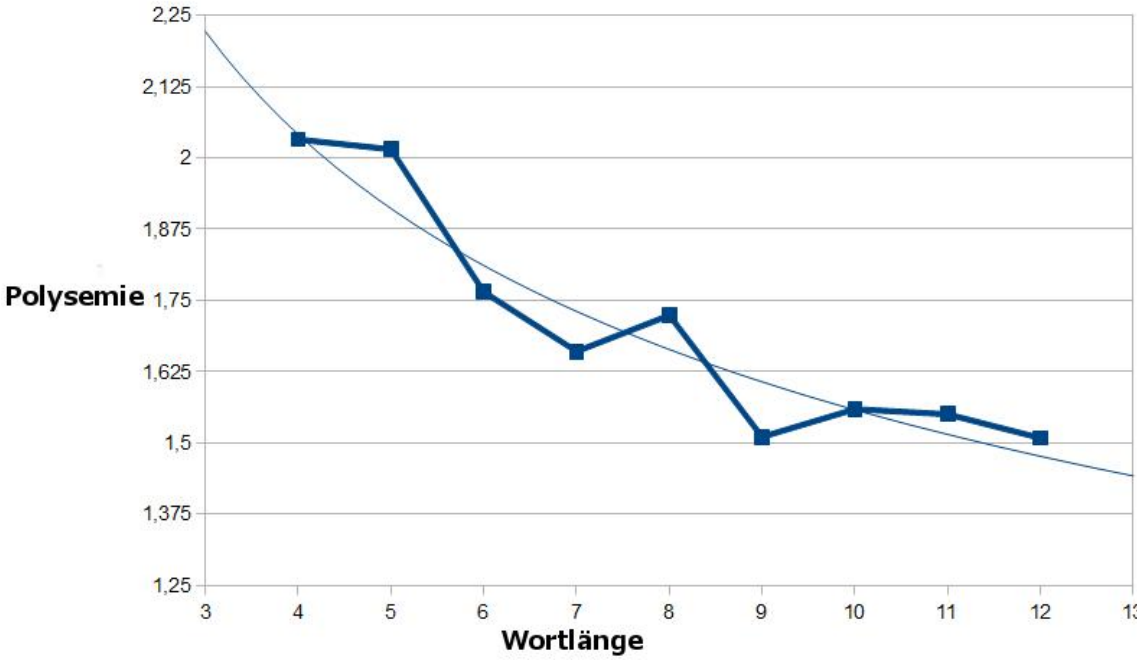


Abbildung 5: Zusammenhang zwischen der Wortlänge und der Polysemie

6. Polysemie und Anzahl der Satzbaupläne

Neben den Untersuchungen zur Valenz wurde auch überprüft, welchen Einfluss die Polysemie auf die Anzahl möglicher Satzbaupläne pro Wort ausübt. Eine Möglichkeit zur Bedeutungsunterscheidung polysemer Wörter ist die Verwendung unterschiedlicher Satzbaupläne, durch die ein grammatikalischer sowie ein semantischer Kontext festgelegt wird, in den die Wörter eingebettet werden. Aus diesem Grund war anzunehmen, dass mit steigendem Grad der Polysemie auch eine steigende Anzahl möglicher Satzbaupläne einhergeht, die für die semantische Disambiguierung notwendig ist.

Da die semantischen Kategorien der Satzbauplan-Elemente bei der Datenerhebung nicht berücksichtigt wurden, war eine feine Unterscheidung der syntaktischen Elemente nicht möglich. Aus diesem Grund konnten maximal vier syntaktisch verschiedene Satzbauplan-Strukturen beobachtet werden. Trotz dieser fehlenden Unterteilung zeichnete sich in den Daten ein deutlicher Verlauf ab, der unserer Vermutung entsprach (Abbildung 6).

