International Quantitative Linguistics Conference 2014

20th anniversary
of International Quantitative
Linguistics Association (IQLA)
& Journal of Quantitative
Linguistics (JQL) foundation

Olomouc

Czech Republic

May 29 – June 1

2014

Book of Abstracts

QUALICO 2014

IQLA

International Quantitative Linguistics Association

Released with the support of the projects Innovation of General Linguistics and Communication Theory Studies in Cooperation with Natural Sciences, reg. no.: CZ.1.07/2.2.00/28.0076 and Linguistic a lexicostatistic analysis in cooperation with linguistics, mathematics, biology, psychology, grant no. CZ.1.07/2.3.00/20.0161. Both of these projects are financed by the European Social Fund and the national Budget of the Czech Republic.

esf european social fund in the czech republic · EUROPEAN UNION · MINISTRY OF EDUCATION, YOUTH AND SPORTS · OP Education for Competitiveness

INVESTMENTS IN EDUCATION DEVELOPMENT

Olomouc
Czech Republic
May 29 – June 1, 2014

Book of Abstracts

IQLA
International Quantitative Linguistics Association

Philosophical faculty, Palacký University Olomouc

ISBN: 00-11-22-33-44-55-66

Department of **General Linguistics**

esf european social fund in the czech republic
EUROPEAN UNION
MINISTRY OF EDUCATION, YOUTH AND SPORTS
OP Education for Competitiveness
INVESTMENTS IN EDUCATION DEVELOPMENT

QUALICO 2014

Olomouc, Czech Republic
May 29 – June 1, 2014

QUALICO 2014

BOOK OF ABSTRACTS

Editors:
Martina Benešová (Palacký University, Czech Republic)
Ján Mačutek (Comenius University, Slovakia)
Emmerich Kelih (University of Vienna, Austria)

Program Comittee:
Emmerich Kelih (University of Vienna, Austria)
Ján Mačutek (Comenius University, Slovakia)
Radek Čech (University of Ostrava, Czech Republic)
Relja Vulanović (Kent State University, USA)
George Mikros (National and Kapodistrian University of Athens, Greece)
Reinhard Köhler (University of Trier, Germany)
Peter Grzybek (University of Graz, Austria)
Hermann Moisl (University of Newcastle, UK)
Ivan Obradović (University of Belgrade, Serbia)
Sheila Embleton (York University, Canada)
Jan Andres (Palacký University, Czech Republic)

Local Organizers:
Radek Čech (University of Ostrava, Czech Republic)
Petra Martinková (Palacký University, Czech Republic)
Martina Benešová (Palacký University, Czech Republic)
Nela Urbaniková (Palacký University, Czech Republic)
Dan Faltýnek (Palacký University, Czech Republic)

Students Organizing Team from Palacký University:
Ludmila Lacková, Juliana Zmetáková, Michaela Roubínková, Kristýna Bajerová, Lenka Spáčilová, Jana Sčigulinská, Denisa Schusterová, Tereza Motalová

Copies: 100
Pages: 148
Publisher: Philosophical Faculty of Palacký University, Olomouc
ISBN: 00-11-22-33-44-55-66

Content

Preface 11

Welcome to Olomouc! 13

Quantitative analysis of poetic space:
discrimination of loci in Eugene Onegin by Pushkin 15
Sergey Andreev

The Menzerath-Altmann Law revisited 17
Jan Andres

The dependency between the variance of word length and word frequency 19
Gemma Bel-Enguix, Kirsten Bohn, Ramon Ferrer-i-Cancho

Distributional models of the verbal predicate-argument structure 21
Andrei Beliankou

Menzerath-Altmann Law in differently segmented text 23
Martina Benešová, Dan Faltýnek, Lukáš H. Zámečník

Application of the Menzerath-Altmann Law to contemporary
written Japanese – short story style 25
Denis Biryukov, Martina Benešová

Impact of semantics on case diversification 27
Radek Čech, Emmerich Kelih, Ján Mačutek

Factors of readability of Polish texts: a psycholinguistic study 29
Edyta Charzyńska, Łukasz Dębowski

Word length distribution in Chinese dialogue and prose texts 32
Heng Chen

Using language network characteristics to do text classification 34
Xinying Chen

A new universal code helps to distinguish natural language from random texts 36
Łukasz Dębowski

Large-scale stylometry using network analysis 38
Maciej Eder

Quantitative studies: the advantages for dialectology 40
Sheila Embleton, Dorin Uritescu, Eric S. Wheeler

Quantitative data on monosyllabism: a cross-linguistic study 42
Gertraud Fenk-Oczlon, August Fenk

Towards a mathematical theory of word order evolution 45
Ramon Ferrer-i-Cancho

The study of text clustering based on Chinese dependency treebank 47
Song Gao

Modelling multidimensional polysemy networks. The case of /over/. 48
Dylan Glynn

Stylistic fingerprints, POS tags and inflected languages: a case study in Polish 51
Rafał L. Górski, Maciej Eder, Jan Rybicki

The Arens-Altmann Law:
a matter of boundary conditions or an ostensible success story? 54
Peter Grzybek

On the exact computation of resampled mean size of paradigm (MSP) 57
Guillaume Guex, Aris Xanthos

Cross-linguistic transference of authorship attribution 59
Belinda Hasanaj, Erin Purnell, Patrick Juola

Distribution pattern of given and new information in written English 61
Wang Hua

Word frequency distribution in genres of modern Chinese 63
Wei Huang

Loan words: a quantitative linguistics perspective 65
Emmerich Kelih

The Menzerath-Altmann Law in film analysis 67
Veronika Koch, Peter Grzybek

Quantitative index text analyser (QUITA) 69
Miroslav Kubát, Vladimír Matlach

Quantitative psycholinguistic analysis of formal parameters of Czech text 71
Dalibor Kučera, Jiří Haviger

Towards generalization of sociolinguistic distributions: 73
English loanwords in contemporary written Japanese
Aimi Kuya

Type-token relation for length motifs in Ukrainian texts 76
Ján Mačutek

Gender identification in Modern Greek tweets 78
George Mikros, Kostas Perifanos

Three models for the Menzerath's Law 80
Jiří Milička

Sentence semantics, word meaning, and nonlinear dynamics 81
Hermann Moisl

Testing language units of written Chinese via Menzerath-Altmann Law 83
Tereza Motalová, Lenka Spáčilová

Structural versus morphological coding. A cross-linguistic study. 85
Sven Naumann

An exploration of the "Golden Section" in Chinese contemporary poetries 86
Xiaxing Pan, Hui Qiu

Modelling proximity in a corpus of literary retranslations: 89
a methodological proposal for clustering texts based on
systemic-functional annotation of lexicogrammatical features
Adriana Pagano, Giacomo Figueredo, Annabelle Lukin

The Krylov Law as a tool for comparative lexicology. 92
The example of Polish 19th century dictionaries.
Adam Pawłowski

Evolutionary derivation of laws for polysemic and age-polysemic 94
distributions of language signs ensembles
Vasiliy Poddubny, Anatoly Polikarpov

Quantitative studies in the corpus of Nko periodicals 97
Andrij Rovenchak

Translations in networks: the (in)visibility of translator styles 99
Jan Rybicki

A co-occurrence and an order of the valency in Japanese sentences 102
Haruko Sanada

Authorship attribution using political speeches 104
Jacques Savoy

Testing language units of spoken Chinese via Menzerath-Altmann Law 106
Jana Ščigulinská, Denisa Schusterová

Comparative rates of change as a diagnostic of vowel phonologization 108
Betsy Sneller

Diversification in the noun inflection of Old English 111
Petra Steiner

Quantitative verification of constancy measures of texts 113
Kumiko Tanaka-Ishii, Shunsuke Aihara

History of words II – types of historical developments 115
Arjuna Tuzzi, Reinhard Köhler

Defying Zipf's Law 117
Marjolein Van Egmond, Sergey Avrutin

Opinion target identification using thematic concentration of the text 119
Kateřina Veselovská, Radek Čech

Grammar efficiency and the idealization of parts-of-speech systems 121
Relja Vulanovič, Tatjana Hrubik-Vulanovič

Polyfunctionality and polysemy in Chinese 123
Lu Wang

Structural complexity of Chinese characters and Zipf's Law 126
Yanru Wang, Xinying Chen

The influences of the word unit and the sentence length on the ratio of the parts of speech in Japanese 129
Makoto Yamazaki

Addresses of authors 131

Register 146

Dedicated to Gabriel Altmann in honour of his 83th birthday
and to Luděk Hřebíček in honour of his 80th birthday.

Preface

Over the last decades, Quantitative Linguistics has undergone a rapid and fruitful development and already reached the status of a mature and fully developed branch of linguistics and text analysis. The International Quantitative Linguistics Association (IQLA) was founded on November 28, 1994 and celebrates this year its 20th anniversary. The same holds true for the *Journal of Quantitative Linguistics* (*JQL*), the official journal and flagship of IQLA. Additionally Quantitative Linguistics is growing in regard to the members of the association as well as to the number of journals and book series which are explicitly devoted to problems of quantitative methods in linguistics and quantitative text analysis (*Glottometrics*, *Glottotheory*, *Quantitative Linguistics*, *Studies in Quantitative Linguistics* etc.).

One of the most important, main institutional contributions of IQLA to the development and propagation of Quantitative Linguistics is the regular organisation of International Congresses on Quantitative Linguistics (QUALICO). A tradition, which already started in Germany (Trier, 1991) and continued in Russia (Moscow, 1994), Finland (Helsinki, 1997), Czech Republic (Prague, 2000) and the United States (Athens, GA, 2003). At the 5th Trier Conference on Quantitative Linguistics in 2007 the regular organisation of Qualico was confirmed and re-established, and the following conferences took place in Austria (Graz, 2009) and Serbia (Belgrade, 2012). QUALICO 2014 is hosted by the recently established Department of General Linguistics at the Palacký University Olomouc (Czech Republic) and the organization of QUALICO would not have been possible without a highly motivated and encouraged team of local organizers. Due to the help and substantial financial support of ESF projects it was possible to organize QUALICO 2014 and to celebrate the anniversary of IQLA and *JQL* in an appropriate and worthy way. Without any doubt QUALICO 2014 will remain as a further important step in establishing and the dissemination of Quantitative Linguistics.

Emmerich Kelih

President of IQLA
on behalf of the IQLA-Council

Welcome to Olomouc!

Dear QUALICO 2014 participants and friends of Quantitative Linguistics, if we wanted to understand the word quantitative and consulted a dictionary, we would find that it means *relating to, measuring, or measured by the quantity of something rather than its quality.* Therefore, with this respect, we wish to launch this book of abstracts and the conference itself by *quantities* related to Quantitative Linguistics and to some of its representatives.

Firstly, we would like to warmly welcome you to Olomouc and to Palacký University in Olomouc, which is the second oldest in the Czech Republic. Last year, the University celebrated the 440th anniversary of its foundation. The University is here represented by the Department of General Linguistics. The team of the department and the wider cooperating team consist of linguists, philosophers, statisticians and mathematicians; i.e. we promote team work and interdisciplinary approach which, we believe, would be an unavoidable feature for future Quantitative Linguistics and general linguistics too. The tradition of our department is quite short, it dates three years back, yet we are very eager and proud to be a co-organizer of QUALICO 2014.

We would be happy if the conference assissted in indicating at least tracks or possibilities of the future direction of Quantitative Linguistics, its position inside the wide stream, answers to questions like e.g. what the laws are, what the fundamental questions are, what appropriate methodology to use, to what extent to use sciences outside linguistics, i.e. mathematics, statistics etc. We would like to do our bit by co-organizing the conference to contribute to the process of revealing and maintaining the position of Quantitative Linguistics.

Last but not least, Gabriel Altmann, one of the founders of international Quantitative Linguistics, celebrated last week his birthday and in a few days Luděk Hřebíček, one of the pioneers of Quantitative Linguistics in the Czech Republic, will celebrate his 80th birthday. Despite they are not physically present, we would like to dedicate the conference to them. Above all to Luděk Hřebíček. On the other hand Luděk Hřebíček regards the contribution of Gabriel Altmann as one of the milestones of the 20th century linguistics. He goes as far as to say in his contemplations that he calls the linguistics of the last century as Altmannian linguistics.

Welcome to Olomouc and enjoy your stay here!
Local organizers
Department of General Linguistics
Palacký University Olomouc, Czech Republic

Quantitative analysis of poetic space: discrimination of loci in Eugene Onegin by Pushkin

Sergey Andreev

Keywords: LOCUS, DISCRIMINANT ANALYSIS, VERSE TEXT, EUGENE ONEGIN, CLASSIFICATION, NOMINALITY

The representation of Time and Space (Chronotope) in novels is one of the most important and popular subjects of many philological studies. Special attention is usually paid to the latter of the above-mentioned components of Chronotope – Space. It has been analyzed from literary, cultural, semiotic, semantic, cognitive points of view. The purpose of this paper is to carry out a quantitative comparison of text extracts referring to different loci. By a locus we understand a specific locality of the poetic world whose borders can be crossed in most cases only by the main characters of the novel and very often with dramatic results (Lotman 1970). The database of this study is the novel in verse Eugene Onegin by Pushkin which is considered to be one of the highest (if not the highest) achievements in Russian literature. Traditionally the following main loci are singled out in this novel: Saint-Petersburg, Village, Moscow. As Saint-Petersburg is present in two rather distant chapters – the first and the last – we divided it into two: SPb-1 and SPb-2. The same was done for Village (Village-1 – in the second chapter and Village-2 – in the last but one). The formal characteristics which were used for the analysis include features which show certain deviation from the standard: three types of enjambement (rejet, contre-rejet and double-rejet), syntactic pause in the verse line, inversion (full), emphatic line (its formal markers are exclamation, interrogation marks or dots at the end of the line), emphatic sentence (the same markers as for emphatic line, but at the end of the sentence). Besides, coordinate and subordinate sentences, direct speech and three parts of speech – nouns,

verbs and adjectives were also counted. Each of these three part-of-speech characteristics was subdivided into three groups depending on the position of the word in the line: initial (first metrically strong position), final (last strong metrical position) and intermediate (all other positions in the middle of the line).

5 main loci (SPb-1, SPb-2, Moscow, Village-1, Village-2) were compared with the help of the multivariate discriminant analysis. It gave unexpectedly good results of discrimination (post hoc test showing 100% correctness). The main characteristics, forming the discriminant model, are subordinate clauses, inversion, emphatic sentences, double-rejet, nouns in the initial and final positions. On the other hand, parts of speech in the middle of the line were found obscuring the discrimination. This can be interpreted as a proof of high relevance of features reflecting vertical organization of verse. Using these results the other loci (Way, The Tatyana's dream, Duel, Tatyana's letter, Onegin's letter) were classified. The other stage of analysis included the comparison of the distribution of nouns, verbs and adjectives in different loci according to the method, suggested in (Naumann et al. 2012). The results showed differences of the degree of nominality and type of description (static vs. dynamic).

References

Lotman, Y. M.
(1970) *The structure of the literary text*. Moscow.

Naumann S., Popescu I.-I., Altmann G.
(2012) Aspects of nominal style, *Glottometrics,* 23, 23–55.

Acknowledgment

The research was supported by Russian Humanitarian Fund, Project №14-04-00266.

The Menzerath-Altmann Law revisited

Jan Andres

Keywords: Menzerath-Altmann Law, optimization of parameters, accuracy of calculations

One of the milestones of quantitative lingustics, the Menzerath-Altmann Law (MAL), will be examined from two perspectives:

(i) a possible accordance of the verbal form of MAL vs. an optimal related mathematical formula,

(ii) the accuracy of calculations to the shape parameters in MAL (which is crucial for the fractal analysis of texts).

For both goals, we will compare the results concerning the data without averaging, the averaged weighted data and those without weights. The concrete illustrative examples will be supplied. A suitability of various formulas of the Menzerath-Altmann Law will be examined from the point of view of its verbal version (i.e. the tendency of the approximated functions to be decreasing) as well as of quantitative statistical criteria characterizing the accuracy of fitted data. In particular, we will concentrate on the optimal calculation of the shape parameter *b* and we explain its fundamental role for the fractal analysis of texts.

References

Andres, J., Kubáček, L., Machalová, J., Tučková, M.

(2012) Optimization of parameters in the Menzerath-Altmann Law. *Acta Universitatis Palackianae Olomucensis, Facultas rerum naturalium, Mathematica*., 51, 1, 5–27.

Andres, J., Benešová, M., Chvosteková, M., Fišerová, E.
(2014) Optimization of parameters in the Menzerath-Altmann Law, II. *Acta Universitatis Palackianae Olomucensis, Facultas rerum naturalium, Mathematica*., submitted.

The dependency between the variance of word length and word frequency

Gemma Bel-Enguix, Kirsten Bohn, Ramon Ferrer-i-Cancho

Keywords: language evolution, compression, dependency word frequency / variance of length, English, Catalan, French, German, Latin

It is well-known that the length of words tends to decrease as their frequency increases. This phenomenon corresponds to Zipf's Law of abbreviation (Zipf 1949). G. K. Zipf attributed the Law to the minimization of a cost function consistent with the mean code length of information theory (Ferrer-i-Cancho et al, 2013). The Law of abbreviation can be seen as an epiphenomenon of a general principle of compression (Chater and Vitányi, 2003). Here, the problem of abbreviation is investigated from the perspective of the dependency between frequency and the variance of the length of words having that frequency. It is found that an estimator of the variance of word length tends to decrease as frequency increases in English, Catalan, French, German and Latin texts. These results suggest that the pressure for reducing word length increases as word frequency increases and that models of the dependency between frequency and length (e.g., Strauss et al 2007) should take into account heteroscedasticity.

References

Chater, N., Vitányi, P. M. B.
(2003) Simplicity: a unifying principle in cognitive science? *Trends in Cognitive Sciences*, 7, 19–22.

Ferrer-i-Cancho, R., Hernández-Fernández, A., Lusseau, D., Agoramoorthy, G., Hsu, M. J., Semple, S.
(2013) Compression as a universal principle of animal behavior. *Cognitive Science*, in press. doi: 10.1111/cogs.12061.

Strauss, U., Grzybek, P., Altmann, G.
(2007) Word length and word frequency. In: P. Grzybek (ed.), *Contributions to the science of text and language*, Dordrecht: Springer, 277–294.

Zipf, G. K.
(1949) *Human behaviour and the principle of least effort*. Cambridge, MA: Addison-Wesley.

Distributional models of the verbal predicate-argument structure

Andrei Beliankou

Keywords: PREDICATE-ARGUMENT STRUCTURE, VERB VALENCY, FULL VALENCY, SEMANTIC FRAMES, DISTRIBUTIONAL MODELS

The predicate-argument structure is a longstanding topic for the research community. We are going to describe some quantitative properties of the verbal argument structure both on the syntactic and semantic levels. On the syntactic level, many works were devoted to dependency structures and valency properties of verbs. We are going to investigate full valency of verbs in Russian, German, Latin, and English and to compare our models to previous works on full valency (e.g. Čech et al. 2010). We plan to compare the approaches distinguishing optional and obligatory arguments with approaches not doing such a distinction. Proven hypotheses about the functional dependency of verb frequency and number of arguments as well as about the type of argument number distribution should be tested on new data sets for other European languages. On the semantic level, we want to prove the hypothesis about a tight interrelation of semantic roles and syntactic arguments for coarse- and fine-grained annotation frameworks (e.g. VerbNet vs. SALSA/FrameNet). We expect to find a similar distribution of full valency frames and coarse-grained semantic frames and develop a method for mapping coarse- and fine-grained semantic annotations in order to find a quantitative approach for semantic description of verbal predicate-argument structures.

References

Čech, R., Pajas, P., Mačutek, J.
(2010) Full Valency. Verb Valency without Distinguishing Complements and Adjuncts. English. *Journal of Quantitative Linguistics.* 17, 4, 291–302.

Liu, H.
(2011) Quantitative Properties of English Verb Valency. J*ournal of Quantitative Linguistics.* 18, 3, 207–233.

Menzerath-Altmann Law in differently segmented text

Martina Benešová, Dan Faltýnek, Lukáš H. Zámečník

Keywords: Menzerath-Altmann Law, segmentation of linguistic sample, construct, constituent, synergetic linguistics

The aim of our paper is to discuss one the fundamental assumptions of synergetic linguistics that a language system tends to balance itself due to possessing self--regulating and self-organizing mechanisms, and to postulate an inquiry when it happens, if it ever does. We chose the well-known and often quoted Menzerath-Altmann Law (MAL) as a model whose assumed validity confirms that it does. The MAL original verbal formulation was extended to become a relationship between the generalized units of the construct and the constituent. A unit of a particular linguistic level, therefore, acts as a construct toward the unit of the immediately lower neighbouring level and a constituent to the immediately higher neighbouring level. This has to lead inevitably to the precise definition of either on every linguistic level. Hence, the first discussed pitfall to be commented on is the choice of suitable linguistic units and the rules of a linguistic sample segmentation. One text was chosen and additionally tested using different segmentation criteria and units. The results are to be compared and interpreted.

Altmann, G.
(1980) Prolegomena to Menzerath's Law. *Glottometrika* 2, 1–10.

Andres, J., Benešová, M., Kubáček, L., Vrbková, J.
(2011) Methodological note on the fractal analysis of texts. *Journal of Quantitative Linguistics*, 18, 4, 337-367.

Andres, J., Kubáček, L., Machalová, J., Tučková, M.
(2012) Optimization of parameters in Menzerath-Altmann Law II. *Acta Universitatis Palackianae Olomucensis, Facultas rerum naturalium, Mathematica,* 51, 1, 5–27.

Andres, J., Benešová, M., Chvosteková, M., Fišerová, E.
(2014) Optimization of parameters in Menzerath-Altmann Law. *Acta Universitatis Palackianae Olomucensis, Facultas rerum naturalium, Mathematica,* 53, 1, 3–23.

