NINJAL
National Institute for Japanese Language and Linguistics

IQLA

QUALICO 2021 Tokyo

BOOK OF ABSTRACTS

QUALICO 2021

Quantitative Approaches to Universality and Individuality in Language

Book of Abstracts

Tokyo 2021

National Institute for Japanese Language and Linguistics

Ver. 1.1 (7 September 2021)

QUALICO 2021 Book of Abstracts

QUALICO 2021 is supported by The International Quantitative Linguistics Association (IQLA), National Institute for Japanese Language and Linguistics (NINJAL) and Center for Corpus Development, NINJAL.

Long Memory in Natural Language

Kumiko Tanaka-Ishii
The University of Tokyo, Japan

In this talk, I present a frontier of quantification of the long memory underlying natural language. Long memory is a common quality of complex systems that also exists in natural language. In language, long memory appears as clustering of the word occurrences in a word sequence and is caused mainly by context shifts. Thus far, there have been two different quantification techniques for long memory: long-range correlation and fluctuation analyses. Both produce power laws that suggest the self-similar nature of the clustering phenomena to underlie natural language sequences. After explaining the state-of-the-art quantification methods, their outputs, and the understandings gained, I argue the further signification of long memory with respect to machine learning, as studied in the field of computational linguistics.

Automated Essay/Speech Scoring: A Stylometric Approach to Language Assessment

Yuichiro Kobayashi
Nihon University, Japan

Automated scoring, in which computer technology evaluates and scores written or spoken content (Shermis and Burstein, 2003), aims to sort a large body of data, which it assigns to a small number of discrete proficiency levels. Objectively measurable features are used as *exploratory variables* to predict scores defined as *criterion variables*. It is a form of language assessment and an application of stylometric methods, such as authorship attribution or chronological stylistic analysis. This talk will outline the basic concepts and technologies for automated essay and speech scoring. In particular, some useful feature sets for automated grading will be discussed in the context of statistical prediction of

learners' proficiency levels. Furthermore, the results of automated English speech grading will be reported in order to show the possibilities and limits of statistical language evaluation in educational settings. For the speech grading, I used two corpora of Japanese English language learners' spoken utterances, the NICT JLE Corpus (Izumi, Uchimoto, and Isahara, 2004) and the Longitudinal Corpus of Spoken English (Abe and Kondo, 2019), which are coded into nine oral proficiency levels. The nine levels, which were manually assessed by professional raters and pertained to such aspects of examinees' speech as vocabulary, grammar, pronunciation, and fluency, were used as criterion variables, and 67 linguistic features analyzed in Biber (1988) were considered as explanatory variables. The random forest algorithm (Breiman, 2001) was employed to predict oral proficiency.

Abe, M., & Kondo, Y. (2019). Constructing a longitudinal learner corpus to track L2 spoken English. *Journal of Modern Languages, 29,* 23-44.

Biber, D. (1988). *Variation across Speech and Writing.* Cambridge University Press.

Breiman, L. (2001). Random forests. *Machine Learning, 45*(1), 5-32.

Izumi, E., Uchimoto, K., & Isahara, H. (2004). *A Speaking Corpus of 1,200 Japanese Learners of English.* ALC Press.

Shermis, M. D., & Burstein, J. C. (Ed.) (2003). *Automated essay scoring: A cross-disciplinary perspective.* Routledge.

Dynamic Properties of Sign Language Structures

Jan Andres and Jiri Langer
Palacký University Olomouc, Czech Republic

Persistence, predictability and stability of sign language texts (speeches) will be considered in terms of the Hurt exponents, the Liapunov exponents and fractal dimensions. Theoretical results, applicable also to natural spoken languages, will be demonstrated by an illustrative example of a concrete example of a concrete sign language speech.

Keywords: persistence, predictability, fractality, stability, Hurst exponent, Liapunov exponent, fractal dimension, sign language

Dynamics of the Integration of Foreign Vocabulary into the Old Anatolian-Türkic Language System in the XIII - XV Centuries (on the Material of Phonological and Morphological Subsystems of Monuments in Old Anato-Türkic Language)

Apollinaria Avrutina
St. Petersburg State Univeristy, Russia

Counting in the study of the history of the linguistic and literary written tradition of the Turkic-speaking population of Asia Minor is taken from the first existing monument in Old Anatolian Türkic, mesnevî of the religious and philosophical content, "The Wheel of Fate", "Çarh-name", created by thinker and religious activist Ahmed Fakih (presumably, died before 1252).

Written tradition, founded in this monument, is continued in the "Tale of Malik Dānishmand" (Kıssa-i Melik Danışmend). The so-called "Leningrad" manuscript of this monument dates back to 1622. A detailed description of this manuscript was made by V.S. Garbuzova, I. Melikof and V.G. Guzev. However, the edition of the text of the legend of this list

dates back to the XVth century. I. Melikof refers the text of the monument to the XIV century.

Both texts have its own specificity: they demonstrate a phenomenon that was unusual for the ancient Turkic languages, i.e. monuments of the early eras, for example Turkic runic monuments, but which, due to the strengthening of the role of Islam and the cultural role of Iran, is manifested in all Turkic languages of the Near and Middle East, i.e. Central Asia and Asia Minor. Foreign languages (mostly Arabic & Persian) are beginning to predominate in the vocabulary of the Turkic languages of this period. Lexical borrowings from the Arabic and Persian languages gave some of theirs distinctive features, for example, phonemes, in the process of assimilation by the Turkic language system (and phonological subsystem as well). Here it is impossible not to recall the famous theses formulated by E.Sepir that there are no purely inflectional languages, purely agglutinative languages or purely isolating languages - in every language one can find elements of inflexibility, agglutination, isolation; even within a particular paradigm, typologically different word forms can co-exist.

If we are talking about some lexical/ morphological /morphophonological / phonological units, then we can talk about quantitative characteristics of these units, about the degree of predominance. This factor allows us to consider the mathematical method of investigation as the most effective among all known methods of studying the language, since it allows obtaining much more accurate and objective results than other methods. In addition, this method is practically not much used on the Turkic phonological material.

Only in diachrony it is possible to trace the process of assimilation of foreign language units by the Turkic language system. In this study, we will use quantitative methods to show the place and amount of borrowings at a particular moment in time.

Keywords: Turkic languages, Turkic phonology, Old-Anatolian-Turkic, quantitative linguistics, Kıssa-i Melik Danışmend, Çarh-name

Analyzing the Effectiveness of Using Character n-grams to Perform Authorship Attribution on Informal Documents in the English Language

David Berdik
Duquesne University, United States of America

Authorship attribution is a subfield of natural language processing which can be applied to practical issues such as copyright disputes. While there are many different methods that can be used to perform such an analysis, the effectiveness of these methods varies depending on the material that is being analyzed as well as the parameters chosen for the selected methods. One of these methods involves using groups of n consecutive characters, called character n-grams, where n refers to the number of characters in the gram. It is expected that n-grams of different lengths will vary in their accuracy of attributing the correct author to a questioned document. Specifically, it is expected that as n-grams become larger, performance will improve, reach a peak, and then begin to degrade.

Using Juola's Java Graphical Authorship Attribution Program, we performed seven different types of analyses on Schler's blog corpus by taking all corpus entries with at least 300 sentences, separating their first 100 sentences and last 100 sentences into separate entries, and running n-gram tests from 1 to 50 to determine what an ideal size would be for performing authorship attribution using character n-grams. Based on the results of the seven types of tests, we showed that contrary to the bell curve-like performance that was anticipated, n-gram accuracy peaks much earlier before beginning its decline on the first type of test and is closer to what was expected but is still not bell curve-like for the second and third types of tests. Additionally, we showed that the fourth type of experiment performs very similarly to the first type and the fifth and sixth types of experiments perform with extremely high accuracy.
Furthermore, we demonstrate that the seventh type of test performs very similarly to the first type of test. Future work will involve performing character n-gram analyses on different types of documents as well as different languages to determine how much variance, if any, is present between languages.

Keywords: authorship attribution, natural language processing, machine learning, character n-grams

In Search for Russian Low-Frequency Words

Olga V. Blinova and Valeriya V. Modina
St. Petersburg State University, Russia

One of the parameters for text complexity assessing is words frequency, see, e. g. [1], [2]. The presumption that the reader is having difficulty meeting low-frequency or unfamiliar words, is used in assessing the readability of texts. For example, the Dale-Chale readability formula takes into account the number of unfamiliar words [3]. There are relatively simple solutions to the problem of determining which words are to be considered unfamiliar to the reader. For example, if we are talking about an educational text for second-language learners, words that are not included in the lexical minimum, may be considered unfamiliar.

Our current study is related to complexity assessment for the texts of Russian official documents [4]. We conduct a survey to determine the perceptual («subjective») complexity. In addition, we created the Corpus of Russian Internal Documents and Acts (CorRIDA, 1.5 mln words) and use quantitative corpus techniques to describe the «objective» complexity. We face the task of determining the share of low-frequency words in the texts. It is unclear what units we can consider as words with a low general-language frequency.

There are various approaches to the empirical solution to the problem of forming a list of low- frequency words. For example, one can use threshold frequency values (for both relative and even raw frequency; in particular, such threshold is usually taken to be 5 ipm). Sometimes, researchers use the ranks or dispersion values, etc. For example, in [5] the words chosen from the bottom quartile of the frequency range are considered as low-frequency ones.

This paper is aimed at forming the list of low-frequency words by comparing frequency lists, obtained on the material of three Russian corpora. We use the RuTenTen11 web corpus (14,553 mln. words), Araneum Russicum Russicum Maius (859 mln. words) and Russian National Corpus, whose frequency vocabulary was formed on the basis of

a subcorpus of 92 mln. words [6]. The data on general-language frequency will be obtained by comparing the indices of the frequency and the word ranks that are taken from the frequency lists of all the above-mentioned corpora. In particular, we will calculate the average frequency values, compare the corpora in pairs by computing Damerau's relative frequency ratio, and estimate the difference between the corpora using the Spearman's rank correlation coefficient.

Keywords: Russian, linguistic corpora, web corpus, frequency lists, general-language frequency, lexical complexity

References

[1] Chen X., Meurers D. (2016), Characterizing text difficulty with word frequencies // Proceedings of the 11th Workshop on Innovative Use of NLP for Building Educational Applications, pp. 84-94.

[2] Solovyev V., Ivanov V., Solnyshkina M. (2018), Assessment of reading difficulty levels in Russian academic texts: Approaches and metrics // Journal of Intelligent & Fuzzy Systems, 34(5), pp. 3049-3058.

[3] Chall J. S., Dale E. (1995), Readability Revisited: The New Dale-Chall Readability Formula, Brookline Book, Cambridge, MA.

[4] Belov S.A., Blinova O.V., Gulida V.B., Zubov V.I., Larionova E.Ju., Tolstikova P.S. (2018), Korpus russkih lokal'nyh dokumentov i aktov CorRIDA: celi formirovanija, sostav, struktura [Corpus of Russian Internal Documents and Acts CorRIDA: Goals, Composition and Structure] // Komp'juternaja lingvistika i vychislitel'nye ontologii (Trudy XXI Mezhdunarodnoj objedinennoj konferencii «Internet i sovremennoe obshhestvo, IMS-2018) [Computational linguistics and computational ontologies (Proceedings of the XXI International Joint Conference «Internet and modern society», IMS-2018)]. St. Petersburg, University ITMO, Issue 2, pp. 131-147.

[5] Zhao Y., Jurafsky D. (2009), The effect of lexical frequency and Lombard reflex on tone hyperarticulation // Journal of Phonetics 37, pp. 231-247.

[6] Lyashevskaya O., Sharov S. (2009), Chastotnyj slovar′ sovremennogo russkogo jazyka na materialach Nacional′nogo korpusa russkogo jazyka [The frequency dictionary of modern Russian language], Moscow, available at: http://dict.ruslang.ru/freq.php.

A Quantitative Study of Pragmatic Markers in Everyday Spoken Russian

Natalia V. Bogdanova-Beglarian, Olga V. Blinova and Tatiana Y. Sherstinova
St. Petersburg State University, Russia

The term "pragmatic markers" is used to denote specific items of spoken speech, which are not part of the propositional content of the sentence, but which perform various (often multiple) pragmatic functions [1]. Pragmatic markers help speakers to initiate/close discourse; to attract the attention of the hearer; to mark a boundary in discourse (i. e., to indicate a new topic, a partial shift in topic, resumption of an earlier topic, etc.); to aid the speaker to find the proper words; to repair one's own or others' discourse; to serve as a filler; to express a response/reaction/attitude towards the discourse (including as well "back-channel" signals of understanding); to effect cooperation, sharing, or intimacy between speaker and hearer, and so on [2]. They may be as single words (such as the English "well") or multi-word expressions (e. g., "or whatever you want to call it"). It turned out that these elements are frequently used in spoken discourse of quite different languages, they much depends on the context and are usually difficult to translate. During the last years the notion of pragmatic markers has become rather popular among linguists and a great amount of related works has appeared. However, there is still no generally accepted understanding of what elements should be considered to be pragmatic markers. In this study we stick to understanding of pragmatic markers as it was proposed in [3].

Current research aims at statistical description of pragmatic markers usage in spoken Russian in a systematic way based on a representative amount of speech data. The study is carried out on the basis of two representative speech corpora— the corpus of Russian everyday speech "One Day of Speech" (ORD corpus) and "Balanced Annotated Text Collection" (SAT corpus) [4]. The report will contain quantitative data on the frequency of individual pragmatic markers usage in oral speech, comparative data of pragmatic markers used in different types of discourse (monologue and dialogue), correlation analysis of pragmatic markers functioning in different communicative situations, and depending on speakers' social and psychological characteristics.

Keywords: spoken Russian, everyday speech, oral discourse, pragmatic markers, pragmatics, corpus linguistics, sociolinguistics, quantitative methods

References

[1] Fraser, B. (1996): Pragmatic markers, Pragmatics, vol. 6, is. 2, 1996, pp. 167-190.

[2] Brinton, L. J. (1996): Pragmatic markers in English – Grammaticalization and discourse functions. Berlin/New York: Mouton de Gruyter.

[3] Bogdanova-Beglarian, N. V. (2014): Pragmatemy v ustnoj povsednevnoj rechi: opredelenie ponyatia i obshchaja tipologia ["Pragmatic markers in spoken everyday speech: Definition and general typology"]. In Vestnik Permskogo universiteta. Rossijskaja i zarubezhnaja filologia [Perm University Herald. Russian and Foreign Philology], iss. 3 (27), pages 7-20.

[4] Bogdanova-Beglarian, N., Sherstinova, T., Blinova, O., Martynenko, G., Baeva, E. (2018): Towards a description of pragmatic markers in Russian everyday speech, In: Potapova R., Jokisch O., Karpov A. (eds.) Speech and Computer (SPECOM 2018), Lecture Notes in Computer Science (including subseries Lecture Notes in Artificial Intelligence and Lecture Notes in Bioinformatics), Volume 11096, LNAI, Springer Verlag, p. 42-48.

Dependency Distance Through the Lifespan of Two English Writers

Neus Català[1], Brita Elvevåg[2] and Ramon Ferrer-i-Cancho[1]
[1]Universitat Politècnica de Catalunya, Catalonia, [2]University of Tromsø, Norway

The distance between syntactically related words in sentences is known to be smaller than expected by chance. A pressure to reduce dependency distances is believed to arise to counteract interference and decay of activation during sentence processing (Liu et al 2017). Interestingly, such dependency distances are reduced in cases of mild cognitive impairment (Roark et al 2011), but in the case of second language speakers it increases as linguistic competence increases (Ouyang and Jiang 2018).

We investigated dependency distances as expressed in the novels written by two renowned contemporary English female authors: Irish Murdoch, who late in life was diagnosed with Alzheimer disease, and a control female writer. We find that dependency distances are consistently smaller in Irish Murdoch throughout her whole lifespan. We discuss the implications of our findings for Murdoch's diagnosis and previous research on this writer involving dependency distance and other quantitative measures (e.g., Pakhomov et al 2011). In addition to traditional measures of dependency distance, we investigate a new z-scoring technique from the theory of spatial networks (Ferrer-i-Cancho 2019) as a possible way to control for factors such as sentence length and the syntactic structure of sentences.

Keywords: dependency grammar, dependency distance, English writers, Alzheimer

References

Ferrer-i-Cancho, R. (2019). The sum of edge lengths in random linear arrangements. Journal of Statistical Mechanics, 053401.

Liu, H., Xu, C. & Liang, J. (2017). Dependency distance: A new perspective on syntactic patterns in natural languages, Physics of Life Reviews 21, 171-193.

Ouyang, J. & Jiang, J. (2018). Can the probability distribution of dependency distance measure language proficiency of second language learners? Journal of Quantitative Linguistics 25 (4), 295-313.

Pakhomov, S. V., Chacon, D. A., Wicklund, M., & Gundel, J. K. (2011). Computerized assessment of syntactic complexity in Alzheimer's disease: A case study of Iris Murdoch's writing. Behavior Research Methods, 43(1), 136-144. https://doi.org/10.3758/s13428-010-0037-9

Roark, B., Mitchell, M., Hosom, J., Hollingshead, K. & Kaye, J. (2011). Spoken Language Derived Measures for detecting mild cognitive impairment. IEEE Transactions on Audio, Speech, and Language Processing 19, 2081-2090.

Why Does Negation Shorten a Clause?

Radek Čech[1], Barbora Benešová[1] and Ján Mačutek[2]
[1]University of Ostrava, Czech Republic, [2]Comenius University, Slovakia

The Menzerath-Altmann law (MAL, Cramer 2005) expresses relations between sizes of constructs and their constituents, with the constructs and the constituents being direct neighbours in a language unit hierarchy (such as e.g. phonemes – syllables/morphemes – words – clauses - sentences). It says that longer constructs tend to consist of, on average, shorter constituents.

We investigate a particular aspect of the relation between clause length and word length. This relation, as opposed to other ones like word – syllable, word – morpheme, or sentence – clause, has never been corroborated directly, probably because of difficulties connected with finding boundaries between clauses in sentences (the relation between sentence length and word length, a consequence of the MAL on the levels of sentence – clause and clause – word, was discussed in several papers, see Grzybek 2011 and references therein). Provided that the MAL is valid also on the level of clause and word, longer clauses should consist of shorter words, and vice versa.

From principles of synergetic linguistics (Köhler 2005) it follows that a change of word length should cause a change of clause length. We focus on negation in Czech which is (mostly, although not exclusively) realized by adding the prefix *ne-* to the beginning of the word:

(1) *Marie přišla*
[Mary came]
(2) *Marie* ***ne****přišla*
[Mary did **not** come]

This prefixation makes the word one syllable (and morpheme) longer, which should shorten the clauses containing the negated forms.

As the MAL is of stochastic character, it describes general tendencies. Thus, some instances which are "against the law" are admissible. In other words, validity of a stochastic law is manifested on large samples and some counterexamples do not violate the law (as opposed to deterministic laws).

The aim of the study is to test the following hypothesis:
Clauses with the negative form of the predicate are, on average, shorter than clauses with the affirmative predicate.

Only predicates were chosen for the analysis, because according to dependency grammar (Mel'čuk 1988) predicates constitute roots of clause structures, with the decisive impact on clause properties.

Data and the annotation of both clause and word from the Prague Dependency Treebank 3.0 were used. We tested 59 verb forms, with the minimal number of occurrences of each form (affirmative and negative) in the PDT being 20. Clauses with negative verb forms are on average shorter for 48 (out of 59) verb forms which were analysed.

Keywords: Menzerath-Altman law, clause length, negation, Czech

References

Cramer, I.M. (2005). Das Menzerathsche Gesetz. In R. Köhler et al. (eds.), *Quantitative Linguistics*. An International Handbook (pp. 659–688). Berlin, New York: de Gruyter.

