QUALICO 2016

Trier 24-28 August

Abstracts

Universität Trier

Stream-based *RF*-Measure in Text Classification

M. F. Ashurov and V. V. Poddubny
National Research Tomsk State University, Tomsk, Russian Federation

A new system for performing and evaluating stream-based text classification with using frequencies of substring repetition is described. Stream-based text classification considers any text as a stream of symbols. This approach differs from the feature-based approach, which relies on traditional extracting of text features. The comparison of stream- based approaches and the feature-based SVM approach, described by Thomas and Teahan (2007), shows that all stream-based classifiers are better than SVN one. Count-based *R*-measure (from repetition), a stream-based approach proposed by Khmelev and Teahan (2003), counts all test text substrings that are present in a class supertext. The supertext is formed by free-order concatenating all texts of the particular class. *R*-measure is normalized by the number of used substrings of the test text. Truncated *R*-measure similarity counts all substrings of test text length n in the supertext, but having a range of substring lengths, from *k1* to *k2*, unlike classical *R*-measure where $k1 = 1$ and $k2 = n$. The *k1* and *k2* parameters selection used in truncated R-measure calculation is based on using natural language features. The minimum length *k1* is equal to 10 symbols and the maximum length *k2* is equal to 45 that allows us to avoid matching with many common words in Russian language and repeating long phrases. A comparison of stream-based approaches based on *R*-measure and PPMD compression shows that an accuracy of classification by *R*-measure is less than accuracy by the PPMD approach in some cases. These situations could be explained by the fact that PPMD approach uses frequencies of substring repetition. So the system implements a symbol- or character-based natural language model based on the *R*-measure modification, named *RF*-measure (from repetition frequency), using a substring repetition frequencies. *RF*-measure is adopted through the same arrangement of truncation, but it combines counting text substrings and calculating a degree of similarity of substring frequencies between a test text and a supertext. The similarity measure is a ratio of a frequency of test text substring repetitions and an average frequency of supertext substring repetitions. Accordingly, the similarity measure should be normalized and be in $[0,1]$ range. The results of analyzing of the system demonstrate that the *RF*-measure approach outperforms *R*-measure in classification tasks such as authorship ascription. In case of genre mixing free into author's text classes, accounting frequency of test text substring repetition in supertexts increases the classification accuracy. The classification accuracy (calculated by Van Rijsbergen's *F*-measure) of the system using *RF*-measure is more than 0.9 in average on Russian fiction text data which consists of two samples based on authors grouping by 19th and 20th centuries. This accuracy is on the several percent better than one for the *R*-measure.

Keywords: Stream-based classification, R-measure, RF-measure, frequency of substring repetition

REFERENCES

Khmelev, D. V., & Teahan, W. J. (2003). Verification of text collections for text categorization and natural language processing. *Technical Report AIIA 03.1*. Bangor. School of Informatics, University of Wales.

Thomas, D. L., & Teahan, W. J. (2007). Text categorization for streams. In *Proceedings of the 30th Annual International ACM SIGIR Conference on Research and Development in Information*

Retrieval (pp. 907). New York: ACM Press.

Functional Explanation in Quantitative Linguistics

Martina Benešová, Dan Faltýnek and Lukáš Zámečník
Palacký University Olomouc, Czech Republic

One of the central methodological problems of quantitative linguistics is the question of the nature of scientific explanation which is available in this linguistic approach. Recently these questions have provoked many debates (e.g. Meyer, Zámečník, Milička) above all in the context of synergetic linguistics, which is an example of a successful quantitative linguistics theory. Although such a debate does not conclude unambiguously, the need for a more accurate delimitation of the concept of functional explanation in quantitative linguistics can be regarded as an important mutual result.

The paper aims at demonstrating the form and problems of existing functional explanations in quantitative linguistics (Köhler, etc.) and at introducing variants of functional explanations which are immune to the mentioned problems. The main support of new forms of functional explanation is the systemic dynamic description of economizing principles which are normally applied in functional explanations. The suggested approach to functional explanation, thus, represents a possible alternative to the synergetic hypothesis of quantitative linguistics (Köhler, etc.).