Givón, T.
(2001) *Syntax. An Introduction*. Amsterdam: John Benjamins.

Haspelamth, M., Sims, A. D.
(2010) *Understanding Morphology*. London: Hodder Education, an Hachette UK Company.

Hewlett, N., Beck, J.
(2006) *An Introduction to the Science of Phonetics*. Mahwah, NJ: Lawrence Erlbaum.

Hřebíček, L.
(1997) *Lectures on Text Theory*. Prague: The Academy of the Sciences of the Czech Republic (Oriental Institute).

Laver, J.
(1994) *Principles of Phonetics*. Cambridge: Cambridge University Press.

Levinson, S. C.
(1983) *Pragmatics*. Cambridge: Cambridge University Press.

Tallerman, M.
(1998) *Understanding Syntax*. London: Arnold.

Application of the Menzerath-Altmann Law to contemporary written Japanese – short story style

Denis Biryukov, Martina Benešová

Keywords: Menzerath-Altmann Law, written Japanese, segmentation, short story

The main objective of this experiment is verification of the applicability of the Menzerath-Altmann Law to contemporary written Japanese. The sample text style chosen for the analysis is a short story by a Japanese author. The methodology, i.e., sample text segmentation methods used in this experiment, partially stem from an earlier experiment on contemporary written Chinese, particularly the language unit of the component (island) proposed by Motalová and Spáčilová. Simultaneously, new language units and processes of segmentation of Japanese texts are also proposed and tested in this experiment as the Japanese writing system otherwise differs considerably from Chinese in many aspects. Analysis results will be presented and interpreted in the paper.

References

Habein, Y. S.,
(2000) Decoding Kanji. *A Practical Approach to Learning Look-Alike Characters*. 1st ed. Tokyo: Kodansha International, 2000.

Hřebíček, L.
(2002) *Vyprávění o lingvistických experimentech s textem*. Praha: Academia.

Kamada, T., Torataroō Y., Y.
(2004) *Shin Kangorin. Tōokyoō: Taishuūkan Shoten*, 用例プラス Electronic Dictionary Edition.

Koike, S.
(2010) *Úvod do gramatiky moderní japonštiny*. 1st edition. Brno: Tribun EU.

Kraemerová, A.
(2000) *Úvod do japanologie: jazyk a literatura*. 1st edition. Olomouc: Univerzita Palackého, Filozofická fakulta.

Motalová, T., Spáčilová, L.
(2013) *Aplikace Menzerath-Altmannova zákona na současnou psanou čínštinu*. Olomouc, 2013. thesis (Mgr.). Univerzita Palackého v Olomouci. Filozofická fakulta.

Nitsuú, N. Satoó, F.
(2004) 留学生のための論理的な文章の書き方. Tokyo: 3A Corporation.

Spahn, M., Hadamitzky, W., Fujie-Winter, K.
(1989) *Kan-Ei Jukugo Ribāsu Jiten. Shohan*. Tōkyō: Hatsubai Kinokuniya Shoten.

Švarný, O. (et al.)
(1967) *Úvod do hovorové čínštiny: Příručka pro vys. šk. 2.* Praha: SPN.

常用漢字表.
(2010) http://www.bunka.go.jp/kokugo_nihongo/pdf jouyoukanjihyou_h22pdf

文章の書き方・ととのえ方: 5, 段落の作り方. 三省堂
(2014) Web Dictionary [online]. 2014 [accessed 2014-01-04]. http://www.sanseido.net/main/words/hyakka/howto/05.aspx

文部省教科書局調査課国語調査室. くぎり符号の使ひ方.
(1946) [accessed 2014- 01-04]. http://www.bunka.go.jp/kokugo%5Fnihongo/joho/kijun/sanko/pdf/kugiri.pdf

Impact of semantics on case diversification

Radek Čech, Emmerich Kelih, Ján Mačutek

Keywords: CASE DISTRIBUTION, NOUN, SEMANTICS, CZECH

The relationship between semantics of noun and the frequency of particular grammatical cases in highly inflected languages is well known, e.g. Greenberg (1990). For instance, noun denoting person tends to occur most commonly in nominative – because of its agentive role – while place noun in locative. In our analysis, however, we try to interpret the relationship between semantics of noun and case distribution of particular nouns in a rather more general way. Based on both the Wimmer-Altmann theory (2005) and the idea of a diversification process (Altmann 2005), we assume that the distribution of cases of particular nouns in general should be ruled by an underlying mechanism which can be viewed as a manifestation of complex mutual interrelations in language system, in the sense of synergetic linguistics approach (Köhler 1996, 2005). Further, we assume that semantics of noun influences not only the ranking of particular cases but the distribution of cases as a whole. Therefore, hypotheses on a general relationship between the semantics of nouns and case diversification will be presented. The analysis will be focused on case frequency and case frequency distribution in contemporary Czech corpora (lemmatized and morphologically tagged synchronic corpus data from Czech National Corpus are used). Particularly, we will consider several noun categories, as for instance a) gender (masculine, feminine and neutral), b) animate/inanimate, c) number (singular, plural) and d) general semantic features (concrete vs. abstract nouns) and e) specific semantic features (animals, body parts, place nouns, nouns referring to persons and professions etc.). In addition to descriptive statistics, first steps in mathematical modelling and the integration of particular hypotheses into synergetic linguistics will be discussed.

References

Altmann, G.
(2005) Diversification processes. In: Köhler, R., Altmann, G., Piotrowski, R.G. (eds.), *Handbook of Quantitative Linguistics.* Berlin: de Gruyter, 649–659.

Greenberg, J. H.
(1990) The Relation of Frequency to Semantic Feature in a Case Language (Russian). In: Denning, K., Kemmer, S. (eds.), *On language: selected writings of Joseph H. Greenberg.* Stanford: Stanford University Press, 207–226.

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

Köhler, R.
(2005) Synergetic Linguistics. In: Köhler, R., Altmann, G., Piotrowski, R.G., *Handbook of Quantitative Linguistics.* Berlin: de Gruyter, 760–775.

Těšitelová, M.
(1996) On quantification in grammar and semantics. Partee, B. N., Sgall, P. (eds.): *Discourse and meaning. Papers in honour of Eva Hajičová.* Amsterdam/ Philadelphia: Benjamins, 369–378.

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

Factors of readability of Polish texts: a psycholinguistic study

Edyta Charzyńska, Łukasz Dębowski

Keywords: READABILITY, POLISH, PSYCHOLINGUISTICS, WORD LENGTH, SENTENCE LENGTH

Readability of a given text is the degree of difficulty of understanding the text. This degree can be measured using psycholinguistic methods such as multiple choice tests or cloze tests. It can be partly predicted by statistical properties of the text such as the mean sentence length or the percentage of difficult word tokens. Many formulae for predicting readability from such statistics have been proposed for English (Flesch 1948, Dale and Chall 1948, Gunning 1952, McLauglin 1969, Caylor et al. 1973, Kincaid et al. 1975). Subsequently these formulae have been transferred to other languages, e.g., Slovak (Mistrík 1968) and Polish (Pisarek 2007), but have not been verified experimentally. The aim of the present paper is to report the results of the first psycholinguistic study investigating readability formulae for Polish. For the study, a sample of 15 differentiated texts (2 scientific texts, 2 texts from the secondary school handbooks, 2 enactments, 2 official letters, 2 instructions, 2 Law brochures, 3 journalistic articles from various fields) has been selected and their readability has been evaluated by 3 different tests (multiple choice, cloze and open-ended question tests) on a sample of 1,309 subjects. The studied sample represents heterogeneity in sociodemographics (sex, age, education level, place of residence). Therefore the standard deviation of the test results is quite large. Despite that, the regression between the measured readability and text statistics has been investigated. The analyzed text statistics included: the mean sentence length (in words), the mean word length (in syllables) and the mean word entropy (-log frequency of a word in the National Corpus of Polish).

In the study two of three used psycholinguistic tests have shown a strong negative correlation between the text comprehension and the mean sentence length (respectively, cloze test: $\rho = -.436$; $\rho < .001$; open-ended question test: $\rho = -.386$; $p < .001$; multiple-choice test: $\rho = -.016$; $\rho > .05$). The mean word length and the mean word entropy were much worse predictors of text comprehension, regardless of the type of the test. A probable reason for that is the peculiar property of the text sample where the mean sentence length and the mean word length were strongly negatively correlated (rho_MC = -.508; $\rho < .001$; rho_Cloze = -.423; $\rho < .001$; rho_Open = -.460; $p < .001$). We suppose that this negative correlation is too strong to be explained by the Menzerath--Altmann Law (Altmann 1980) but rather is an accidental property of the selected sample of texts. Further studies, using more homogeneous text samples, are required to verify this hypothesis.

References

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

Caylor, J. S., Stitch, T. G., Fox, L. C., Ford, J. P.
(1973) *Methodologies for determining reading requirements of military occupational specialties, Technical Report*. 73–5, Alexander, Virginia: Human Resources Research Organization.

Dale, E., Chall, J. S.
(1948) A formula for predicting readability. *Educational Research Bulletin*, 27, 1–20, 27, 37–54.

Flesch, R.
(1948) A new readability yardstick. *Journal of Applied Psychology*, 32, 221–233.

Gunning, R.
(1952) *The Technique of Clear Writing*. New York: McGraw-Hill.

Kincaid, J.P., Fishburne, R. P., Rogers, R. L., Chissom, B. S.
(1975) Derivation of new readability formulas (Automated Readability Index, Fog Count, and Flesch Reading Ease Formula) for Navy enlisted personnel. *CNTECHTRA Research Branch Report*, 8–75.

McLaughlin, G. H.
(1969) SMOG grading-a new readability formula. *Journal of Reading*, 22, 639–646

Mistrík, J.
(1968) Meranie zrozumiteľnosti prehovoru. S*lovenská re*č, 33, 3, 171–178.

Pisarek, W.
(2007) Jak mierzyć zrozumiałość tekstu? *O mediach i języku*, 245–262, Kraków.

Word length distribution in Chinese dialogue and prose texts

Heng Chen

Keywords: Chinese, word length distribution, measuring unit, style variation, synergetic linguistics

Word length measuring needs reasonably defined linguistic units, so if we want to investigate a language's word length distribution, a key premise is to find an appropriate word length measuring unit of the language, in addition to defining the „word". Grotjahn and Altmann (1993) stated that „there are three basic types of units to measure the word length, namely (a) graphical, (b) phonetic, and (c) semantic ones" (1993: 142). Yet, regretfully, just as what Grzybek (2006) said, "yet, even today, there are no reliable systematic studies on the influence of the measuring unit chosen, nor on possible interrelations between them." An appropriate word length measuring unit is not fixed, it varies with language types and text styles. We will explore these questions in this paper. This paper investigates the word length distribution of Chinese spoken and written languages using 20 dialogue texts (spoken language) and 20 prose texts (written language) as data sources, in which the length of words is potentially determined in terms of pinyin letter, phoneme, syllable and stroke, component, character respectively. Results show the syllable is the most appropriate word length measuring unit of the Chinese spoken language, and the component is the most appropriate word length measuring unit of Chinese written language; the Chinese word length distribution can be described with the poisson or binomial distribution series, among which extended logarithmic and mixed poisson are the most generally accepted models. The results also show that different measur-

ing units are highly related with one another. Further the reasons are given to explain why the syllable and component are the most appropriate word length measuring units for the Chinese spoken and written languages respectively by referring to synergetic linguistic theories and the Menzerath's Law. In addition, the paper scrutinizes the application of the Chinese word length distribution in language style variations studies: written language has more word length classes than spoken language; it also has more word types in word length classes 1 and 2 compared with spoken language. These two differences can also be observed through the model's parameter changes, which may make it useful for literary studies such as genre discrimination.

References

Grotjahn, R., Altmann, G.

(1993) Modeling the distribution of word length: some methodological problems. In: Köhler, R., Rieger, B. (eds.), *Contributions to Quantitative Linguistics: Proceedings of the First International Conference on Quantitative Linguistics, QUALICO, Trier.* Dordrecht: Kluwer, 141–153.

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.* Dordrecht: Springer, 15–90.

Using language network characteristics to do text classification

Xinying Chen

Keywords: NETWORK CHARACTERISTICS, GENRE CLASSIFICATION, AUTHOR IDENTIFICATION, CLUSTERING, SYNTACTIC COMPLEXITY

Linguistic research using modern network analysis tools is an upcoming domain. The key interest of the network approach in linguistic research is that it provides a new way to analyze language systems. A central assumption of modern linguistic theories is that language is a system (Kretzschmar 2009). This widely accepted point of view, however, has remained on a purely theoretic level due to the absence of an operational methodology, until corpora and modern network analysis tools appeared. Modeling language as a network provides an operational way for observing the macroscopic features of the language system, such as text styles or genres. Some research has been done on language clustering (Liu and Li 2010, Liu and Cong 2013, Abramov and Mehler 2011) and all of these studies show good results in identifying genealogical groups. The common aspect of these studies is that they all focus on the clustering or classification of different languages. So for testing the effectiveness of using network characteristics identifying different text groups in a single language, we applied a similar strategy to text classification of a single language, more specifically to automatic genre classification and author identification. For genre classification, we constructed six Chinese syntactic dependency networks based on six dependency treebanks of different styles. Then we did a comparative analysis of network characteristics, including the number of edges, the number of nodes, the average degree, the clustering coefficient, the average path length, the centralization, the diameter, the index of

power-law, the coefficient of determination. After that, we used different clustering methods, with characteristics as variables, to do clustering analysis of these networks parameters. The results showed that, using some of the main parameters of the networks, namely the number of the nodes, the clustering coefficient, the average path length, the centralization and the index of power-law, we can get very good automatic clustering results on texts and the result is better than simple text clustering based on word frequency distributions. For author identification, we constructed four-word co-occurrence networks based on 48 English articles, separated into two topics: anecdote and film, all written by 24 Chinese students. Each of the 24 students was asked to write two articles. We split up the 24 anecdotal articles into two groups just as the 24 film articles. The first and the second group of each genre were written by the same students. The same process was repeated with the genre classification: The results show that, using some main parameters of networks, namely the number of the nodes, the clustering coefficient, the average degree, the average path length, the centralization, and the index of power-law, we can get fair results on identifying author groups. The results, however, are not as good as the results we got in genre classification. This work shows that network features offer a new source of typological information about texts and this information can contribute to a better understanding of difference of styles in a language.

References

Kretzschmar, W. A.
(2009) *The Linguistics of Speech*. New York: Cambridge University Press.

Liu, H.
(2010) Language Clusters based on Linguistic Complex Networks. *Chinese Science Bulletin*, 55, 30, 3458–3465.

Liu, H., Cong, J.
(2013) Language Clustering with Word Cooccurrence Networks based on Parallel Texts. *Chinese Science Bulletin*, 58, 10, 1139–1144.

Abramov, O., Mehler, A.
(2011) Automatic Language Classification by Means of Syntactic Dependency Networks. *Journal of Quantitative Linguistics*, 18, 4, 291–336.

A new universal code helps to distinguish natural language from random texts

Łukasz Dębowski

Keywords: UNIVERSAL CODING, ENGLISH, RANDOM TEXT, ZIPF'S LAW, HILBERG'S CONJECTURE

It is widely agreed that texts in natural language differ statistically from „monkey-typing", i.e., the output of producing characters at random (Zipf 1965). The difference can be revealed using various experimental methods. For example, there is a prominent difference between texts in natural language and random texts with respect to the rank-frequency distribution of words (Ferrer-i-Cancho and Elvevåg 2010). The latter result is somewhat unexpected since both texts in natural language and random texts obey some versions of Zipf's law (Mandelbrot 1954, Miller 1957).

In this paper we wish to demonstrate a difference between random texts and texts in natural language which can be detected by means of universal coding. We consider two universal codes: the Lempel-Ziv code (Ziv and Lempel 1977) and the plain switch distribution (Dębowski 2013b). These two codes are applied to two texts: „20,000 Leagues under the Sea" by Jules Verne and a unigram model of this novel. In case of the Lempel-Ziv code we observe no substantial difference between these two texts. For both texts, the compression rate decreases hyperbolically, whereas the mutual information grows like a power Law. If we apply the plain switch distribution, however, there arises a huge difference. The compression rate for the text in natural language decreases hyperbolically, whereas the compression rate for the random text almost stabilizes. Moreover, the mutual information for the text in natural language grows like a power Law, whereas for the random text it grows logarithmically. This observation provides a new

stronger support for Hilberg's conjecture (Dębowski 2013a), an important hypothesis concerning the entropy of natural language. By the „no hypercompression" inequality (Grunwald 2007, p. 103), we can also use the aforementioned observation for proving that the probability that „20,000 Leagues under the Sea" was generated by the unigram model is less than $2^{(-1,000,000)}$.

References

Dębowski, Ł.
(2013a) Empirical evidence for Hilberg's conjecture in single-author texts. In: Obradović, I., Kelih, E., Köhler, R. (eds.), *Methods and Applications of Quantitative Linguistics – Selected papers of the 8th International Conference on Quantitative Linguistics* (*QUALICO*). Belgrade: Academic Mind, 143–151.

Dębowski, Ł.
(2013b) Preadapted Universal Switch Distribution for Testing Hilberg's Conjecture. http://arxiv.org/abs/1310.8511.

Ferrer-i-Cancho, R., Elvevåg, B.
(2010) Random texts do not exhibit the real Zipf's law-like rank distribution. *PLoS ONE* 5, 3, e9411.

Grunwald, P. D.
(2007) *The Minimum Description Length Principle*. Cambridge, MA: MIT Press.

Mandelbrot, B.
(1954) Structure formelle des textes et com munication. *Word*, 10, 1–27.

Miller, G.
(1957) Some effects of intermittent silence. *American Journal of Psychology*, 70, 311–314.

Zipf, G. K.
(1965) *The Psycho-Biology of Language: An Introduction to Dynamic Philology*, 2nd ed. Cambridge, MA: The MIT Press.

Ziv, J., Lempel, A.
(1977) A universal algorithm for sequential data compression. *IEEE Transactions on Information Theory*, 23, 337–343.

Large-scale stylometry using network analysis

Maciej Eder

Keywords: STYLOMETRY, AUTHORSHIP ATTRIBUTION, COMPUTATIONAL STYLISTICS, BIG DATA, NETWORK ANALYSIS

Stylometric methodology, developed to solve authorship problems, can easily be extended and generalized to assess different questions in the field of text analysis. Namely, the underlaying idea of tracing similarities between (anonymous) texts can be extended to map textual relations in large-scale approaches to literature. Explanatory multidimensional methods, relying on distance measures and supported with visualization techniques, are particularly attractive for this purpose. However, they are very sensitive to the number of features (usually: frequent words) analyzed. Even worse, they are either unable to fit dozens of texts on a single scatterplot (e.g. Multidimensional Scaling), or highly dependent on the choice of a linkage algorithm (e.g. Cluster Analysis). The technique introduced in this study combines the concept of network as a way to map large-scale literary similarities (Jockers 2013), the concept of consensus (Lancichinetti and Fortunato 2012), and the assumption that textual relations usually go beyond mere nearest neighborship. Particular texts can be represented as nodes of a network, and their explicit relations as links between these nodes. The procedure of linking is twofold. One of the involved algorithms computes the distances between analyzed texts, and establishes, for every single node, a strong connection to its nearest neighbor (i.e. the most similar text), and two weaker connections to the 1st and the 2nd runner-up (i.e. two texts that get ranked immediately after the nearest neighbor). The other algorithm performs a large number of tests for similarity with different number of features to be analyzed (e.g. 100, 200, 300, ..., 1,000 MFWs). Finally, all the connections produced in particular "snapshots" are added, resulting in a consensus

network. Weights of these final connections tend to differ significantly: the strongest ones mean robust nearest neighbors, while weak links stand for secondary and/or accidental similarities. The validation of the results – or rather self-validation – is provided by the fact that consensus of many single approaches to the same corpus sanitizes robust textual similarities and filters out apparent clusterings. The idea discussed in this paper can be applied to map large collections of texts, such as the corpus provided by the "Perseus Project" database (1,127 texts), but also to represent similarities in smaller corpora. One of the examples includes an ad-hoc collection of 124 Ancient Greek texts assessed with the aforementioned technique. The nodes and edges of the network have been computed using the "stylo" package for R (Eder et al. 2013) and visualized with Gephi. The ForceAtlas2 layout (Bastian et al. 2009) has been used to establish the spacial relations between the nodes, and the modularity detection algorithm (Blondel et al. 2008) to mark distinctive clusters with different colors. The obtained network reveals a clear genre separation (prose, epic poetry, drama), as well as a chronological development of style.

References

Bastian, M., Heymann, S., Jacomy, M.
(2009) Gephi: an open source software for exploring and manipulating networks. *International AAAI Conference on Weblogs and Social Media*.

Blondel, V. D., Guillaume, J.-L., Lambiotte, R., Lefebvre, E.
(2008) Fast unfolding of communities in large networks. *Journal of Statistical Mechanics*: *Theory and Experiment*, 10, P1000.

Eder, M., Kestemont, M., Rybicki, J.
(2013) Stylometry with R: a suite of tools. In: *Digital Humanities 2013: Conference Abstracts*. Lincoln: University of Nebraska-Lincoln, 487–489.

Jockers, M.
(2013) *Macroanalysis: Digital Methods and Literary History*. Champaign: University of Illinois Press.

Lancichinetti, A., Fortunato, S.
(2012) Consensus clustering in complex networks. *Scientific Reports*, 2, 336, 1–7.