Grzybek, P. (2011). Der Satz und seine Beziehungen I: Satzlänge und Wortlänge im Russischen (Am Beispiel von L.N. Tolstojs "Anna Karenina"). *Anzeiger für Slavische Philologie* 39, 39-74.

Köhler, R. (2005). Synergetic linguistics. In R. Köhler et al. (eds.), *Quantitative Linguistics. An International Handbook* (pp. 760–774). Berlin, New York: de Gruyter.

Mel'čuk, I.A. (1988). *Dependency Syntax. Theory and Practice.* Albany: SUNY Press.

The Co-Effect of Menzerath's Law and Heavy Constituent Shift in Natural Languages

Xinying Chen[1], Kim Gerdes[2], Sylvain Kahane[3] and Marine Courtin[2]
[1]Xi'an Jiaotong University,China, [2]Université Sorbonne Nouvelle - Paris 3, France, [3]Université Paris Nanterre, France

We aim to link Menzerath's law to the Heavy Constituent Shift phenomena and discuss its co-effect in natural languages.

Heavy Constituent Shift (Stallings et al., 1998) is a commonly

observed language phenomenon. It assumes that constituents that contain several words have more linguistic information and therefore are heavy. These constituents have a tendency to shift from their normal positions to the end of the sentence.

Similar to Menzerath's law, people believe that Heavy Constituent Shift is also associated with "language information", and this information is measured in terms of constituent size. This suggests that Heavy Constituent Shift is potentially compatible with Menzerath's law since both of them are focusing on the 'size' of linguistic components. Our hypothesis then is built based on these two premises. To be more specific, we investigate the sizes of constituents in terms of words and compare the difference between the sizes of these different constituents in all sentences that have either one or two constituents to the right of the main predicate X (and any number of dependents to the left of X):

1. XAB (the main predicate X has two constituents A and B to its right, and A precedes B)
2. XC (the main predicate X has only one constituent C to its right)

According to dependency grammar, the predicate is usually the main verb of a sentence. a, b, c are the sizes (the number of words/nodes) of the constituents A, B, and C, which are dependent on the main predicate.

First, according to the Menzerath's law, we can expect that:

I. $(a+b) / 2 < c$

And then, according to the Heavy Constituent Shift, we also expect that:

II. $a < b$

When we combine I & II, we can get that:

III. $a = (a+a) / 2 < (a+b) / 2 < c$

$\Rightarrow a < (a+b) / 2 < c$

$\Rightarrow \boldsymbol{a < c}$

We thus claim that "$\boldsymbol{a < c}$" is the co-effect of Menzerath's Law and Heavy Constituent Shift in natural languages.

We tested the "a < c" hypothesis with the Surface-Syntactic version (SUD 2.4) of the Universal Dependencies treebank set, which includes 83 different languages from various typological groups, with a majority of Indo-European languages. Our results show that our empirical analysis confirms the "$\boldsymbol{a < c}$" inequality across these typologically different languages.

To conclude, making use of the recently available coherently annotated

multi-lingual Universal Dependencies, we can bridge Menzerath's law with classic linguistic notions such as the Heavy Constituent Shift and expand the scope of studies on Menzerath's law.

Keywords: Menzerath's Law, Heavy Constituent Shift, co-effect

Reference

Stallings Lynne M., MacDonald Maryellen C., O'Seaghdha Padraig G. (1998) "Phrasal ordering constraints in sentence production: Phrase length and verb disposition in heavy-NP shift." *Journal of Memory and Language* 39 (3): 392-417.

Does the Century Matter? Machine Learning Methods to Attribute Historical Periods in an Italian Literary Corpus

Michele A. Cortelazzo[1], Franco M. T. Gatti[1], George K. Mikros[2] and Arjuna Tuzzi[1]
[1]University of Padova, Italy, [2]Hamad Bin Khalifa University, Qatar

The purpose of this study is to analyse an Italian literary corpus from a diachronic perspective, by means of machine learning methods. With reference to a basis of novels written between the XVIII and the XXI Century, the aim is to apply two machine learning algorithms, i.e. Support Vector Machine (SVM) and Random Forest (RF), in order to see whether it is possible to place a novel of unknown date in its right age. The corpus includes 200 Italian novels: 11 novels from XVIII Century, 105 from XIX, 80 from XX and 4 from XXI. Through a model called Author's Multilevel N-gram Profile (Mikros and Perifanos, 2013; Cortelazzo, Mikros and Tuzzi, 2018), which takes into account a set of linguistic features (i.e. the occurrences of bigrams, trigrams, words and word bigrams) each novel is splitted in text chunks of 2000 words in length, and then it is submitted to the statistical learning process of each algorithm. While SVM aims at finding the ideal hyperplane that identifies the greatest distance between two classes, RF builds multiple trees to assess the overall contribution of the classification task (James et al. 2013; Vapnik, 1995).

The results of this preliminary research have shown an impressive

accuracy in classification, since the precision has reached the 96.2% with SVM and the 92.8% with RF. More specifically, considering SVM , 420 out of 425 text chunks written in XIX Century and 198 out of 201 written in XX Century have been correctly classified. By means of RF the amount of correctly classified text chunks has been 416 out of 425 in the XIX Century, and 181 out of 201 in the XX. As expected, lower precision has been registered in XVIII and XXI Century, since those subcorpora included a lower number of novels and, consequently, a lower number of text chunks the algorithms could base their learning process on. However, all misclassification cases concerned contiguous centuries (e.g. 5 text chunks of the XIX Century were incorrectly assigned to the XX Century). This first experiment will be replicated with different arrangements of novels and corpora to assess the robustness of the adopted methods.

Keywords: machine learning, classification, diachronic corpora, Italian literature, Support Vector Machine, Random Forest

References

Cortelazzo, M.A., Mikros, G.K., Tuzzi, A. (2018), Profiling Elena Ferrante: a Look Beyond Novels, in: Iezzi D. F., Celardo L., Misuraca M. (eds), JADT '18 Proceedings of the 14th international conference on statistical analysis of textual data, vol. 2, Universitalia, Roma, pp.165-173.

James, G., Witten, D., Hastie, T., Tibshirani, R. (2013). An introduction to statistical learning: with applications in R. New York: Springer.

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

Vapnik, V. (1995). The nature of statistical learning theory. New York: Springer-Verlag.

Unpacking Lexical Intertextuality – Number of Types Shared Among Texts

Jiří Milička, Václav Cvrček and David Lukeš
Charles University, Czech Republic

Mainstream quantitative linguistics believes that quantitative laws are to be derived only from within texts as naturally occurring and homogeneous units of language (see Köhler et al. 2008). It is presumably due to the fact that linguistic laws applicable to isolated texts are blurred or distorted (cf. Berg 2014) when observed on aggregates of texts (e.g. in a language corpus).

However, we propose that insisting on conceptualizing language as a sum of individual texts would be a significant oversimplification. It has been pointed out several times in qualitative studies that intertextual links are an integral part of the textual ecosystem (e.g. Teubert 2005) and play an important role in facilitating communication. A quantitative example of this link can be found on various levels, e.g. the frequency of a word in one text can be estimated from its frequencies in other texts (which enables comprehension in situations when the word is missing or cannot be recognized). We would thus argue that texts are not independent entities but rather an interconnected ecosystem which influences even basic linguistic characteristics such as frequencies of units.

In this paper, we will focus on lexical intertextuality in relation to the variability of texts. Our starting hypothesis is that (H1:) the number of word-types shared by all texts in a collection (a corpus) is significantly smaller than what can be expected from a random model. This quantitative model is supposed to provide a baseline for comparison with authentic linguistic data by calculating the expected number of types shared by all texts in a hypothetical corpus with all types randomly shuffled among texts. The model will be tested against natural textual data – middle sized corpora (i.e. hundreds of millions of running words) in several languages, including English, German, Arabic and Czech.

Furthermore, this model is supposed to provide a baseline for testing other related hypotheses:

- (H2:) a corpus harvested from one domain (i.e. a random sample of the domain) shows a significantly higher relative number of

shared types than a heterogeneous corpus sampled from various domains; or

- (H3:) a corpus of text excerpts is necessarily more heterogeneous than a corpus of identical length consisting of whole texts etc.

Finally, this model can be used for contrastive purposes to compare the number of shared types in corpora of different languages and to formulate general quantitative laws controlling intertextuality on the lexical level.

Keywords: lexicon, intertextuality, corpus, random model, number of types

References

Berg, T. (2014): On the Relationship between Type and Token Frequency, Journal of Quantitative Linguistics, 21:3, 199-222, DOI: 10.1080/09296174.2014.911505

Köhler, R. – Altmann, G. – Piotrowski, R. (Eds) (2008). *Quantitative Linguistik / Quantitative Linguistics. Ein internationales Handbuch / An International Handbook*. Berlin, Boston: De Gruyter Mouton.

Teuber, W. (2005): My version of corpus linguistics. International Journal of Corpus Linguistics 10:1, 1-13.

A Quantitative Analysis of Queen Elizabeth II's and American Presidents' Christmas Messages' Syntactic Features

Zheyuan Dai
Zhejiang University, China

Christmas messages have been evolved as new traditions, namely Royal Christmas Broadcasts for U.K. and Presidents' Speeches with Lighting the National Christmas Tree for U.S. Queen of United Kingdom – Elizabeth II and American presidents deliver annual messages for their national. Several studies focus on different Heads of State's speeches at lexical level, especially on vocabulary richness or connotations (Čech, 2014; Dai & Liu, 2019; Savoy, 2010).

Nevertheless, sentences – linear syntactic structures – received little attention. A few researches employed qualitative analyses for holistic stylistic studies (Kredátusová, 2009) while in few can find quantitative methods.

Based on materials over 50 years (1967-2018), the present study aims at employing quantitative methods to compare the deep syntactic structures of messages delivered by two different country's Heads of States. Dependency Distance (MDD) as an effective metric to measure syntactic complexity (Hudson, 1995; Liu, et al. 2017) was adopted. Results show that Mean Dependency Distance (MDD) values are significantly different between two groups. However, MDD can be greatly affected by Sentence Length (SL) (Jiang & Liu, 2015). After the frequency analysis towards SL, the present study chose the sentences within the range of 10-30 words (10 sentences for each SL, 420 in total). DD's distribution of most randomly- selected sentences well fits to Right truncated modified Zipf-Alekseev distribution (Wimmer & Altmann, 1999) by Altmann-Fitter (2013). Within the SL span, there exists no significant difference of MDD values between two groups, denoting from the perspective of syntactical complexity, Christmas messages of two groups are similar. This results further may indicate the difference between British English and American English in terms of syntax is not significant, and suggest that DD is a universal human-cognitive index rather than a factor can be affected by social-specific reasons.

Keywords: Syntactic features, Zipf-Alekseev distribution, Dependency Distance, Christmas messages, Queen Elizabeth II, American presidents

References:

Altmann-Fitter, (2013). Altmann-Fitter User Guide. The Third Version. Downloadable at. http://www.ram-verlag.eu/wp-content/uploads/2013/10/Fitter-User-Guide.pdf (2014-11-29).

Čech, R. (2014). Language and ideology: Quantitative thematic analysis of New Year speeches given by Czechoslovak and Czech presidents (1949–2011). *Quality & Quantity*, 48(2), 899–910.

Dai, Z., & Liu, H. (2019). Quantitative Analysis of Queen Elizabeth II's and American Presidents' Christmas Messages. *Glottometrics*,

45, 63-8.
Hudson, R. (1995). *Measuring Syntactic Difficulty (Manuscript)*. London: University College.
Jiang, J., & Liu, H. (2015). The effects of sentence length on dependency distance, dependency direction and the implications–based on a parallel English–Chinese dependency treebank. *Language Sciences*, 50, 93-104.
Kredátusová, M. (2009). Queen's Christmas Speeches 1952–2007: Discourse Analysis. Brno: Masaryk University.
Liu, H., Xu, C., & Liang, J. (2017). Dependency distance: a new perspective on syntactic patterns in natural languages. *Physics of life reviews*, 21, 171-193.
Savoy, J. (2010). Lexical Analysis of US Political Speeches. *Journal of Quantitative Linguistics*, 17(2), 123–141.
Wimmer, G., & Altmann, G. (1999). *Thesaurus of univariate discrete probability distributions*. STAMM Verlag, Essen.

Too Much of a Good Thing

Sheila Embleton, Dorin Uritescu and Eric S. Wheeler
York University, Canada

"More is better", it is said, and especially when it is more data. It seems reasonable that the more data we have about a subject, the more understanding we can gain. So-called "Big Data" is a popular concept just now and there is much truth to this. Quantitative studies have always relied on having enough data to make a meaningful model. From early counts of words (e.g. Zipf 1949; Herdan 1956) to modern studies over digitized collections of millions of tokens, having enough data seems essential.

It is no less the case in our work (Embleton, Uritescu & Wheeler 2013 and elsewhere) on language variation vs geographic distribution. It is essential to have enough geographic locations, and enough linguistic items at any location to be able to assess how much the linguistic variation correlates with geographic factors such as distance, travel time, and even the scale of distance (cf. distances across China vs. distances across a small set of communities in Africa).

However, in at least one of our studies, we have found that too much detail hides the patterns we expect to find. In particular, the data set from the Mambila languages found along the Nigeria-Cameroon border region (Connell p.c.) shows every location with some differences to every other location -- some minor, others more basic, but in general all different. Hence, the multidimensional scaling (MDS) pictures that we use to show linguistic similarities and differences are, at best, uninformative.

A more meaningful approach will group linguistic variants according to their status as cognates. The labour and expertise required to do this over the more than 800 data sets is prohibitive unless we can use an automated process -- and the process we investigate is one of eliminating some differences (such as tone variation, vowel differences, etc.; we try various combinations) until we have a sequence of, say, consonants that will match likely cognates. By hiding detail, we bring out patterns that are of interest. Less information does more.

Having identified this challenge in one case, we realize that there have been other cases where we got better results with less information (or more precisely, with small subsets of more homogeneous data). Grouping linguistic variation according to grammatical types has proven to give better correlations between language and geography. Using meta-data labels on data items greatly enhances our ability to search for and identify such groupings.

Our theoretical conclusion, then, is that the principle of "More is better" needs to be counter-balanced by a principle that says "Less is better when less is more informative". And automatic processes, applied axiomatically, do not necessarily produce the best results; the researcher must (creatively) find or select the model that provides the most insightful results.

Keywords: quantitative study of language variation, quantitative methods, big data, Mambila languages, Chinese languages, multidimensional scaling

References

Connell, Bruce. p.c. Data for Mambila languages. unpublished.
Embleton, Sheila, Dorin Uritescu & Eric S. Wheeler. 2013. *Literary and Linguistic Computing* 28(1):13-22. DOI: 10.1093/llc/fqs048
Herdan, G. (1956) Language as Choice and Chance. Groningen, Netherlands: P. Noordhoff Ltd. Zipf, G.K. 1949. Human Behavior and

the Principle of Least Effort; An introduction to human ecology. facsimile 1965. New York: Hafner Pub. Co.

A Quantitative Study of the Semantic Integration Between the Verbs of Sound Emission and the *Way*-Construction Paradigm in English: The *Way* Construction and the Intransitive Motion Construction Taken as the Cases

Zhanfeng Fang
Ningbo University, China

This paper is concerned with a frequency-based quantitative study of the semantic integration between the verbs of sound emission and the way-construction paradigm on the basis of British National Corpus. The way-construction paradigm centers on the way construction and the intransitive motion construction. The paper finds that: (1) different from Levin & Song (1997) and Rohde (2001), the verbs of sound emission participating in the intransitive motion construction outnumber those in the way construction in terms of the token and type frequency; (2) the fitting results of the data of the rank and the token frequency of the verbs in the two constructions are compatible with Zipf's Law, indicating that a few high-frequency verbs are often used, and most of the low-frequency verbs are seldom used. This first finding results from different semantic integrations between the above verbs and the two constructions. The way construction always offers the motion meaning to the verbs, and the sound contributed by the verbs occurs concomitantly with motion. The intransitive motion construction always licenses the verbs with the motion meaning at the lexical level, and the emission of sound is always caused by the instigation of motion. It is shown that the intransitive motion construction is more typical than the way construction when expressing the constructional meaning of concrete motion. The second finding comes from the fact that the high-frequency verbs have more semantic and syntactic freedom to enter the different superordinate constructions. I hope that this paper will be conducive to the studies of the relation between the subordinate verb and the superordinate construction, and the interplay of quantitative linguistics and cognitive construction grammar.

Keywords: the verbs of sound emission; the way construction; the intransitive motion construction; form and meaning; quantitative study

References

Goldberg, A. (1995). *Constructions: A Construction Grammar Approach to Argument Structure*. Chicago: University of Chicago Press.

Levin, B. & Song, G. (1997). Making sense of corpus data: A case study of verbs of sound. I*nternational Journal of Corpus Linguistics*, 2(1), 23-64.

Liu, H. (2017). *An Introduction to Quantitative Linguistics*. Beijing: The Commercial Press.

Rohde, A. R. (2001). *Analyzing Path: The Interplay of Verbs, Prepositions and Constructional Semantics*. Ph. D Dissertation. Houston: Rice University.

Zipf, G. K. (1935). *The Psycho-Biology of Language: An Introduction to Dynamic Biology*. Boston: Houghton-Mifflin Company.

A Quantitative Investigation of Translator Stylometry in Modern Chinese-English Literary Translation

Jing Gao and Jingyang Jiang
Zhejiang University, China

As one of the important issues in quantitative stylistics, the specific problem of attributing translated texts to their translator and original author has received little attention and remained relatively under-studied in the field of stylometry (Hedegaard & Simonsen, 2011; Lynch, 2014). Though application of computational and statistical approaches in the analyses of literary translations has gained considerable ground in recent years, there is still a good deal of progress to be made (El-Fiqi Petraki & Abbass, 2011; Lynch & Vogel, 2018). Previous studies in translator stylometry have mostly focused on European language pairs(Lee, 2018; El-Fiqi, Petraki & Abbass, 2019; Volansky Ordan & Wintner, 2015). Only a small number of studies have been conducted on Chinese-English cannon translation and they all followed the corpus methodology

proposed by Mona Baker. (Huang & Chu, 2014; Hou, 2015).

Employing the quantitative indexes and statistic approaches from stylometry, this study focuses on the works of Howard Goldblatt, one of the most prolific English-language translators of Chinese literature, and used multivariate analysis to detect and identify the translator's stylistic fingerprint and to track different authorial voices in the translation of Goldblatt. Word and POS n-gram features (specifically unigram, bigram, trigram and four-grams) are first used to discriminate between translation and non-translation, and then to discriminate between different authorial styles by the same translator in the translated texts. The discriminations are achieved by means of hierarchical cluster analysis and validated by bootstrap consensus tree analysis with a consensus strength at 0.5. Then distribution tests (Student's f-test or Mann-Whitney test) are employed to identify linguistic patterns which contribute significantly to the stylistic fingerprint of the translator. The results suggest that both word and POS n-gram features prove useful for distinguishing between translation and non-translation texts, between different translator's works and then between different authorial voices in the comparable corpus. Attribution efficacy increases with the growth of frequency vector size. Word unigram generally perform better than other word and POS n-grams in discrimination. 150 MFW can give most accurate attribution results whereas other word and POS n-grams need much larger vector size for correct attribution. The results confirmed a high attribution effectiveness of word unigram in previous translation studies in English translation corpus.

Keywords: quantitative stylistics, translator stylometry, multivariate analysis

References

El-Fiqi, H., Petraki, E., & Abbass, H. A. (2011). *A computational linguistic approach for the identification of translator stylometry using Arabic-English text*. Paper presented at the International Conference on Fuzzy Systems, Taipei, Taiwan.