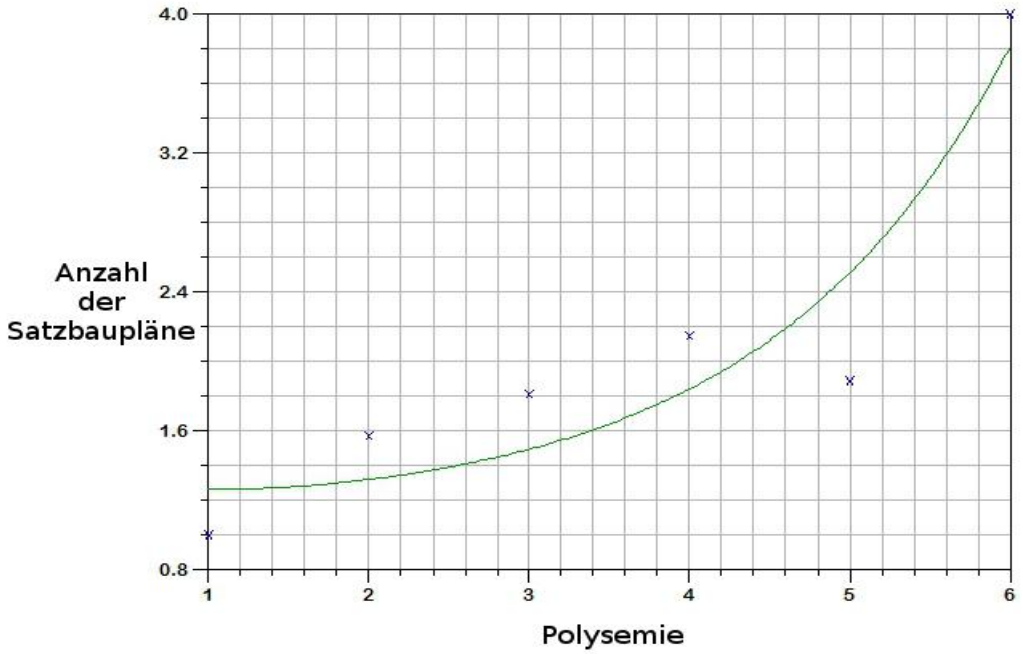


Abbildung 6: Zusammenhang zwischen der Polysemie und der Anzahl der Satzbaupläne

Der Verlauf dieser Abhängigkeit kann durch $y = 0{,}1234\ x^{-0{,}9435}\ e^{0{,}7730} + 1$ mit $r^2 = 0{,}86$ ausgedrückt werden.

7. Anzahl der Satzbaupläne und Polysemie

Interessant schien es zu überprüfen, ob der in Abschnitt 6 gezeigte Zusammenhang zwischen Polysemie und der Anzahl der Satzbaupläne auch umgekehrt bestätigt werden kann. Tatsächlich konnte durch die Funktion $y = 0.3956x^{1.379} * e^{0.099} + 1$ der Verlauf sehr gut angenähert werden, was durch ein Bestimmtheitsmaß von $r^2 = 0.99$ bestätigt wurde (s. Abbildung 7). Allerdings ist hier zu beachten, dass insgesamt nur zwei Substantive die maximale Anzahl von vier Satzbauplänen aufwiesen. Aufgrund der Seltenheit wurde diese Klasse aus der Stichprobe entfernt. Für die repräsentativen Klassen von eins bis drei Satzbauplänen, konnte der Zusammenhang zwischen der Anzahl der Satzbaupläne und der Polysemie durch folgende lineare Gleichung angenähert werden $y = 0{,}9890x + 0{,}3755$, die ein Bestimmtheitsmaß von $r^2 = 0{,}998$ aufwies.