Quantitative studies: the advantages for dialectology

Sheila Embleton, Dorin Uritescu, Eric S. Wheeler

Keywords: DIALECTOLOGY, DIALECT, ROMANIAN, QUANTITATIVE, QUANTIFIED GRADATIONS, ADVANTAGES, QUANTITATIVE STUDIES, PALATALIZATION, DENTAL, VULGAR LATIN, BORROWING

Quantitative studies of texts have been established for a long time, but the advantages of making such studies over large data sets hold in dialectology as well. Using some of the digital tools now available, it is possible to find and select a wide range of data subsets, to count occurrences within the selected subsets, and to display the results as maps or other visible charts in such a way that it is easy to understand the selected phenomena. We offer several examples where quantitative studies show results that could not readily be found in traditional printed formats. In one case, the phenomenon (dentals not palatalizing before high front vowels in some geographic regions) was known and attributed to borrowing, but the quantitative study showed how much more widely it occurred, even in non-borrowed words, and how the traditional view was a matter of degree. In another case, we were able to select only those items that came from Vulgar Latin into Romanian, and uncover a relationship that was lost in other dialects. The phenomenon was not obvious except through a count of a large number of separate lexical items. These and other cases show that a large-scale quantitative approach could help uncover new territory even in comparatively well-studied Romanian dialectology. Quantitative dialectology is able to make much stronger claims than traditional dialectology because it can count *all* the occurrences of a phenomenon over a wider range of data; it can disprove claims that are based only on some or a few examples; and it can uncover relationships

that are not obvious until one looks at large amounts of data. Even the definition of "dialect" changes when it is expressed in terms of quantified gradations of change.

More quantitative studies of this sort, using current and future tools and data sets, will advance the study of dialects.

References

Embleton, S., Uritescu, D., Wheeler, E. S.
(2008a) *Digitalized Dialect Studies: North-Western Romanian*. Bucharest: Romanian Academy Press.

Embleton, S., Uritescu, D., Wheeler, E. S.
(2008b) Identifying Dialect Regions: Specific features vs. overall measures using the Romanian Online Dialect Atlas and Multidimensional Scaling. Leeds, UK: Methods XIII Conference. August 2008. In: Heselwood, B., Upton, C. (eds.), (2009). *Proceedings of Methods XIII. Papers from the Thirteenth International Conference on Methods in Dialectology*, 2008. Frankfurt/Main: Lang, 79–90.

Quantitative data on monosyllabism: a cross--linguistic study

Gertraud Fenk-Oczlon, August Fenk

Keywords: MONOSYLLABISM, SYLLABIC COMPLEXITY, TYPOLOGY, CROSS-LINGUISTIC CORRELATIONS

Monosyllabism has since long been considered a typologically relevant phenomenon. Because of its gradual character it does not permit any clear-cut classification. Here, the interesting point is the distribution of languages along the continuum of "proportion of monosyllables" (A) and the interaction of that parameter with other linguistic features (B).

(A) The proportion of monosyllables: In order to determine this parameter, Stolz (2007) analysed the so-called core lexicon of the extended version of the Swadesh lists. In a sample of 50 languages he found a proportion ranging from 2% in Greenlandic to 87% in English and a strong dependency on areal factors. Maddieson (2009) points out several reasons for complementing such analyses of lexical entry forms by analyses of the word forms in matched wordlists and texts. He found a distribution ranging from 3% (Tamil) to 80% (Thai) and a mean of roughly a quarter.

(B) Interactions with other parameters: A previous study by the authors (2008) revealed significant positive cross-linguistic correlations between the number of monosyllables, the number of syllable types, syllable complexity, and phonemic inventory size. In that study the data were collected and calculated from Menzerath's (1954) description of 8 Indo-European languages (English, German, Romanian, Croatian, Catalan, Portuguese, Spanish and Italian).

The present study differs from the one mentioned above in both the method and the sample.

Instead of statistical descriptions by Menzerath we now analyse "matched texts", i.e., the translations of a set of 22 English or German sentences produced by native speakers of 32 different languages (13 Indo-European, 19 Non-Indo-European).

The main hypothesis: The larger a language's proportion of monosyllables, the higher its syllabic complexity in terms of the number of phonemes per syllable.

Results: The assumption of a positive correlation between the number of monosyllables and syllabic complexity could be corroborated ($r = + .62$, $p < .001$) in our now extended and typologically more widespread sample, and despite the use of a different method. This indicates high robustness of that cross-linguistic correlation. Some more results: English (73), Welsh (74), Dutch (67) and German (57) showed the highest number of monosyllables in our textual material, while languages often classified as most typically monosyllabic had far less monosyllables (Mandarin 41, Vietnamese 49).

Discussion: Further interactions will be discussed between the tendency to monosyllabism and phenomena such as a tendency to homophony (Ke 2006), a tendency to a high proportion of idioms, formulaic speech and rigid word order, and a tendency to stress-timed rhythm.

References

Ammann , A. (ed.)
(2006) *Linguistic festival*. Bochum: Brockmeyer, 96–133.

Fenk-Oczlon, G., Fenk, A.
(2008) Complexity trade-offs between the subsystems of language In: Miestamo, M., Sinnemäki, K., Karlsson, F. (eds.), *Language Complexity*: Typology, Contact, Change, 43–65. Amsterdam: John Benjamins.

Ke, J.
(2006) A cross-linguistic quantitative study of homophony. *Journal of Quantitative Linguistics*, 13, 129–159

Maddieson, I.
(2009) *Monosyllables and syllabic complexity*. Paper presented at the conference Monosyllables: from Phonology to Typology, Festival of Languages, University of Bremen, Sept. 2009.

Menzerath, P.
(1954) *Die Architektonik des deutschen Wortschatzes*. Bonn: Dümmler. (= Phonetische Studien, 3).

Stolz, T.
(2007) Being monosyllabic in Europe: an areal typological project in statu nascendi.

Towards a mathematical theory of word order evolution

Ramon Ferrer-i-Cancho

Keywords: WORD ORDER, LANGUAGE PRINCIPLES, EVOLUTION OF LANGUAGE, PERMUTATION RING

The study of the evolution of word order has been fueled by the finding of a preference for subject-object-verb or its semantic correlate (actor-patient-action) in recently emerging languages and gestural experiments (Goldin-Meadow et al 2008, Langus and Nespor 2010). This raises two fundamental questions: (a) why is this ordering appearing? (b) why do all languages on earth not have it as dominant? Here we argue that the maximization of mutual information yields the subject-object-verb orderings and that this ordering is lost when sentences become more complex due to a conflict with a principle of online memory minimization promoting that the verb is placed at the center. We address various puzzles surrounding this general approach. First, among two orderings with the verb at the center, subject-verb-object is very frequent while object-verb-subject is rare, apparently contradicting an attraction of the verb towards the center. Second, the preference for subject-verb-object arises in experiments with verbs corresponding to intensional events (Schouwstra et al 2011), which apparently contradicts the principle of mutual information maximization. Third, modifiers tend to precede nominal heads in languages having subject-verb-object as dominant, which apparently contradicts the principle of online memory minimization. Fourth, the rather uniform distribution of possible orderings of subject, object, verb that is naively expected from a fair battle among word order principles, which is contradicted by the rather skewed distribution of dominant word orders. The solutions to these puzzles requires viewing word orders as

products of cultural evolution and noticing that the most likely word order transitions from a given word order follow a ring structure (Ferrer-i-Cancho 2013). Interestingly, the number of languages showing a certain dominant word order decays perfectly as one moves away from the initial subject-object-verb in a particular sense throughout that ring.

References

Ferrer-i-Cancho, R.
(2013) The placement of the head that minimizes online memory: a complex systems approach. *Language Dynamics and Change.* [in press]

Goldin-Meadow, S., So, W. C., Özyürek, A., Mylander, C.
(2008) The natural order of events: how speakers of different languages represent events nonverbally. *PNAS*, 105, 27, 9163–9168.

Langus, A., Nespor, M.
(2010) Cognitive systems struggling for word order. *Cognitive Psychology,* 60, 4, 291–318.

Schouwstra, M., Van Leeuwen, A., Marien, N., Smit, M., De Swart, H.
(2011) Semantic Structure in Improvised Communication. In: Carlson, L., Hoelscher, C., Shipley, T. (eds.), *Proceedings of the 33rd Annual Conference of the Cognitive Science Society.* Cognitive Science Society, 1497–1502.

The study of text clustering based on Chinese dependency treebank

Song Gao

Keywords: text clustering, clustering feature, Dependency Treebank, dependency relation, part of speech

Text clustering is an important part of information retrieval. The method of applying the information of syntactic distribution to research text clustering is presented, in order to avoid using the complex clustering algorithm and interpret clustering features and the results of clustering from the aspect of linguistics. Ten dependency relations, which distribute distinctively between oral and written Chinese, are investigated based on Chinese Dependency Treebank. The similarity of spoken and written classes achieves 81.98% and 83.13% respectively by using 5 dependency relations as a clustering feature. The experiment result shows that this method using dependency relations to research text clustering is feasible and effective.

Modelling multidimensional polysemy networks. The case of /over/.

Dylan Glynn

Keywords: CORPUS LINGUISTICS, LOGISTIC REGRESSION, MANUAL ANNOTATION, MULTIVARIATE STATISTICS, POLYSEMY

The statistical modelling of a semasiological structure, which is also sensitive to morpho-syntactic and socio-pragmatic context, remains a fundamental goal of semantics. To these ends, a promising method under development is the multifactorial usage-feature analysis (Geeraerts et al. 1999, Gries 2003, Glynn and Fischer 2010, Glynn and Robinson 2013) (also termed the profile-based analysis or sentiment analysis). This method has been used to model near-synonymous relations successfully (Glynn 2009, Divjak 2010, et al.) and constructional alternations (Heylen 2005, Bresnan et al. 2007 et alii), but its application to lexical polysemy has, thus far, been limited to an exploratory analysis (Gries 2006, Glynn 2009). The principle problem lies in the fact that, in usage-profile polysemy research, the lexical 'sense' is neither discrete nor measurable, instead its profile is operationalised as a cluster of usage-features. To date, no confirmatory statistical method has been successfully applied to this research problem. The current study proposes a method to overcome this hurdle in a case study on the polysemy of the lexeme *over*. The proposal consists of adding two statistical steps to the analysis of the results of the usage-feature analysis. In the previous research, a combination of multiple correspondence analysis and hierarchical cluster analysis has been used to identify the complex usage-patterns that are believed to represent the semasiological structure. The first step is to delineate the clusters of features in such a way that discrete senses can be proposed. The study uses *k*-means cluster analysis to determine the best number of clusters of these features and principle components

analysis upon the results of the correspondence analysis in order to determine which of the features contribute the most to structuring those clusters. This produces a small set of lexical 'senses', determined by key features. Importantly, these senses are relative to a range of morpho-syntactic and sociolinguistic contexts. The second step is to submit these senses to confirmatory modelling. Here, the senses are treated as the independent variable in the multinomial logistic regression analysis. This step determines the predictive accuracy of the analysis and the model based upon it. In order to test the proposal, 800 occurrences of the preposition *over* are extracted from British and American Literature (BNC and ANC) and on-line personal diaries (LiveJournal Corpus). The semantic structure of the preposition *over* represents a contentious example of a polysemy research within Cognitive Linguistics (Lakoff 1987 et alii). This rich tradition of an introspection-based research offers an extensive range of semantic (conceptual-functional) dimensions to include in the feature analysis. The results will offer a falsifiable, quantitative, and frequency-based description of the semantic structure of the lexeme. More over, the goodness of fit and predictive accuracy of regression models, based on the analytical categories proposed by the introspection studies listed above, will determine which of these 'theoretical' models is the most accurate.

References

Bresnan, J., Cueni, A. Nikitina, T., Baayen, H.
(2007) Predicting the dative. In: Bouma, G., Krämer, I., Zwarts, J. (eds.), In: *Cognitive foundations of interpretation alternation*. Amsterdam: Royal Netherlands Academy of Arts and Sciences, 69–94.

Divjak, D.
(2010) *Structuring the lexicon: A clustered model for near-synonymy*. Berlin, New York: Mouton de Gruyter.

Geeraerts, D., Grondelaers, S., Speelman, D.
(1999) *Convergentie en Divergentie in de Nederlandse Woordenschat. Amsterdam:* Meertens Instituut.

Glynn, D., Fischer, D. (eds.)
(2010) *Quantitative Cognitive Semantics: Corpus-driven approaches*. Berlin, New York: Mouton de Gruyter.

Glynn, D., Robinson, J.
(2013) *Corpus Methods for Semantics. Quantitative studies in polysemy and synonymy.* Amsterdam: John Benjamins.

Glynn, D.
(2010) Synonymy, lexical fields, and grammatical constructions. A study in usage-based Cognitive Semantics. In: Schmid, H.-J., Handl, S. (eds.), *Cognitive foundations of linguistic usage-patterns: Empirical studies.* Berlin, New York: Mouton de Gruyter, 89–118.

Glynn, D.
(2009) Polysemy, syntax, and variation. A usage-based method for Cognitive Semantics. In: Evans, V., Pourcel, S. (eds.), In: *New directions in Cognitive Linguistics.* Amsterdam, Philadelphia: John Benjamins, 77–106.

Gries, St. Th.
(2003) *Multifactorial analysis in corpus linguistics*: *A study of particle placement*. London: Continuum Press.

Gries, St. Th.
(2006) Corpus-based methods and Cognitive Semantics: The many senses of *to run*. In: Gries, St. Th., Stefanowitsch, A. (eds.), Corpora in Cognitive Linguistics: *Corpus-based approaches to syntax and lexis.* Berlin, New York: Mouton de Gruyter, 57–99.

Heylen, K.
(2005) A quantitative corpus study of German word order variation. In: Kepser, St., Reis, M. (eds.), In: *Linguistic evidence: Empirical, theoretical and computational perspectives.* Berlin, New York: Mouton de Gruyter, 241–264.

Lakoff, G.
(1987) *Women, fire, and dangerous things: What categories reveal about the mind.* London: University of Chicago Press.

Stylistic fingerprints, POS tags and inflected languages: a case study in Polish

Rafał L. Górski, Maciej Eder, Jan Rybicki

Keywords: STYLOMETRY, AUTHOR ATTRIBUTION, GRAMMAR, CLASSIFICATION

In classical approaches to authorship attribution, frequencies of the most frequent words (MFWs) and character n-grams are claimed to outperform other types of style-markers (Koppel et al. 2009, Stamatatos 2009), even if their performance varies significantly across different languages (Eder 2011, Eder and Rybicki 2013). Also, it has been proven that letter n-grams reveal a very high resistance to untidily prepared corpora (Eder 2013). Defined as simple strings of letter and non-letter characters, all these style-markers are easily extracted from input texts – in practice, attribution tests can be flawlessly applied to any web-scraped plain text file. Very attractive as they are, these simple style-markers have also some limitations. Firstly, in inflected languages, different forms of the same word cannot be recognized using generic text processing (e.g. regular expressions); secondly, word endings play a prominent role (they are the equivalent of function words) and should not be ignored; thirdly, countless inflected word forms make word frequencies sparse, and this complicates most statistical procedures. The aim of this paper is to examine the possibilities and limitations of recognizing authorship in grammatical features of a text. It is true that extracting syntactic markers from parsed grammatical trees is still beyond our capabilities; however, a straightforward insight into grammar can be obtained using Part-of-Speech tags combined into n-grams. Attempts to solve this problem have already yielded promising results (Baayen et al. 1996, Hirst and Feguina 2007); yet, once again, mostly in English. Morphologically-rich languages with a relatively free word-order, such as Polish, are radically different from the

grammatical point of view. Rich morphology results with a much higher number of tag-types (over a thousand, as compared to ca. 60 for English); free word order makes one expect a considerably higher number of possible POS-tag trigrams. The basic question underlying this paper is not whether this approach outperforms authorial attribution based on lexis; rather, it tries to establish the degree of freedom of choice within lexis and grammar. Does a highly inflected language allow for more variation (which in turn allows for individual style) compared to a positional language as English? Morphologically rich languages are usually annotated with so-called positional tags, i.e. sequences of codes for all the values of grammatical categories which pertain to a word, where only one segment of a tag stands for Part of Speech. A secondary aim of this study is to test the extent to which particular segments of positional tags (analyzed separately and combined into n-grams) are the best features for authorial recognition. Thus, apart from the entire tags, their segments have also been assessed, e.g. n-grams of POS codes, n-grams of single categories, combinations of two tag segments, etc. The corpus consists of several dozen Polish novels. All the texts have been automatically annotated, and the words removed, to operate exclusively on grammatical tags. The texts have been analyzed using standard multidimensional methods, including distance-based explanatory analyses (Cluster Analysis, Principal Components Analysis), and supervised techniques such as Burrows's Delta and Support Vector Machines. Even if using POS-tags n-grams did not improve the attribution performance, interesting differences between different segments (types) of assessed POS tags could be observed.

References

Baayen, H., van Halteren, H., Tweedie, F.
(1996) Outside the cave of shadows: using syntactic annotation to enhance authorship attribution. *Literary and Linguistic Computing*, 11, 121–132.

Eder, M.
(2011) Style-markers in authorship attribution: a cross-language study of the authorial fingerprint. *Studies in Polish Linguistics*, 6, 99–114.

Eder, M.
(2013) Mind your corpus: systematic errors in authorship attribution. *Literary and Linguistic Computing*, 28, 4, 603–14.

Eder, M., Rybicki, J.

(2013) Do birds of a feather really flock together, or how to choose training samples for authorship attribution. *Literary and Linguistic Computing*, 28, 2, 229–36.

Hirst, G., Feiguina, O.

(2007) Bigrams of syntactic labels for authorship discrimination of short texts. *Literary and Linguistic Computing*, 22, 4, 405–17.

Koppel, M., Schler, J., Argamon, S.

(2009) Computational methods in authorship attribution. *Journal of the American Society for Information Science and Technology*, 60, 1, 9–26.

Stamatatos, E.

(2009) A survey of modern authorship attribution methods. *Journal of the American Society for Information Science and Technology*, 60, 3, 538–556.

The Arens-Altmann Law: a matter of boundary conditions or an ostensible success story?

Peter Grzybek

Keywords: Arens-Altmann Law, Menzerath-Altmann Law, sentence length, word length

The Arens-Altmann Law concerns the relation between sentence length and word length. It has been formulated by Altmann (1983), based on earlier observations by Arens (1965). In his book *Verborgene Ordnung*, Arens calculated average sentence length and average word length for 117 literary prose texts and observed an increase of word length to go along with a regular increase of sentence length. After Arens had interpreted this relation to be of a linear kind, Altmann re-interpreted the findings in analogy with the Menzerath-Altmann Law, arguing in favor of a non-linear relation. Subsequent studies have not unanimously been able to corroborate these findings. One reason for this unclear situation is that the two Laws have not always been consequently distinguished, the Menzerath-Altmann Law having been designed for intra-textual relations, the Arens-Altmann Law, in contrast, being of inter-textual relevance. Those studies which have concentrated on the inter-textual relation of the Arens-Altmann Law, arrived at contradictory results. Hug (2004), for example, on the basis of 103 French journalistic texts, found the expected tendency, the results appearing to be of a linear kind. As compared to this, the re-analysis of data from 380 text books from schools, published by Bamberger and Vanecek (1980), showed the expected tendency non-linear tendency. In contrast, the re-analysis of data by Fucks from the 1950s yielded no convincing results (Grzybek and Stadlober 2007); but with $N = 54$, the authors suspected the overall small sample to be a possibly disturbing factor. Yet, also the analysis of 199 Russian texts

also yielded no consistent results (Grzybek et al. 2007). Summarizing the state of the art, the conditions under which the Arens-Altmann Law functions, and the reasons why it may take different shape or even prove to fail under others, seem to be far from conclusive, the only possible reason thus far discussed being data heterogeneity which may eventually be an important factor, the role of which has hitherto remained unclear, however. The present contribution makes an attempt to shed new light on the uncertain situation. Analyzing and re-analyzing more than 500 texts (including Arens's original data, 199 Russian and 180 Croatian texts), some boundary conditions of the Arens tendency will be studied in detail, concentrating on possibly intervening third variables, which thus far have never been systematically controlled, such as (i) text type, (ii) time of origin, or (iii) grade level for which the texts under study were written.

References

Altmann, G.

(1983) H. Arens «Verborgene Ordnung» und das Menzerathsche Gesetz. In: Faust, M., Harweg, R., Lehfeldt, W., Wienold, G. (eds.), *Allgemeine Sprachwissenschaft, Sprachtypologie und Textlinguistik*. Tubingen: Narr., 31–39.

Arens, H.

(1965) *Verborgene Ordnung: Die Beziehungen zwischen Satzlänge und Wortlänge in deutscher Erzählprosa vom Barock bis heute*. Dusseldorf: Schwann.

Bamberger, R., Vanecek, E.

(1984): Lesen – Verstehen – Lernen – Schreiben. Die Schwierigkeitsstufen von Texten in deutscher Sprache. Wien: Jugend & Volk, Frankfurt/M: Diesterweg.

Grzybek, P., Stadlober, E., Kelih, E.

(2007) The Relationship of Word Length and Sentence Length. The Inter-Textual Perspective. In: Decker, R., Lenz, H.-J. (eds.), *Advances in Data Analysis*. Berlin, Heidelnerg: Springer, 611–618.

Grzybek, P., Stadlober, E.

(2007) Do we have problems with Arens' Law? A new look at the sentence-word relation. In: Grzybek, P., Köhler, R. (eds.), *Exact Methods in the Study of Language and Text.* Berlin, New York: de Gruyter, 205–217.

Hug, M.

(2004) La loi de Menzerath appliquée à un ensemble de textes, *Lexicometrica*, 5, 1–10.