El-Fiqi, H., Petraki, E., & Abbass, H. A. (2019). Network motifs for translator stylometry identification. *PLOS ONE*, 14(2), e211809.

Hedegaard, S., & Simonsen, J. G. (2011). Lost in Translation: Authorship Attribution using Frame Semantics, *Proceedings of the 49th Annual Meeting of the Association for Computational Linguistics* (pp. 65-70). Portland, Oregon: Association for Computational Linguistics.

Hou, Y. (2015). A corpus-based study of nominalization in English translations of Chinese literary prose. *Digital Scholarship in the Humanities*, 30(1), 39-52.
Huang, L., & Chu, C. (2014). Translator's style or translational style? A corpus- based study of style in translated Chinese novels. *Asia Pacific Translation and Intercultural Studies*, 1(2), 122-141.
Lee, C. (2018). Do language combinations affect translators' stylistic visibility in translated texts? *Digital Scholarship in the Humanities*, 33(3), 592-603.
Lynch, G. (2014). A Supervised Learning Approach Towards Profiling the Preservation of Authorial Style in Literary Translations, *COLING* (pp. 376-386).
Lynch, G., & Vogel, C. (2018). The translator's visibility: Detecting translatorial fingerprints in contemporaneous parallel translations. *Computer Speech & Language*, 52, 79-104.
Volansky, V., Ordan, N., & Wintner, S. (2015). On the features of translationese. *Digital Scholarship in the Humanities*, 30(1), 98-118.

The Paragraph Through the Lense of Menzerath-Altmann Law

Volker Gröller
Austria

In this paper the relation between paragraph and sentence length is examined.

The former often being regarded as an intuitive next entity of higher order after the sentence, we investigate its interconnection with the length of sentences. Starting off with a short history of the paragraph and Menzerath-Altmann's law, we proof a solid interconnection abiding said law. This not only brings the paragraph closer to the sentence from a hierarchical perspective of language categories, but opens the door to new research, reaching for even bigger concepts, like text length, chapter length etc. The material used for this research is Fedor Dostoyevsky's novel "Crime and Punishment" in Russian.

Keywords: Menzerath-Altmann's law, paragraph, sentence

Phonetic Features Affecting the Naturalness of Pitch Accent: Result From a Web-Based Survey

Naoki Hayashi
Nihon University, Japan

Objectives
This paper reports the results of an online survey which was conducted using synthesized speech in order to shed light on speech characteristics in regard to the "naturalness" of accents.

Methods
First, in order to record sample voices for the survey, we asked the people who used to be news anchors to record voices. Next, out of the recorded voices, we picked three words: *mizu* ("water," an accentless word), *yama* ("mountain," accented on the second mora), and *mado* ("window," accented on the first mora). Then, we manipulated the steepness of the drop-in pitch and relative peak positions of the pitch accent and created 30 synthesized speech patterns.

Following this, we conducted an online survey where we asked participants to listen to such synthesized speech and judged whether the voice is natural or not based on the perspective of their local dialects. The subjects of the survey were those who grew up in the Tokyo metropolitan area and in their 20s-60s, with a total of 784 respondents participating in the survey.

Results
The results showed that for the accentless word, smaller drops in pitch were judged more natural, whereas larger drops in pitch were judged more natural. With respect to the relative peak position, among the accented words, *yama* (accented on the second mora) was judged more natural with a later peak, while *mado* (accented on the first mora) was judged more natural with an earlier peak.

Furthermore, the size of the drop-in pitch reinforced the effect of the

naturalness judgment due to the relative peak position, which suggests that the two features (drop in pitch and relative peak position) interact in mutually reinforcing ways to affect naturalness.

Following the considerations mentioned above, this paper attempts to analyze as well the effect the fall width and relative peak position have on the determination of "naturalness" using statistical methods.

Keywords: Tokyo Metropolitan Area, accent, perception

Linguistic Laws in Catalan

Antoni Hernández-Fernández[1], Juan María Garrido[2], Bartolomé Luque[3] and Iván González Torre[3]
[1]Universitat Politècnica de Catalunya, Spain, [2]UNED, Spain,
[3]Universidad Politécnica de Madrid, Spain

Catalan is a Romance language spoken in the Western Mediterranean by more than ten million people, with other small communities of speakers spread around the world. In a previous work, a methodology that does not require a previous segmentation of the signal was applied to directly study acoustic speech signals in sixteen different languages, including Catalan, and successfully recovering at levels even below the phoneme some well-known regularities of human communication [1]. These patterns, also called linguistic laws, are statistical regularities emerging across different linguistic scales that can be formulated mathematically, and have been postulated and widely researched mainly in written texts [2].

It is known that human languages have an acoustic origin and consequently linguistic laws may fit better in the case of oral corpus than in written texts, as seen in a previous study for oral English [2]. It has been argued that linguistic laws in written texts are not an inherent property of written language (symbolic hypothesis) but are in turn a by-product of similar structures already present in oral communication, thereby pointing to a physical origin of linguistic laws (physical hypothesis)[2].

In this work we have analyzed in depth the Catalan speech at different levels, following a methodology previously developed [2], providing new

empirical evidence in favor of the physical hypothesis. To this end, we have studied the Glissando oral corpus [3] and compared the results found for linguistic laws (Zipf's law, Brevity law, Herdan's law, Menzerath-Altmann's law and Lognormality law) in Catalan, with those existing for English and Spanish. This new evidences reinforce the idea that linguistic laws would be universal patterns in human languages inherent in the physical production, and encourage researchers to directly explore the acoustic signals, recovering the experimental approach of some pioneers of Quantitative Linguistics [4].

The question is if these laws are the result of acoustics processes inherent to human physiology, shared with the acoustic communication systems of other species, or due to universal principles of animal behavior [5]. Future work is necessary to increase empirical evidences by analyzing other languages and to extend this approach to other acoustical communication systems, or to signals of unknown code.

Keywords: Catalan, linguistic laws, physical hypothesis, Zipfian laws, Menzerath-Altmann's law

References

[1] González Torre, I., Luque, B., Lacasa, L., Luque, J. & Hernández-Fernández, A. (2017). Emergence of linguistic laws in human voice. *Scientific Reports* 7, 43862.

[2] González Torre, I., Luque, B., Lacasa, L., Kello, C. & Hernández-Fernández, A. (2019). On the physical origin of linguistic laws and lognormality in speech, *Royal Society Open Science* 6, 191023.

[3] Garrido, J. M., et al. (2013). Glissando: a corpus for multidisciplinary prosodic studies in Spanish and Catalan. *Language Resources and Evaluation* 47(4): 945-971.

[4] Menzerath, P. & De Oleza, J.M. (1928). *Spanische lautdauer: eine experimentelle untersuchung*. Berlin, Germany: Mouton De Gruyter.

[5] 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* 37, 1565-1578.

On the Application of Quantitative-Linguistic Models and Methods in Audiovisual Studies

Marek Holan
Palacký University Olomouc, Czech Republic

This research crosses the border between two disciplines of humanities: audiovisual studies and quantitative linguistics. The principal question is whether there can be found any quantifiable regularities in the area of audiovisuality with particular regard to the television genre series field. The models derived from quantitative linguistics (for example Menzerath-Altmann law) are applied on the source material for the purpose of solving this issue. A special question is whether these regularities – if found – have any impact on the construction of specific genres (apart from the content-related criteria). The initial opinion is that audiovisual material shares some structural similarities with language (or that there can be found some textual qualities); therefore the application of exact quantitative-linguistic methods can be tested.

The area of audiovisual studies is a source of the empirical material. Firstly – when defining the genre itself –the concept of genre as an open-ended category dependent on pragmatic usage of its various forms in specific situations of cultural communication is held. Secondly defining and measuring the exact units of television genre series follows. The segmentation is rather complicated because it requires some speculative definitions too. There is a wide variety of units ranging from the average shot length (related to the scene length) up to a point of sequence level. The semantic criteria are necessary as well as the rigid ones because some units (for example the shot) are more precise and some are more fluid.

The area of quantitative linguistics provides the methodology and procedures which are then used for the interpretation of the data gained in the empirical phase. There are already some basic outcomes of measurements from the preliminary part of the research. The crime fiction genre (specifically two series from different historical periods) was chosen as a case study. It is clear from the first results that there is a major difference between measuring the average shot length related to the length of a whole episode and dividing the source material into several interrelated units as mentioned above. In the first case the relational complexity is

reduced and the Menzerath-Altmann law seems to be disqualified even at first sight. The second case, however, provides a better way because the segmentation from less to more complex units is derived from the text segmentation and offers a more complex and precise view on the whole matter.

Keywords: television genre series, Menzerath-Altmann law, semantic criteria, average shot length, regularity

References

Frow, J. (2006). *Genre*. London, England: Routledge

Grzybek, P (Ed.). (2006) *Contributions to the Science of Text and Language. Word Length Studies and Related Issues*. Dordrecht, Netherlands: Springer

Grzybek, P., Koch, V. (2012). Shot Length: Random or Rigid, Choice or Chance? An Analysis of Lev Kulešov's Po zakonu. In E. Hess-Lüttich (Ed.). *Sign Culture. Zeichen Kultur* (pp.169-188). Würzburg, Germany: Königshausen & Neumann

Koch, V. (2014). *Quantitative Film Studies: Regularities and Interrelations exemplified by Shot Lengths in Soviet Feature Films*. Graz, Austria: The University of Graz

Building and Exploring Networks of Components of Chinese Characters

Wei Huang and Junting Li
Beijing Language and Culture University, China

According to traditional Chinese philology, there is a dominant configuration system which governs both the formation and the evolution of tens of thousands of characters in Chinese scripts. Chinese characters can be deconstructed into one or more components. From the viewpoint of system theory, 560 elementary components composed more than 20 thousand characters in modern Chinese scripts through only 11 construction modes. It is natural to investigate the Chinese character configuration system by using the approach of complex network.

In this article, four Chinese character configuration networks are proposed preliminarily, and a series of characteristics of each network are quantitatively measured.

The first basic network takes components as nodes, which are linked by only the co- occurrence relationship within a character. For example, the components '讠' (yan, speak),
'五' (wu, five) and '口' (kou, mouth) form a triangle since they compose a character '语' (yu, speech). Within a definite collection of Chinese characters, a component network is formatted by deconstructing all of the characters, such as the most common 3500 characters in modern Chinese.

If the frequency of components used in the character formation is considered, a weighted network will be renewed by adding the occurrence frequency on the links between component nodes. Moreover, in the formation of characters the components always occur in proper order. Some of them almost appears in certain position, while others don't have positional orientation. If the sequence of occurrence is also concerned, the first network can be transformed into a directed network. Here the direction between components can be technically defined according to the order of strokes in handwriting.

The above networks take advantage of only the properties in planar structure of Chinese characters, neglecting the hierarchical dimension. In the above example, '五' and '口' composed '吾' (wu, myself) firstly and then '讠' and '吾' formed '语'. Thus lots of characters composed by more than 2 components have a hierarchical structure in the formation. The hierarchical characteristics in character formation is examined by a deleting operation, i.e., the co-occurrence links between components which are not in the same structural level are deleted from the basic network.

To describe and compare the four networks above, some common characteristics of networks, including but not limited to average (in-/out-)degree, (in-/out-)degree distribution, degree centrality, density, etc., will be computed.

Keywords: Chinese character, complex network, co-occurrence network

Genitive Inversion in Norwegian

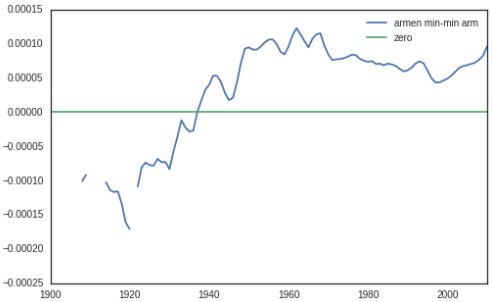
Lars G. Johnsen
National Library of Norway, Norway

We show how the Norwegian nominal possessive construction is subject to an ongoing change, using data from the National Library's ngram database, spanning two centuries of books and newspapers (Breder Birkenes et.al. 2015). The possessive construction changes with respect to a choice between pre- or postnominal pronominal possessor, where the data suggests that a postnominal possessor position is preferred over a prenominal for a large class of nouns. Norwegian differs in this construction from its closely related neighbouring languages Swedish and Danish.

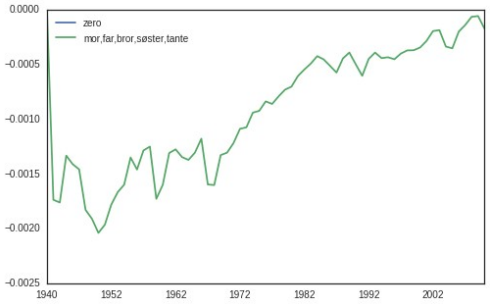
The Norwegian possessive construction comes in two variants, prenominal and postnominal, (cf. Hellan, 1980; Lødrup, 2011).

Armen hans/hans arm er sterk
Arm#def his/his arm is strong

The statistics, or frequency patterns, of these two variants shows some notable properties.



Here is a plot for each year of the 20th century, of "armen hans" (arm#def his) and "hans arm" (his arm)), shown as a difference between the two (postnominal minus prenominal).



The illustration shows that there is a clearly defined crossing between the two constructions around 1940, when the postnominal is moving ahead and continues to stay ahead.

The trend shown in the figure above is pervasive throughout Norwegian nouns, but different classes of nouns behaves differently. We show this with reference to body parts, posessions, and family relations among other. Interestingly, an aggregate of family words (mor-mother, far-father and so on) shows that while they seem to approaching a shift, still has some time before crossing the zero line.

The theory of possessives that we employ here is taken from (Partee & Borschev, 2003; Barker, 2011), where the distinction between inherently relational nouns and implicational relational nouns are made. We will try to establish that inherently relational nouns are late changers while implicational change early on with body parts on one end of the spectrum, and family relations on the other.

The result is established through an analysis of the change rate (derivate) of the development curves of different classes. Change rates and the shapes of the curves (convexity, concave) differ between individual words and across noun classes.

Keywords: Inversion, N-grams, trends, large data

References

Barker, Chris. 2011. Possessives and relational nouns. In *Semantics. An International Handbook of Natural Language Meaning*, Vol. I, ed. by C. Maienborn, K. von Heusinger, & P. Portner. Berlin: De Gruyter Mouton

Breder Birkenes, Magnus & Johnsen, Lars G. & Lindstad, Arne Martinus

& Ostad, Johanne, 2015. “From digital library to n-grams: NB N-gram” in *Proceedings of the 20th Nordic Conference of Computational Linguistics*, pp.293-295, Linköping University Electronic Press, Sweden

Delsing, Lars Olof,1998. “Possession in Germanic”, in *Possessors, Predicates and Movement in the Determiner Phrase*, ed. Artemis Alexiadou & Chris Wilder, John Benjamins, Amsterdam.

McKinney, Wes, 2012. *Python for Data Analysis, Data Wrangling with Pandas, NumPy, and IPython*. O'Reilly Media.

Partee, Barbara H. & Vladimir Borschev. 2003. Genitives, relational nouns, and argument-modifier ambiguity. In *Modifying Adjuncts*, ed. by E. Lang, C. Maienborn, & C. Fabricius-Hansen, 67–112. Berlin/New York: Interface Explorations 4, Mouton de Gruyter.

The Information Content of Machine Translation

Patrick Juola
Duquesne University, United States of America

It is a commonplace observation that “translation adds information” (Juola, 1997; Ivir, 2002- 2003; House, 2015); the translation process adds information necessary in the cultural context of the new reader that would have been implicit to the original reader. Examples of this marking a “brother” (in English) as older or younger in Chinese, or expanding references and providing context that would not necessarily make sense. [Baker (1992) provides an excellent example: Who is “Truman” to an Arabic speaker, and what political metaphor is being used?]

This observation has been confirmed by empirical, information-theoretic studies (Juola, 1997; Juola, 2005), where standard compression technology (e.g. gzip, bzip2, etc.) is used to estimate the Komogorov complexity or the “information contained” in a given document. Juola (1997) showed that there was generally a substantial information gain between the original version and translated versions of a given text.

In a pilot study, Juola (in press) analyzed a small ad hoc collection of 33 documents based on repeated translations of the Oxford University Bodleian oath and showed that back-translating these documents to English using Google Translate did not increase the amount of

information (as measured by gzip), and, in fact, reduced it. We extend and replicate this finding with a larger collection of documents, using several different machine translation methods, and several different types of compression programs. We show that, irrespective of the translation method used, the amount of information post-translation is significantly less than the amount pre-translation. This finding also holds using target languages other than English.

There are several potential explanations: machine translation may simply be "lossy," or the nature of the original documents may have significant properties that produce this counterintuitive finding.

Keywords: information theory, machine translation, computational linguistics, cognitive linguistics

References:

Juola, Patrick (1997). A Numerical Analysis of Cultural Context in Translation. Proceedings of the Second European Conference on Cognitive Science, Manchester, United Kingdom.

Ivir, Vladimir. Translation of Culture and Culture of Translation - SRAZ XLVII-XLVIII, 117-126 (2002-2003)

House, Julian. (2015) Translation Quality Assessment: Past and Present. London:Routledge Baker, Mona. (1992). "In Other Words: A Coursebook on Translation." London:Routledge

Juola, Patrick (2005). Compression-based Analysis of Language Complexity. Proceedings of Approaches to Complexity in Language, Helsinki, Finland.

Dating and Geolocation of Medieval and Modern Spanish Notarial Documents Using Distributed Representation

Yoshifumi Kawasaki
The University of Tokyo, Japan

This paper proposes a method to probabilistically date and geolocate medieval and modern Spanish notarial documents. They generally bear date and place of issue, but there do exist those without explicit provenance. We use as corpus *CODEA+*

2015 containing nearly 2500 documents composed during the period from 1100 to 1800 across the present-day Spain. Among them, some 300 are undated. We intend to assign these undated documents an estimated year and location as text classification task.

Our model is inspired by distributed representation of words (Mikolov *et al*. 2013). We construct a neural network for learning spatio-temporal similarity among words so that the words that appear in documents written within a chrono-geographically close area have similar embedding vectors. The spatio-temporal similarity is learned in a framework of multi-task learning to acquire more suitable representation than when learned independently (Goldberg 2017). Once we have obtained word embeddings, a document embedding is simply computed as mean over all the words therein. Then we can convert this document embedding through the learned weight matrices into two probability distributions: chronological and geographical ones. Finally, the document is ascribed to the period and place with the highest probability in respective distributions.

We performed an experiment by splitting around 2200 documents with known date and place into 2000 training data and 200 test data. We obtained a mean absolute error of 22 years and 140 km for dating and geolocation, respectively. The result seems promising. However, we believe there is still room for improvement. Future investigation should be directed toward leveraging character-level embeddings to be acquired by using Recurrent Neural Network. This approach would enable us to exploit orthographic variation as well as morphological commonality that are not addressed in the present study.

One of the advantages of our model is its ability to detect the most contributing words to estimation. The degree of contribution corresponds to vector norm. As the document embedding is calculated as mean over all the word therein, the words with larger norm become more decisive. We observed that most of the words with largest norm turn out to be those closely related with certain chrono-geographical area, well-known to Hispanic Philology (Menéndez Pidal 1999, Zamora Vicente 1967). Our proposed model constitutes the first step toward developing a quantitative method of dating and geolocation accompanied by empirical evidence.