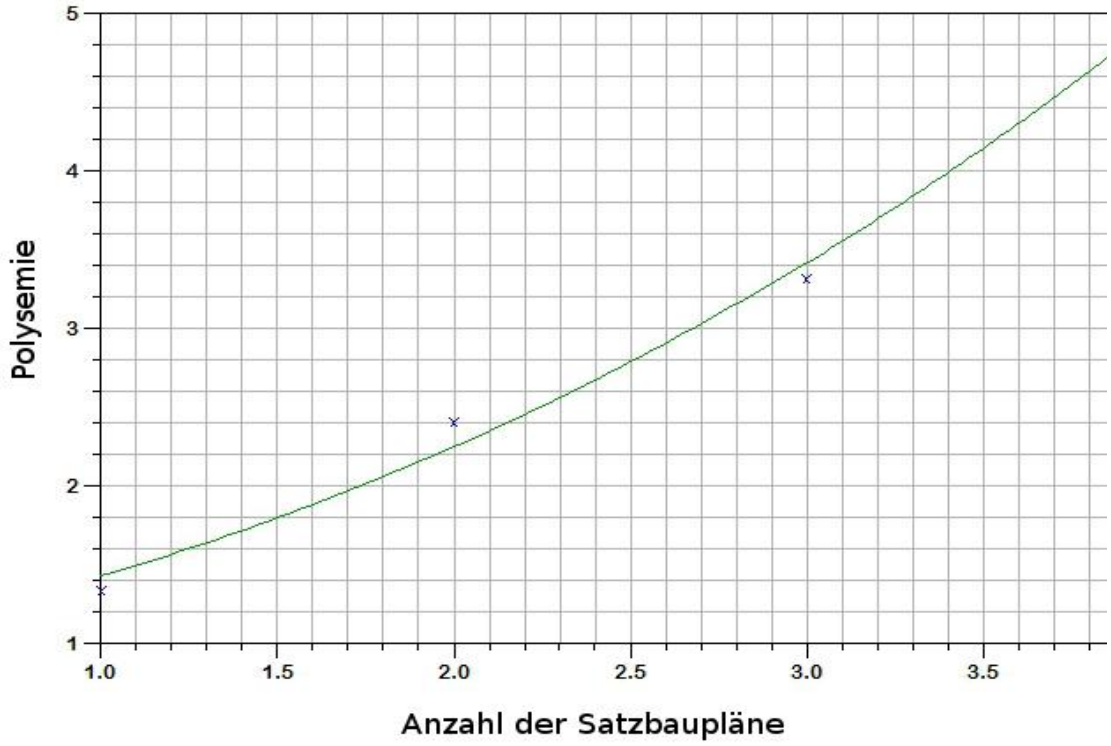


Abbildung 7: Zusammenhang zwischen der Anzahl der Satzbaupläne und der Polysemie

8. Fazit

Aufgrund der vorliegenden Resultate lässt sich sagen, dass der Bereich der Substantiv-Valenz in jedem Fall ein lohnenswertes Feld ist, um nach Gesetzmäßigkeiten in der Sprache zu suchen. Die hier erzielten Ergebnisse geben einen Hinweis darauf, dass die Valenz eines Substantives durchaus eine Größe sein kann, die für Vorhersagen in Bezug auf den Grad der Polysemie und Länge eines Wortes herangezogen werden kann. Darüber hinaus scheinen auch die Polysemie und die Anzahl möglicher Satzbaupläne, die ein Substantiv aufweist, in einem berechenbaren Zusammenhang zu stehen.

Selbstverständlich ist die Gültigkeit der in dieser Untersuchung erzielten Ergebnisse nur in Bezug auf Qualität und Repräsentativität der verwendeten Quelle zu interpretieren. Mit Sicherheit wäre es interessant zu sehen, ob die Ergebnisse auch durch eine Stichprobe deutlich größeren Umfangs oder durch die Einbeziehung weiterer Informationen bestätigt werden können. Die Berücksichtigung der semantischen Kategorien von Satzbauplan-Elementen bietet die Möglichkeit, die Ergebnisse, insbesondere in Bezug auf den Zusammenhang zwischen der Polysemie und der Verteilung der Satzbaupläne, weiter zu validieren.

Literatur

Köhler, R. (1986):.*Zur linguistischen Synergetik. Struktur und Dynamik der Lexik.* Bochum: Brockmeyer.

Köhler, R. (1999). Der Zusammenhang zwischen Lexemlänge und Polysemie im Maori. In: Ondrejovič, S., Genzor, J. (Hrsg.): *Pange lingua. Zbornik na poèest´ ViktoraKrupu*: *27-33.* Bratislava, Veda.

Köhler, R.., Altmann, G. (1986). Synergetische Aspekte der Linguistik. *Zeitschrift für Sprachwissenschaft 5, 253-265.*

Sommerfeld, K., Schreiber, H. (1980). *Wörterbuch zur Valenz und Distribution der Substantive.* Leipzig: VEB Bibliographisches Institut.

Anhang

Tabelle 1: Abkürzungen

Symbol	Volle Form	Erklärung
A 1	Aktant 1	-
Adj	Adjektiv	-
Akk	Akkusativ	-
Attr	Attribut	-
Dat	Dativ	-
Inf	Infinitiv	Infinitiv bzw. Infinitivkonstruktion
NS	Nebensatz	-
p	Präposition	-
pS	Präpositionales Substantiv	-
S	Substantiv	-
Sg	Substantiv im Genitiv	-
Sm	Substantiv im Monoflexiv	-
V 1	Variante 1 usw.	Bedeutungsvariante