On the exact computation of resampled mean size of paradigm (MSP)

Guillaume Guex, Aris Xanthos

Keywords: INFLECTIONAL DIVERSITY, MEAN SIZE OF PARADIGM, LEXICAL DIVERSITY, ROBUSTNESS, RANDOM SAMPLING

Lexical diversity measures are notoriously sensitive to variations of the sample size. In order to deal with this issue, most recent approaches involve the computation of the resampled variety of lexical units, i.e. their average variety in random subsamples of a fixed size drawn from the corpus. This technique, which has been shown to effectively reduce the influence of sample size variations, has been further applied to measures of inflectional diversity such as the average number of wordforms per lemma, also known as the mean size of paradigm (MSP) index. In this contribution we argue that, while random sampling can indeed be used to increase the robustness of inflectional diversity measures, setting a fixed subsample size is only justified under the hypothesis that there are no variations of lexical diversity between the corpora that we compare. In a more general case where they may have differing degrees of lexical diversity, a more sophisticated strategy can and should be adopted. This claim is supported by theoretical as well as empirical arguments. A novel approach to the measurement of inflectional diversity is proposed, aiming to cope not only with sample size variations, but also with lexical diversity variations. Experimental evaluation suggests that although there is still room for improvement, the proposed methodology attenuates the impact of lexical diversity discrepancies.

References

Johnson, W.
(1944) Studies in language behaviour: I. A program approach. *Psychological Monographs*, 56, pp. 1–15.

Malvern, D., Richards, B., Chipere, N., & Durán, P.
(2004) *Lexical diversity and language development: Quantification and assessment*. Basingstoke: Palgrave MacMillan.

Tweedie, F.J. & Baayen, R.H.
(1998) How Variable May a Constant be? Measures of Lexical Richness in Perspective. *Computers and the Humanities*, 32, 5, 323–352.

Xanthos, A.
(2013) L'évaluation de l'évaluation de la diversité lexicale. In: A. Prikhodkine & A. Xanthos (eds.), *Mélanges offerts en hommage à Remi Jolivet*. Cahiers de l'ILSL, 36, pp. 231–252.

Xanthos, A. & Gillis, S.
(2010) Quantifying the development of inflectional diversity. *First Language*, 30, 2, 175–198.

Cross-linguistic transference of authorship attribution

Belinda Hasanaj, Erin Purnell, Patrick Juola

Keywords: AUTHORSHIP ATTRIBUTION, CROSS-LINGUISTIC TRANSFERENCE

Cross-linguistic transference is an important problem in statistical authorship attribution. Many studies have focused on attributing authorship in major languages but not on some of the less common languages. In this study, we performed experiments in Albanian (sole modern survivor of a subgroup of Indo--European languages) and Finnish (polysynthetic) and compared the performance of chosen methods with the methods' performance on well-studied languages: English, Italian, and Greek. We had to build a corpus to be able to do this study. The Albanian corpus consists of forty 20th century prose documents by four authors of Albanian literature. The Italian corpus contains sixty-one 19th century novels by six authors from the Liber database. The Finnish corpus contains four student thesis papers from Finnish universities and consists of eighteen partial-documents. We used the English problems (A-H) from the Ad-hoc Authorship Attribution Competition (AAAC) corpus (Juola 2004). The corpus consists of thirteen problems of mixed genre and size. For Greek, we used a collection of one hundred blog posts by ten authors (Mikros 2013). The documents were divided into training/testing sets at roughly an 80/20 ratio. We used the Java Graphical Authorship Attribution Program with two different chosen analysis methods: Support Vector Machines using Sequential Minimal Optimization (SMO) and Nearest Centroid with three distances: Intersection, Manhattan and Cosine. These were combined with the following textual features: Words, Rare Words, Characters and CharN-Grams (N equal to 3, 4, 6 and 8) (JGAAP, 2005). The average accuracy obtained in the Albanian Corpus was 82.14% and the highest accuracy was obtained using SMO

combined with CharN- Grams. The average accuracy obtained for the Finnish Corpus was 87% and the highest accuracy 100% was obtained with multiple tests including each distance function with Rare Words and CharN-Grams, and SMO with CharN-Grams. The performances obtained in Albanian, English, Finnish, and Italian were not independent. This is supported by the Spearman's product--moment correlation. We display the Spearman correlation since the assumptions of normality, linearity, homoscedasticity (required for Pearson correlation) seem to be violated in some of the cases): 0.6895 (Albanian & Italian), 0.5605 (Albanian & English), 0.5029 (English & Italian), 0.4851 (Finnish & Albanian), 0.3005 (Finnish & Italian), 0.7996 (Finnish & English). The correlations between Albanian, Italian, Finnish, and English were highly significant ($p << 0.01$). The correlation is higher between Albanian and Italian, which might be a result of both being rich in vocabulary and morphology. The interesting results were from Greek and Finnish experiments. The analysis methods did not perform as well in Greek as in the other languages, shown with the Spearman's product-moment correlation: 0.4281 (Albanian & Greek), 0.6952 (Italian & Greek), 0.1586 (English & Greek), -0.0505 (Fin & Greek). Finnish performed extremely well which might be due to the small sample size or the fact that the samples came from the same thesis papers but did not overlap. This is a work in progress; we will be running more experiments in Greek and Finnish and hope to add an additional polysynthetic language.

References

JGAAP,
(2005) Documention: http://evllabs.com/jgaap/w/index.php/Main_Page.

Juola, P.
(2004) Ad-hoc Authorship Attribution Competition, *ALLC/ACH 2004 Conference Abstracts,* Gothenburg: University of Gothenburg.

Mikros, G. K.
(2013) Authorship Attribution and Gender Identification in Greek Blogs. In: Obradović, I., Kelih, E., Köhler, R. (eds.), In: *Selected papers of the VIIIth International Conference on Quantitative Linguistics in Belgrade, Serbia,* April 16–19, 2012. Belgrade: Academic Mind, 21–32.

Distribution pattern of given and new information in written English

Wang Hua

Keywords: INFORMATION DISTRIBUTION, QUANTITATIVE LINGUISTICS, SENTENCE LENGTH, SYNTACTIC COMPLEXITY

"Given information" and "new information" are two vitally important notions in the information structure of language. They have been investigated from different perspectives in many linguistic fields. Among these fields, the distribution and position of them are the most basic and interesting one. However, the current investigations merely provide qualitative methods and explanations with several examples or some experiments but without any quantitative description and evidence from large corpus. This paper investigates the distribution of given and new information using the method of quantitative linguistics and the written part of British Component of the International Corpus of English (ICE-GBW), associating it with length and syntactic complexity of the sentence-initial and sentence-final constituents. Length and complexity of a linguistic construct play a very important role in quantitative linguistics and are two essential components in synergetic linguistics. Results show that, no matter if in the overall calculation of all sentences into the separate calculation of sentences with different length and complexity, the sentence-final constituents are both longer and more complex than the sentence-initial constituents or in the comparison of each sentence, which confirms quantitatively that the new information is often at the end of the sentence and the given is at the beginning of the sentence

References

Arnold, J. E., Losongco, A., Wasow, T. and Ginstrom, R.
(2000) Heaviness vs. newness: The effects of structural complexity and discourse status on constituent ordering. *Language* 76, 28–55.

Halliday, M. A. K.
(1994) *An introduction to functional grammar*. London: Edward Arnold.

Ji, S.
(2007) A textual perspective on Givón's quantity principle. *Journal of Pragmatics*, 39, 2, 292–304.

Köhler, R.
(1999) Syntactic structures: Properties and interrelations. J*ournal of Quantitative Linguistics* 6, 1, 46–57.

Quirk, R., Greenbaum, S., Leech, G. N., Svartvik, J.
(1972) *A grammar of contemporary English*. London: Longman.

Vallduví, E., Engdahl, E.
(1996) The linguistic realization of information packaging. *Linguistics,* 34, 3, 459–520.

Word frequency distribution in genres of modern Chinese

Wei Huang

Keywords: WORD FREQUENCY DISTRIBUTION, CHINESE, GENRES

Word frequency has been an essential discussion topic in quantitative linguistics for a long time. It has been widely studied in many languages, in different ways and angles. The researches presented by Popescu and others (Popescu et al. 2009) are good examples of a multidimensional study. Based on the quantitative analysis that was carried on 20 languages, they discussed and concluded some quantitative characteristics of word frequency in different languages. For instance, they found that the word frequency distributions of these 20 languages follow, all or some of, the Zipf, Zipf-Mandelbrot, zeta, geometric, negative hypergeometric distribution while all parameters have been given for reference. Two questions were brought out and highlighted after their research: the conclusion needed to be tested in more languages; the parameters they found showed differences in texts that were written by different authors or belong to different genres. However, the word frequency distribution of modern Chinese has been studied by few linguists in China. Some researches show that the word frequency distribution follows Zipfian Laws. Nevertheless, the diversity of the texts which these studies used should be enhanced. Moreover, the question whether there is some difference of the word frequency distributions between genres of Chinese is still left not replied. Therefore, we collected texts and took some quantitative exploration of modern Chinese. 360 texts within 12 genres in modern Chinese, including both the written texts, such as news report, news review, academic articles, and the transcriptions of an oral speech, were collected for exploring the word frequency distribution.

Each text was segmented automatically and revised manually. And then the word frequency (spectrum) distribution of each text was examined and every parameter of the distribution formulas was estimated. The mean values of these genres were compared statistically. On the explorations and comparison in texts we found that: 1) the word frequency in texts follow the Zipf distribution, zeta distribution, Zipf-Mandelbrot distribution or negative hypergeometric distribution, while the word frequency spectrum abides by the Waring distribution, zeta distribution, Johnson-Kotz distribution, Yule distribution or Zipf-Mandelbrot distribution, which shows that there is no difference in the word frequency (spectrum) distribution between Chinese and other languages. 2) The parameters of the word frequency (spectrum) distribution as a quantitative characteristic of a text have statistically significant difference in certain genres. For example, the parameter 'a' of the zeta distribution in word frequency distinguishes the news reports from other written texts fairly. We suggest that the quantitative characteristics of the word frequency (spectrum) distribution should be applied in the Chinese text classification as genre indicators.

References

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.

Loan words: a quantitative linguistics perspective

Emmerich Kelih

Keywords: Piotrowski's Law, language contact, loan words, borrowing

In quantitative linguistics language contact is mainly devoted to problems of the diachronic development of the adaption and adoption of loan words and borrowings. In this context the Piotrowski's Law is a well-known tool for the modelling of the frequency of loan words. Furthermore, Piotrowski's Law is used for the statistical description of different development processes in languages (changes at all levels from phonemics to semantics, language acquisition with children etc.). Complementary to this fundamental question the presentation tackles the question of the frequency distribution of loan words (borrowings) within particular predefined semantical classes (body parts, religion, animals, plants etc.). Alongside with descriptive findings about the distribution of the loan words within languages the question of the theoretical modelling the rank-frequency distribution will be discussed in detail. The empirical findings are based on Mandarin Chinese, Slovene, Iraqw, Sakha (Yakut), Lower Sorbian, Thai, Kanuri, British English, Tarifiyt Berber, Selice Romani, which have, according to "Loanwords in the world's languages" (Haspelmath and Tadmor 2009) quite different borrowing rates, resp. loan word frequency profiles.

References

Altmann, G.

(1983) Das Piotrowski-Gesetz und seine Verallgemeinerungen. In: Best, K.-H., Kohlhase, J. (eds.), *Exakte Sprachwandelforschung*. Göttingen: Herodot, 54–90.

Altmann, G.

(1985) On the dynamic approach to language. In: Ballmer, T.T. (ed.), *Linguistic Dynamics.* Berlin, New York: de Grutyter, 181–189.

Haspelmath, M., Tadmor, U. (Hg.)

(2009) *Loanwords in the world's languages. A comparative handbook*. Berlin: de Gruyter.

The Menzerath-Altmann Law in film analysis

Veronika Koch, Peter Grzybek

Keywords: FILM ANALYSIS, MENZERATH-ALTMANN LAW, SHOT LENGTH/DURATION, FILM LENGTH/DURATION

The Menzerath-Altmann Law, initially formulated by Altmann (1980), has been repeatedly corroborated in linguistics for units of adjacent linguistic levels (cf. Cramer 2005); and although the statistical formulation of this Law does not imply transitivity, a relation between units from non-adjacent levels (e.g., sentence and word length) could empirically be shown to exist as well (cf. Grzybek 2013). However, for this condition (i.e., leapfrogging one level) a more complex level of Wimmer and Altmann's Unified Theory is needed, of which the Menzerath--Altmann Law is but a special case (cf. Wimmer/Altmann 2005, 2006). The present contribution attempts to transfer the theoretical framework outlined to the realm of quantitative film analysis, concentrating on the relation between shot length and film length. A shot represents a basic unit of a moving picture, its length usually being defined as the time interval between two transitions (cut, wipe, fade out, etc.). In the field of film analysis, shot length is usually measured in time units (i.e., seconds, or deciseconds), or, more rarely, in the number of frames per shot; as compared to this, film length is measured either in minutes, or in the number of shots per film. At closer sight, it might be reasonable to conceptually and terminologically juxtapose shot / film duration (namely, when time units are used for measuring) to shot / film length (referring to the number of constituting units). Since both measures can, in principle, be mutually translated, the present contribution will initially follow the ambiguous, but established usage of the term 'length'. An attempt will then be made to analyze the relation between shot length / duration to film length / duration. It will be tested if a systematic

relation exists, and if so, if it follows the Menzerath-Altmann Law or, by way of an alternative, a more complex model from the Unified Theory. It will be of particular importance to see if the definition of length vs. duration (and the choice of measuring units resulting from this definition) is relevant for modeling the relations under study. By way of an example, 70 Soviet feature films from the 1920s till the end of the 1980s will be analyzed.

References

Altmann, G.
(1980) Prolegomena to Menzerath's law. In: *Glottometrika* 2. Bochum: Brockmeyer, 1–10.

Cramer, I. M.
(2005) Das Menzerathsche Gesetz. In: Köhler, R.; Altmann, G.; Piotrowski, R.G. (eds.), *Quantitative Linguistik – Quantitative Linguistics. Ein Internationales Handbuch – An International Handbook*. Berlin, New York: de Gruyter, 650–688.

Grzybek, P.
(2013) Close and Distant Relatives of the Sentence: Some Results from Russian. In: Obradović, I., Kelih, E., Köhler, R. (eds.), *Methods and Applications of Quantitative Linguistics*. Belgrade: Academic Mind, 44–58.

Grzybek, P., Koch, V.
(2012) Shot Length: Random or Rigid, Choice or Chance? An Analysis of Lev Kulešov's Po zakonu [By the Law]. In: Hess-Lüttich, E. W. B. (ed.) *Sign Culture. Zeichen Kultur.* Würzburg: Königshausen & Neumann, 169–188.

Wimmer, G., Altmann, G.
(2005) Unified derivation of some linguistic laws. In: Köhler, R., Altmann, G., Piotrowski, R. G. (eds.): *Quantitative Linguistik: Ein internationales Handbuch // Quantitative Linguistics: An International Handbook.* Berlin, New York: Walter de Gruyter, 791–807.

Wimmer, G., Altmann, G.
(2006) Towards a unified derivation of some linguistic laws. In: Grzybek, P. (ed.): *Contributions to the Science of Text and Language. Word Length Studies and Related Issues.* Dordrecht, NL: Springer (Text, Speech and Language Technology, 31), 329–337.

Quantitative index text analyser (QUITA)

Miroslav Kubát, Vladimír Matlach

Keywords: TEXT ANALYSIS, STYLOMETRY, FREQUENCY STRUCTURE, SOFTWARE, STATISTICAL TESTING, GRAPHICAL VISUALIZATION

The poster will present a new software for quantitative text analysis. Quantitative Index Text Analyser (QUITA) covers the most common indicators, especially those connected with the frequency structure of a text. In addition to computing results of the indicators, QUITA provides also statistical testing and graphical visualization of the obtained data. QUITA is a versatile tool with many uses designed for researchers from various disciplines (linguistics, literary criticism, history, sociology, psychology, politics, biology, etc.). The program enables basic text processing functions like creating word lists, text lemmatizing or creating n-grams. The program also provides more advanced tools such as a random text creator or a binary file translator. However, the main part of the software is an indicator computing. Although the authors focused mainly on the indicators connected to the frequency structure of a text (e.g. h-point, entropy, repeat rate, adjusted modulus, Gini's coefficient, lambda), there are also several other characteristics such as thematic concentration, activity and descriptivity or writer's view. The main purpose of QUITA is to provide a user-friendly tool of the quantitative text analysis for researchers (especially from the humanities) without deeper knowledge of quantitative linguistics, statistics and programming. Apart from generating results, QUITA also enables simple statistical comparisons and creating charts. There is no need to use any additional software such as spreadsheet applications or special statistical programs. In sum, QUITA is a program that combines all important parts of any quantitative research: obtaining results,

statistical testing and graphical visualization. The indicators included in QUITA were mostly selected in accordance with the following books: Word Frequency Studies (Popescu et al. 2009), Aspects of Word Frequencies (Popescu et al. 2009) and Metody kvantitativní analýzy (nejen) básnických textů (Čech et al. 2013).

References

Čech, R., Popescu, I. I., Altmann, G.
(2013) *Metody kvantitativní analýzy (nejen) básnických textů*. Olomouc: Univerzita Palackého v Olomouci. (in press)

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.

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

Quantitative psycholinguistic analysis of formal parameters of Czech text

Dalibor Kučera, Jiří Haviger

Keywords: PSYCHOLINGUISTICS, TEXT, PSYCHODIAGNOSTICS, COMPUTATIONAL LINGUISTICS, QUANTITATIVE TEXT ANALYSIS

The paper deals with the use of metric analysis of texts for the processing of Czech verbal communication within psycholinguistic disciplines and its description. It focuses on the application of a computer-assisted procedure, which aims to classify and quantify formal characteristics (FPT) of recorded texts. This system enables the analysis of vast amounts of content data and it also facilitates the comparison of the results with other data, in particular with information about the communicators, such as their personal characteristics. The paper also briefly presents the research P-QPA-FPT, which was carried out in 2013-2014 at the University of South Bohemia in Ceske Budejovice (Department of Pedagogy and Psychology) and which aims at enriching the current state of knowledge of the psycholinguistic analysis of texts, to enhance computer-assisted methods of text analysis, and to provide data for a possible application of the procedure within diagnostic and psychological disciplines.

References

Cheng, K. H. C.
(2011) Further Linguistic Markers of Personality: The Way We Say Things Matters. *International Journal of Psychological Studies*, 3, 1, 2–16.

Mehl, M. R.
(2006) Quantitative text analysis. In: Eid, M., Diener, E. (eds.), *Handbook of Multimethod Measurement in Psychology*. Washington: American Psychological Association, 141–156.

Pennebaker, J. W.
(2011) The secret life of pronouns. *New Scientist*, 3, 50–53.

Towards generalization of sociolinguistic distributions: English loanwords in contemporary written Japanese

Aimi Kuya

Keywords: LOANWORDS, LEXICAL VARIATION, LANGUAGE--EXTERNAL FACTORS, CORPUS STUDY, SOCIOLINGUISTICS

This paper examines the influence of language-external factors on lexical alternation between English loanwords and their Japanese equivalents found in contemporary written Japanese. Adopting the variationist approach, this study recognizes a loanword as a lexical variant of its native equivalents when they share the same meaning. Kuya (2013) identified several statistically significant factors (i.e. author's age, level of education, type of media) that play distinctive roles where the loan noun *keesu* ('case') rivals with its native equivalents. This paper attempts to prove that these factors are also important in predicting the occurrence of other loanwords.

Several most frequently adopted loanwords are examined with respect to their distributions according to the author's age, the gender and level of education and the type of media. The data are collected from the corpora of books and magazines published between 2001 and 2005 (30 million words) in the *Balanced Corpus of Contemporary Written Japanese* (National Institute for Japanese Language and Linguistics 2011).

The statistical analyses show that (i) the author's age, (ii) the author's level of education and (iii) the type of media all have an impact on the occurrence of most loanwords surveyed as is expected from my previous work. First, the tendency for the younger generation to use the loanwords more often proves that there has been a lexical change over time in favor of the loanwords as opposed

to their Japanese counterparts. The 'apparent time' approach has succeeded here in grasping a clearer picture of the language change in favor of loanwords than the 'real time' approach. Second, the loanwords are disfavored by more educated authors contrary to a popular assumption, or the claim in other opinion-based sociolinguistic studies (cf. NINJAL 2004, Tanaka 2007), that more educated people have more positive attitudes towards loanwords and therefore are heavier users of them. Third, loanwords are used less often in books than in magazines, because books seem to be stylistically more formal than magazines in general. Lastly, the gender factor is newly attested with respect to half of the lexical items investigated, with female authors employing loanwords more frequently.

Overall, among the factors examined, the statistical significance of the author's level of education and the type of media are more consistent than that of their age and gender. This implies that the use of loanwords well attested across the generation or gender is still influenced by educational factors and/or stylistic factors such as formality/casualness. In that sense, the status of loanwords in contemporary written Japanese remains peripheral, although some studies point out that many loanwords have been fully developed in terms of grammar and frequency to be recognized as part of 'basic' Japanese vocabulary (cf. Kim 2011).

References

Haugen, E.
(1950) The analysis of linguistic borrowing. *Language*, 26, 2, 210–231.

Kim, E.
(2011) *Shift of the loanwords to basic words in the Japanese newspaper vocabulary in the second half of the 20th century: Monograph on Handai Japanese Studies*, 3. PhD thesis. Osaka University.

Kuya, A.
(2013) Synchronic Distribution of Loanwords in Contemporary Written Japanese: A Case Study of *keesu* ('case'), *NINJAL Research Papers, 6*, 45–65.

Lavandera, B.
(1978) Where does the sociolinguistic variable stop? *Language in Society*, 7, 171–183.

Loveday, L.
(1996) *Language contact in Japan: A socio-linguistic history*. Oxford: Oxford University Press.

National Institute for Japanese Language and Linguistics (NINJAL)
(2004) *An opinion survey on the use of loanwords (nationwide*) (in Japanese). Tokyo: NINJAL.