Keywords: dating, geolocation, distributed representation

References

GITHE (Grupo de Investigación Textos para la Historia del Español): *CODEA+ 2015 (Corpus de Documentos Españoles Anteriores a 1800)*. http://www.corpuscodea.es/

Goldberg, Y. (2017). *Neural Methods for Natural Language Processing. In Synthesis Lectures on Human Language Technologies*, Vol.10. San Francisco: Morgan Claypool.
Mikolov, T., Chen, K., Corrado, G. & Dean, J. (2013). Efficient Estimation of Word Representations in Vector Space. https://arxiv.org/pdf/1301.3781.pdf
Menéndez Pidal, R. (1999). *Manual de gramática histórica española* (Vigésima tercera ed.). Madrid: Espasa- Calpe.
Zamora Vicente, A. (1967). *Dialectología española* (Segunda edición muy aumentada). Madrid: Gredos.

Quantitative Loanword Studies: A New Synergetic Perspective

Emmerich Kelih
Universität Wien, Austria

This contribution will give an overview on quantitative loanword studies. In quantitative linguistics in this respect the so called Piotrowski law has a very prominent role, which is used for the modelling of the frequency of loanwords in its diachronic development. Our aim is to go beyond the solid ground of the Piotrowski law (which is, in addition to Zipf's and Menzerath's law without any doubt a cornerstone of QL) and to implement a new synergetic view on loanword studies in general. Here first of all usage- and frequency-based approaches will be discussed. Secondly, a particular emphasis is laid on general requirements, which can influence and shape the integration of loanwords (where a birth- and-death process seems to be relevant) into a language. In a third step we will discuss some interrelations between the frequency, the age, polysemy, synonymy, morphological productivity, and idiomatic potential of loanwords. Finally some preliminary empirical results from the languages of the world are presented.

Keywords: loanword, Piotrowski law, synergetic approach

References

Kelih, Emmerich (2014): Zur quantitativen Lehn- und Fremdwortforschung:

Eine Einleitung. In: Best, Karl- Heinz, Kelih, Emmerich (eds.): *Entlehnungen und Fremdwörter – Quantitative Aspekte*. Lüdenscheid: Ram-Verlag (Studies in Quantitative Linguistics, 15), S. 1–6.

Does Sentence Length Matter in MDD and MHD to Measure Syntactic Development?: In the Case of Japanese Learners' Essays

Saeko Komori[1], Masatoshi Sugiura[2] and Wenping Li[3]
[1]Chubu University, Japan, [2]Nagoya University, Japan, [3]Dalian Maritime University, China

The purpose of this study is to examine the effects of sentence length (SL) in measuring syntactic development using mean dependency distance (MDD) and mean hierarchical distance (MHD) with Japanese intermediate learners essays.

Ouyang and Jiang (2018) conducted a study of MDD using Chinese L1 English L2 learners' compositions in eight grades and reported the MDD increased as their grades proceeded. Komori et al. (2018) examined the MDD with Chinese L1 Japanese L2 advanced learners' written data and reported that there was no significant differences in the MDD among three levels of advanced learners. Komori et al. (2019), examining the MDD and MHD with Chinese L1 Japanese L2 intermediate learners' essays, reported that there was no significant dereferences in the MDD between two levels of the intermediate learners, but they found a gradual increase in the MHD from lower to higher intermediate learners.

Ferrer and Liu (2014) recognized the risks of dependency lengths from sequences of different sentence length, and advocated investigating the distribution of dependency lengths in sentences of the same length.

In this study, we analyze the relationships among SL, MDD and MHD in order to clarify the effects of SL for the two measures of syntactic development. First, we looked at the data as a whole to see how MDD and MHD change as SL gets longer. Second, we scrutinized the distribution of MDD and MHD for each SL individually from SL 5 to 16. We compared the results of the two levels of the intermediate learners to see if the measures increase as SL gets longer. Lastly, we examined the

relationships between MDD and MHD for each SL.

As a result, we found the following three points:

1) There were correlations between SL and MDD, and between SL and MHD in both lower and higher level learners.
2)MDD did not show differences between the two levels of learners. MHD, however, showed differences for longer sentences.
3)MDD and MHD showed negative correlations in all SL in both level learners' data.

Based on the results, we may conclude that MDD could not measure syntactic development, but MHD could.

Keywords: dependency length, leaners' syntactic development, Japanese analysis, reexamination by each sentence length

References

Ferrer-i-Cancho, R., & Liu, H. (2014). The risks of mixing dependency lengths from sequences of different length. *Glottotheory*, 5, 143-155.

Komori, S., Sugiura, M., & Li, W. (2018). Examining the applicability of the mean dependency distance (MDD) for SLA:A case study of Chinese learners of Japanese as a second language. *Proceedings of the 4th Asia Pacific Corpus Linguistic Conference (APCLC 2018)*, 237-239.

Komori, S., Sugiura, M., & Li, W. (2019). Examining MDD and MHD as syntactic complexity measures with intermediate Japanese learner corpus data. *Proceedings of Syntax Fest* 2019.

Ouyang, J., & Jiang, J. (2018). Minimization and probability distribution of dependency distance in the process of second language acquisition, *QUALICO*, 2018.

Attributivity and Syntactic Subjectivity in Contemporary Written Czech

Miroslav Kubát[1], Xinying Chen[2], Kateřina Pelegrinová[1] and Radek Čech[1]
[1]University of Ostrava, Czech Republic, [2]Xi'an Jiaotong University, China

The submitted study focuses on two syntactic features of various text

types and genres in contemporary written Czech language based on an analysis of a large corpus.

The Czech stylistics is mainly focused on the lexical features of styles. Phonetic, morphological, and syntactic features are usually rather out of the main interest of scholars (cf. Čechová et al. 2008; Hoffmannová et al. 2016). The exception is Bečka (1992) who paid extraordinary attention to syntax. Meanwhile, Czech stylistics is rather based on qualitative than quantitative approaches. There are only a few quantitative studies dealing with syntactic functions or parts of speech in Czech from a stylistic point of view (e.g. Kubát 2016, Těšitelová 1985, Uhlířová 1974).

Since these studies are usually limited to (a) small samples and (b) few analyzed styles, we aim to tackle these issues in a different manner. First, our samples are collected from a large corpus. Second, we analyze not only the main style groups such as fiction and non-fiction, but we focus also on particular genres such as novels, short stories etc.

To be more specific, our data source is the largest corpus of contemporary written Czech language with syntactic annotation - SYN2015 (Křen et al. 2016). This corpus, provided by Czech National Corpus, has one hundred million tokens and consists of written texts of three main text types/styles (fiction, non-fiction, newspapers & magazines). These three types/styles are then sorted into 16 categories/genres such as novels, short stories, poetry, drama, humanities, natural science, autobiographies, administrative texts, nationwide newspapers, leisure magazines.

Two indicators are measured for different text types as well as for selected genres:
(a) Index of Attributivity defined as the ratio of the frequency of attributes and the sum of frequencies of nouns and pronouns; (b) Index of Syntactic Subjectivity defined as the ratio of frequencies of subjects and predicates. The general ideas behind these indicators are (a) sentences with higher complexity are generally longer and prefer using more attributes, thus higher Attributivity might appear in more formal texts with longer sentences; (b) since Czech has rich morphological features, subjects can be omitted in expressions, therefore, higher Syntactic Subjectivity would also appear in more formal texts which are more syntactically well-formed.

This study enriches the previous research on the style of Czech texts with new perspectives.

Keywords: attributivity, subjectivity, style, genre, corpus

References

Bečka, J. V. (1992). Česká stylistika.
Čechová, M., Krčmová, M., Minářová, E. (2008). Současná stylistika.
Hoffmannová, J. et al. (2016). Stylistika mluvené a psané češtiny.
Křen, M. et al. (2016). SYN2015: Representative Corpus of Contemporary Written Czech. In: Proceedings of the Tenth International Conference on Language Resources and Evaluation (LREC'16), 2522-2528. Portorož: ELRA.
Kubát, M. (2016). Kvantitativní analýza žánrů.
Těšitelová, M. (1985). Kvantitativní charakteristiky současné češtiny.
Uhlířová, L. (1974). O frekvenci větných členů v souvislém textu. Slovenska Reč, 39(3), 141-146.

Predicting the Probability of Adopting Homophonic Translation: English Words in Public Media in Japan

Aimi Kuya
Ritsumeikan University, Japan

A large influx of English words into public media could cause miscommunications and misunderstandings. There has been concern among linguistics experts about whether technical/unfamiliar English terms could be adopted as they are, i.e., as a form of HOMOPHONIC TRANSLATION (HT), or as a form of LOAN TRANSLATION (LT). This paper attempts to apply the multivariate logistic regression model to this problem. The dependent variable, named [ADOPTION], is the probability that people prefer HT (for a given English word) to its LT counterpart. The independent variables include [AGE], [GENDER], and [YEAR OF SURVEY]. We examine tworesearch questions.

The first question is whether [ADOPTION] for the general public corresponds with [ADOPTION] for public sector workers. To answer this question, we revisit two previous national surveys conducted by the National Institute for Japanese Language and Linguistics (NINJAL) in 2003: (a) NINJAL (2004a) involving approximately 3,000 randomly sampled members of the general public in Japan, and (b) NINJAL (2004b) involving approximately 14,000 randomly sampled

regional government workers. The statistics reveals that [AGE] has a negative impact on [ADOPTION] in both surveys, and that [ADOPTION] for the general public often does not correspond with [ADOPTION] for public sector workers regardless of [AGE].

The second question is whether [ADOPTION] for the general public corresponds with what we call [COMPREHENSION], i.e., the probability that they comprehend the meaning of the English word in question. To answer this, several previous national surveys on comprehension of English words (e.g. Agency for Cultural Affairs 2003), involving 500 to 1,500 randomly sampled members of the general public, are added to the analysis. However, [ADOPTION] and [COMPREHENSION] are not immediately comparable because the former is based on the survey conducted in 2003 and the latter in 2002-2009. Here, the logistic regression analysis is applied 'as a means of predicting probabilities' (Yokoyama & Sanada 2007) for [COMPREHENSION] as of 2003, so that it becomes comparable to [ADOPTION]. The result shows [ADOPTION] tends to stay below [COMPREHENSION]. This suggests that the probability of adopting HT for a given English word in public media should be estimated to be lower than [COMPREHENSION].

These findings will need special attention from institutions when they put English words into practice in publicmedia.

Keywords: logistic regression, language contact, English as a global language, language change, welfare linguistics

References

Agency for Cultural Affairs (2003). *Kokugo ni Kansuru Yoron Choosa: Heisei 14* (in Japanese) [*Opinion Poll on the National Language: Heisei 1*4]. Tokyo: Agency for Cultural Affairs.

NINJAL (2004a). *Gairaigo ni Kansuru Ishiki Choosa* (in Japanese) [*The National Survey onAttitudes to Loanwords*]. Tokyo: NINJAL.

NINJAL (2004b). *Gyoosei Joohoo o Wakariyasuku Tsutaeru Kotobazukai no Kufuu ni Kansuru Ishiki Choosa: Jichitai Choosa* (in Japanese) [*The Survey on Attitudes to Word Choice for Smoother Communication: Regional Government Workers*]. Tokyo: NINAJAL.

Yokoyama, S., & Sanada, H. (2007). Multiple Logistic Regression Analysis for Formulating a Change in Language (in Japanese). *Mathematical Linguistics*, 26(3), 79-93.

Linearization of Simultaneities in Sign Language

Jiri Langer and Jan Andres
Palacký University Olomouc, Czech Republic

The signs of the Deaf can be regarded as analogous to words. On the other hand, the relationship between the respective morphemes seems to be, in view of their simultaneous character, quite sophisticated and unclear. In my talk, I will address this problem and propose a simple idea of possible linearization for the calculation of lengths to simultaneously occurring morphemes. This idea will be illustrated on an example, when testing the Menzerath-Altmann law.

Keywords: sign language; Menzerath-Altmann law; simultaneous constituents; linearization; sign language fractals

A Quantitative Analysis of Syntactic Complexity Development in German Learners' Interlanguage: A Dependency Syntactically-Annotated Corpus Study

Yushan Li
Zhejiang University, China

As a kind of natural language, studying the interlanguage development helps to understand the law of human language acquisition. In the meantime, measures of syntactic complexity are frequently introduced as an objective approach to characterize the development of interlanguage. Based on the self-built, multi-layer annotated and currently the largest L2 German longitudinal corpus in China that contains more than 1,000 texts of different genres from German learners at 10 proficiency levels, this study aims to explore the syntactic complexity development in the writings L2 German through quantitative and corpus-based methods. Within the framework of dependency grammar, the data are examined by dependency distance, the mean dependency distance and the probability distribution of dependency distance. It is found that (1) the overall

syntactic complexity in German learners' interlanguage increases fluctuatingly along with the growth of grades, and this upward trend is affected by cognitive ability; (2) language acquisition is more influenced by cognitive ability than learning time; (3) genre is also an influence factor of the syntactic complexity in interlanguage; (4) like the other nature languages, the probability distribution of dependency distance of interlanguage can be captured by the Zipf-Alekseev distribution, and the Zipfian parameters, besides the mean dependence distance of certain specific syntactic forms, can distinguish as well as predict proficiency levels. This study can contribute to determining the effects of pedagogical interventions on second language teaching.

Keywords: quantitative analysis; interlanguage development; syntactic complexity; dependency grammar; L2 German corpus

Word Length and Word Length Frequency in German Texts During the 17th-19th Century

Fei Lian
Zhejiang University, China

Since Mendenhall (1887) published his study on words in Oliver Twist, the relation between word length and word length frequency has become a key research issue in the field of quantitative linguistics, especially in the framework of synergetic linguistics in recent decades (Köhler, 2005). Various languages have been analyzed and new findings have been gained from different perspectives (Kelih, Grzybek, & Stadlober, 2003; Best, 2006; Strauss, Grzybek, & Altmann, 2007, etc.). Taking an overall view of the existing studies, however, they are mainly synchronic research and the era-specific or genre-dependent conditions have been not taken into account in many cases.

How does the word length evolve in written German within hundreds of years? Exert boundary conditions such as the genre and the time factor an impact on the word length evolution? Can the rank-frequency distribution in diachronic texts be fitted with power law function as well? To answer these questions, a multi-genre corpus with 360 German texts

originated between the 17th and 19th centuries was built based on data from the biggest annotated balanced corpus German Text Archive. This paper provides a time- and genre-based analysis of word length from both diachronic and synchronic perspectives. The diachronic study is going to explore the change of word length in terms of syllable numbers as well as its relation with word length frequency during 300 years, and look into the reasons of language evolution in cultural and social context; while the synchronic study is aimed to reveal the influence on word length exerted by the genre.

Keywords: German, word length, word length frequency, quantitative linguistics

References

Best, K.-H. (2006). Wortlängen im Deutschen. Göttinger Beiträge zur Sprachwissenschaft, 13, 23- 49.

Kelih, E., Grzybek, P., & Stadlober, E. (2003). Das Grazer Projekt zu Wortlängen(häufigkeiten). *Glottometrics* 6, 94-102.

Köhler, R. (2005). Synergetic linguistics. In R. Köhler, G. Altmann, & R. Piotrowski (Eds.), *Quantitative Linguistik. Ein internationales Handbuch* [Quantitative linguistics. An international handbook] (pp.760-774). Berlin: Walter de Gruyter.

Mendenhall, T. (1887). The characteristic curves of composition. *Science* (Supplement), 9(214), 237-249.

Strauss, U., Grzybek, P., & Altmann, G. (2006). Word length and word frequency. In P. Grzybek (Ed.), *Contributions to the science of text and language - Word length studies and related issues* (pp.15-90). Dordrecht: Springer.

Qualitative vs. Quantitative Approach to Dialect Affinity: A Case Study of Japanese Loans in Taiwan Hakka

Chihkai Lin
Tatung University, Taiwan

Introduction: This study investigates how Taiwan Hakka dialects are

classified in qualitative and quantitative approaches by looking into prosodic adaptation in Japanese loans. There are five major subdialects in Taiwan Hakka: Sixian, Hailu, Dapu, Raoping and Zhaoan (Chung 2004, 2017). To distinguish the five subdialects, there is lengthy research concerning segmental differences between the five subdialects. This paper adopts the opposite perspective, seeking the internal affinity of the five subdialects, an issue that is seldom investigated in the literature. This paper relies on phonological similarity among the five subdialects in Taiwan Hakka. In particular, this paper is interested in how prosodic features reflect phonological affinity from a corpus-based approach.

Corpus and data selection criteria: This paper follows Luo's (2013) method using Japanese loans as the primary data. The data include 121 disyllabic phrases, 113 trisyllabic phrases, and 46 quadrisyllabic phrases. As the original data are presented by Romanization, all the data are converted to tonal values. After the conversion, the contour patterns are further analyzed in phonological notation: register and locus (Shimabukuro 2007).

Results: When the number of syllables increases, the initial syllable prefers to be low. In disyllabic phrases, the register is not predicable. In trisyllabic and quadrisyllabic phrases, the register is low. As for register, Sixian, Dapu, and Zhaoan prefer penultimate locus, and Hailu and Raoping prefer final locus. In particular, Sixian disfavors final locus.

Phonological affinity of Taiwan Hakka: The affinity is discussed in conventional classification and statistical analyses. The data are first analyzed according to register and locus. The results reveal that register outweighs locus in conventional classification. As prominent in penultimate locus, Sixian, Dapu, and Zhaoan are grouped together; as prominent in final locus, Hailu and Raoping are classified as one unit. The data are also analyzed in Chi-square tests. The results suggest that Dapu and Zhaoan do not significantly differ from each other and so do Hailu and Raoping.

Combined approach to the affinity: The qualitative or quantitative approach does not provide a complete picture of the phonological affinity of the five dialects. This paper proposes a new phonological affinity of the five dialects based on a combined approach that employs both phonological notation and statistical analyses. Phonological notation suggests that there are two main branches. Sixian, Dapu and Zhaoan belong to the same branch, and Hailu and Raoping to the other branch. When the statistical analyses are also taken into account, Sixian is distant

from Dapu and Zhaoan.

Keywords: Taiwan Hakka, dialect affinity, Japanese loans, qualitative approach, quantitative approach

References

Chung, Rung-fu. 2004. *Taiwan Khjia yuyin daolun* [An Introduction to Taiwan Hakka phonetics.] Taiepi: Wunan.

Chung, Rung-fu. 2017. *Taiwan Khjia yuyin daolun*, 2nd edition [An Introduction to Taiwan Hakka phonetics, 2nd edition] Taiepi: Wunan.

Luo, Ji-li. 2013b. Gandai Hakkago no Nihongo shayou- sono onsei, onyindetki tougou [Lexical Borrowing from Japanese-A Case Study of Phonetic Changes]. *Journal of Japanese Language Education in Taiwan* 21: 301-326.

Shimabukuro, Moriyo. 2007. *The Accentual History of the Japanese and Ryukyuan Languages- A reconstruction.* London: Global Oriental.

Communicative Efficiency in Conversational English

Maja Linke and Michael Ramscar
University of Tuebingen, Germany

Does variation in the articulation of phonetic contrasts serve a communicative function? To address this question, we analyze the distributional properties of word initial phonetic contrasts in Buckeye Corpus of conversational speech (Pitt et al. 2005), framing our interpretation in terms a recently proposed, discriminative theory of human communication (Ramscar, 2019).