Tabelle 2: Häufigkeiten der Satzbauplanarten

Satzbauplan	**Vorkommen**
Sg, pS	563
Sg	222
Sg, pS1, pS2	152
Sg, pS / NS	43
Sg, pS/ns/inf,	32
Sg, pS / Inf	31
Sg, pS1, pS2 / NS	24
pS	20
Sg / pS	19
Sg, pS1, pS2 / NS / Inf	13
Sg / pS1, pS2	9
Sg, pS1, pS2 / Inf,	6
Sg, pS1 / NS, pS2	6
Sg, NS / Inf	5
Sg, Inf	4
Sg, pS / Inf / NS	3
Sg, pS1 / Inf, pS2	3
Sg, pS1, pS2, pS3	3
Adj,Sg	3
Sg/pS1, pS2, pS3,	2
Sg, pS1 / NS / Inf, pS2	2
Sg, pS, NS / Inf	2
va-Sg, pS1, pS2 / NS	1
Sg, + pS, NS	1
Sg, pS / NS / Ins	1
Adj,Sg, pS	1
Sg, pS, Inf	1
Sg, pS / Adj	1
Adj	1
pS	1
Sg, pS / Inf / Als	1
Sg, pS, NS	1
Sg, Adj / pS	1
Sg / pS / NS	1
Sg, Adj / Numerale	1
Sg, pS, Adj	1
Sg / pS1, pS2 / Inf,	1
Sg / pS / NS / Inf	1
Sg, pS1/ns/inf, pS2 / NS / Inf	1
Inf	1
pS / NS	1
Sg / Adj	1
Sg / pS / SMD	1
Sm / pS	1
Sm	1
Sg / Sm	1
Sg, pS1, NS / Inf	1

Tabelle 3: Polysemie und mittlere Valenz

Polysemie	mittlere Valenz	erwarteter Wert	Abweichung
1	2,1734	2,2365	0,0631
2	1,8982	2,0155	0,1173
3	1,8357	1,8163	0,0195
4	1,95	1,6368	0,3132
5	1,6667	1,4750	0,1916
6	1,6667	1,3293	0,3374
7	1	1,19790	0,19790

Tabelle 3: Anzahl der Satzbaupläne und mittlere Valenz

Satzbauplan	mittlere Valenz	erwarteter Wert	Abweichung
1	2,1120	2,1405	0,0285
2	1,8460	1,8637	0,0177
3	1,8500	1,6542	0,1958
4	1,3333	1,4954	0,1621

Tabelle 4: Wortlänge und Polysemie

Wortlänge	Polysemie	erwartet	Abweichung
4	2,0323	2,0410	0,0088
5	2,0149	1,9110	0,1039
6	1,7647	1,8110	0,0463
7	1,6604	1,7305	0,0702
8	1,7237	1,6637	0,0599
9	1,5102	1,6069	0,0967
10	1,5591	1,5578	0,0014
11	1,55	1,5146	0,0354
12	1,5082	1,4763	0,0319

Tabelle 5: Wortlänge und Valenz

Wortlänge	Valenz	erwartet	Abweichung
4	1,7226	1,7327	0,0101
5	1,7935	1,8053	0,0118
6	1,8909	1,8752	0,0157
7	2,0318	1,9437	0,0881
8	1,9121	2,0115	0,0994
9	2,1156	2,0792	0,0364
10	2,1102	2,1472	0,0370
11	2,2103	2,2158	0,0055
12	2,3087	2,2851	0,0237

Tabelle 7: Polysemie und Mittelwert der unterschiedlichen Satzbaupläne

Polysemie	Satzbaupläne	erwartet	Abweichung
1	1,0000	1,2673	0,2673
2	1,5701	1,3227	0,2474
3	1,8116	1,4966	0,3150
4	2,1500	1,8440	0,3060
5	1,8889	2,5147	0,6258
6	4,0000	3,8135	0,1865
7	1,0000	6,3516	5,3516

Tabelle 6: Wortlänge und Mittelwert der unterschiedlichen Satzbaupläne

Wortlänge	Satzbauplananzahl	erwartet	Abweichung
3	1,4444	1,5657	0,1212
4	1,7419	1,4998	0,2421
5	1,4328	1,4456	0,0127
6	1,3382	1,3987	0,0605
7	1,2642	1,3570	0,0929
8	1,3684	1,3191	0,0493
9	1,2551	1,2842	0,0291
10	1,1935	1,2517	0,0582
11	1,2333	1,2212	0,0121
12	1,2459	1,1924	0,0535
13	1,2593	1,1650	0,0942
14	1,0000	1,1389	0,1389
15	1,1667	1,1140	0,0527
16	1,1667	1,0900	0,0767
17	1,0000	1,0669	0,0669