National Institute for Japanese Language and Linguistics (NINJAL)
(2011) *The Balanced Corpus of Contemporary Written Japanese: Version.1.0* [DVD]. Tokyo: NINJAL.

Tanaka, M.
(2007) Attitudes towards loanwords as opposed to Kango and Wago. *NINJAL Report 126: Loanwords in public media: research in favor of the proposal for changing the wording of loanwords* (in Japanese). 302–310.

Weinreich, U.
(1968) *Languages in contact: Findings and problems*. The Hague: Mouton.

Type-token relation for length motifs in Ukrainian texts

Ján Mačutek

Keywords: LENGTH MOTIF, TYPE-TOKEN RELATION, UKRAINIAN

The type-token relation is a ratio of the number of different units to the number of all units used in a text. There are several mathematical expressions of the relation (cf. Wimmer 2005). In this presentation it will be applied to length motifs (a length motif is a relatively new linguistic unit described shortly by Köhler 2008). Ukrainian texts from seven different genres (prose, drama, blog, sport reportage, sermon, scientific papers - humanities, scientific papers - physics) will be used (the data are described by Kelih et al. 2009). It will be shown that data from almost all texts can be modelled by a simple power Law. The parameter values seem to be dependent on a text genre, hence they could perhaps be used for an automatic text classification (which was successfully applied to German and Russsian texts in Köhler 2006 and Köhler and Naumann 2008). A relation of parameter values to text lengths will be discussed.

References

Wimmer, G.
(2005) The type-token relation. In: Köhler, R., Altmann, G., Piotrowski, R.G. (eds.), *Handbook of Quantitative Linguistics*. Berlin: de Gruyter, 361–368.

Kelih, E., Buk, S., Grzybek, P., Rovenchak, A.
(2009) Project description: Designing and constructing a typologically balanced Ukrainian text database. In: Kelih, E., Levickij, V., Altmann, G. (eds.), *Methods of Text Analysis*. Chernivtsi: ChNU, 125–132.

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

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

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.* Berlin, Heidelberg: Springer, 635–646.

Gender identification in Modern Greek tweets

George Mikros, Kostas Perifanos

Keywords: GENDER IDENTIFICATION, AUTHOR PROFILING, AUTHORSHIP ATTRIBUTION, TWITTER, MACHINE LEARNING

Since its launch in 2006 Twitter has expanded rapidly and become one of the most active social networking sites worldwide. Last usage statistics for 2013 revealed that approximately 5,700 tweets per second are sent from more than 231 million registered accounts. Twitter, has transformed radically the ways information is spread over the internet and created a new language genre with specific linguistic conventions and usage rules. Users form messages in 140 characters or less producing text that is semantically dense, has many abbreviations and often carries extralinguistic information using specific character sequences (smileys, interjections etc.) (Crystal 2008).

Authorship attribution methods have already been applied to tweets, since Twitter is an extremely popular service and cybercrime frequently uses it for illegal activities (Boutwell 2011, Layton et al. 2010, Mikros and Perifanos 2013, Sousa-Silva et al. 2011). However, little attention has been paid to author profiling including gender identification. The present study aims to explore appropriate methods for gender identification in Twitter data. We utilized the Greek Twitter Corpus (Mikros and Perifanos 2013) and selected 10 users based on their popularity (number of followers) and their activity (number of tweets in a month). The selected subcorpus contained 12,973 tweets (4,473 from males / / 8,500 from women) totaling 130,918 words (59,134 words from males / 71,784 words from women). All data were preprocessed removing all @replies, #tags and manual retweets (RT's) in order to focus on the clear text content of the tweets. We counted a vast array of stylometric features (including classic features like

lexical "richness" indices, word and sentence length statistics, frequent words and Author's Multilevel Ngram Profiles (AMNP). The resulting data were used for the training of various Machine Learning Algorithms (Support Vector Machines - SVM, Random Forests and Naïve Bayes) using cross-validations procedures for overfitting avoidance. Preliminary results indicate that gender identification is highly accurate in Twitter data. Our SVM trained algorithm achieved a mean accuracy of 0.97 (s.d. 0.07) across different text-size blocks leading us to the conclusion that gender language differences, observed in everyday communication, persist even on micro-texts and can be reliably utilized for automatic author's gender detection.

References

Boutwell, S. R.
(2011) *Authorship attribution of short messages using multimodal features.* (MSc), Naval Postgraduate School, Monterey, California.

Crystal, D.
(2008) *Txting: The Gr8 Db8*. Oxford: Oxford University Press.

Layton, R., Watters, P., Dazeley, R.
(2010) *Authorship Attribution for Twitter in 140 Characters or Less 2nd Workshop on Cybercrime and Trustworthy Computing Workshop (CTC)*, 19-20 July 2010, Ballarat, Australia, 1–8.

Mikros, G. K., Perifanos, K.
(2013) Authorship attribution in Greek tweets using multilevel author's n-gram profiles. In: Hovy, E., Markman, V., Martell, C.H., Uthus, D. (eds.), *Papers from the 2013 AAAI Spring Symposium „Analyzing Microtext", 25-27 March 2013*, Stanford, California, 17–23. Palo Alto, California: AAAI Press.

Sousa-Silva, R., Laboreiro, G., Sarmento, L., Grant, T., Oliveira, E., Maia, B.
(2011) 'twazn me!!! ; ('Automatic Authorship Analysis of Micro-Blogging Messages. In: Muñoz, R., Montoyo, A., Métais, E. (eds.), *Natural Language Processing and Information Systems* (Vol. 6716). Berlin, Heidelberg: Springer, 161–168.

Three models for the Menzerath's Law

Jiří Milička

Keywords: Menzerath's Law, Menzerath-Altmann Law, structure information

The paper consists of four parts:

(1) Short introducing to the Menzerath's Law and the Gabriel Altmann's models.
(2) Presentation of the new model which is based on the idea that constructs contain regular constituents and structure information.
(3) Presentation of another model which assumes that the Menzerath's Law is related to the inhomogeneities in text.
(4) The last part is dedicated to the reconciliation of these three models and further generalizing.

These models are both theoretically derived and tested against empirical data.

References

Altmann, G.
(1980) Prolegomena to Menzerath's law. In: Grotjahn, R. (ed.), *Glottometrika* 2, 1–10. Bochum: Brockmeyer.

Milička, J.
(2014) Menzerath's Law: The whole is greater than the sum of its parts. *Journal of Quantitative Linguistics*, 21, 85–99.

Sentence semantics, word meaning, and nonlinear dynamics

Hermann Moisl

Keywords: SENTENCE SEMANTICS, WORD MEANING, NONLINEAR DYNAMICS

Sentence semantics in Turing-computational models of language such as generative linguistics is based on Frege's Principle of Compositionality: that the meaning of a sentence is a function of the meanings of its constituent words and of the manner of their combination. The 'manner of combination' has been extensively studied for decades in linguistic modelling of natural language syntax. Word meaning has been far less well developed, but current understanding is that it is a semiotic relation between linguistic objects, that is, words, and non-linguistic ones, that is, mental representations of relevant aspects of the physical world and of mind-internal concepts derived from them. Because these representations are external to its domain, linguistics has not proposed well-defined models of them and therefore has little to say about how the relation between them and words is established. This limits linguistic semantics both scientifically and technologically. The scientific limitation is that a topic at the core of the study of language – how language encodes and communicates information about perceptions of the natural world – is only partially understood. The technological one is that the usefulness of linguistic models in such applications as language understanding components of artificial intelligence systems is compromised by the current vestigial understanding of how the (word, representation) relation can be implemented, as the slow progress of such applications demonstrates. The present paper proposes a nonlinear dynamical word meaning model in which representations and their associations with words are fixed-point attractors in mental phase

space learned from a physical environment. The discussion is in four parts: the first part gives a brief introduction to nonlinear dynamics, the second motivates the application of dynamics to word meaning, the third describes the architecture of the model and presents results derived from an implementation of it, and the fourth articulates the implications of the model for how the 'manner of combination' in sentence semantics should be understood.

Burton-Roberts, N.
(2013) Meaning, semantics, and semiotics. *Perspectives on Philosophy and Pragmatics*, ed. Capone, A., Lo Piparo, F., Carapezza, M., Berlin: Springer.

Fekete, T.
(2010) Representational systems. *Minds and Machines,* 20, 69–101.

Fresco, N.
(2012) The explanatory role of computation in cognitive science. *Minds and Machines,* 22, 353–80.

Geeraerts, D.
(2009) *Theories of Lexical Semantics*. Oxford: Oxford University Press.

Jackendoff, R.
(2003) *Foundations of Language: Brain, Meaning, Grammar, Evolution*. Oxford: Oxford University Press.

Kolen, J., Kremer, S.
(2001) *Dynamical Recurrent Networks.* New York: IEEE Press.

Strogatz, S.
(2000) *Nonlinear Dynamics and Chaos.* Cambridge MA: Perseus Books.

Tino, P., Horne, B., Giles, C., Collingwood, P.
(1995) Finite state machines and recurrent neural networks – automata and dynamical systems approaches. *Technical Report UMIACS-TR-95-1*, Institute for Advanced Computer Studies, University of Maryland.

Van Gelder, T.
(1998) The dynamical hypothesis in cognitive science, B*ehavioral and Brain Sciences,* 21, 615–28.

Van Leeuwen, M.
(2005) Questions for the dynamicist: the use of dynamical systems theory in the philosophy of cognition, *Minds and Machines,* 15, 271–333.

Testing language units of written Chinese via Menzerath-Altmann Law

Tereza Motalová, Lenka Spáčilová

Keywords: LANGUAGE UNITS, WRITTEN CHINESE, MENZERATH-ALTMANN LAW

Language units were determined on the basis of the graphical criterion in particular. The segmentation employing these units was tested via the Menzerath-Altmann Law (MAL), more precisely parameters A, b of the MAL were calculated. The MAL was applied to several sample texts which were written in the simplified form of Chinese characters and in different stylistic styles. The outcomes of the analyses will be discussed and compared to those obtained by analysing of spoken Chinese.

References

Ping, Ch.
(1999) *Modern Chinese: History and Sociolinguistics*. New York: Cambridge University Press.

Hřebíček, L.
(1997) *Lectures on Text Theory.* Prague: Academy of Sciences of the Czech Republic.

Norman, J.
(2012) *Chinese*. New York: Cambridge University Press.

Vochala, J.
(1986) *Chinese Writing System: Minimal Graphic Units*. Prague: Charles University.

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

Andres, J., Benešová, M., Kubáček, L., Vrbková, J.
(2012) Methodological Note on the Fractal Analysis of Texts. *Journal of Quantitative Linguistics.* 19, 1, 1–31.

GB/T 15834 – 2011. *Zhongguo renmin gongheguo guojia biaozhun: Biaodian fuhao yongfa*《中国人民共和国国家标准: 标点符号用法》 *National Standards of People's Republic of China: General rules for punctuation* (2012). Beijing: Zhongguo biaozhun chubanshe (in Chinese).

Acknowledgment

This contribution could be realized thanks to specific university research MŠMT support allotted to Palacký University in Olomouc. The contribution is a is a part of the collective work IGA projects S*egmentation for testing the Menzerath-Altmann law and the hypotheses related to it I (IGA_FF-2012_035)* and *Segmentation for testing the Menzerath-Altmann law and the hypotheses related to it II (IGA_FF_2014_083).*

Structural versus morphological coding. A cross--linguistic study.

Sven Naumann

Keywords: SYNTACTIC COMPLEXITY, CODING REQUIREMENTS, TREEBANKS, ENGLISH, GERMAN, HUNGARIAN

Köhler (2012) discusses a number of hypotheses addressing various properties (such as complexity, depth, length and position) of the syntactic structures found in human languages and their distribution. One of the requirements or principles used in several hypotheses he calls the requirement for minimization of the complexity of a syntactic construction (R or minX for short). The complexity of an expression or constituent is taken to be the number of its immediate (sub) constituents. This principle, he argues, helps to limit the size of the (short-term) memory needed to process the constituent. But as there are quite a number of languages (like Japanese, Walpiri, etc.) which according to standard linguistic analysis demonstrate a rather flat structure with few, but complex constituents, it seems that there are other factors which have to be taken into consideration.

A simple hypothesis is that the size of minX depends on the way syntactic information is coded in a given language. Syntactic information can be coded structurally or morphologically. We should expect that the value of minX is much higher in languages where morphology does not play a vital role in the coding of syntactic information. This hypothesis is explored using syntactic data from various treebanks for English, German and Hungarian.

References

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

An exploration of the "Golden Section" in Chinese contemporary poetries

Xiaxing Pan, Hui Qiu

Keywords: Chinese contemporary poetry, golden section, h-point, word-frequency distribution, Zip-Alekseev model, syllable

In mathematics, the ancient Greek letter φ represents one of the world's most astonishing numbers: 1.61803…, which is named as Golden Section (GS). It has been proved that this number is all around us (Fett 2006). The present study tends to explore the GS of Chinese contemporary poetries. Poetries are visually realized onto paper as texts. Many studies (Martináková et al. 2008, Popescu et al. 2009, 2012, Tuzzi et al. 2010) on GSs of texts concentrate on the so-called "h-point" in the sequence of the word-frequency curves. Cited from Hirsch (2005), the "h-point" is introduced into linguistics by Popescu and Altmann (2007, 2009). The three points P_1 (1, f(1)), P_2 (1, f(V)) or (V, f(V)), H (h, h) on any rank-frequency curve can form a triangle, in which the ∠α (∠P_1HP_2) of "h-point" is metaphorically called as the "writer's view", whose value converges on the GS (i.e. 1.618). The collected data in Tuzzi (2010: 95-106) fit this trend well. For instance, the value of ∠α on the distribution curve of the 60 Italian texts fits the function:

$y=1.618+8.7094/x^{0.5}$ ($R^2=0.833$). However, GS in Chinese contemporary poetry texts acts differently. The data in the present study consists of 297 poems published since 1917. Fitting the corresponding data about the 297 poems into the function:

$y=1.618+a/x^{0.5}$, we get $a=4.0934$, $R^2=0.39534$, which means an unsatisfactory result. And then we fit the data into this function: $y=a+b/x^{0.5}$, and get

a=1.58916, b=5.02352, R^2=0.40518, which still says an unsatisfactory result. Accordingly, we may draw a conclusion that the radian of ∠α sitting on the "h-point" cannot tell the GS of Chinese contemporary poetry texts well.

We then turn to the study of rhythm and find out that the word-frequency distribution of the poems act differently when we fit them into the Zipf-Alekseev model: $x^{(a+b*\ln x)}$. When the lines of a poetry are less than 20, and word types less than 76 at the same time, it fits the model quite well, but fails if any one of the condition changes. Meanwhile, the number of syllables plays an important role in the fitting process. To measure the Chinese contemporary poems' rhythm, the basic unit "foot" is discussed. Mostly, one "foot" is considered to be composed by two to three Chinese syllables (generally, a syllable is a Chinese character), and one line of one poem is composed by three to five "feet". Thus, we suppose that the GS of Chinese contemporary poetry texts may hides in the arrangement of the syllables, especially the proportion between the monosyllables and multisyllables (including the disyllables), the so-called "foot" in the verses, etc.

We are looking forward to the certification in the study.

Even Tuzzi (2010: 40) concludes: "The assumption that the tendency to the GS does exist but except for text size no other causes of differences could be found", when we are seeking the GS of Chinese contemporary poems, it is a good attempt to take rhythm into account.

References

Tuzzi, A., Popescu, I.-I., Altmann G.
(2010) *Quantitative Analysis of Italian Texts*. Lüdenscheid: RAM-Verlag.

Fett, B.
(2006) An In-depth Investigation of the Divine Ratio. *The Montana Mathematics Enthusiast*, 3, 2, 157–175.

Hirsch, J. E.
(2005) An index to quantify an individual's scientific research output. *PNAS*, 102, 46, 16569–16572.

Popescu, I.-I., Altmann, G.
(2007) Writer's view of text generation. *Glottometrics*, 15, 71–81.

Popescu, I.-I., Altmann, G.
(2009) A modified text indicator. In: Kelih, E., Levickij, V., Altmann, G. (eds.), *Problems of Quantitative Text Analysis*. Černivci: RUTA. 208–229.

Popescu, I.-I., Mačutek, J., Altmann, G.
(2009) *Aspects of Word Frequencies*. Lüdenscheid: RAM-Verlag.

Popescu, I.-I., Čech, R., Altmann, G.
(2012) Some Geometric Properties of Slovak Poetry. *Journal of Quantitative Linguistics,* 19, 2, 121–131.

Martináková, Z., Mačutek, J., Popescu, I.-I., Altmann, G.
(2008) Some problems of musical texts. *Glottometrics*, 16, 80–110.

Modelling proximity in a corpus of literary retranslations: a methodological proposal for clustering texts based on systemic-functional annotation of lexicogrammatical features

Adriana Pagano, Giacomo Figueredo, Annabelle Lukin

Keywords: CLUSTER ANALYSIS, TRANSLATED TEXT, LITERARY TRANSLATION, RETRANSLATION, SYSTEMIC FUNCTIONAL GRAMMAR, LEXICOGRAMMATICAL FEATURES

This paper seeks to contribute to a model for quantitative exploration of literary translations by adopting clustering techniques to search for patterns of comparability in a corpus of retranslations (Oakes and Ji 2012). Drawing on systemic functional grammar (Halliday and Matthiessen 2004) as a framework for text analysis, it reports on an exploratory study aimed at investigating source – target text relations as computed through statistical methods for a manually annotated representative text sample. Unlike approaches based on frequency of the most frequent words, the frequency of values attributed to text samples for pre-defined number of variables was used in order to obtain clusters to map proximity relations as visualized in a dendrogram (Ke 2012). The main purpose of our study was to develop a useful analytical framework for comparing source and target texts on the basis of theory-informed categories of functions realized by choices in the grammar of each language system. As the categories refer to functions description under a common general theory, they apply to variation in the functional organization across language systems, each language having its particular lexicogrammatical realizations. Comparability was thus ensured between texts written in different languages to which equivalence is assigned

because they stand in a relation of translation to one source text. Additionally, our study sought to explore the claim that retranslations – literary works translated more than once into the same language by different authors over a period of time – tend to be more source-oriented than first translations (Berman 1990). Orientation was interpreted as proximity between texts in a cluster relation or, in other words, the distance between the source and target texts as computed through cluster analysis. The results point to the productivity of the methodology used. The clusters obtained at each step point out that similarities between texts computed on the basis of categories ascribed to the lexicogrammatical choices made by each author within the grammatical systems analysed. These pertain to the metafunctional organization of the meanings construed and enacted, which we argue is relevant to the theme of the short story. As regards the retranslation hypothesis, our results corroborate findings by researchers who used other methodologies (Desmidt 2009, O'Driscoll 2009). While they seem to confirm the relative distance of a first translation from the source text, retranslations show varied degrees of proximity to the source text and are sometimes further away from it than first translations. In the case of Katherine Mansfield, from the perspective of norms or the factors conditioning text rewriting, more recent retranslations would be expected to be closer to the source text in that literary criticism has brought rereadings of Mansfield's work and the significance of her use of language to craft her texts. However, there are other factors at play in retranslation (publishing house agendas, target readers) and the retranslation hypothesis did not prove sufficiently sensitive to them.

References

Berman, A.
(1990) La retraduction come espace de la traduction. *Palimpsestes,* 4, 1-7.

Desmidt, I.
(2009) (Re)translation Revisited. *Meta,* 54, 669-683.

Halliday, M. A. K., Matthiessen, C. M. I. M. (eds.)
(2004) *An Introduction to Functional Grammar.* London: Arnold.

Ke, S.-W.
(2012) Clustering a translational corpus. In: Oakes, M.P., Meng, J. (eds.) *Quantitative Methods in Corpus-Based Translation Studies: A Practical Guide to Descriptive Translation Research*. Amsterdam: John Benjamins.

O'Driscoll, K.
(2009) *Around the World in Eighty Changes: A diachronic study of the multiple causality of six complete translations (1873-2004), from French to English, of Jules Verne's novel Le Tour du Monde en Quatre-Vingts Jours*. PhD diss., Dublin City University.

Oakes, M. P., Meng J. (eds.)
(2012) *Quantitative Methods in Corpus-Based Translation Studies: A Practical Guide to Descriptive Translation Research*. Amsterdam: John Benjamins.

The Krylov Law as a tool for comparative lexicology. The example of Polish 19th century dictionaries.

Adam Pawłowski

Keywords: Krylov Law, Polish lexicography, Polish dictionaries, quantitative lexicology, polysemy, homonymy

Statistical laws of language can be divided into two categories: those describing dependencies between frequencies of linguistic units and other measurable variables observed in texts (e.g. Menzerath-Altmann or Zipf Laws), and those defining dependencies of measurable variables in the system of language. The Krylov Law belongs to the latter group. In general terms it describes systemic polysemy in the lexicon of a given language (Krylov 1982). It stipulates that if in the lexicon of a given language *y* is the number of lexemes having *x* meanings, the relationship *y=f (x)* will be a discrete and strictly decreasing, monotonic function, approximated by some continuous function. It is of interest that this relationship had been noted by Zipf, who, however, did not develop it into a fully-fledged law. There have been several theoretical models of the Krylov Law (Kułacka 2009, cf. also Ruszkowski 2010). In one of its most basic versions, it was hypothesized that if x_1 is the number of single-sense lexemes, the number x_i of lexemes having *i* meanings will follow the geometric series: *x_(i+1)=x_1/2^i*

Very few researchers have noticed, however, that the fundamental difficulty in interpreting the Krylov Law is the way lexeme meanings are defined and quantified. The majority of available analyses have relied on the data selected from contemporary dictionaries. Unfortunately, each dictionary represents a network of meanings that reflects its author's lexicographical knowledge and approach to lexicon, but not necessarily the real polysemy of language, observed in specific

utterances. One should also observe that the difference between polysemy and homonymy is often unclear. For these reasons, at this stage of research, the Krylov Law should be regarded as an efficient quantitative tool for comparative lexicology rather than as a universal law of language.