Ramscar (2019) observes that while aggregated probability distributions of language tokens follow power laws, such distributions do not maximize the efficiency of variable length codes. Rather, a series of analyses of the empirical distributions that are discriminated by context in language use (defined in terms of patterns of systematic invariance in word co-occurrences that empirically discriminate lexical subcategories) shows that the empirical lexical distributions that are actually encountered from moment to moment in human communication (first names in Western and Sinosphere languages, and contextually determined sub-

categories of English nouns and verbs), are geometric, thereby fitting the pattern actually predicted by information theory. These results suggest that linguistic power laws distributions simply reflect the result of mixing functional communicative distributions (Newman, 2005; e.g., while first names across the USA have a power-law distribution, by-state US names are geometrically distributed). It thus appears that empirical distributions of lexical tokens incrementally serve to discriminate between possible messages and reduce uncertainty about upcoming parts of signals in a near optimal way. This study extends this analysis to sub-lexical patterns in spoken English.

An initial analysis of the phonetic label distributions over both observed and citation forms in the corpus revealed poor fits to both power law and exponential distributions, suggesting that the aggregated distribution of the phonetic labels observed in our corpus may reflect a mixture of the underlying communicative distributions. To examine this, part of speech classes were employed to provide a first simple, objective method for contextually disaggregating individual communicative distributions from the mixed distribution of phonetic labels in our corpus (part of speech tagging is determined by patterns of invariance in word co-occurrences).

Analysis of the distribution of word initial phonetic labels discriminated by these contexts confirmed they were geometrically distributed. While word initial variance is observable in all part of speech categories, we find that the extent to which tokens vary is closely correlated to the type/token ratio and the distributional properties of the category. Importantly, despite large differences in the extent to which initial tokens deviate from the citation form, the probability distributions of tokens arising from this variance converge on nearly identical distributional properties across part ofspeech.

These results show how the variance amongst pronounced forms across part of speech classes serves to structure the information that is provided by communicative context, and thus supports the suggestion that these distributions are components of a larger, similarly structured linguistic communication system.

Keywords: communication efficiency, speech, lexical distributions, sublexical distributions

References

Newman, M. E. (2005). *Contemporary Physics*, 46(5), 323-351.
Pitt, M. A., et al (2005). S*peech Communication*, (45), 89-95.
Ramscar, M. (2019). *arXiv preprint arXiv*:1904.03991.

Cross-Modal Authorship Attribution in Russian Texts

Tatiana A. Litvinova and Olga A. Litvinova
Voronezh State Pedagogical University, Russia

The problems of authorship attribution (AA), i.e. determining the author of an anonymous text, and authorship profiling, which is the task of revealing author characteristics, are of great importance for national security, marketing, etc. Authorship attribution is a hot topic, however, a lot of problems remain unsolved. For instance, a cross-domain scenario, i.e. the one where texts with the known authorship and disputed texts come from different domains (genres, topics, mode, etc.), though very common in real world, remain difficult for classifiers [3]. Meanwhile, some aspects of cross-domain AA, e.g., cross-modal type where texts come from different modes (oral/written) are surprisingly understudied, in part due to the difficulties in obtaining appropriate corpora [5]. Meanwhile, as some studies show [1; 4], even in a topic-controlled scenario a mode shift causes dramatic changes in linguistics styles of the speakers. To address the problem of cross- modal AA, we made use of RusIdiolect Corpus which is being developed in Corpus Idiolectology Lab (https://rusidiolect.rusprofilinglab.ru/). RusIdiolect Corpus contains both multiple texts by the authors and metadata describing both the authors (gender, age, for some – personality test scores) and communication situation where the texts were produced (experimental vs. natural, genres, etc.) We performed experiments in cross-modal AA according to two scenarios: 1) using oral and written texts produced under experimental conditions, which allowed us to control for topic; 2) using oral and written texts produced under real-world conditions (thus texts differed not only in topics, but also in genre, degree of formality, etc.). Oral texts were transcribed manually by the annotators. We used several types of features (function word-based; complexity-based; part-of- speech-based; discourse-based) and both supervised and unsupervised techniques as implemented in R package Stylo [2] separately for different types of

features as well as for the whole feature set. We describe the results of our experiments with a special focus on the feature analysis. Our research contributes to the understanding of the level of stability of idiolectal features during a mode change.

Acknowledgment. The work is supported by the grant of Russian Science Foundation No 18-78-10081 "Modelling of the idiolect of a modern Russian speaker in the context of the problem of authorship attribution".

Keywords: authorship attribution, idiolect, Russian language

References

1. Aragón, G. J.: An analysis of authorship attribution: Identifying linguistic variables in oral and written discourse. (2016).
2. Eder, M., Rybicki, J. and Kestemont, M.: Stylometry with R: a package for computational text analysis. R Journal, 8(1): 107-121 (2016).
3. Kestemont, M. et al.: Overview of the Cross-Domain Authorship Attribution Task at PAN 2019. In: CLEF 2019 Labs and Workshops, Notebook Papers (2019).
4. Litvinova, T., Litvinova, O., Seredin, P.: Assessing the Level of Stability of Idiolectal Features across Modes, Topics and Time of Text Production. In: Proceedings of FRUCT 2018 (2018).
5. Stewart, J. & Winder, R. & Sabin, R.: Person Identification from Text and Speech Genre Samples. EACL 2009 – 12th Conference of the European Chapter of the Association for Computational Linguistics (2009).

Does Menzerath-Altmann Law Hold True for Translational Language: Evidence From English Literary Translated Texts

Ruimin Ma and Yue Jiang
Xi'an Jiaotong University, China

Menzerath-Altmann Law (MAL), a functional law of the relation between

language construct and its immediate constituents, is regarded as one of the fundamental linguistic laws due to its wide validity for language at various linguistic levels and in different types of registers.

Translational language, seen as different from both the source language and the original (non-translated) language due to "the nature of translated text as a mediated communicative event", is thus assumed as "the third code" or "translation universals" (Baker, 1993). Nonetheless, the validity of MAL for translational language as a natural language has scarcely been studied. The little relevant research goes to a study (Maria Roukk, 2007) that testified the validity of MAL in translated Russian and English texts, but with some limitations. Therefore, this study delved into the validity of MAL in English translational language by exploring the relationship between sentence length (in number of clauses) and clause length (in number of words), aiming to broaden the applicability of MAL and to view "the third code" from quantitative linguistics. The following two research questions are addressed.

(1) Does MAL hold true for translational language in each separate translated text and the whole corpus?

(2) If it is so, can the fitting parameters a, b of the MAL formula gauge the typological difference in sentence-clause relation between translated texts and original texts?

A Chinese-English translational corpus and a reference corpus of original texts comparable in genre, time span and text size were set up. To address the existing problematic and rough description of clauses, we didn't simply take the number of finite verbs as the number of clauses as previous studies did. Instead, we counted in both finite and infinitive ones so as to identify clauses exhaustively and reasonably and justified the method. The sentence-clause relation was fitted with $y = ax^{-b}$ by NLREG program and fitting parameters were then obtained.

Results show that the sentence-clause relation can be fitted by $y = ax^{-b}$ with the fitting R square obtained within an acceptable range for the translated texts. This corroborates that MAL is not only applicable to original language but also to translational language. Also, the average clause length for both translated and original texts range from 5 to 7, which is consistent with Miller's (1956) 7 ± 2 span for cognitive capacity of human mind. This suggests that human language production functions within basic cognitive constraints, which reciprocally proves that our method for defining clauses is viable because it reflects a reasonable range of information flow.

Statistical analysis of the difference in fitting parameters a, b answers the second question and shows that the average clause length of translational language is slightly shorter than that of the original language. Besides, the translated texts have more sentences with shorter mean sentence length (measured in clauses) than the original texts. All of these are probably because translators have to shorten the length of components to achieve cognitive balance by self-adaptation and self-regulation in the translating process, which echoes the underlying cognitive principle of MAL.

Keywords: Menzerath-Altmann Law, translational language, sentence-clause level

Free or Not So Free? On Stress Position in Russian, Slovene and Ukrainian

Ján Mačutek[1] and Emmerich Kelih[2]
[1]Comenius University in Bratislava, Slovakia, [2]Universität Wien, Austria

This contribution provides some quantitative insight into the stress position of free stress languages. In opposition to languages with fixed stress, stress in these languages can be placed on any syllable and is therefore often described as free or unpredictable (Hyman 1977). While such a statement is true in the sense that there are no deterministic rules for stress position, the stress position displays some statistical tendencies.

Based on our selection of three Slavic languages (Russian, Slovene, and Ukrainian) which are considered to be free stress languages (Krüger 2009), we will in the first step summarize some results from older quantitative stress studies in Russian (cf. the overview by Kempgen 1995, pp. 32-35, who re-analyzed data by Moiseev 1976). The data indicate that stress in Russian occurs predominantly in the second half of words, and that it gravitates towards the middle of the word. Thus, based on these findings, a tentative assumption about a general "tendency towards the center" can be made. In the second step we will present new results from two other Slavic languages with free stress, namely from Slovene and Ukrainian. In particular, we will present data on the stress position in (a)

dictionaries and texts, and (b) different parts of speech. Some (preliminary) mathematical models, which predict the stress position as a function of word length, will be developed.

Finally, our findings allow to formulate some preliminary rules for free stress (for the time being valid for the analyzed Slavic languages): it is free, but not entirely random (the distribution of stressed syllables is not uniform, but certain medium positions are preferred), and it is predictable (albeit not deterministically, but only stochastically).

Keywords: free stress, stress position, word length, Slavic languages

References

Kempgen, S. (1995). *Russische Sprachstatistik*. München: Sagner.

Hyman, L.M. (1977): On the nature of linguistic stress. In: Hyman, L.M. (ed.): *Studies in Stress and Accent*. Los Angeles: Univ. of Southern California (Southern California occasional papers in linguistics, 4), pp. 37–82.

Krüger, K. (2009). Freier Akzent (Flexion). In: Berger, T. et al. (eds.), *The Slavic Languages. An International Handbook of their History, their Structure and their Investigation. Volume 1* (pp. 86-100). Berlin, New York: de Gruyter.

Moiseev, A. (1976). Mesto slovesnogo udarenija v sovremennom russkom literaturnom jazyke. *Studia Rossica Posnaniensia* 7, 77–87.

Big-Five Personality Author Prediction in Modern Greek Essays Using Stylometric Features

George Markopoulos[1], George K. Mikros[2] and Sofia Gagiatsou[1]
[1]National and Kapodistrian University of Athens, Greece, [2]Hamad Bin Khalifa University, Qatar

Our research is focused on the automatic prediction of the author's personality features based on a corpus of essays written in Modern Greek by high-school students. The participating students have been profiled with the use of a personality questionnaire based on the model of Big-Five factor markers (Goldberg, 1992). Personality detection and prediction

under the general research framework of author profiling, i.e. the stylometric research that infers the author's metadata using textual quantitative features (Argamon et al., 2005; Mairesse et al., 2007; Rangel et al., 2015).

The feature set we employed was normalized by the text length and it was based on a combination of the most frequent part-of-speech tags, character/word bigrams and trigrams, most frequent words, hapax legomena, as well as word and sentence length. Since personality prediction represents a complex multidimensional research problem, we decided to exploit a number of different machine learning (ML) algorithms in order to find the best approach in terms of model performance. We compared seven ML methods, i.e. Support Vector Machines, Naive Bayes, Generalised Linear Model, Logistic Regression, Decision Trees, Deep Neural Networks, Random Forest and ranked them according to their cross-validated accuracy.

The best results were obtained by the Generalised Linear Model. The author's personality prediction accuracy reached 86% on Openness, 71% on Conscientiousness, 68% on Extraversion, 70% on Agreeableness and 66% on Neuroticism, according to the Big Five personality classification. The reported results present a competitive approach to the personality prediction problem and validate the use of stylometric features sets for tackling this kind of research questions.

Keywords: personality prediction, Big-Five factor markers, stylometric features, machine learning.

Keywords: personality prediction, Big-Five factor markers, stylometric features, Machine Learning

References

1. Argamon, S., Dhawle, S., Koppel, M., & Pennebaker, J. W. (2005). Lexical Predictors of Personality *Type. Proceedings of Joint Annual Meeting of the Interface and the Classification Society of North America*.
2. Mairesse, F., Walker, M. A., Mehl, M. R., & Moore, R. K. (2007). Using Linguistic Cues for the Automatic Recognition of Personality in Conversation and Text. *Journal of Artificial Intelligence Research*, 30, 457-500.
3. Rangel, F., Celli, F., Rosso, P., Potthast, M., Stein, B., & Daelemans, W. (2015). Overview of the 3rd Author Profiling Task at PAN 2015. In

L. Cappellato, N. Ferro, J. Gareth, & E. San Juan (Eds.), *CLEF 2015 Evaluation Labs and Workshop – Working Notes Papers*.
4. Goldberg, L. R. (1992). The development of markers for the Big-Five factor structure. *Psychological Assessment*, 4(1), 26-42.

Utilization of Quantitative Linguistics in Cypher Breaking

Vladimir Matlach
Palacký University Olomouc, Czech Republic

In this contribution, several quantitative and computational linguistic methods are presented in a case of deciphering historical pre World War 2 era postcards as demanded by historians.

Cryptanalysis has been known since Al-Kindi first described language-dependent letter frequency analysis in order to break substitution cyphers in 9th century. Such use of letter frequencies and their mutual relationships were common knowledge, as purported in a 17th century book by John Wilkins (2003).

However, substitution cyphers get more difficult to crack when specific tricks are applied, e.g.: leaving out spaces, using multiple substitution symbols for one letter or intentional misspellings. Encrypted texts (or *cryptograms*) are also getting more difficult to crack with decreasing text length: an encrypted text of 100 characters and less is considered *very difficult* to crack (Singh 2002, 24).

For the cryptanalyst, the situation is a lot harder when there is no certainty about the language of the cryptogram. Such was the case in the featured example where 146 unknown symbols in an encrypted message were *probably* in German.

Quantitative linguistic methods were used to identify the type of encryption by examining letter distributions, which should remain natural-like when using monoalphabetic ciphers (in contrary to polyalphabetic or homophonic ciphers which tend to distribute the letter frequencies more uniformly). Afterwards, the task of language identification based only on single letters and their n- grams was successfully carried out: Gini's coefficient (Gastwirth 1972), Entropy (Shannon 1949) and *simple type-to-token ratio* were used within a multidimensional approach to finding the nearest neighbor among

multiple language samples. This method identified only one language candidate with high certainty, leading in our specific case to the abandonment of the hypothesis that the postcard text was in German.

The next step was to find language specific words inside the cryptogram while avoiding the possibility of failing due to missing spaces, intentional or unintentional misspellings and use of obsolete words, and, at the same time, being plausible with the substitution cypher mechanics. Considering the cryptogram's short length and possible homophony, we applied a computational approach based on genetic algorithm (Goldberg and Holland 1988) and sample language corpora.

After several thousands of attempts, few complex words appeared inside the cryptogram. From this point onward, substitution guessing based on word context allowed us to successfully decypher of the rest of the historical postcard, which explains the fate of a real person.

Keywords: cryptanalysis, cypher breaking, n-grams, indices, ginis coefficient, entropy, type-to-token ratio, optimization

Keywords: cryptanalysis, cypher breaking, n-grams, indices, ginis coefficient, entropy, type-to-token ratio, genetic algorithm

References

Gastwirth, Joseph L. The Estimation Of The Lorenz Curve And Gini Index. *The Review Of Economics And Statistics*, str. 306-316, 1972.

Goldberg, D. E., & Holland, J. H. (1988). Genetic algorithms and machine learning. *Machine learning*, 3(2), 95-99.

Shannon, C. E. (1949). Communication theory of secrecy systems. *Bell system technical journal*, 28(4), 656-715.

Singh, S. (2000). *The code book: the science of secrecy from ancient Egypt to quantum cryptography*. Anchor.

Wilkins, J. (2003). *Mercury Or the Secret and Swift Messenger*, 1694. Kessinger Publishing.

SemioGraphs: Visualizing Topic Networks as Mulit-Codal Graphs

Alexander Mehler, Tolga Uslu and Daniel Baumartz
Goethe University Frankfurt am Main, Germany

In this article, we introduce *multicodal graphs* henceforth called SemioGraphs, that is, graphs whose vertices and edges are simultaneously mapped onto different systems (or codes) of types or labels. To this end, we present a technique for visualizing SemioGraphs in which information units of different provenance can be interactively browsed within the same visualization. As an application scenario for exemplifying this technique, we utilize word embedding net- works. Word embeddings have become indispensable in the field of natural language processing, as they allow for significantly improv- ing the outcome of many machine learning tasks. In our application scenario we experiment with a range of different embeddings that have been computed for a set of different training corpora. The aim is to demonstrate the informativity of our technique for visu- alizing SemioGraphs that even allow for interactively comparing different lexical neighborhoods of the same seed word. The paper describes the functional spectrum of our visualization technique for SemioGraphs. This is done by means of the SemioGraphwebsite on which users have free access to our implementation of SemioGraphs in the context of visualizing word embeddings.

For the initial creation of SemioGraphs we use *force-graph*, a framework to represent a graph data structure in a 2-dimensional canvas using a force-directed iterative layout algorithm. In order to informationally enrich this structure, we expand it in several ways.

In a SemioGraph a node can represent several numerical parameters. This is achieved, by encoding information into its height, width and transparency. Thus, unlike traditional graph visualizations, we do not use circles to represent nodes, but ellipses, which can vary in height and width (see Figure 1). The color of a node encodes its membership to certain groups or classes. In addition, we provide a special mode that transforms the nodes into pie charts and can thus encode multiple classifications and their membership values. This even allows us to display class

membership distributions at node level. Each node can have a specific label, which is displayed next to the node. The line color and thickness of a node can also be parameterized to encode certain information.

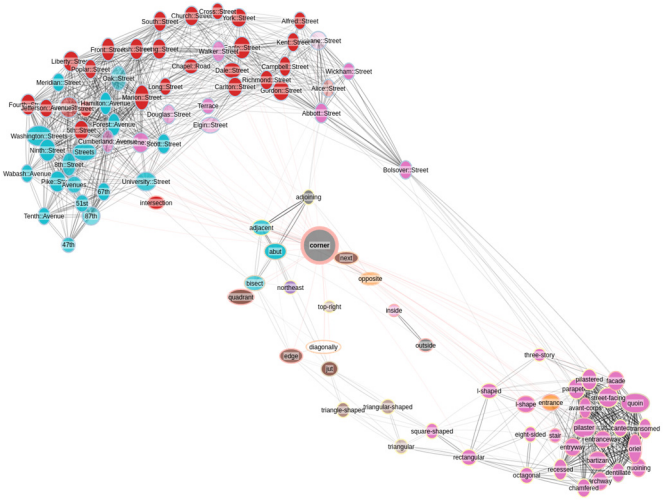


Figure 1: Example of an SemioGraph.

As far as the edges of a SemioGraph are concerned, we also encode and display various information units. Here we refer to the thickness and transparency of the edges to display numerical infor- mation. Furthermore, the color of the edges can be used to represent different edge classes. In addition, the edges can be orientated to obtain arrows in a directional graph.

Various interaction options are also available to further refine the graph analysis. One can hover a node or an edge to perform a function like getting an tooltip for more information. With various sliders, it is possible to filter a SemioGraph according to user criteria even after its creation.

With this functional spectrum it is possible to visualize, analyze and interact with highly complex (directed and undirected) graphs.