The objective of this presentation is to compare, using the Krylov Law, four general dictionaries of Polish, published in the 19th and 20th centuries. Since the overall number of entries in these dictionaries is enormous, the representative method will be applied and approximately 4,000 random entries per dictionary will be analysed. The hypothesis to be tested is that every dictionary represents its own, specific structure of polysemic words.

References

Krylov, J. K.
(1982) Eine Untersuchung statistischer Gesetzmäßigkeiten auf der paradigmatischen Ebene der Lexik natürlicher Sprachen. In: Guiter H., Arapov M.V. (eds.), *Studies on Zipf´s law.* Bochum: Brockmeyer, 234–262.

Kułacka, A.
(2009) Procedura weryfikacji prawa Kryłowa [Verification procedure of the Krylov Law]. *LingVaria* 2, 8, 10–20.

Ruszkowski, M.
(2010) Prawo Kryłowa w polskich badaniach językoznawczych [Krylov Law in Polish linguistic research]. *Respectus Philologicus* 18, 23, 58–63.

Evolutionary derivation of laws for polysemic and age-polysemic distributions of language signs ensembles

Vasiliy Poddubny, Anatoly Polikarpov

Keywords: ASSOCIATIVE SEMANTIC POTENTIAL OF LINGUISTIC SIGNS, DISSIPATIVE STOCHASTIC MODEL, DERIVATION OF LAWS FOR POLYSEMIC AND AGE-POLYSEMIC DISTRIBUTIONS

The proposed continuous dissipative stochastic model is based on the assumption of the dissipative character of language signs' polysemy development. (Some other aspects of the model one may see in works (Poddubny and Polikarpov 2011, 2013; 2013; Polikarpov, 2013). This means that each linguistic sign at the moment of its birth has a personal limit G (called associative semantic potential, ASP) of the possible maximum number of meanings for acquisition during its life-span. ASP is gradually wasted in the process of meanings acquisition by a sign. The rate of formation of new meanings (as a rule, each next being relatively more abstract) is assumed to be at each moment proportional to the still untouched part of the ASP. Due to this the emergence rate of new meanings is gradually slowing down. At the same time, but with a lag time τ_0, starts to flow a similar and significantly slower process of losses of initially acquired relatively more specific meanings. The actual polysemy of the sign at any moment of time t of its life cycle is expressed by the difference $x(t) = x_1(t) - x_2(t)$ of the processes of acquiring $x_1(t)$ and loss of meanings $x_2(t)$. Continuous model assumes that these processes are continuous and subject to linear differential equations of the type:

$dx_1(t)/dt = (G - x_1(t))/\tau_1$, $x_1(t \le 0) = 0$, $dx_2(t)/dt = (G - x_2(t))/\tau_2$, $x_2(t - \tau_0 \le 0) = 0$,

where $\tau_1 = a_1/G$, $\tau_2 = a_2/G$ – are inversely proportional to the ASP time constants

of the growth and decline of polysemy, respectively, and $\tau_1 << \tau_2$. The model assumes further that the values G in the signs' ensemble and lag τ_0 are independent exponentially distributed random variables with parameters (mathematical expectations) $<G>$ and $<\tau_0>$ respectively. The flow of signs' births is assumed to be a stationary Poisson process with intensity $\lambda = 1/<\tau>$, so that the intervals of time between the emergence of adjacent signs in the flow are assumed to be independent exponentially distributed random variables with a parameter (mathematical expectation) $<\tau>$. The value of $<\tau>$ in the model is considered as a unit of time. Positions of signs in the stationary Poisson thread have the uniform distribution. Thus, the proposed model is described by four parameters: $<G>$, $<\tau_0>$, $a1$, $a2$. The individual life cycle curve of sign's polysemy development $x(t, G, \tau_0)$ in the time t, dependent also on G and τ_0, is obtained from the solution of the above differential equations. Conditional density distribution of polysemy x for an individual sign, depending on time (on a sign's age) t, G and a lag τ_0, is present by Dirac delta-function $p(x|t, G, \tau 0) = \delta(x - x(t, G, \tau_0))$. Averaging the δ-shaped conditional density on distributions of lag $\tau 0$ and G we obtain the conditional distribution density $p(x|t)$ of polysemy x for an ensemble of signs of the same age t. By subsequent averaging on the uniform distribution of t we can get unconditional density $p(x)$ for the distribution of signs' polysemy in an ensemble regardless to the age of signs. By comparing theoretically derived polysemy distribution with the empirical distributions for Russian and English we can define the parameters of the model $<G>$, $<\tau_0>$, a_1, a_2. The relevance of the model was tested by the polysemyage data.

References

Poddubny V. V., Polikarpov A. A.

(2011) Dissipative Stochastic Dynamic Model of Development of Linguistic Signs. *Computer Research and Simulation,* 3, 2, 103–124. (In Russian).

Poddubny V. V., Polikarpov A. A.

(2013) Stochastic Dynamic Model of Evolution of Language Sign Ensembles. In: Obradović, I., Kelih, E., Köhler, R. (eds.), *Methods and Applications of Quantitative Linguistics - Selected papers of the 8th International Conference on Quantitative Linguistics (QUALICO).* Belgrade: Academic Mind, 69–83.

Polikarpov A. A.

(2013) Life-cycle Model for a Sign: Theoretical Foundations of Historical Lexicology and Word Formation In: Chernyshova M.I., Azbukovnik, M. (eds.), *Slavic Lexicography. International collective monograph*, 679–702. (In Russian).

Quantitative studies in the corpus of Nko periodicals

Andrij Rovenchak

Keywords: Nko script, Maninka language, Manding languages (West Africa), text corpus, Zipf's Law, Menzerath-Altmann Law

Nko is an alphabet created in 1949 by Solomana Kantɛ as a writing system for the Manding languages of West Africa. Presently, the script is mainly used for the Guinean Maninka (Maninka-Mori) promoting the establishment of the literary norm. Nko is the only modern indigenous African script, in which a significant body of literature exists (not counting the Ethiopic script, which is of an ancient origin). Moreover, a lot of materials are available electronically facilitating the task of automatic text processing. The corpus of Nko periodicals has been compiled. The following journals and newspapers were included: Dalu Kɛndɛ (ca. 225 thousand words), Yereya fɔbɛ (ca. 122 thousand words), Sinjiya fɔɔbɛ (ca. 76 thousand words), Jansan (ca. 40 thousand words), Yelen (ca. 23 thousand words), Mandenka (ca. 9 thousand words), and Sekutureya Kibaro (ca. 4 thousand words). The total size of the corpus is above 500,000 running words.

The Nko script is now a part of the Unicode standard but most of the periodicals were produced using pre-Unicode fonts, often with different encodings. At least four encodings were discovered and convertors to the Unicode were created for them. Statistical properties of Nko texts are analyzed. In particular, two regimes in the rank–frequency dependence are observed, which suggest a possible fuzzy border between core and marginal vocabulary. Preliminary analysis demonstrates the size of the core vocabulary of about 1,000 most frequent words. In particular, the Zipf exponent for ranks 10–1,000 is (–1.132 ± 0.003),

while for ranks 1,000–10,000 its value is already (–1.225 ± 0.001). Note that the final numbers can change slightly as the frequency list is checked for some most typical errors. Frequency studies are made for phonemes, tonal patterns, syllable co-occurrences, etc. The Menzerath–Altmann Law is checked and suggestions towards a proper unit to measure syllable length in the Maninka language are made. The present study is the first quantitative analysis of a Manding language, where an electronic corpus of such a large size is used. On the basis of the obtained list of types the work towards the creation of the Maninka dictionary in the Toolbox format has been started. The work on the Maninka corpus, with periodicals as its part, is planned within the program LabEx EFL, Strand 6, opération "Corpora for Manding Languages".

References

Ferrer-i-Cancho, R., Solé, R. V.
(2001) Two regimes in the frequency of words and the origins of complex lexicons: Zipf's law revisited. *Journal of Quantitative Linguistics*, 8, 3, 165–173.

Rovenchak, A.
(2011) Phoneme distribution, syllabic structure, and tonal patterns in Nko texts. *Mandenkan*, 47, 77–96.

Vydrin, V.
(2012) Une bibliographie préliminaire des publications maninka en écriture N'ko. *Mandenkan,* 48, 59–121.

Translations in networks: the (in)visibility of translator styles

Jan Rybicki

Keywords: STYLOMETRY, AUTHORSHIP ATTRIBUTION, AUTHORIAL SIGNAL, CLUSTER ANALYSIS, BURROWS'S DELTA, NETWORK VISUALIZATION

Methods based on statistical analyses of the most frequent word frequency series are well-established in authorship attribution; it is usually enough to compare frequencies of several dozen most frequent words in a corpus to order the texts comprised therein by author. When the same stylometric methods are applied to corpora of translations, the results continue to present the domination of the authorial signal, somewhat counterintuitively when one considers the obvious fact that they are now based on very different frequencies of the most frequent words in another language, and that corresponding most-frequent-word lists in the source and target languages exhibit little one-on-one correspondence between individual lexical items. The translatorial signal is sometimes visible in two cases: 1. when a given translation is known to depart from others in its style, quality and/or chronology; and/or 2. when translations of the same work of the same author are compared. Also, individual variations in a translator behaviour can also be observed. This phenomenon is presented in this paper using the stylo package for the statistical programming environment R. The package processes the texts to produce word frequencies; these are converted into text size-independent distance measures (in this case, Burrows's Delta), which in turn serve as an input for the cluster analysis; finally, the results obtained for different values of several parameters (word frequency list length; pronoun deletion; selectiveness of the wordlist) are pooled for better consistency and processed in the GEPHI

network visualization programme. This approach allows a more complete image of patterns of similarity and difference between individual texts, original authors and translators, with numerous linkages possible, in contrast to the simple binary character of nearest-neighbour associations in traditional cluster analysis trees. The material analysed here includes novel-size literary texts in English, French and Polish; texts in one language are compared to their translations into the other two. Apart from the more immediate insight into patterns of similarity and difference between texts and authors of interest to literary scholars, this phenomenon is of significance for translation studies as an unexpected vindication of Venuti's "translator's invisibility." After all, most practitioners of literary translation (including the author of this paper) would probably agree that they suspected their own "stylistic fingerprint" to be visible rather than concealed in the usage of the most frequent words, most of them "synsemantic." While there is no theory yet behind the empirical success of most-frequent-word-based authorship attribution methods, the possible explanations might include references to the model of language and meaning as proposed by cognitive linguistics.

References

Bastian, M., Heymann, S., Jacomy, M.
(2009) Gephi: an open source software for exploring and manipulating networks. *International AAAI Conference on Weblogs and Social Media*.

Burrows, J.
(1987) *Computation into Criticism: a Study of Jane Austen's Novels and an Experiment in Method*. Oxford, Clarendon Press.

Burrows, J.
(2002) Delta: a measure of stylistic difference and a guide to likely authorship. *Literary and Linguistic Computing*, 17, 3, 267–87.

Eder, M.
(2011) Style-markers in authorship attribution: a cross-language study of the authorial fingerprint. *Studies in Polish Linguistics*, 6, 99–114.

Eder, M., Kestemont, M., Rybicki, J.
(2013) Stylometry with R: a suite of tools, *Digital Humanities 2013 Conference Abstracts. Lincoln: University of Nebraska-Lincoln,* 487–489.

Hoover, D. L.
(2004a) Testing Burrows's delta. *Literary and Linguistic Computing*, 19, 4, 453–75.

Hoover, D. L.
(2004b) Delta prime? *Literary and Linguistic Computing*, 19, 4, 477–95.

Jockers, M. L., Witten, D. M.
(2010) A comparative study of machine learning methods for authorship attribution. *Literary and Linguistic Computing*, 25, 2, 215–23.

Luyckx, K., Daelemans, W.
(2011) The effect of author set size and data size in authorship attribution. *Literary and Linguistic Computing*, 26, 1, 35–55.

Mosteller, F., Wallace, D.
(1964) *Inference and Disputed Authorship: The Federalist.* Addison-Wesley, Reading.

R Core Team.
(2013) *A language and environment for statistical computing.* Wien: R Foundation for Statistical Computing, http://www.R-project.org/ [1.1.2014].

Rybicki, J.
(2009) Translation and Delta revisited: when we read translations, is it the author or the translator that we really read? *Digital Humanities 2009: Conference Abstracts.* University of Maryland, College Park, MD, 245–47.

Rybicki, J.
(2012) The Great Mystery of the (Almost) Invisible Translator: Stylometry in Translation. In: Oakley, M. Ji, M. (eds.) *Quantitative Methods in Corpus-Based Translation Studies.* Amsterdam: John Benjamins, 231–248.

Rybicki, J., Eder, M.
(2011) Deeper delta across genres and languages: do we really need the most frequent words? *Literary and Linguistic Computing*, 26, 3, 315–21.

Rybicki, J., Heydel, M.
(2013) The stylistics and stylometry of collaborative translation: Woolf's Night and Day in Polish. *Literary and Linguistic Computing*, 28, 4, 708–717.

Venuti, L.
(1995) *The Translator's Invisibility. A history of translation.* London-New York: Routledge.

A co-occurrence and an order of the valency in Japanese sentences

Haruko Sanada

Keywords: VALENCY, CO-OCCURRENCE, WORD ORDER, POSTPOSITION, JAPANESE

Japanese is known as the language with less restrictions of the word order. Many postpositions express a role of the case or the complement in the sentence, and ellipses of any valencies can be possible to avoid a redundancy. However, in the recent studies it is discussed that there are broad rules for co-occurrences of postpositions and for an order of valencies.

In this paper we use a valency database which shows sample sentences with (a) cases and complements as the deep case and (b) postpositions as the surface case. We studied following problems from the point of view of the deep case and the surface case respectively:

1) Is there a tendency of co-occurrences of postpositions?

2) Is there a tendency of the order of postpositions in the sentence?

2) Method and data

Employing the valency database (Ogino et al. 2003), we focused on the 6 verbs in 1,051 sentences of which the following are the same as employed in our last studies (Sanada 2012, 2014) „au“ (meet), „hataraku“ (work), „yaburu“ (tear, break), „umareru“ (be born, arise), „ugoku“ (move), and „ataeru“ (give). We extracted 621 pairs of cases and complements as the deep case and 601 pairs of postpositions as the surface case and analyzed them.

3) Results

We observed the following points for the co-occurrence of complements or postpositions:

- The frequency of the pairs of complements or postpositions is individual by verb.

- Considering the order of complements or postpositions, the frequency of the pairs is not symmetric. For example, the frequency for the pair of „ga“ (subject) and „ni“ (direct object) for the verb „au“ (meet) is 38, while the frequency for the pair of „ni“ (direct object) and „ga“ (subject) is 0.

To measure the tendency of the order of complements or postpositions in the sentence, we give the score of 2 points to the first complement or postposition of the pairs and a score of 1 point to the other one, and calculate the average scores by complement or postposition. We obtained the order of complements or postpositions sorted by their average scores, and used t-test to evaluate the significance of the difference. It is confirmed that postpositions for the theme and the subject have a „higher“ position in the sentence and the postposition for the direct object has a „lower“ position. However, the order of postpositions is individual by verb.

References

Ogino, T., Kobayashi, M., Isahara, H.
(2003) *Nihongo Doshi no Ketsugoka* (Verb valency in Japanese). Tokyo: Sanseido.

Sanada, H.
(2012) Joshi no Shiyo Dosu to Ketsugoka ni Kansuru Keiyoteki Bunseki Hoho no Kento (Quantitative approach to frequency data of Japanese postpositions and valency) (in Japanese). *Rissho Daigaku Keizaigaku Kiho (The quarterly report of economics of Rissho University)*, 62, 2, 1–35.

Sanada, H.
(2014, to appear) The choice of postpositions of the subject and the ellipsis of the subject in Japanese. In: Uhlířová, L., Altmann, G., Čech, R., Mačutek, J. (eds.), *Issues in Quantitative Linguistics*. Lüdenscheid: RAM-Verlag.

Authorship attribution using political speeches

Jacques Savoy

Keywords: AUTHORSHIP ATTRIBUTION, DISCOURSE ANALYSIS, POLITICAL SPEECHES

In this communication we describe a lexical study over the State of the Union addresses from 1790 through 2014 corresponding to 225 speeches uttered by 41 US presidents. We view each address as a compound signal that includes the author's style and the topics (speech content). Moreover, since all these speeches have the same audience, context, and genre, they form an interesting corpus for authorship attribution. They are also freely available and correctly spelled. However, if we can assume that many US presidents who lived in the 18th or 19th century were in fact the real authors of their own speeches, this hypothesis is wrong for the last presidencies (e.g., T. Sorensen was the ghost writer for J. F. Kennedy, and J. Fraveau is behind B. Obama). In the latter case, we can assume that the same person was in charge of the State of the Union addresses for a given president.

After applying a POS tagger, we represent each speech by its lemmas (entry in the dictionary) instead of directly using the surface words. According to previous studies in authorship attribution, we have taken into account the k most frequent lemmas (with k varying from 100 to 500) to define the style. We then present a set of experiments on authorship attribution trying to assign each speech to the assumed author achieving a success rate around 85% (using KLD (Zhao and Zobel 2007), Labbé's (2007) distance, naïve Bayes (Mitchell 1997), or the Delta method (Burrows 2002)). Using only the POS tags, the resulting success rate was lower and the combination of lemmas with POS information does not always present better performance levels. In these experiments, we found a set of speeches that

were always difficult to assign correctly. A deeper analysis reveals interesting reasons explaining the underlying difficulty to assess these cases.

In the second set of experiments, we have assigned each address based on their contents. In this perspective, we assume that each presidency wants to impose its own political agenda and thus will focus on some particular questions distinct from problems presented by other presidents. These objectives will therefore be present in their speeches and in the choice made to express its intents (e.g., in the vocabulary). In this perspective, we represent each speech by their lemmas, ignoring the most frequent ones (e.g., 100 to 500), and the lemmas appearing once or twice. The success rate was interesting (around 75%) but lower than when considering only the author's style. This high success rate tends to indicate that our underlying assumption is correct and each presidency tends to have a set of distinctive issues that they focus on (e.g., health care reform, new jobs, or clean energy with Obama).

Finally, we propose to combine the style and the content to improve the quality of the authorship attribution scheme. The resulting performance shows mixed results, depicting in some cases a small improvement, and degradation in other cases. The idea that combining different sources of evidence will always provide a better result is not always true.

References

Burrows, J. F.
(2002) Delta: A Measure of Stylistic Difference and a Guide to Likely Authorship. *Literary and Linguistic Computing*, 17, 3, 267–287.

Labbé, D.
(2007) Experiments on Authorship Attribution by Intertextual Distance in English. *Journal of Quantitative Linguistics*, 14, 1, 33–80.

Mitchell, T. M.
(1997) *Machine Learning*. New York: McGraw-Hill.

Zhao, Y., Zobel, J.
(2007) Entropy-Based Authorship Search in Large Document Collection. In: *Proceedings ECIR*. Berlin: Springer, 381–392 (LNSC 4425).

Testing language units of spoken Chinese via Menzerath-Altmann Law

Jana Ščigulinská, Denisa Schusterová

Keywords: language units, spoken Chinese, Menzerath--Altmann Law

Language units were primarily determined on the basis of the phonetic criterion. The segmentation employing these units was tested via the Menzerath-Altmann Law (MAL), more precisely parameters *A*, *b* of the MAL were calculated. The MAL was applied to several samples which were produced by native speakers who represented different musical genres. The outcomes of the analyses will be discussed and compared to those obtained by analysing of written Chinese.

References

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

DeFrancis, J.
(1990) *The Chinese Language Fact and Fantasy.* Taipei: The Crane Publishing co., LTD.

Hřebíček, L.
(1997) *Lectures on Text Theory.* Prague: Oriental Institute.

Ping, Ch.
(1999) *Modern Chinese*. Cambridge: Cambridge University Press.

Švarný, O. (et al.)
(1998) *Hovorová čínština v příkladech 3*. Olomouc: Vydavatelstvo Univerzity Palackého.

Acknowledgment

The contribution is a part of the collective work IGA projects *Segmentation for testing the Menzerath-Altmann Law and the hypotheses related to it I (IGA_FF-2012_035)* and *Segmentation for testing the Menzerath-Altmann Law and the hypotheses related to it II (IGA_FF_2014_083).*

Comparative rates of change as a diagnostic of vowel phonologization

Betsy Sneller

Keywords: LANGUAGE CHANGE, SPOKEN CORPORA, SOCIOLINGUISTICS

Problem: The mechanism of vowel change, particularly of phonological changes like vowel splits and mergers, has largely been a matter of ideological speculation. Whether phonological reanalysis is the result of incremental sound changes (e.g., Ohala 1981) or the result of phonologization ex nihilo (e.g., Janda and Joseph 2003, Fruehwald 2013) has been more of a theoretical debate than an empirical one. Using large-scale corpora of spoken interviews, it becomes possible to actually test these two competing hypothesis.

Background: It is common to examine the formant space of different allophones, as a diagnostic of distinct phonological targets. When two allophones have similar long-term trajectories across F1-F2 space, they are often analyzed as being a single phonological target. This paper aims to expand on the traditional analysis, adding comparative rates of change as an additional dimension for analysis.

Methods: Data for this study are drawn from two large-scale corpora of spontaneous sociolinguistic interviews: the Philadelphia Neighborhood Corpus (PNC), which consists of 30-60 minute interviews with 214 speakers with dates of birth ranging from 1888-1996, and the Origins of New Zealand English (ONZE) corpus, which consists of 30-60 minute interviews with 512 speakers with dates of birth ranging from 1851-1987. Vowel tokens from each speaker were extracted using the FAVE suite (Rosenfelder et al. 2012), and normalized using z-scores. Using these two corpora which are stylistically similar but from distinct

populations allows us to make stronger generalizations about the usefulness of rate of change as a diagnostic of phonologization. Rates of change were calculated as the first derivative of a locally weighted smoothed regression model of the individual vowel tokens.