Keywords: word embeddings, visualization, visualization techniques, semantics

1 SCOPE OF THE SOFTWARE DEMONSTRATION
A beta version of SemioGraph's web application can be found at:
http://semiograph.texttechnologylab.org

A Mouth Full of Names: Anthroponymy-Based Text Concentration in the Czech Novels Featuring the Character of Švejk

Michal Místecký, Jaroslav David and Jana Davidová Glogarová
University of Ostrava, Czech Republic

The goal of the contribution is twofold – first, to present, on a larger scale, the proprial thematic concentration of text, a modified calculation of thematic words with a focus on proper names (cf. Čech – Kubát, 2016; Místecký, 2019), and second, to apply the method on various texts featuring the character of Švejk. Švejk is a grotesque, slow-witted, and talkative WWI soldier, which appeared, for the first time, in *Osudy dobrého vojáka Švejka za světové války* ("The Good Soldier Švejk"), a four-volume novel by Jaroslav Hašek. The volumes were published in 1921, 1922, 1922, and 1923, respectively. As Hašek's book attracted both domestic and foreign readers, during the Second World War, a new, two-volume novel with the same protagonist was anonymously published (1941, 1945). Since the original text is a patchwork of Švejk's never-ending anecdotes with a prominent use of personal names, the contribution will compare the first four books with the fifth and the sixth ones on the basis of anthroponyms. The counts of the anthroponymy-based thematic concentration will be carried out for individual chapters, and the averaged values for the books will be compared on the grounds of statistical tests. The resulting scatterplots will be able to show the differences among the books, and will possibly confirm the idiosyncratic position of the last ones, which were not authored by Jaroslav Hašek. The outcomes of the research may be of help for literary scholars centred on the Švejk texts, but they may also give impetus to onomastics and the study of authorship attribution.

Keywords: Švejk, Czech literature, proprial thematic concentration,

thematic concentration, stylometry, authorship attribution, personal names, onomastics

References

Čech, Radek – Kubát, Miroslav (2016): Text length and the thematic concentration of text. *Mathematical Linguistics*, 2(1), 5–13.

Místecký, Michal (2019): Využití rankové frekvenční distribuce při výzkumu antroponym v literárních dílech [Using the Rank-Frequency Distribution in Analysing Personal Names in Fiction]. *Linguistica Brunensia*, 67(1), 27–38. In print.

The Menzerath-Altmann Law in Syntactic Relations of Chinese Language Based on the Universal Dependencies (UD)

Tereza Motalova
Palacky University Olomouc, Czech Republic

The Menzerath-Altmann law (MAL) describes the relationship of two language units – construct and constituent – based on the inverse proportionality of their lengths. It says that the mean size of constituents is a function of the size of the construct. The MAL has been applied to various languages and Chinese is no exception. Researchers tested the MAL on various language units of both spoken and written Chinese, e.g. Hou et al. (2017) and Chen (2018) tested sentences, clauses and words where commas or semicolons defined the clause.

This study explores whether or not the MAL is valid for respective language units of contemporary written Chinese, such as

(character) – word – direct dependent element of main the predicate – sentence.

The study primarily tries to verify the validation of the MAL in the syntactic dependency structure of written Chinese. Particular emphasis is placed on the language unit called "d*irect dependent element of the main predicate*". The length of this unit is measured by the number of core arguments and non- core dependents which are in direct dependent relation

to the main predicate of a sentence and which are measured by the number of words in it. The determination of core arguments and non-core dependents will be based on the annotation provided by Universal Dependencies for Mandarin Chinese. The study will examine how this language unit behaves as a construct and constituent in relation to language units that are immediately adjacent to it. In terms of the MAL – we assume the following: the longer a sentence, the shorter the *direct dependent element of its main predicate* (measured in its constituents, i.e. words); furthermore, the longer the *direct dependent element of the main predicate*, the shorter its words (measured in its constituents, i.e. characters.). The approach of using dependency grammar was already applied to Czech, and the research yielded that the MAL is valid in syntactic dependency structure for this language (Mačutek et al, 2018). The contribution will present both the applied methodology and the results obtained.

Keywords: Menzerath-Altmann Law, syntax, universal dependencies (UD), Chinese language

References

Hou, Renkui; Huang, Chu-Ren; Do, Hue San; Liu, Hongchao (2017). A Study on Correlation between Chinese Sentence and Constituting Clauses Based on the Menzerath-Altmann Law. *Journal of Quantitative Linguistics*, 24:4, 350-366.
DOI: 10.1080/09296174.2017.1314411

Chen, Heng (2018). Testing the Menzerath-Altmann Law in the Sentence Level of Written Chinese. *Open Access Library Journal*, 5: e4747. https://doi.org/10.4236/oalib.1104747

Mačutek, Ján; Čech, Radek; Milička, Jiří (2017). Menzerath-Altmann Law in Syntactic Dependency Structure. In Montemagni, S., Nivre, J. (eds.). *Proceedings of the Fourth International Conference on Dependency Linguistics (Depling 2017)*. Linköping Electronic Conference Proceedings No. 139, 100-107.

Text as Time Series, Time Series as Function: Information Theory and Signal Processing in Text Analysis

Cosimo Palma
Italy

This paper represents the cornerstone of a broader research focus, which sets as its long-term goal the **synaesthetic translation**, i.e. the consistent and sound conversion of any stream of signs into a different one, for instance from text to music, conveying same meanings, feelings and nuances.

The semantic consistency of the proceeding will be further proved by composing the logical abstractions underlying the two realizations and by demonstrating that they respect the l**ogic amalgamation property**.

The first step to accomplish in order to correctly build up this pipeline (the only one fully addressed in the present article) is the reduction of the text into a stream of **self-information values**, which could either refer to clauses or to paragraphs, labelled after Fillmore's and Propp's models, and obtained by measurements based on the **Kullback-Leibler divergence**.

The text, deployed as **time series** on cartesian axes, i.e. rendered as a function interpreted in turn as signal, will be decomposed in its constituent frequencies through Fast Fourier transform in order to detect causal relationships between clauses distant from each other.

For this purpose **Recurrent Neural Networks** (RNN) will be utilised as disambiguation tool.

The final discussion will open to the possibility of replacing the above mentioned self-information values with empirically collected data, such as heart rate measurements collected over a large sample of readers.

Keywords: Information theory, Self-information, Signal processing, Fourier Transform, Multifractal analysis, Time series analysis, Kullback-Leibler divergence, Recurrent Neural Networks

References

Hamilton, J. D. 1994. *Time series analysis*, volume 2. Princeton university press Princeton

Palus, M.; Vejmelka, M.; and Bhattacharya, J. 2007. *Causality detection*

based on information-theoretic approaches in time series analysis. Physics Reports 441(1): 1-46.
Lutkepohl, H. 2005. *New introduction to multiple time series analysis*. Springer Science & Business Media.
Matsubara, Y.; Sakurai, Y.; Prakash, B. A.; Li, L.; and Faloutsos, C. 2012. *Rise and fall patterns of information diffusion: model and implications*. In KDD, 6-14.
Mirza, P., and Tonelli, S. 2014. *An analysis of causality between events and its relation to temporal information*. In Proceedings of COLING, 2097-2106.
Mirza, P. 2014. *Extracting temporal and causal relations between events*. ACL 2014 10.
Reagan A.J., Mitchell L., Kiley D., Danforth C. M., Dodds P. S., *The emotional arcs of stories are dominated by six basic shapes* , EPJ Data Science, 5. Jg., Nr. 1, 2016, S. 31. Chambers N., Jurafsky D., Unsupervised Learning of Narrative Event Chains. Proceedings of ACL-08: HLT, 2008, S. 789-797.

Statistical Tools, Automatic Taxonomies, and Topic Modelling in the Study of Self-Promotional Web Texts of Polish Universities

Adam Pawłowski[1] and Tomasz Walkowiak[2]
[1]University of Wrocław, Poland, [2]Wrocław University of Technology, Poland

The websites of higher education institutions fulfill informational and promotional functions. They are designed to create a desired image of a university in the consciousness of the audience using various means: graphic, linguistic, and sometimes also audiovisual. They are addressed to potential stakeholders, in particular candidates for studies and business representatives. The importance of these texts results from the fact that the recognition and identity of a university is one of the measures of its competitiveness and attractiveness.

The presentation deals with the content in the form of continuous texts, lists or bulletins, which are placed on the universities' website in the "Mission", "Vision" or "About us" tabs. In justified cases, we also include

other fields (e.g. "History", "Our University"). The corpus of promotional texts provided by universities was also analysed using the methods of statistics, topic modelling and automatic taxonomy.

The automatic taxonomy methods applied in this study use algorithms trained by machine learning on large corpora of general language. They make it possible to classify huge sets of texts, reducing them to the form of dendrograms or other infographics in 2D space. The method of topic modelling allows to extract the content of texts and group them as semantic clusters (topics). This results in a very synthetic, but quite accurate representation of the contents of large data sets. Statistical methods allow to create quantitative characteristics of the corpus (descriptive statistics, univariate analysis), to measure its lexical richness, to draw histograms of the frequency of vocabulary, as well as to describe its stereotypicality, often occurring in functional texts belonging to the same genre.

The empirical part of the work consists of four parts. In the first one we present the quantitative characteristics of the corpus, including descriptive statistics, frequency histograms and statistical distributions of vocabulary. The second one is devoted to the extraction of repeatable word n-grams and their quantitative evaluation. In the third part we present automatic taxonomies of texts and verify the hypothesis that the taxonomy of texts representing individual universities should reproduce (with a reasonable accuracy) the classification based on explicit profiles of these universities. In the fourth part we use the method of topic modelling to reconstruct the main content of the promotional message of Polish universities.

Keywords: quantitative methods, university mission, topic modelling, automatic taxonomy

Quantitative Characteristics of Phonological Words (Stress Units) in Czech

Kateřina Pelegrinová
University of Ostrava, Czech Republic

A pilot study investigating the behavior of phonological words (also called prosodic words or stress units) in Czech will be presented. First, we examine several quantitative characteristics of phonological words and, second, we compare these results with the same quantitative characteristics of words in written texts. Specifically, the size of inventory, rank-frequency distribution, distribution of length, and entropy is analyzed in this study.

Both the phonological word and the word belong to the same linguistic level. The former is a phonological unit of the spoken language. In this analysis, we adopt the formal rules for the segmentation of text into phonological words developed by Palková (2004), who describes several segmentation principles (including, for example, the position of the syllable in the clause and on the number of syllables in the word). The word is a morphosyntactic unit of the written language defined as a chain of characters between spaces.

Studies analyzing the behavior of phonological words, as opposed to those analyzing the behavior of words, are rare in quantitative linguistics. Some preliminary assumptions follow from the very nature of the phonological word. For particular characteristics, we have the following assumptions: First, the inventory of phonological words will be larger than the inventory of words because the former can consist of combinations of the latter. Second, the rank-frequency distribution of phonological words will fit one of the Zipfian distributions. In comparison to the rank-frequency of words, we expect that more hapax legomena will be observed. The third characteristic is the distribution of lengths of phonological words. We assume that it follows some modification or generalization of the Poisson distribution, in analogy to the distribution of length of word. We expect the phonological word to be longer, because it can consist of more than one word. The last, characteristic to be analyzed is entropy, which describes how diversified the text is with respect to the occurrence of different phonological words.

keywords: phonological word, stress unit, length distribution, rank-frequency distribution, entropy, word

References

Palková, Z. (2004). The Set of Phonetic Rules as a Basis for the Prosodic Component of an Automatic TTS Synthesis in Czech. *Phonetica Pragensia* 10, 2004, 33–46.

Grzybek, P. (ed.) (2005). *Contributions to the Science of Language. Word Length Studies and Related Issues*. Kluwer, Dordrecht.

Popescu, I-I et al. (2009). *Word Frequency Studies*. Berlin, New York: de Gruyter.

Těšitelová, M. (1985). *Kvantitativní charakteristiky současné češtiny*. Praha: Academia.

Explorative Study on the Menzerath-Altmann Law Regarding Style, Text Length, and Distributions of Data Points

Haruko Sanada
Rissho University, Japan

Study aim and hypothesis:
This study used text from Japanese newspapers to conduct an empirical investigation of the Menzerath-Altmann Law (MAL) (Altmann 1980, Menzerath 1954). Previous studies on this law by Köhler (1984) and Cramer (2005) interpreted parameters of the MAL formula. The present study empirically investigated the MAL for two data sets consisting of newspaper text, not used in former studies, which addressed the same topics but had different readerships (i.e. adults and children). Our hypothesis was that the differences in newspaper readerships could be expressed through parameters of the MAL formula.
Data:
Two data sets with paired topics were prepared from newspaper articles written for either adults or elementary school children. Both were published by the same company.
Methods of analyses:

We investigated four relationships with the following three linguistic levels for the two newspapers:

(1) Text length measured in sentences and average sentence length in clauses;
(2) Sentence length in clauses (SL) and average clause length in morphemes (CL); and
(3) Clause length in morphemes and average morpheme length in the number of characters using *kanji*, or average morpheme length in the number of characters when *kanji* is converted to *hiragana* (phonetic Japanese).

Results and conclusions:

Articles in the two data sets have a different style, with the lengths of text, sentences, and clauses for adults almost two times greater than for children. However, relationship tendencies for (1) to (3) are similar to each other in both data sets. The relationship tendencies are individualised according to linguistic levels, as also posited by Köhler (1980) and Cramer (2005).

The present study also observed how a scatter of data points affects distribution of the total number of data points, sums, or averages of linguistic properties as dependent variables of the MAL. As such, the following assumptions can be made: (1) a distribution of the data points (DP) is systematically related to a distribution of the sum of CL; (2) a distribution of DP seems to be related to averages of CL if DP is a function of SL; and (3) averages of CL do not directly depend on text length because the average is determined by a ratio of DP and a sum of CL. Although these results are explorative, this is the first research conducted on the number of data points as a function of linguistic properties for the MAL.

Keywords: Menzerath-Altmann Law, frequency, newspaper, Japanese, Synergetic Linguistics.

References

Altmann, Gabriel. (1980). Prolegomena to Menzerath's law. In: Grotjahn, Rüdiger. (Ed.) *Glottometrika*, 2 (pp.1-10). Bochum: Brockmeyer.

Cramer, Irene M. (2005). The Parameters of the Altmann-Menzerath Law. *Journal of Quantitative Linguistics* 12 (1), 41-52.

Köhler, Reinhard. (1984). Zur Interpretation des Menzerathschen Gesetzes. In: Boy, Joachim.; Köhler, Reinhard. (Eds.), *Glottometrika* 6 (pp.177-183). Bochum: Brockmeyer.

Menzerath, Paul (1954). *Die Architektonik des deutschen Wortschatzes*.

Bonn: Dümmler.

Designing a Corpus-Driven Resource for Teaching Prepositional Patterns to Advanced English Learners

Denisa Šebestová
Charles University Prague, Czech Republic

This study aims to design corpus-informed teaching materials for advanced English students, reflecting differences in native as opposed to non-native phraseologies. We are building on previous studies suggesting that even advanced L2 learners tend to use a limited repertory of phraseological sequences in ways which differ considerably from native usage (e.g. Granger 2017; or Hasselgård 2017, referring back to Hasselgren 1994, terms these 'phraseological teddy bears'). This presents a potential issue as limited phraseological choices hinders the students´ language production in terms of accuracy.

We focus on phraseological patterning involving prepositions. As pointed out by Hunston (2008), focus on function words in corpus analyses can be beneficial as phraseological patterns containing e.g. prepositions are involved in text structuring and can help point out larger-scale patterning within discourse, highly relevant to advanced students.

We employ data from the BNC (100 mw) complementing these by a COCA sample (3.6 mw). First, a list of the 10 most frequent prepositions is compiled for each corpus. For each, we extract 3-5-grams containing the preposition in any slot, using Engrammer (Milička 2019). Engrammer enables searches for sequences of different lengths at once, comparing collocation strength between grams using various collocation measures. It reveals alternations in the selected slot, detecting variants where applicable. We retrieve a list of patterns for each preposition and examine it in context. We adopt a constructional approach, identifying its textual functions and semantic features. This functional-semantic analysis results in the compilation of a list of prepositional constructions. These are compared between British and American English to control for potential dialectological variance.

Preliminary results suggest that some prepositional constructions have particular semantic prosodies. E.g.as a result of correlates with negative

consequences (collapse, shortage, breakdown), while in the face of presents the actor in a positive light.

Keywords: corpus-driven phraseology, n-grams, prepositions, advanced learners of English, corpus-driven TEFL resources, collocation

References

Granger, S.: Academic phraseology. A key ingredient in successful L2 academic literacy. In: Fjeld, H., Henriksen, J. (eds.) *Academic Language in a Nordic Setting–Linguistic and Educational Perspectives*. Oslo Studies in Language, vol.9, no.3, pp.9–27. Olsen&Prentice (2017).

Hasselgård, H.: Phraseological teddy bears: frequent lexical bundles in academic writing by Norwegian learners and native speakers of English. In: Mahlberg, M., Wiegand, V. (eds) *Corpus Linguistics, Context and Culture*. De Gruyter Mouton, Berlin (forthcoming).

Hasselgren, A.: Lexical teddy bears and advanced learners: a study into the ways Norwegian students cope with English vocabulary. Int. J. *Appl. Linguist*. 4, 237-259 (1994).

Hunston, S.: Starting with the small words. In: Römer, U., Schulze, R. (eds.) *Patterns, Meaningful Units and Specialized Discourses. Int. J. Corpus Linguist*. 13(3), 271-295 (2008).

Sources

The British National Corpus, version 3 (BNC XML Edition). 2007. Distributed by Bodleian Libraries, University of Oxford, on behalf of the BNC Consortium. URL: http://www.natcorp.ox.ac.uk/

Davies, Mark. (2008-) *The Corpus of Contemporary American English* (COCA): 560 million words, 1990- present. Available online at https://corpus.byu.edu/coca/.

Milička, J.: *Engrammer*. Software, available from http://milicka.cz/en/engrammer/.

Modeling Lexical and Syntactic Variety of Russian Prose Language in 1900-1930

Tatiana Y. Sherstinova
St. Petersburg State University, Russia

The current research is made within the project titled "The Russian language on the edge of radical historical changes: the study of language and style in prerevolutionary, revolutionary and post-revolutionary artistic prose by the methods of mathematical and computer linguistics (a corpus-based research on Russian short stories)" supported by the Russian Foundation for Basic Research (# 17-29-09173). The initiator and the head of this project was the outstandi ng Russian linguist, the founder of Russian School of stylometrics, Prof. Gregory Martynenko [1; 2]. The aim of the project is to study and model the system of linguistic and stylistic variables in dynamics during the first three decades of the 20th century, and to identify and describe the changes that occurred in the Russian language in the chain of dramatic events of the World War I (1914–1918), the February and October Revolutions of 1917, and the Russian Civil War (1917–1922).

To achieve these tasks, the unique literary corpus of Russian short stories is being created. The choice of the genre is justified by the fact that the short story is the most common literary genre, which allows comparative studies of a large number of texts and writers. When creating the corpus, we strive to include in the corpus texts by the maximum number of prose writers who wrote in 1900-1930 [3]. Such approach provides an opportunity for modeling the literary systems [4] of the historical periods in concern (prerevolutionary, revolutionary and post-revolutionary) and analyze lexical and syntactic variety both in synchrony and diachrony.

In order to perform quantitative analysis, the subcorpus of short stories by 300 different writers is being annotated on lexical and syntactic levels. As a result we obtain the following quantitative data: frequency word lists, POS-distribution, data on word formation, typical syntactic constructions, etc. for each text, each historical period and subcorpus in the whole. For example, frequency word lists are analyzed by means of the following parameters: dictionary size, hapax, entropy, rank average [1], mode, median, upper and lower quartiles, quartile deviation, frequency of

the first word in the list, concentration index, etc. Having obtained the given quantitative data, it become possible to model lexical and syntactic variety of literary language within each historical period and to reveal those linguistic parameters which changed the most from 1900 to 1930.

Keywords: stylometrics, quantitative linguistics, literary corpus, literary system, Russian language, Russian prose, diachronic language changes

References

[1] Martynenko G.Ya. Foundations of Stylometrics [Osnovy stilemetrii]. Leningrad State University, Leningrad (1988).

[2 Martynenko G.Ya. The methods of mathematical linguistics in stylistic studies [Metody matematicheskoj lingvistiki v stilistichestikh issledovaniyakh]. Nestor-Istoriya, St. Petersburg (2019).

[3] Martynenko, G., Sherstinova, T., Popova, T., Melnik, A., Zamirayilova, E.: On the principles of the Creation of the Russian Short Story Corpus of the First Third of the 20th Century

[O principakh sozdaniya korpusa russkogo rasskaza pervoj treti XX veka]. In: Proc. of TEL Conference on Computational Linguistics-2018, vol. 1, 180-198, Kazan (2019).

[4] Tynyanov, Yu.N.: Archaists and Innovators [Arkhaisty i novatory]. Priboj, Leningrad (1929).

Comparing the Effectiveness of SVM and Deep Learning in Stylometry: The Case of The Dream of the Red Chamber

Jianjun Shi
Shanghai International Studies University, China

Applying deep learning to stylometric analysis is regarded as a development in computational linguistics. So far, however, such study is relatively rare. With deep learning, can we set aside previous knowledge on traditional machine learning and stylistic features? Focusing on this question, this paper aims at depicting the differences between LSTM

model and SVM model in stylometric analysis. In this paper, we collect text of "*A Dream of Red Mansions*"(*the Cheng-Gao version, containing 120 chapters in total*) for a specialized corpus, trying to investigate the application value of traditional machine learning and that of deep learning. By analyzing frequency of 44 function words in classical Chinese as stylometric feature, Shi (2011) made a classification study of "*A Dream of Red Mansions*" by means of SVM. The results of this study are as follows:

First, after implementing SVM model, a discernible gap is found between chapter 80 and chapter 81 of the book, which confirms a popular hypothesis in the academic circles that the first 80 chapters and the last 40 chapters of the book were written by two different authors. Another finding worth mentioning is that for some chapters whose authorship are disputed, the result is consistent with previous studies, e.g. the authorship of Chapter 67 is attributed to the author of Chapter 81-120.

Second, despite that the overall accuracy rate of authorship attribution increases as more samples are included, applying LSTM model fails to detect style variation in this novel, which makes it unable to identify its authorship accurately. Usually, LSTM model would perform better if large volumes of data were provided—which is nearly impossible in the case of ancient classics where resources are finite.

Therefore, we conclude that SVM model has an edge over LSTM model in authorship identification of ancient classics.

Keywords: SVM, LSTM, authorship attribution, stylometry, The Dream of the Red Chamber

References

Chen, D. (1987). Authorship attribution of Chapter 81-120 in "A Dream of Red Mansions": evidence from computational linguistics. *Studies on "A Dream of Red Mansions"*, (1), 293-318.

Chen, D. (1988). Untenable: a revisit to "New introduction to the disputed 'A Dream of Red Mansions'", *Journal of East China Normal University(Humanities and Social Sciences)*, (1), 3-13.

Li, X. (1987). New introduction to the disputed "A Dream of Red Mansions". *Fudan Journal(Social Sciences Edition)*, (5), 3-16.

Shi, J. (2011). Authorship attribution of "A Dream of Red Mansions" with support vector machines. *Studies on "A Dream of Red Mansions"*, (5), 35-52.

Ren, Z., Xu, H., Feng, S., Zhou, H., & Shi, J. (2017). Sequence labeling Chinese word segmentation method based on LSTM networks. *Application Research of Computers*, 34(5), 1321-1324.

Hypotheses on Morphological Complexity

Petra Steiner
Friedrich-Schiller-Universität Jena, Germany

In this paper, two hypotheses and their functional interplay in word formation are investigated. These are a. Hawkins Principle of Early Immediate Constituents and b. the impact of the polylexy of lexemes on their word-formation activity. The second hypothesis has been corroborated for free and bound constituents of words (Krott, 2004; Steiner, 1995). The first hypothesis by Hawkins (1994) has been widely examined for syntactic levels (Hoffmann, 1999; Hawkins, 1999; Köhler, 2012, p. 138ff.), however, its transfer to morphology is a new step.

German is a language with a rich derivational and compositional morphology whose combination produces a high amount of deep-level constructions. Therefore, it yields an interesting foundation for testing hypotheses whose units are defined on grounds of the lengths and depths of complex word constructions and their constituents. As data, we use the German morphological trees database built by the tools of Steiner (2017). It combines the analyses of the German part of the CELEX database (Baayen et al., 1995) which were exploited by Steiner and Ruppenhofer (2018), and the annotated compounds from the GermaNet database (Henrich and Hinrichs, 2011). Figure 1 presents a typical example for which the hypothesis of the relationship between position and depth/complexity does not hold. Only 9 percent of the sample of 54,000 complex morphological analyses show strong tendencies towards right branching. The investigation indicates towards multiple factors for the construction of complex words.

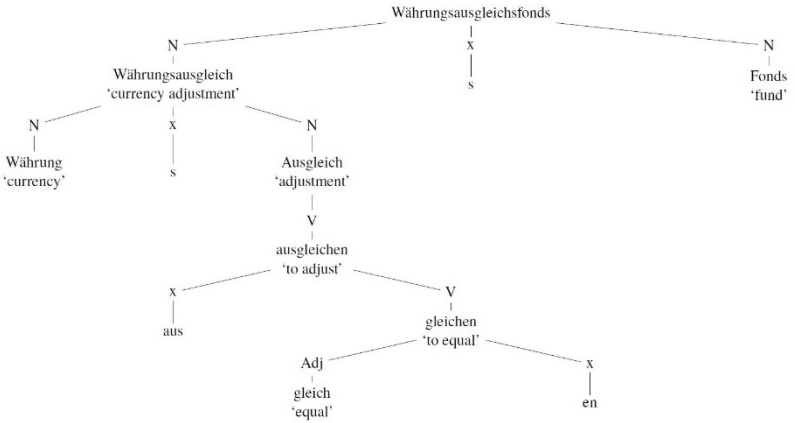


Figure 1: Morphological analysis of Währungsausgleichsfonds 'currency adjustment fund'

Keywords: morphological complexity, Principle of Early Immediate Constituents, polylexy, compounding, derivation, German morphology

References

Harald Baayen, Richard Piepenbrock, and Léon Gulikers. 1995. The CELEX lexical database (CD-ROM).

John A. Hawkins. 1994. *A performance theory of order and constituency*. Cambridge studies in linguistics. Cambridge Univ. Press, Cambridge u.a. Literaturverz. S. 470-482.

John A. Hawkins. 1999. The relative order of prepositional phrases in English: Going beyond MannerPlaceTime. *Language Variation and Change* 11(3):231266. https://doi.org/10.1017/S0954394599113012.

Verena Henrich and Erhard Hinrichs. 2011. Determining Immediate Constituents of Compounds in GermaNet. In *Proceedings of the International Conference Recent Advances in Natural Language Processing, Hissar, Bulgaria, 2011*. Association for Computational Linguistics, pages 420-426. http://www.aclweb.org/anthology/R11-1058.

Christiane Hoffmann. 1999. Word Order and the Principle of "Early

Immediate Constituents" (EIC). Journal of *Quantitative Linguistics* 6(2): 108-116. https://doi.org/10.1076/jqul.6.2.108.4133.

Andrea Krott. 2004. Ein funktionalanalytisches Modell der Wortbildung [A functional analytical model of word formation]. In Reinhard Köhler, editor, *Korpuslinguistische Untersuchungen zur Quantitativen und Systemtheoretischen Linguistik [Corpus-linguistic Investigations of Quantitative and System-theoretical Linguistics]*, Elektronische Hochschulschriften an der Universität Trier, Trier, pages 75-126. http://ubt.opus.hbz- nrw.de/volltexte/2004/279/pdf/04_krott.pdf.

Reinhard Köhler. 2012. *Quantitative Syntax Analysis*. Quantitative linguistics. de Gruyter Mouton, Berlin/Boston.

Petra Steiner. 1995. Effects of Polylexy on Compounding. *Journal of Quantitative Linguistics* 2(2):133-140. https://doi.org/10.1080/09296179508590042.

Petra Steiner. 2017. Merging the Trees-Building a Morphological Treebank for German from Two Resources. In *Proceedings of the 16th International Workshop on Treebanks and Linguistic Theories, January 23-24, 2018, Prague, Czech Republic*. pages 146-160. https://aclweb.org/anthology/W17-7619.

Petra Steiner and Josef Ruppenhofer. 2018. Building a Morphological Treebank for German from a Linguistic Database. In *Proceedings of the Eleventh International Conference on Language Resources and Evaluation (LREC 2018)*, Miyazaki, Japan. European Language Resources Association (ELRA). https://www.aclweb.org/anthology/L18-1613.

Word Frequency Distributions: A Comprehensive Bayesian Approach

Trudie Strauss[1], Damián Blasi[2], Sean van der Merwe[1] and Michael von Maltitz[1]

[1]University of the Free State, South Africa, [2]Harvard University, United States of America

Many parametric models and statistical laws have been suggested to model word frequency dis- tributions, the most famous being Zipf's law (the inverse proportionality between the rank of a word and its frequency).

Zipf's law has been generalised and extended to model the relationship between the frequency spectrum and its groups, and also to include more parameters. As such, there are currently several proposed models of Zipf-like distributions that seem to describe word frequency distributions relatively well. These models, with their respective advantages and disadvantages in particular settings, have been shown to hold in different contexts: particular ranges of sample size, certain genres of text, etc.

In this study, we implement a Bayesian approach by fitting a class of general models based on Zipfian distributions that encompasses these current propositions for word frequency distributions. We further identify several linguistically relevant features that may be calculated from word frequency distributions of languages and express natural language as a multidimensional array of these measures. Ultimately, through a Bayesian analysis, we compare the theoretical model with empirical data based on 200 languages from 22 language families, to determine the space that these measures occupy within the posterior distributions. Comparing the patterns and distributions of empirical natural languages to what may be expected from these theoretical models yields interesting results about the nature of word frequency distributions. From this, we are able to determine how some of these linguistically relevant measures behave in a universal manner across languages and we investigate the linguistic implication.

Keywords: word frequency distributions, LNRE models, Zipf's law, Bayesian modeling

Stylometric Features of PhD Theses: A Quantitative Analysis of Text Activity

Shuyi Sun[1] and Wei Xiao[2]
[1]The University of Queensland, Australia, [2]Chongqing University, China

Stylometry, also known as computational stylistics, studies textual styles and writing habits based on the application of quantitative methodology to linguistic features (Liu & Xiao, 2018). As a result of self-regulation, style reveals the internal rules of language (Hawkes, 2003; Melka & Místecký, 2019), which are in turn reflected in stylometric indicators. In this study, the main attention will be paid to the stylometric dyad, namely

activity and descriptivity, which is based on the verb-adjective ratio Q (Kubát & Čech, 2016; Zörnig et al., 2015). By conducting a quantitative analysis on text activity of PhD theses, we aim to find out:

(1) Among the three major disciplines (natural sciences, social sciences and humanities), what oscillations of text activity are displayed in each discipline?
(2) Is text activity an effective indicator for discriminating the three major disciplines?

According to our preliminary analysis, text activity Q increases steadily within the first 10,000 tokens, whereas this indicator tends to remain in the interval <0.65-0.7> with tokens counted to more than 15,000, which might reveal the "active" or "strongly active" regularity of PhD theses (Zörnig et al., 2015, p. 4). Besides, the overall activity of humanities is the highest, although the three major disciplines do not differ significantly in text activity. Based on the findings, we might draw a conclusion that PhD theses, as an academic genre and a systematically organized entity, share some universal patterns with respect to text activity. Nevertheless, the outcome might as well reveal the writing conventions of individual disciplinary branch, such as the more narrative nature of humanities compared with social sciences and natural sciences, which could quantitatively support or reject the previous interpretation of disciplinary knowledge construction (e.g., Hyland & Bondi, 2006). On the whole, this study is expected to have implications for both PhD theses instructions and a wider application
of quantitative methods.

Keywords: stylometry, activity, quantitative analysis, PhD theses

References

Hawkes, T. (2003). *Structuralism and semiotics* (2nd ed.). New York, NY: Routledge.

Hyland, K., & Bondi, M. (Eds.). (2006). *Academic discourse across disciplines*. Frankfurt, Germany: Peter Lang.

Kubát, M., & Čech, R. (2016). Quantitative analysis of US presidential inaugural addresses. *Glottometrics*, 34, 14-27.

Liu, Y., & Xiao, T. (2018). A stylistic analysis for Gu Long's Kung Fu novels. *Journal of Quantitative Linguistics*, 1-30.

Melka, T. S., & Místecký, M. (2019). On stylometric features of H. Beam Piper's *Omnilingual*. *Journal of Quantitative Linguistics*, 1-40.

Zörnig, P., et al. G. (2015). *Descriptiveness, activity and nominality in formalized text sequences*. Lüdenscheid, Germany: RAM-Verlag.

Lexical Diversity in Creative Writing of L2: A Quantitative Approach

Ioanna Tyrou
National and Kapodistrian University of Athens, Greece

Evaluating L2 creative writing in language classrooms is not only debated but it is also an essential part of language teaching. The aim of our research is to create a quantitative model in order to evaluate L2 creative writing in a systematic way. For this reason, we compiled a corpus based on short texts written by undergraduate Greek students learning Italian as L2. The corpus collection took place at the Department of Italian Language and Literature at the University of Athens.

The same group of these undergraduate students that wrote these essays was "transformed" to judges-raters in order to peer-evaluate the essays produced during this research. Their task was to evaluate the essays using an analytic rubric (seven criteria in the scale from 1 to 4), that captures a broad range of creative writing evaluation criteria. More specifically, the criteria used are: a) quantity of ideas, b) types of ideas, c) rarity of ideas, d) image, e) voice f) characterization, g) narration. (Mozaffari, 2013; Vaezi & Rezaei, 2019). We then calculated the median of the assessment scores for each one of the seven assessment sub-categories in order to have a measure of the central tendency of the evaluators' scores. We also processed all essays using the QUITA tool (Kubát, Matlach, Čech, 2014) for measuring a broad number of vocabulary differentiation indices and creative a broad quantitative vocabulary profile for each text. In the last step of the experimental procedure we fitted 8 different linear regression models, each for a different assessment criterion and one for the median of the total scores in the evaluation task. The models were finally evaluated using their R^2 values using them as an indication of their prediction efficiency using as predictor variables the vocabulary differentiation indices calculated for each document.

The results provide new evidence that creative writing can be

systematically evaluated using quantitative vocabulary indices. Most linear regression models exhibited excellent fit with specific models approaching R^2 scores near 1. More specifically, the R^2 values for each assessment criterion are: quantity of ideas 0.65, types of ideas 0.96, rarity of ideas 0.9, image 0.7, voice 0.56, characterization 0.29, narration 0.44, and the total evaluation reached 0.6. The above results confirm our initial hypothesis that creative writing can be fairly successfully evaluated using quantitative text profiles. Moreover, our research results can be further exploited in the development of automatic essay scoring systems and automated essay content analysis.

Keywords: creative writing, second language, quantitative vocabulary indices

References

Kubát, M., Matlach, V., & Čech, R. (2014). *QUITA: Quantitative Index Text Analyzer*. Lüdenscheid: RAM-Verlag.

Mozaffari, H. (2013). An Analytical Rubric for Assessing Creativity in Creative Writing. Theory and Practice in Language Studies, 3:12, pp. 2214-2219, ACADEMY PUBLISHER Manufactured in Finland. doi:10.4304/tpls.3.12.2214-2219

Vaezi, M & Rezaei, S. (2019). Development of a rubric for evaluating creative writing: a multi- phase research, New Writing, 16:3, pp 303-317, doi: 10.1080/14790726.2018.1520894

Probabilistic Frequency Applied to Diachronic Data of Spanish

Hiroto Ueda[1] and Antonio Moreno Sandoval[2]
[1]University of Tokyo, Japan, [2]Autonomous University of Madrid, Spain

In 2017 we formulated a reliable and robust frequency type based on the binomial probability. On this occasion, we present its mathematical foundations along with its applications to the diachronic data of Spanish.

In corpus linguistics, with several search patterns and multiple attributes (for example, years), we get two-dimensional frequency

tables composed of linguistic forms and variables (years) in absolute frequency (AF) and relative frequency (RF). However, both frequencies are not suitable for comparing the figures with different bases. For example, 3 in 3 (RF: 1,000, 100%) presents the value greater than 8 in 12 (RF: 0.667, 66.7%), although we intuit that 3 in 3 is less important than 8 in 12 and much less important than 80 in 120. This means that the relative frequency does not serve for numerical comparison, for the reason that, for example, 3 goals in 10 matches do not guarantee 30 goals in 100 games, which says precisely 30%. We believe that the percentage serves to describe the proportion that each case occupies within a set. However, it does not help to compare each case between several sets with different bases (populations). We will look for the solution to this problem of the numerical evaluation, typical of the relative frequency. The same can be said of the Normalized frequency (NF) obtained on the totality of words.

To overcome the lack of comparability in absolute, relative and normalized frequencies, we have introduced the concept of binomial probability to the calculation of the new type of frequency: Probabilistic frequency (P F). First, the expected Probability is calculated from the absolute frequency, the base, and the desired significance, for example, of 99%. The Probabilistic Frequency (FP) formula is FP = Expected Probability * Rounded Multiplier.

We apply the method of Probabilistic frequency to two diachronic questions of Spanish. Firstly, the three variant spellings, <u>, <b> and <v> of the lemma «voz» 'voice', which is interesting for the history of the Spanish language in the sense that current spelling <v> does not represent a labiodental consonant but a bilabial consonant: «voz» [boθ] 'voice'. Secondly, we study combinations of preposition and definite article in forms of <del>, <dela>, <al>, <ala>, etc., of which only <del> 'of the' and <al> 'to the' have been maintained in the current Spanish. For explaining these historical changes, to know the numerical vicissitudes observed in ancient documents through centuries is essential. We propose the use of the new Probabilistic frequency (FP) instead of the three traditional frequencies (AF, RF, and NF), due to its high statistic reliability, significance and robustness.

Keywords: absolute frequency, relative frequency, normalized frequency, probabilistic frequency, significance, expected probability, Spanish

A Quantitative Approach to Literature Research: Thematic Co-Occurrence Networks

Lieke Verheijen and Lidwien van de Wijngaert
Radboud University, the Netherlands

Scientists tend to be somewhat selective in choosing which literature to include in their theoretical framework, prioritizing previous studies that are in line with their own narrative. The paper presented here reports on a novel approach to literature research using network analysis. This study differs from existing applications of co-citation networks in focusing on content rather than research fields, and builds on prior research which allowed the creation of networks based on hypotheses (Van de Wijngaert, Bouwman & Contractor, 2014). The present research uses networks for identifying themes and offers the possibility to include publications without specific hypotheses. The topic of our analysis was inspired by the request of a Dutch municipality for a systematic review of literature on inclusiveness of online government services. Our corpus of 46 relevant papers on this topic was imported to ATLAS.ti (Friese, 2019) for data analysis, thereby innovatively using software that is intended for qualitative analysis in a truly quantitative fashion. We conducted a search for which content words occurred most frequently (at least 100×), excluding research terminology. The top words were scrutinized and where necessary combined to provide a list of top lemmas. Next, the papers were coded for these top lemmas and their co-occurrences computed. The resulting co-occurrence table, with scores for each combination of lemmas, was exported via Excel to Gephi (Bastian, Heymann, & Jacomy, 2009). By means of this visualization software, we generated a network showing the co-occurrences of the top lemmas in sentences in the corpus. This reveals which content words relevant to the domain of inclusiveness of online government services often co-occur in the literature. Using the strongest relationships only, we created a graph with Force Atlas as a layout algorithm. Node size was based on weighted

degree; the Modularity algorithm (Lambiotte, Delvenne & Barahona, 2008) was used to determine clusters. Based on the graph, we distinguished four large clusters in the co- occurrence network: (a) people's access to the web, (b) information and communication via digital technologies, (c) services to the public, and (d) citizens' social inclusion due to government policies. These clusters closely connect to Orlikowski's (1992) structurational model, which includes the pillars of technology, institutional characteristics, and human agents. Thus, this case study presents an approach for efficiently and systematically detecting themes in a corpus of scientific literature.