Predictions: If two allophones are different only due to phonetic coarticulation, their rates of change will be identical (if coarticulation has a linear effect on vowel quality) or linearly related (if coarticulation has a quadratic effect on vowel quality). Similarly, if two allophones are becoming distinct from each other due to incremental phonetic changes, then their rates of change should also be related, since the same phonological process is acting on both phonological targets. When reanalysis occurs, however, each target is open for distinct phonological targets. In this case, we should expect the rates of change for each phoneme to be unrelated.

Results: We test this hypothesis on several sound changes in the PNC which are known to be phonetic (the fronting of [o][n] compared to [o]; [t][u] compared to [u]), as well as several which are known to be phonological ([ey] in closed syllable compared to open; [ay] before a voiceless segment compared to voiced) and find that comparing rates of change is a useful diagnostic of phonological distinctness. We then use this as a tool for showing a new phonological reanalysis in the ONZE corpus of [u] following [j], as well as corroboration of the [ir]-[er] merger in New Zealand English. We find that in cases of phonological split or merger, the comparative rates of changes suggest that reanalysis occurs almost immediately, supporting a `Big Bang' theory of phonologization.

References

Fruehwald, J.

(2013) *The phonological influence on phonetic change*. PhD dissertation: University of Pennsylvania.

Janda, R., Joseph, B.

(2003) Reconsidering the Canons of Sound-Change: Towards a „Big Bang" Theory. In: Blake, B., Burridge, K. (eds.), *Historical Linguistics. Selected Papers from the 15th International Conference on Historical Linguistics*, Melbourne. Amsterdam, Philadelphia: John Benjamins, 205–219. John Benjamins.

Ohala, J.
(1981) The listener as a source of sound change. In: Masek, C., S., Hendrick, R., A. and Miller, M., F. (eds.), *Papers from the Parasession on Language and Behavior.* Chicago: Chicago Ling. Soc, 178–203.

Rosenfelder, I., Fruehwald, J., Evanini, K., Jiahong Y.
(2011) *FAVE* (*Forced Alignment and Vowel Extraction*) Program Suite. http://fave.ling.upenn.edu.

Diversification in the noun inflection of Old English

Petra Steiner

Keywords: DIVERSIFICATION, INTEGER PARTITIONS, OLD ENGLISH, INFLECTIONAL PARADIGMS, SYSTEM, LANGUE

The inventory of the means for inflection in synthetic languages is relatively small, compared to the means of derivation. For most Indo-European languages, the set of inflectional morphemes can be listed on one page, while derivational affixes usually take up more space. If the number of inflectional morphemes is restricted, the number of paradigms is so too. Additionally, the frequency distributions of inflectional morphemes within inflectional paradigms are even more limited due to the principles of diversification. Evidence for this was found for both German and Icelandic noun inflection in Steiner and Prün (2007) and Steiner (2009). Hence, investigations of inflectional morphology in Quantitative Linguistics can provide interesting insights into the language system. In addition, they are feasible, as the data can be easily obtained. The hypotheses and methods can be transferred from other usage-oriented investigations. This study is concerned with the inflectional paradigms of Old English in regard to the frequency distributions of their inflectional affixes. For Modern German and Icelandic, it could be shown in the above-mentioned investigations that typical distributions occur with a small number of forms which occurred often and many forms that occurred rarely, leading to typical distributions showing steep beginnings and long tails. The distribution of *case syncretism* is another interesting aspect. Counting and sorting the different forms of inflectional suffixes lead to so-called *unordered integer partitions* of the numbers of the elements of an inflectional paradigm. For example, the integer partition for the Old English noun cyning (see Table 1) comprising the suffixes *-, -as, -e, -um, -es* and *-a,* would be (*2 2 2 2 1 1*).

Table 1: Inflectional suffixes of the Old English noun cyning

	Singular	**Plural**
Nominative	-	-as
Accusative	-	-as
Genitive	-es	-a
Dative	-e	-um
Instrumental	-e	-um

These integer partitions can be described as urn models, leading to typical distributions.

References

Steiner, P.

(2009) Diversification in Icelandic Inflectional Paradigms. In: Köehler, R., (ed.) I*ssues in Quantitative Linguistics*. Lüdenscheid: RAM-Verlag, 126–154. (= Studies in Quantitative Linguistics, 5).

Steiner, P., Prün, C.

(2007) The effects of diversification and unification on the inflectional paradigms of German nouns. In: Grzybek, P., Köhler, R. (eds.), *Exact methods in the study of language and text: dedicated to Professor Gabriel Altmann on the occasion of his 75th birthday*. Berlin: de Gruyter, 623–631.

Quantitative verification of constancy measures of texts

Kumiko Tanaka-Ishii, Shunsuke Aihara

Keywords: Yule's K, entropy, Zipf's Law

In this talk, we will present mathematical and empirical verifications of computational constancy measures for natural language text. A constancy measure characterizes a given text by having an invariant value for any size larger than a certain amount. The study of such measures has a 70-year history dating back to Yule's *K*, with the original intended application author identification. We examine various measures proposed since Yule and reconsider reports made so far, thus overviewing the study of constancy measures. We then explain how *K* is essentially equivalent to both an approximation of the second-order Renyi entropy and the correlation integral, thus indicating its profound significance within language science. We then empirically examine the constancy candidates within this new, broader context. The approximated higher-order entropy exhibited stable convergence across different languages and kinds of text. We also show, however, that it could not identify authors, contrary to Yule's intention. Lastly, we apply *K* to two unknown scripts, of the Voynich manuscript and Rongorongo, and show how the results support previous hypotheses about these scripts.

Rényi, A.
(1961) On measures of entropy and information. *Proceedings of the Fourth Berkeley Symposium on Mathematics, Statistics and Probability*, 1, 547–561.

Shannon, C.
(1948) A mathematical theory of communication. *Bell System Technical Journal*, 27, 379–423, 623–656.

Tweedie, F. J., Baayen, R. H.
(1998) How variable may a constant be? measures of lexical richness in perspective. *Computers and the Humanities*, 32, 323–352.

Yule, G. U.
(1944) *The Statistical Study of Literary Vocabulary*. Cambridge University Press.

History of words II – types of historical developments

Arjuna Tuzzi, Reinhard Köhler

Keywords: CHRONOLOGICAL CORPORA, PIOTROWSKI-ALTMANN LAW, PRESIDENTIAL ADDRESSES, WORD FREQUENCIES

The development of the frequencies of Italian words was observed on data from a corpus including the end-of-year speeches of the 10 Presidents of the Italian Republic (1949-2012).

The data used for this study were organised in two ways: For each word (lemma-types), the frequencies were collected in (1) 10 subcorpora (one for each president) and in (2) 64 subcorpora (one for each year within the period, i.e. one for each presidential address).

The Piotrowski-Altmann Law, which was developed and used so far as a model also of the diffusion of new elements in a linguistic population, was considered as an appropriate model of the frequency dynamics over time. Fitting the corresponding function to the data sets yielded very good results in most cases after smoothing the data by calculating moving averages. The words could be ascribed to several categories of dynamics and the parameters of the Piotrowski-Altmann Law proved a good way to cluster words portraying a similar temporal evolution.

References

Altmann, G.

(1983) Das Piotrowski-Gesetz und seine Verallgemeinerungen. In: Best, K.-H., Kohlhase, J. (eds.): *Exakte Sprachwandelforschung*. Göttingen:

Edition Herodot, 59–90.

Trevisani, M., Tuzzi A.
(2013) Shaping the history of words. In: Obradović I., Kelih E., Köhler R. (eds), *Methods and Applications of Quantitative Linguistics: Selected papers of the VIIIth International Conference on Quantitative Linguistics (QUALICO)*, Belgrade, Serbia, April 16–19, 2012. Belgrade: Akademska Misao, 84–95

Tuzzi, A., Popescu, I.-I., Altmann G.
(2010) Quantitative analysis of Italian texts. *Studies in Quantitative Linguistics,* 6, Lüdenscheid, RAM-Verlag.

Defying Zipf's Law

Marjolein Van Egmond, Sergey Avrutin

Keywords: Zipf's Law, word frequency distributions, lexical words, text judgments

All natural language texts are claimed to conform to Zipf's Law. But would it be possible to create a text that does not conform to Zipf's Law? And, more importantly, how would people judge such a text? These are the questions that we addressed in the current study. We chose to adapt an existing text, namely the first 958 words of the Dutch novel Ik ben omringd door debielen en ik voel me goed by Stevan Nieuwenhuis (2005). The original text conformed to Zipf's Law both in all words and in lexical and grammatical words separately. An important criterion while adapting the text was to keep it grammatical. We therefore chose to change the word frequencies of the lexical words but not the grammatical words to prevent the creation of a text with a deviant syntax. The word frequency distribution of the adapted text had the same number of tokens as the original. The number of types per frequency class, however, did not follow a log-linear distribution, but a normal distribution, in which the average frequency of the original text had the highest probability. The adapted text did not differ from the original text in terms of sentence length, clause length, number of clauses or number of sentences.

Two groups of students were recruited. Each group judged one version of the text by filling out a questionnaire addressing a wide range of textual aspects. In general, the adapted text was judged lower on aspects such as how pleasant, natural and varied it was. Also, reading ease and structure were rated lower. Interestingly, on the question ‚this text contains aberrant choices of words' and ‚the author uses words that I normally wouldn't use in that way' no difference was found. In addition, readers judged the original text as having a better rhythmical

pattern. Finally, respondents were explicitly asked for judgments of the variation in verbs, nouns, and adverbs and adjectives. For all, the variation in choice of words in the adapted text was rated lower than in the original text, in line with the adaptations that were made. From this study, we can draw two conclusions. The first is that it is indeed possible to construct a readable and grammatically correct text that does not conform to Zipf's Law, at least for lexical words and to a lesser extent to all words. For a law that is claimed to apply to any natural language text of sufficient length this is no obvious outcome. The second conclusion is that readers are sensitive to this adaption, and prefer a text that does conform to Zipf's Law: they have an unconscious intuition about Zipf's Law in texts. We argue that this finding provides evidence that Zipf's Law in language is a result of the organization of the human language faculty, and not just a statistic idiosyncrasy.

References

Nieuwenhuis, S.

(2005) *Ik ben omringd door debielen en ik voel me goed*. Groningen, Uitgeverij Passage.

Opinion target identification using thematic concentration of the text

Kateřina Veselovská, Radek Čech

Keywords: THEMATIC CONCENTRATION, OPINION TARGET IDENTIFICATION, SENTIMENT ANALYSIS, WORD FREQUENCY

Opinion target identification, the task in which the evaluated entities need to be identified in natural language texts, is in the long term one of the crucial problems in the field of sentiment analysis, i.e. the task which aims to determine the attitude of a speaker or a writer with respect to some topic (cf. Pang and Lee 2008). In this paper we apply the analysis of thematic concentration of text (cf. Popescu et al. 2009, Popescu and Altmann 2011) to detect particular targets of evaluation within specific domains. The promising results show that our approach has a potential to improve the state-of-the art sentiment classifiers significantly. The paper is based on the assumption that the target of the evaluation is of a hierarchical nature, i.e. that every target is composed of many different aspects (notebook: display, battery etc.), see Liu (2010). The purpose of this contribution is to construct a short list of such aspects for given evaluative text. Since the current methods fail to find these aspects within a random text span (see e.g. Ruppenhofer et al. 2008), it might be beneficial to consider target identification within narrowly focused domains, namely various product reviews. The goal of the paper is to implement and evaluate the quantitative method for automatic identification of the targets of evaluative texts within such domains. For this purpose, we decided to use thematic concentration of the text. This method enables us, besides other things, to detect words representing the main topic (or topics) of the text and, moreover, to quantify its (or their) thematic weight. The method is based on a frequency structure of text and h-point (cf., Popescu et al. 2009, chapter 6).

The actual valuation of the target is extracted using available resources, namely various retail servers. Since reviews of consumer products usually fall into two classes, full-text reviews and brief summaries of pros and cons, we can evaluate precision of our target detection method using simple comparison of the former and the latter (e.g. by intersection of sets). In other words, we confront the thematic words (determined by the thematic concentration analysis) obtained from the long reviews with the targets assigned by the customers in the short (usually one--word) reviews. The preliminary results concerning different reviews of cameras, notebooks and PC games indicate that, using thematic concentration of the text, the targets we obtain from the longer descriptions are the same as the ones we get from the users' reviews. Having the opinion target identification task solved, we can easily assign corresponding polarity to the particular targets using existing methods and thus improve the fine-grained sentiment classification. Moreover, the outputs of the presented research can be applied in many other natural language processing tasks, such as e.g. summarization or question answering.

References

Liu, B.
(2010) Sentiment analysis and subjectivity. *Handbook of natural language processing,* 2, 568.

Pang, B., Lee, L.
(2008) Opinion mining and sentiment analysis. *Foundations and trends in information retrieval*, 2, 1–2, 1–135.

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: Mouton de Gruyter.

Popescu, I.-I., Altmann, G.
(2011) Thematic concentration in texts. In: Kelih, E., Levickij, V., Matskulyak, Y. (eds.), *Issues in Quantitative Linguistics*, 2. Lüdenscheid: RAM-Verlag, 110–116.

Ruppenhofer, J., Somasundaran, S., Wiebe, J.
(2008) *Finding the Sources and Targets of Subjective Expressions*. In: Proceedings of LREC.

Grammar efficiency and the idealization of parts-of-speech systems

Relja Vulanovič, Tatjana Hrubik-Vulanovič

Keywords: PARTS-OF-SPEECH SYSTEMS, GRAMMAR EFFICIENCY, SIGN TEST

Any attempt to establish a classification of languages based on some criteria has to involve a certain level of idealization. Hengeveld's (1992) classification of parts-of-speech (PoS) systems is no exception. Regarding the seven PoS system types he proposes, Hengeveld states that "languages at best show a strong tendency towards one of the types." Hengeveld's original classification has been extended to 13 types in Hengeveld, Rijkhoff, and Siewierska 2004 by the addition of six intermediate types. Fifty languages, representing a genetically, geographically, and typologically diverse sample, are classified in that paper. Two 2010 papers by Hengeveld and van Lier consider 20 of those languages and describe their PoS systems in more detail. The description is in some cases the same as in (Hengeveld et al. 2004), but in some other cases there are differences which do not always seem to be negligible. Most of the differences are such that some parts of speech are omitted in the Hengeveld et al. 2004 description. A natural question is then whether the idealization accompanying the classification in Hengeveld et al. 2004 has significantly changed the overall features of the sample. In our paper, we answer this question by calculating grammar efficiency for each PoS system of interest. We choose grammar efficiency to quantitatively represent the structure of PoS systems because it is known that grammar efficiency correlates with PoS system types and, on the other hand, PoS system types correlate with other linguistic properties. Therefore, we discuss 20 languages such that the PoS system of each is described in two, not necessarily different, ways. We calculate the grammar

efficiency of each PoS system described and get two sets of data. Then we analyze statistically whether the two sets of data are significantly different or not. Since the data are neither normally nor symmetrically distributed, we use the non-parametric sign test.

References

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., and Siewierska, A.
(2004) Parts-of-speech systems and word order. *Journal of Linguistics,* 40, 527–570.

Hengeveld, K., van Lier, E.
(2010a) An implicational map of parts of speech. *Linguistic Discovery,* 8, 129–156.

Hengeveld, K., van Lier, E.
(2010b) Parts of speech and dependent clauses in functional discourse grammar. In: Ansaldo, U., Don, J., and Pfau, R. (eds.), *Parts of Speech: Empirical and Theoretical Advances.* Amsterdam, Philadelphia. John Benjamins, 253–285.

Polyfunctionality and polysemy in Chinese

Lu Wang

Keywords: POLYFUNCTIONALITY, POLYSEMY, CHINESE, PARTS OF SPEECH

This paper makes the first attempt to study the relationship between polyfunctionality (the number of parts of speech of a word) and polysemy. When a word acquires a new part of speech, it will have a new meaning, and maybe new synonyms, antonyms and associated words. Then, it has more connections with other words in the concept network. Such words are more activated in developing new meanings than monofunctional words. We hypothesize that: the more polyfunctionality, the more polysemy. This hypothesis is tested on data from Modern Chinese Dictionary (5th Edition), which includes 51,156 words (47,414 monofunctional words, 3,742 polyfunctional words). We obtain polyfunctionality and the corresponding polysemy as shown in the following table. Considering that, if a new part of speech adds only one new meaning to the word, the relationship should be linear. But as seen from the table it is obviously not. Then, we remove the linear-like relation by dividing the polysemy with its polyfunctionality (i.e. polysemy per part of speech) as shown by the 3rd column of the table. The results show that both relationships are non-linear, and abide by power Law: $a = 0.636$, $b = 1.6479$, $R^2 = 0.987$ and $a = 1.1391$, $b = 0.3472$, $R^2 = 0.9272$ respectively. Therefore, we can conclude: the more polyfunctionality, the more polysemy of each part of speech.

Polyfunctionality	Polysemy	Polysemy / polyfunctionality
1	1.1965	1.1965
2	2.6059	1.3030
3	5.2473	1.7491
4	7.5667	1.8917
5	9.7647	1.9529

Further, we investigate the relationship between polyfunctionality and polysemy of each part of speech. Words are classified into 12 parts of speech in Chinese: adjective, adverb, auxiliary, conjunction, interjection, noun, number, onomatopoeia, preposition pronoun, quantifier and verb. The data of adjective, adverb, auxiliary, noun, number and verb show monotonous increasing shape similar with the above result. However, the data of conjunction, pronoun and quantifier demonstrate a convex shape, and that of preposition is concave. Such difference indicates the meaning expansion behaviors of different parts of speech are not uniform. For example, verb polysemy is always dominant in polyfunctional words, while conjunction contributes the least number of meanings. The former 6 parts of speech abide by power law $y = ax^b$. The latter four abide by Zipf-Alekseev function $y = ax^{(-b-c\ln(x))}$, which mathematically can transform to power function when parameter $c = 0$. Therefore, it is reasonable to adopt Zipf-Alekseev function as a general function. Onomatopoeia and interjection gain only 2 data points respectively, which are impossible to fit with.

References

Fan, F., Altmann, G.
(2008) On meaning diversification in English. *Glottometrics,* 17, 69–81.

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

Tuldava, J.
(1998) *Probleme und Methoden der quantitativ-systemischen Lexikologie*. Trier: WVT.

Hřebíček, L.

(2005) Text Laws. In: Köhler, R., Altmann, G., Piotrowski, R.G. (eds), *Quantitative Linguistics. An International Handbook*. Berlin, New York: de Gruyter, 348–361.

Structural complexity of Chinese characters and Zipf's Law

Yanru Wang, Xinying Chen

Keywords: structural complexity of Chinese characters, Zipf's Law, stroke, component, Chinese

The closest Chinese, the default example of an isolating language, gets to morphology is the morphemic combination of characters into words and the placement of aspectual markers. There is, however, another rarely explored aspect of the Chinese writing system: the inner structural complexity of the Chinese characters. The question addressed in this paper is how this inner-character complexity relates to measures that were taken by Köhler et al. (2005) and Wang (2011) the morphology of flectional languages. So far, the structural complexity of Chinese characters has been the subject of little theoretical research. The interest in character complexity is mainly due to the need of applications of computer processing of Chinese characters and Chinese teaching (Bunke and Wang 1997, McBride-Chang et al. 2003). Most of these studies are oriented towards non-linguistics goals and conclusions, such as improvement in algorithms and in teaching methods or the difficulty of characters recognition. The rare studies addressing character complexity from a theoretic point of view lack sufficient data and a sound theoretical standpoint (Wang 2007). The synergetic model proposed by Köhler et al. (2005) described the relationship between different language features, such as language units' structure complexity, number of meanings, and frequency, and it has been proved applicable to many languages. For Chinese, Wang (2011) proved that the relationship between Chinese words' polysemy and word length fits the model. If the same principles work for the characters, the frequency of Chinese characters should decline as the structural complexity of Chinese characters increases, and

the curve should fit Zipf's Law. Based on this hypothesis, we tested the relation bet -ween frequency of Chinese characters and two most common characters complexity measurements. We used the information of the number of strokes and the number of components of the most common 3,061 Chinese characters, which cover about 99.43% of the original corpus in Chinese Characters' Frequency Dictionary. From Dictionary of Chinese Characters' Information and The Table of Basic Components of Chinese Characters, we calculated the average frequency of Chinese characters that share the same number of strokes or components and tried to use Zipf's, Mandelbrot's or Joos' function to do linear fitting. We found that the frequency-strokes and frequency-components relation is fitted into Joos' function (respectively C=0.0096, b=1.8497, R^2=0.2012; C=0.0012, b=1.6784, R^2=0.9554). However, the former fitting is of poor quality with the small R^2, while the latter is of good quality with a significant R^2. The frequency-components curve is well fitted into three functions while the frequency-strokes curve is not. Since we believe that the frequency-structural complexity curve should follow Zipf's Law, we may conclude that the number of components, rather than strokes, is a better measurement for the structural complexity of Chinese characters. The results of this paper can be used or tested for finding the rules of simplification of Chinese characters. We also see that the often neglected inner structure of Chinese characters, even without any direct relation to the phonological form of the character, follows linguistic patterns. This shines new light also on the extension of linguistic approaches to other domains of human cognition.

References

Bunke, H., Wang, P. S.
(1997) *Handbook of character recognition and document image analysis.* World Scientific.

McBride-Chang, C., Shu, H., Zhou, A., Wat, C. P., Wagner, R. K.
(2003) Morphological awareness uniquely predicts young children's Chinese character recognition. *Journal of Educational Psychology,* 95, 4, 743–751.