Keywords: co-occurrences, network analysis, literature research, Gephi, structuration

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.

Friese, S. (2019). *Qualitative data analysis with ATLAS.ti* (third edition). Los Angeles, LA, etc.: SAGE.

Lambiotte, R., Delvenne, J.C., & Barahona, M. (2008). Laplacian dynamics and multiscale modular structure in networks. *arXiv preprint*, 0812.1770.

Orlikowski, W.J. (1992). The duality of technology: Rethinking the concept of technology in organizations. *Organization Science*, 3(3), 398-427.

Van de Wijngaert, L., Bouwman, H., & Contractor, N. (2014). A network approach toward literature review. *Quality & Quantity*, 48(2), 623-643.

A Study on Probability Distributions of Dependency Distances of Web Genres

Yaqin Wang
Guangdong University of Foreign Studies, China.

Web genres have attracted much attention from corpus linguists recently (Sardinha, 2014; Biber & Egbert, 2018), whose studies mainly focused on

genres' lexical and grammatical features instead of syntactic characteristics. Dependency distance, a useful metric in describing syntactic relationship, has been found to well represent syntactic features of a certain genre to some degree (Wang & Liu, 2017). Studies have shown that the parameters of several probability distributions can well separate different genres (Wang & Liu, 2017; Wang & Yan, 2018) and different levels of EFL learners (Ouyang & Jiang, 2018). Based on the Waring distribution, a word frequency distribution (Baayen, 2001), which was once used for studying the distribution of syntactic units (Köhler and Altmann, 2000), the present study examines the probability distribution of dependency distances of web genres, including personal blog, news webpage, email and twitter. Results show that the right truncated Waring distribution can be well fitted to all distributions of dependency distances, which suggests that all web genres display the tendency of dependency distance minimization (Liu et al., 2017). Two parameters, namely, b and n, of the probability distribution show significant differences across web genres, indicating that parameters of the right truncated Waring distribution can be a useful indicator of web genres classification. In addition, the current research also found that parameters are significantly correlated to the mean dependency distance, between which parameter b's correlation effect size is higher than that of parameter n. The greater the value of the parameter b of a text is, the smaller its mean dependency distance is. The findings suggest the feasibility of mathematical models in web genre investigations, which may shed new light on web genre classification and quantitative linguistics.

Keywords: Web genres, Syntactic feature, dependency distance, probability distribution

References

Baayen, R. H. (2001). *Word frequency distributions* (Vol. 18). Springer Science & Business Media.

Biber, D., & Egbert, J. (2018). *Register variation online*. Cambridge University Press.

Köhler, R., & Altmann, G. (2000). Probability Distributions of Syntactic Units and Properties. *Journal of Quantitative Linguistics*, 7(3), 189-200.

Liu, H., Xu, C., & Liang, J. (2017). Dependency distance: a new perspective on syntactic patterns in natural languages. *Physics of life reviews*. 21, 171-193.

Ouyang, J., & Jiang, J. (2018). Can the probability distribution of dependency distance measure language proficiency of second language learners? *Journal of Quantitative Linguistics*. 25(4), 295-313.
Sardinha, T. B. (2014). 25 years later Comparing Internet and pre-Internet registers. In Sardinha, T. B., & Pinto, M. V. (Eds.). *Multi-Dimensional Analysis, 25 years on: A tribute to Douglas Biber*, 60, John Benjamins Publishing Company. 81-105.
Wang, Y., & Liu, H. (2017). The effects of genre on dependency distance and dependency direction. *Language Sciences*. 59, 135-147.
Wang, Y., & Yan, J. (2018). A quantitative analysis on a literary genre essay's syntactic features. In Jiang, J. & Liu, H. (Eds.). *Quantitative Analysis of Dependency Structures*. Walter de Gruyter GmbH & Co KG. pp. 295-314.

Revisiting Zipf's Law: A New Indicator of Lexical Diversity

Yawen Wang and Haitao Liu
Zhejiang University, China

With C as a normalizing factor and β and α as parameters, the Zipf-Mandelbrot law (Zipf, 1949; Mandelbrot, 1953) is formulated as:

$$f(r_i) = \frac{C}{(\beta + r_i)^{\alpha}}$$

where $f(r_i)$ is the frequency of a word of *i* th rank (r_i) in a rank-frequency profile, n is the number of ranks. The parameters in the Zipf-Mandelbrot law have been used to differ systematically between texts and languages (Popescu et al., 2009), to reflect language complexity (Baixeries, Elvevåg, & Ferrer-i-Cancho, 2013), and to quantitatively measure the lexical diversity of languages (Bentz and Kiela, 2014; Bentz et al., 2014).

This article argues that when β=0, α=1, the constant C as the coefficient of proportionality can be used as cross-linguistic, quantitative measure of the lexical diversity of languages. Lexical diversity is defined as the distribution of word types used to express the same information. Parallel corpora are designed with source and translated texts of *Alices Adventures in Wonderland* and *Le Petit Prince* in more than 20 languages.

These parallel translations provide a natural means of controlling for constant information content. The lexical diversity of these texts is quantified with MATTR (Covington and McFall, 2010).

The results show that, with these measures, the constant C observes a great variety of lexical diversities across language families despite constant content of the texts and can be taken as a reliable indicator of lexical diversity. The findings include: a) the Zipf's constant C calculated with the most frequent 200 words can best reveal the closeness among languages; b) the C value measured based on all words in the text fall within a limited range and cannot distinguish them; c) languages with low lexical diversity display higher C values (e.g. Dutch, Romanian), whereas lexically rich languages have lower C values (e.g. Hungarian, Polish).

Keywords: Zipf's law, lexical diversity, parallel corpus, MATTR

References

Baixeries, J., Elvevåg, B., & Ferrer-i-Cancho, R. (2013). The evolution of the exponent of Zipfs law in language ontogeny. *PloS one*, 8(3), e53227.

Bentz, C., & Kiela, D. (2014). Zipfs Law Across Languages of The World: Towards A Quantitative Measure of Lexical Diversity. *In Evolution of Language: Proceedings of the 10th International Conference (EVOLANG10)* (pp.385-386).

Bentz, C., Kiela, D., Hill, F., & Buttery, P. (2014). Zipfs law and the grammar of languages: A quantitative study of Old and Modern English parallel texts. *Corpus Linguistics and Linguistic Theory*, 10(2), 175-211.

Covington, M. A., & McFall, J. D. (2010). Cutting the Gordian knot: The moving average type-token ratio (MATTR). *Journal of quantitative linguistics*, 17(2), 94-100.

Mandelbrot, B. (1953). An informational theory of the statistical structure of language. *Communication theory*, 84, 486-502.

Popescu, I.-I., et al. (2009). *Word frequency studies*. Berlin & New York: Mouton de Gruyter.

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

Quantitative Features of Common Words in Zhuang Language

Aiyun Wei and Haitao Liu
Zhejiang University, China

This study focuses on investigating the quantitative features of the common words of Zhuang language based on a corpus of more than 600,000 Zhuang tokens. It covers the definition of common words of the language, the quantitative characteristics of two groups of common words, i.e. the group with the rank before "h-point" and the one with the rank of top 1000 and the relationship between polytextuality and frequency. The results demonstrate that the high-frequency common words of Zhuang, Mandarin Chinese and English share certain similarities: the high-frequency common words before the "h-point" of the three languages are short in word length, among which the monosyllabic words account for the largest proportion, which means that they follow the Zipf's Law in that the words with higher frequency tend to be shorter; in the high-frequency common words ranking before the "h-point", the nouns account for the highest proportion; in the common words ranking the top 1000, the proportion of notional words in Zhuang language is significantly different from that of Chinese and English, (esp. English), which is as high as 85.63% and 91.5% respectively. Conjunctions and auxiliary words occupy an extremely low proportion, which indicates that Zhuang language does not rely on auxiliary words but by means of word order combination in forming sentences and conveying meanings. Moreover, the relationship between polytextuality and frequency in Zhuang language can also be described by the function $y=ax^b$.

Keywords: common words of Zhuang language; word frequency; word length; h-point; polytextuality; synergetic linguistics

Activity of Translational Chinese: A Multi-Corpora Based Study

Zuohao Xu and Yue Jiang
Hunan Institute of Science and Technology, China

Activity is a stylistic parameter applied in quantitative linguistics, referring to the ratio of verb number to the sum number of verbs and adjectives (Zörnig et al., 2015; Liu, 2017). It discloses the dynamic and synergetic relationship between verbs and adjectives in a text, and has already been applied to the study of genres and language typology (Popescu et al., 2013; Zörnig et al., 2015; Kubat & Cech, 2016; Zörnig & Altmann, 2016; Mistecky et al., 2018), but not to the study of translational Chinese. Owing to this, the present paper compares translational Chinese, original Chinese, and original English in 4 broad and 15 specific genres respectively, in terms of activity and based on three online corpora，namely， LCMC, ZCTC and Brown Family (extended) (Hardie, 2008). It was found that: (1) Activity of the translational Chinese is within the interval from that of the original Chinese to that of the original English, and is closer to the former and more distant to the latter; (2) Activity of the translational Chinese is close to that of the original Chinese in non-fictions, but distant in fictions, but when compared with the original English, the result is in reverse; (3) The distribution curve of the translational Chinese activity in all genres is not as steep as that of the original Chinese and the original English. These findings suggest that:

1) Translational language is a mediated and restricted language variant;
2) The stylistic features of the non-fictional translations tend to be closer to those of the target language while the stylistic features of the fictional translations are closer to those of the source language; and
3) Translational language is not as genre-sensitive as original languages.

Keywords: activity, translational Chinese, original languages, features

References

Hardie, A. (2008). Corpus Query Processor [DB/OL]. https://cqpweb.lancs.ac.uk/, accessed 04/08/2019.
Kubat, M & R. Cech. (2016). Quantitative analysis of US presidential inaugural addresses [J]. *Glottometrics*, 34:14-27.
Liu Haitao. (2017). *An Introduction to Quantitative Linguistics* [M]. Beijing: The Commercial Press.
Mistecky, M., S. Andreev & G. Altmann. (2018). Piotrowski law in sequences of activity and attributiveness: a four-language survey [J]. *Glottometrics*, 42: 21-38.
Popescu, I.-I., R. Čech & G. Altmann. (2013). Descriptivity in Slovak lyrics [J]. *Glottotheory*, 4(1): 92-104.
Zörnig, P. et. al. (2015). *Descriptiveness, Activity and Nominality in Formalized Text Sequences* [M]. Lüdenscheid: RAM-Verlag.
Zörnig, P. & G, Altmann. (2016). Activity in Italian presidential speeches [J]. *Glottometrics*, 35: 38-48.

A Time Series Analysis of Vocabulary in Japanese Texts: Non-Characteristic Words and Topic Words

Makoto Yamazaki
National Institute for Japanese Language and Linguistics, Japan

In this paper, we will analyze how the vocabulary is distributed in the text. How the words are distributed in the text will be related not only to the content of the text, but also to the general characteristics of the text. Through this research, we can clarify the individuality and universality of texts.

The data used was 635 texts which token ranges from 1950 to 2050 words from BCCWJ. The breakdown of the texts were 259 Library book samples, 191 Publication book samples, 52 Magazine samples, 38 Blog samples, 34 White paper samples, 24 Bestseller book samples, 16 School textbook samples, 10 Newspaper samples, 10 Law text samples and one Minute of National Diet sample. Each text was divided into 10 equal-word intervals, and counted in how many intervals each word appears.

From the results and observations obtained, the following conclusions were drawn.

(1) In the token level, the ratio of particles increases and the ratio of

nouns decreases as the number of appearing intervals increases. Also the ratio of auxiliary verbs becomes slightly higher, and the ratio of verbs does not change much depending on the number of intervals.

(2) On the other hand, in the type level, proportion of parts of speech remains almost unchanged.

(3) The average number of words that appear in all intervals was about 10 words per text, and there was no significant difference between the registers.

(4) There were 470 different words that appeared in all intervals, of which 80 were proper nouns (names and places) and 247 were common nouns. These words correspond to the "topic words" pointed out by Tanaka (1973) and may be closely related to the subject in the given text. The remaining 143 words are likely to be non-characteristic words. These words are similar to Yamazaki's (2012) list of non-characteristic words adjusted to short unit word analysis.

Keywords: time series analysis, vocabulary, distribution, topic word, non-characteristic word, Japanese

References

Tanaka, Akio (1973) Key-Words for Automatic Abstracting of Literary Texts, *Studies in Computational Linguistics*, vol.5, pp.141-184.

Yamazaki, Makoto (2012) Measurement of Textual Cohesion Using Similarity between Paragraphs, *Proceedings of the 2nd Corpus Linguistics Workshop*, pp.291-298.

Possibility of Indicating the Evolution of Chinese Language Through Probability Distribution of Dependency Distance

Jianwei Yan
Zhejiang University, China

Popescu et al. (2014) proposed that the length distribution of many linguistic units can well capture the Zipf-Alekseev function. This has been verified by many quantitative studies of natural languages and even extralinguistic phenomenon (Liu, 2009; Guan, 2019). In this article, we focused on the question whether the dependency distance of Classical

Chinese and Modern Chinese can fit this linguistic law and whether the parameter a of the probability distribution can indicate the evolution of Chinese language. We adopted three Chinese treebanks more than 20k tokens from the repository of Surface-Syntactic Universal Dependencies (SUD) (Gerdes et al., 2018) (Classical Chinese of SUD_Classical_Chinese-Kyoto (three subsets), and Modern Chinese of SUD_Chinese-GSD (three subsets) and SUD_Chinese-PUD (one subset)). We fitted these three treebanks (seven subsets) to the Right truncated modified Zipf-Alekseev distribution, and found that all seven subsets can fit this distribution well with C values smaller than 0.02 (six of them have C values even smaller than 0.01), however, the parameter a does not demonstrate any regularity of Chinese evolution. We assumed that it might be due to the uneven sentence length of these three treebanks (Jiang and Liu, 2015) (the mean sentence length of Classical Chinese is 4.92, while it is 24.12 for Modern Chinese). Hence, we constrained the factor of sentence length (two groups for the range from six to 10, and 11 to 15) to further explore the possibility of indicating Chinese evolution through the parameter a of Right truncated modified Zipf-Alekseev distribution. The result shows that the dependency distance of three most commonly used dependency types in these three treebanks, viz, nsubj, comp:obj and mod, can well fit the Right truncated modified Zipf-Alekseev distribution, and the parameter a of comp:obj and mod in the probability distribution demonstrate some patterns, i.e., the comparatively larger parameter corresponds to longer MDD of Modern Chinese, while the smaller one shorter MDD of Classical Chinese, which might be of great significance to reflect the syntactic evolution of Chinese language. This small trial might provide a glimpse into the development of Chinese language; however, more continuous treebanks with more balanced genres are highly recommended for future studies.

Keywords: Chinese language; language evolution; probability distribution; dependency distance; right truncated modified Zipf-Alekseev distribution

Reference

Gerdes, K., Guillaume, B., Kahane, S., & Perrier, G. (2018). SUD or Surface-Syntactic Universal Dependencies: An annotation scheme near-isomorphic to UD. In *Proceedings of Universal Dependencies Workshop 2018*, 66-74. Brussels.

Guan, W. (2019). Probability Distribution of Represented Sources in Conversations of Adults and Children, *Journal of Quantitative Linguistics*, DOI: 10.1080/09296174.2019.1580812

Jiang, J., & Liu, H. (2015). The effects of sentence length on dependency distance, dependency direction and the implications-Based on a parallel English-Chinese dependency treebank. *Language Sciences*, 50, 93-104.

Liu, H. (2009). Probability distribution of dependencies based on a Chinese Dependency Treebank. *Journal of Quantitative Linguistics*, 16, 256-273.

Popescu, I.-I., Best, K.-H., & Altmann, G. (2014). *Unified modeling of length in language*. Lüdenscheid: RAM-Verlag. ISBN 978-3-942303-26-2.

A Quantitative Study of Prominence Models in the Metaphoric and Metonymic N+N Compounding in Chinese and English

Xianwu Zhou
NingboTech University, China

The N+N compounding is represented by the conceptual integration of events or entities that link our cognitive system, thus the probability distribution represented by events or entities in the fossified integrational structures can be quantitatively analyzed. As Sweetser (1991) proposed, the linguistic categorization depends not just on the naming of distinctions but also on structuring of perceptions of the world. Interestingly, we find that the distinctions in the metaphorical and metonymic N+N compounding are diversified in the prominence model (PM) of property, feature or function of category itself. A noun category represents an entity, numerous entities represent nodes, and the relations between nodes form the edges, thus numerous nodes and edges constitute the complex network of metaphoric and metonymic N+N compounds. However, the previous studies seldom examine the PMs of the target corpus. In this article, we addressed the questions of what the similarities and differences of PMs and what the PM distributions are by a case study of body-lexis-patterned N+N compounds (e.g. headlamp, skinhead) from

the perspective of complex network. The corpus database was self-built by sorting out the corpus data from Xiandai Hanyu Cidian (edition 6) and Longman Dictionary of Contemporary English (edition 6). All the PMs (property prominence model, feature prominence model and function prominence model) of target corpus were annotated in the light of Conceptual Metaphor Theory by Lakoff & Johnson (1980) and the Qualia Structure by Pustejovsky (1993). The statistic results show that, firstly, both the highlighted categories and the PMs themselves emerge a probability distribution as that in the Chinese Dependencies Treebank proposed by Liu (2009) , and exhibit prominence hierarchies motivated by the hierarchical feature of categories themselves, which is consistent with the Zipf's Law; secondly, the diversity of prominent categories and PMs is correlated to languages types, both the difference in hierarchical boundaries of the complex network and the difference of PMs perspectivized by metaphoric and metonymic thinkings result from language types, and the mechanism behind is the Principle of Least Efforts. The survey proves that via such a comparatively closed complex network system, it is more feasible to observe and conclude the feature and law of PMs in the conceptual compounding, and more intuitive to examine the similarities and differences of metaphoric and metonymic N+N compounding at a general level.

Keywords: N+N Compounds, Prominence Model, Metaphor, Metonymy

References

[1] Cong Jin. (2018). A network analysis of two-character words in modern Chinese. In Liu, HT (ed.) *Advances in Quantitative Linguistics*. Hangzhou: Zhejiang University Press.

[2] Larsen-Freeman, Diane & L. Cameron. (2008). *Complex Systems and Applied Linguistics*. Oxford: OUP,

[3] Liu, H. (2009). Probability distribution of dependencies based on a Chinese Dependency Treebank. *Journal of Quantitative Linguistics*, 16, 256-273.

[4] Sweetser, E. (1991). *From Etymology to Pragmatics: Metaphorical and Cultural Aspects of Semantic Structure*. Cambridge: Cambridge University Press.

[5] Zipf, G. (1949). *Human Behavior and the Principle of Least Effort*. New York: Addison-Wesley.

Author Index


Trier
1991
Moscow
1993
Helsinki
1997
Praha
2000
Athens
2003
Trier
2007
Graz
2009
Belgrade
2012
Olomouc
2014
Trier
2016
Wrocław
2018
Tokyo
2021