Wang, G. A.
(2007) *A Handbook for 1,000 Basic Chinese Characters.* Chinese University Press.

Köhler, R., Altmann, G., Piotrowski, R.G. (eds.)
(2005) *Quantitative Linguistics*. Berlin, New York: de Gruyter.

Wang, L.
(2011) Polysemy and word length in Chinese. *Glottometrics*, 22, 73–84.

The influences of the word unit and the sentence length on the ratio of the parts of speech in Japanese

Makoto Yamazaki

Keywords: RATIO OF PARTS OF SPEECH, SENTENCE LENGTH, WORD UNIT, REGISTER, JAPANESE

Kabashima (1954) pointed out that the ratio of parts of speech is determined by the register of the text. The details of the measurement were not given in the paper, but we assure that he used only one word unit and he did not include functional words. In this study, we tested his assertion again by a large, multi genre corpus – the balanced corpus of contemporary written Japanese – and following results were obtained.

1) Proportions of parts of speech vary depending on the kinds of a word unit. We measured the ratio of parts of speech both by a short word unit and a long word unit. From the observation by these two types of a word unit, the ratio of noun is lower when measured by a long word unit than by a short word unit.
2) In many registers, the ratio of parts of speech showed a similar curve when sentence length varies between 1 and around 10, after that area the curve becomes stable to a fixed ratio of each register.
3) In the register of best-selling books and minutes of the Diet, the ratio of particle (functional word) is higher than the one of noun.
4) In the correlation of the tokens of parts of speech and the sentence length, particle, noun, verb and auxiliary showed a high correlation coefficient values whereas conjunction showed a low correlation coefficient.

5) The correlation of the tokens of parts of speech and the sentence varies depending on the register. In comparison with newspaper, the best-selling books, most of which are novels, pronoun, adverb, adnominal and adjective have high correlation coefficient values.

References

Kabashima, T.

(1954) On the Ratio of Parts of Speech in Present-day Japanese and the Cause of its Fluctuation. *Studies in the Japanese language,* 18, 15–20.

Adresses Qualico 2014

Aihara, Shunsuke
Kyushu University
Graduate School of Information Science and Electrical Engineering
744 Motooka, Nishi-ku, Fukuoka city
819-0395, Fukuoka, Japan
e-mail: aihara@cl.ait.kyushu-u.ac.jp

Andreev, Sergey
Smolensk State University
Bakuin str.13, apt.3
214000 Smolensk, Russia
e-mail: smol.an@mail.ru

Andres, Jan
Palacký University
Faculty of Science
Dept. of Mathematical Analysis
17. listopadu 12
771 46 Olomouc, Czech Republic
e-mail: jan.andres@upol.cz

Avrutin, Sergey
Utrecht University
UiL OTS
Trans 10
3512 JK, Utrecht, The Netherlands
e-mail: S.Avrutin@uu.nl

Bel-Enguix, Gemma
Aix Marseille Université
CNRS-LIF
UMR 7279
Marseille, France
e-mail: gemma.belenguix@gmail.com

Beliankou, Andrei
Universität Trier
Computational Linguistics and Digital Humanities
Universitätsring 6
54294, Trier, Germany
e-mail: a.beliankou@uni-trier.de

Benešová, Martina
Palacký University
Philosophical Faculty
Dept. of General Linguistics
Křížkovského 10
771 80 Olomouc, Czech Republic
e-mail: martina.benesova@upol.cz

Biryukov, Denis
Palacký University
Philosophical Faculty
Dept. of General Linguistics
Křížkovského 10
771 80 Olomouc, Czech Republic
e-mail: denis.biryu@gmail.com

Bohn, Kirsten
Department of Biology
Texas A&M University
College Station, TX 77843-3258
USA
e-mail: kbohn@bio.tamu.edu

Čech, Radek
University of Ostrava
Faculty of Arts
Department of Czech language
Reální 5
Ostrava 701 03, Czech Republic
e-mail: cechradek@gmail.com

Charzyńska, Edyta
Uniwersytet Śląski w Katowicach
Wydział Pedagogiki i Psychologii
ul. Grażyńskiego 53/420
40-126 Katowice, Poland
e-mail: edyta.charzynska@us.edu.pl

Chen, Heng
No.148 Tianmushan Rd.,
Xihu District, Hangzhou 310028,
Zhejiang Province,
P.R.China.
e-mail: chenheng1003@163.com

Chen, Xinying
Xi'an Jiaotong Universiy
School of International Study
No 28 Xianning West Road
Xi'an 710049, Shaanxi, China
e-mail: chenxinying@mail.xjtu.edu.cn

Dębowski, Łukasz
Instytut Podstaw Informatyki
Polskiej Akademii Nauk
ul. Jana Kazimierza 5
01-248 Warszawa, Poland
e-mail: ldebowsk@ipipan.waw.pl

Eder, Maciej
Pedagogical University in Kraków
Institute of Polish Studies
ul. Podchorążych 2
30-084 Kraków, Poland
e-mail: maciejeder@gmail.com

Embleton, Sheila
York University
Languages, Literatures and Linguistics
S561 Ross Building
4700 Keele Street
Toronto, Ontario, Canada M3J 1P3
e-mail: embleton@yorku.ca

Fenk-Oczlon, Gertraud
Alpen-Adria-Universität Klagenfurt
Universitätsstrasse 65-67
9020 Klagenfurt, Austria
e-mail: gertraud.fenk@uni-klu.ac.at

Fenk, August
Alpen-Adria-Universität Klagenfurt
Universitätsstrasse 65-67
9020 Klagenfurt, Austria
e-mail: august.fenk@uni-klu.ac.at

Ferrer-i-Cancho, Ramon
Complexity and Quantitative Linguistic Lab,
LARCA Research Group,
Departament de Llenguatges i Sistemes Informatics
Universitat Politecnica de Catalunya (UPC)
Barcelona, Catalonia, Spain
e-mail: rferrericancho@lsi.upc.edu

Figueredo, Giacomo P.
Departamento de Letras
Federal University of Ouro Preto, Brazil
Rua do Seminário, s/n
35420-000 Mariana, Minas Gerais, MG, Brazil
e-mail: giacomojakob@yahoo.ca

Gao, Song
School of Chinese Language and Literature
HeiLongjiang University,
China Faculty of Humanities
No.74, XueFu Road, NanGang District
Harbin City, HeiLongjiang Province
China IJKL Road 10
Harbin 150080, China
e-mail: gaos_0808@163.com

Glynn, Dylan
University of Paris VIII
Linguistique Anglaise, Psycholinguistique (LAPS)
2 Rue de la Liberté
93200 Saint-Denis cedex
France
e-mail: dglynn@univ-paris8.fr

Górski, Rafał L.
Institute of Polish Language
Polish Academy of Sciences
Al. Mickiewicza 31, Kraków
Poland
e-mail: rafalg@ijppan.krakow.pl

Grzybek, Peter
Karl-Franzens-Universität Graz
Institut für Slawistik
Merangasse 70/I
8010 Graz, Austria
e-mail: peter.grzybek@uni-graz.at

Guex, Guillaume
University of Lausanne
Faculty of Arts
Department of language and information sciences
CH-1015 Lausanne
e-mail: guillaume.guex@unil.ch

Hasanaj, Belinda
Duquesne University
EVL Lab
600 Forbes Avenue
Pittsburgh, PA 15219, USA
e-mail: hasanajb@duq.edu

Haviger, Jiří
Department of informatics and quant. methods
Faculty of informatics and management
University Hradec Králové
Hradecká 1249/6, Hradec Králové
e-mail: jiri.haviger@uhk.cz

Hrubik-Vulanović, Tatjana
Kent State University at Stark
Department of Mathematical Sciences
6000 Frank Ave NW
North Canton, Ohio 44720, USA
e-mail: thrubik@kent.edu

Hua, Wang
Zhejiang University
School of International Studies
No.866 Yuhangtang Road
Hangzhou, CN-310058, China
e-mail: wanghuazju@163.com

Huang, Wei
Institute of Chinese Proficiency Test and Educational Measurement
Beijing Language and Culture University
No. 15 Xueyuan Rd.
Haidian District
Beijing, 100083, China
e-mail: huangwei@blcu.edu.cn

Juola, Patrick
Duquesne University
EVL Lab
600 Forbes Avenue
Pittsburgh, PA 15219, USA
e-mail: juola@mathcs.duq.edu

Kelih, Emmerich
University of Vienna
Institute for Slavic Studies
Spitalgasse 2, Hof 3
1090 Wien, Austria
e-mail: emmerich.kelih@univie.ac.at

Koch, Veronika
Karl-Franzens-Universität Graz
Institut für Slawistik
Merangasse 70/I
8010 Graz, Austria
e-mail: veronik.koch@edu.uni-graz.at

Köhler, Reinhard
University of Trier
Computerlinguistik und Digital Humanities
Universitätsring 15
Postfach 3825
54286 Trier, Germany
e-mail: koehler@uni-trier.de

Kubát, Miroslav
Palacký University
Department of General Linguistics
Křížkovského 10
Olomouc 771 80, Czech Republic
e-mail: miroslav.kubat@gmail.com

Kučera, Dalibor
Department of Pedagogy and Psychology
Faculty of Pedagogy
Jihočeská University in České Budějovice
Dukelská 9 (U Tří lvů)
e-mail: dkucera@pf.jcu.cz

Kuya, Aimi
Faculty of Linguistics, Philology & Phonetics
University of Oxford
Clarendon Press Centre
Walton Street
Oxford
HG, UK
e-mail: aimi.kuya@new.ox.ac.uk

Lukin, Annabelle
Department of Linguistics
Faculty of Human Sciences
Macquarie University NSW 2109
Australia
e-mail: annabelle.lukin@mq.edu.au

Mačutek, Ján
Comenius University
Department of Applied Mathematics and Statistics
Mlynská dolina
84248, Bratislava, Slovakia
e-mail: jmacutek@yahoo.com

Matlach, Vladimír
Palacký University
Department of General Linguistics
Křížkovského 10
Olomouc 771 80, Czech Republic
e-mail: v.matlach@seznam.cz

Mikros, George K.
National and Kapodistrian University of Athens
Department of Italian
Language and Literature
School of Philosophy
Lemesou 45
GR-15669, Athens, Greece
e-mail: gmikros@isll.uoa.gr

Milička, Jiří
Charles University
Institute of Comparative Linguistics
Faculty of Arts
Velenovského 2
Praha 10 Záběhlice 10600, Czech Republic
e-mail: milicka@centrum.cz

Moisl, Hermann
Newcastle University
School of English
Percy Building
Newcastle upon Tyne NE1 7RU
RU, UK
e-mail: hermann.moisl@ncl.ac.uk

Motalová, Tereza
Palacký University
Philosophical Faculty
Department of General Linguistics
Křížkovského 10
Olomouc 771 80, Czech Republic
e-mail: tereza.motalova@gmail.com

Naumann, Sven
Linguistische Datenverarbeitung
FB II
Universität Trier
D-54286 Trier
e-mail: sven.naumann@me.com

Pagano, Adriana S.
Federal University of Minas Gerais
Faculdade de Letras da UFMG
Av. Antonio Carlos
6627, Pampulha,
31270901, Belo Horizonte, MG, Brazil
e-mail: apagano@ufmg.br

Pan, Xiaxing
Zhejiang University
School of Humanities
Center for the Study of Language and Cognition
Hangzhou, China
e-mail: sumloong@hotmail.com

Pawłowski, Adam
University of Wrocław
Instytut Informacji Naukowej i Bibliotekoznawstwa
pl. Uniwersytecki 9/13
50-137 Wroclaw, Poland
e-mail: apawlow@uni.wroc.pl

Perifanos, Kostas
National and Kapodistrian University of Athens
Department of Linguistics
School of Philosophy
Panepistimiopoli Zografou
GR-15784 , Athens, Greece
e-mail: kostas.perifanos@gmail.com

Poddubny, Vasiliy
Tomsk State University
1 Computer Science Faculty
Tomsk, 634050, Russian Federation
e-mail: vvpoddubny@gmail.com

Polikarpov, Anatoly
Lomonosov Moscow State University
Faculty of Philology
Laboratory for General and Computational Lexicology and Lexicography
Moscow, 119991, Russian Federation
e-mail: anatpoli@mail.ru

Purnell, Erin
Duquesne University
EVL Lab
600 Forbes Avenue
Pittsburgh, PA 15219, USA
e-mail: purnelle@duq.edu

Qiu, Hui
Zhejiang University
School of Humanities
Center for the Study of Language and Cognition
Hangzhou, China
e-mail: sebastianqiu@hotmail.com

Rovenchak, Andrij
Ivan Franko National University of Lviv
Department for Theoretical Physics
12 Drahomanov St.
79005 Lviv, Ukraine
e-mail: andrij.rovenchak@gmail.com

Rybicki, Jan
Jagiellonian University
Institute of English Studies
Ul. Mikołajska 6/12
31-027 Kraków, Poland
e-mail: jkrybicki@gmail.com

Sanada, Haruko
Rissho University
Faculty of Economics.
4-2-16, Osaki
Shinagawaku
Tokyo 141-8602, Japan
e-mail: hsanada@ris.ac.jp

Savoy, Jacques
University of Neuchatel
Computer Science Department
Rue Emile-Argand 11
CH-2000 Neuchâtel, Switzerland
e-mail: Jacques.Savoy@unine.ch

Schusterová , Denisa
Palacký University
Philosophical Faculty
Department of General Linguistics
Křížkovského 10
Olomouc 771 80, Czech Republic
e-mail: schusterovad@gmail.com

Ščigulinská, Jana
Palacký University
Philosophical Faculty
Department of General Linguistics
Křížkovského 10
Olomouc 771 80, Czech Republic
e-mail: jana.scigulinska@gmail.com

Sneller, Betsy
University of Pennsylvania
Linguistics Lab
3810 Walnut St
Philadlephia, PA 19104, United States
e-mail: esnell@sas.upenn.edu

Spáčilová, Lenka
Palacký University
Philosophical Faculty
Department of General Linguistics
Křížkovského 10
Olomouc 771 80, Czech Republic
e-mail: lenka.spacilova@gmail.com

Steiner, Petra
Technische Universität Chemnitz
Philosophische Fakultät
Angewandte Sprachwissenschaft/Technikkommunikation
09107 Chemnitz, Germany
e-mail: petra.steiner@phil.tu-chemnitz.de

Tanaka-Ishii, Kumiko
Kyushu University
Graduate School of Information Science and Electrical Engineering
744 Motooka, Nishi-ku, Fukuoka city
819-0395, Fukuoka, Japan
e-mail: kumiko@cl.ait.kyushu-u.ac.jp

Tuzzi, Arjuna
Università di Padova
Dipartimento di Filosofia, Sociologia, Pedagogia e Psicologia Applicata (FISPPA)
via M. Cesarotti, 10/12
35123 Padova, Italy
e-mail: arjuna.tuzzi@unipd.it

Uritescu, Dorin
York University
Glendon College, Department of French Studies
Linguistics and Language Studies Programme
2275 Bayview Ave.
Toronto, Ontario, Canada M4N 3M6
e-mail: dorinu@yorku.ca

van Egmond, Marjolein
Utrecht University
Utrecht Institute of Linguistics
Trans 10, 3512 JK Utrecht
The Netherlands
e-mail: M.vanEgmond1@uu.nl

Veselovská, Kateřina
Charles University in Prague
Faculty of Mathematics and Physics
Institute of Formal and Applied Linguistics
Malostranské náměstí 25
118 00 Prague 1, Czech Republic
e-mail: veselovska@ufal.mff.cuni.cz

Vulanović, Relja
Kent State University at Stark
Department of Mathematical Sciences
6000 Frank Ave NW
North Canton, Ohio 44720, USA
e-mail: rvulanov@kent.edu

Wang, Lu
University of Trier
Am Trimmelterhof 95
Trier 54296, Germany
e-mail: wanglu-chn@hotmail.com

Wang, Yanru
East China Normal University
School of Psychology and Cognitive Science
No 3663 Zhongshan North Road
Putuo District
Shanghai 200062, China
e-mail: wangyanrupsyer@gmail.com

Wheeler, Eric S.
York University
Faculty of Arts
Department of Languages, Literatures and Linguistics
33 Peter Street
Markham, Ontario, Canada L3P 2A5
e-mail: wheeler@ericwheeler.ca

Xanthos, Aris
University of Lausanne
Anthropole
CH–1015 Lausanne
e-mail: aris.xanthos@unil.ch

Yamazaki, Makoto
National Institute for Japanese Language and Linguistics
Department of Corpus Studies
10-2, Midori-cho
Tachikawa City, 190-8561, Japan
e-mail: yamazaki@ninjal.ac.jp

Zámečník, Lukáš H.
Palacký University
Philosophical Faculty
Dept. of General Linguistics
Křížkovského 10
771 80 Olomouc, Czech Republic
e-mail: lukas.zamecnik@seznam.cz

Register

A

accuracy of calculations 17
Arens-Altmann Law 54, 55
authorship attribution 38, 51, 52, 53, 59, 78, 99, 100, 101, 104, 105

B

borrowing 40, 65, 74
Burrows's delta 101

C

case distribution 101
Catalan 19, 42
Chinese 25, 32, 33, 34, 35, 47, 63, 64, 65, 83, 84, 86, 87, 106, 123, 124, 126, 127, 128, 135, 136
classification 15, 34, 35, 42, 51, 64, 76, 120, 121
clustering 34, 35, 38, 39, 47, 52, 89
coding requirements 85
component 25, 32, 33, 126
compression 19, 36, 37
computational linguistics 71
computational stylistics 38
constituent 23, 62, 81, 85
construct 23, 61, 118, 119
co-occurrence 35, 102
corpus 27, 39, 48, 50, 52, 57, 59, 61, 73, 89, 91, 97, 98, 99, 104, 108, 109, 115, 127, 129
cross-linguistic correlations 42
cross-linguistic transference 59
Czech 1, 2, 13, 24, 27, 71, 83, 131, 132, 137, 138, 139, 142, 143, 144, 145

D

dental 40
dependency relation 47
dialect 41
dialectology 40
discourse analysis 104
discriminant analysis 15, 16
dissipative stochastic model 94
diversification 27, 111, 112, 124

E

English 19, 21, 22, 29, 35, 36, 42, 43, 51, 52, 59, 60, 61, 62, 65, 73, 85, 91, 95, 100, 105, 108, 109, 111, 112, 124, 139, 141
entropy 29, 30, 37, 69, 113
Eugene Onegin 15

F

film analysis 67
French 19, 54, 91, 100, 143
frequency structure 69, 119

G

gender identification 78, 79
genre classification 34, 35
German 19, 21, 42, 43, 50, 76, 85, 111, 112
golden section 86
grammar efficiency 121

H

Hilberg's conjecture 36, 37
homonymy 92, 93
h-point 69, 86, 87, 119
Hungarian 85

J

Japanese 25, 73, 74, 75, 85, 102, 103, 129, 130, 145

K

Krylov Law 92, 93

L

language evolution 19
language change 74, 108
language principles 45
Latin 19, 21, 40
laws for polysemic and age-polysemic distributions 94
length motif 76
lexical diversity 57, 58
lexical variation 73
lexicogrammatical features 73, 89
lexicology 92, 93. 96, 141
literary translation 89, 100
loan words 65
locus 15
logistic regression 48

M

machine learning 78, 101
Manding languages (West Africa) 97
Maninka language 97, 98
manual annotation 48
Menzerath-Altmann Law 17, 18, 23, 24, 25, 54, 67, 68, 80, 83, 97, 106, 107
monosyllabism 42, 43
multivariate statistics 48

N

network analysis 34, 38
network visualization 99, 100
Nko script 97
nominality 15, 16
nonlinear dynamics 81, 82
noun 27, 73, 111, 112, 124, 129

O

Old English 111, 112
opinion target identification 119, 120
optimization of parameters 17

P

palatalization 40
part of speech 47, 123, 124
permutation ring 45
Piotrowski-Altmann Law 115
Polish 29, 51, 52, 92, 93, 100, 101, 133, 135
political speech 104, 105
polyfunctionality 123, 124
polysemy 48, 49, 50, 92, 93, 94, 95, 123, 124, 126
postposition 102, 103
predicate-argument structure 21
presidential addresses 115
psycholinguistics 29, 71

Q

quantified gradations 40, 41

R

random sampling 57
random text 36, 69, 119
readability 29, 30, 31
register 129, 130
retranslation 89, 90
robustness 43, 57
Romanian 40, 41, 42

S

segmentation 23, 25, 83, 106
semantics 21, 27, 28, 48, 65, 81, 82
sentence length 29, 30, 54, 61, 79, 117, 129
semantic potential of linguistic signs 94

sentence semantics 81, 82
sentiment analysis 48, 119, 120
shot length/duration 67
sign test 121, 122
sociolinguistics 73, 108
statistical testing 69, 70
stroke 32, 126
structural complexity of Chinese characters 126, 127
structure information 80
style variation 32
stylometry 38, 51, 69, 99, 101
syllable 32, 33, 42, 43, 44, 86, 87, 98, 109
synergetic linguistics 23, 27, 32, 61
syntactic complexity 34, 61, 85
systemic functional grammar 89

T

text analysis 38, 69, 71, 72, 77, 88, 89
text judgments 117
thematic concentration 69, 119, 120
treebank 47
Twitter 78, 79
type-token relation 76
typology 42

U

Ukrainian 76
universal coding 36

V

valency 21, 102, 103
verse text 15
Vulgar Latin 40

W

word frequency 19, 20, 35, 63, 64, 65, 70, 99, 117, 119
word length 19, 29, 30, 32, 33, 54, 67, 126, 128
word meaning 81, 82
word order 43, 45, 46, 50, 52, 102, 122

Y

Yule's K 113

Z

Zip-Alekseev model 86
Zipf's Law 19, 126