Keywords: Functional explanation, synergetic hypothesis, synergetic linguistics, systemic dynamic description, economizing principles

Segmentation for MAL on Content-Semantic Levels

Martina Benešová and Petra Vaculíková
Palacký University Olomouc, Czech Republic

The paper is based on the hypothesis “thanks to working with authentic spontaneous dialogues the real character of language can be revealed, which can be proven be observing the MAL manifestation”. Some linguists claim “spoken languages become chaotic” (e.g. Hoffmannová, 2010). We presume the spoken language has been developed by thematic structures which are not chaotic. In this context we presume the longer thematic elements are on the higher level, the fewer elements will be included on the immediately lower level (measured in its subconstituents). We presume information structuring in spontaneous languages is a natural phenomenon. We have chosen content-semantic elements because they reflect the utterence content and the content of human minds. The closest concepts which can reveal content-semantic structures of spoken dialogues are various apparatus for thematic development. A suitable approach of thematic development for this purpose seems to be the concept of aboutness (Daneš,1974; Fries, etc.). In agreement with e.g. Firbas (1995), Leong Ping (2004) and others, we do not leave out rhematic elements in speech. Compared to Altmann (2015), we do not aim at themes as basic thematic elements but we aim at thematic-rhematic nexus which is formed from the theme and rheme (sometimes any of these elements is omitted, in some dialogues the whole nexus can be omitted

because it is realized nonverbally, or is interrupted by a communication situation or associations etc.). For this research there is not important to state the number of words in the topic and in the comment in each clause (compare with Beliankou, Köhler & Naumann, 2013), but the development of thematic-rhematic nexus which is called thematic progressions is very important (Daneš set 5 basic types of TP, Dubois, 1987; Crompton, 2004; Leong Ping, 2005 and others add some more progressions). In this paper MAL will be tested on these levels: 1st LEVEL –HYPERTHEMES measured in the number of THEMATIC COMPLEXES 2nd LEVEL –THEMATIC COMPLEXES measured in the number of THEMATIC CHUNKS 3rd LEVEL –THEMATIC CHUNKS measured in the number of THEMATIC PROGESSIONS Here for clarification we provide and summarize brief definitions of terms above. The theme (the terminology topic) means for this study: the main idea that one talks about and lets the reader or listener know what the clause is going to be about. *The rheme* (the terminology comment or focus) – this is what one says about the theme. *The thematic progression* – the theme and rheme pattern makes together the theme-rhematic nexus, and these pairs create the thematic progressions (constant theme/continuous theme, simple liner theme/thematization of rheme, a split rheme etc.). *The thematic chunk* is a part of the dialogue where the speakers become in thematic harmony, it means they speak on one topic and they agree to speak about it, thematic chunk finishes where the topic is shifted. *The thematic complex* brings the themes which correspond one another and are interconnected in cohesive lines. Hypertheme is a unit on the highest level, from which several themes are derived.

REFERENCES

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

Altmann, G. (2015). Topic – Comment. *Problems in quantitative linguistics*, *5* [Studies in Quantitative Linguistics; 21] (pp. 74–76). Lüdenscheid: Ram-Verlag.

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

Beliankou, A., Köhler, R., & Naumann, S. (2013). Quantitative properties of argumenation motifs. In I. Obradovič, E. Kelih & R. Köhler (Eds.), *Methods and Applications of Quantitative Linguistics* (pp. 35–43). Belgrade: Academic Mind.

Crompton, P. (2004). Theme in Discourse. Thematic progressions and method of development re-evaluated. *Functions of language*, *11*(2), 213–249.

Daneš, F. (1974). Functional sentence perspective and the organization of the text. In F. Daneš (Ed.), *Papers on Functional Sentence Perspective* (pp. 106–128). Praha: Academia.

Dubois, B. L. (1987). A reformulation of thematic progression typology. *Text*, *7*(2), 89–116.

Firbas, J. (1995). On the thematic and the rhematic layers of a text. Organization in Discourse: *Proceedings from the Turku Conference* (pp. 59–72), Anglicana Turkuensia 14 (Warwik, Tauskanen and Hiltunen (Eds.).

Fries, P. H. (1995). Themes, methods of development, and texts. In R. Hasan & P. H. Fries (Eds.), *On Subject and Theme* (pp. 316–360). Amsterdam / Philadelphia, PA: John Benjamins.

Hoffmanová, J., & Cmejrková, S. (2010). *Mluvená čeština: hledání funkčního rozpětí*. Praha: Academia.

Hřebíček, L. (1995). T*ext Levels. Language Constructs, Constituents and the Menzerath-Altmann Law*. Trier: Wissenschaftlicher Verlag Trier.

Leong Ping, A. (2004). *Theme and Rheme: An Alternative Account*. Bern: Peter Lang.

Leong Ping, A. (2005). Talking themes: the thematic structure of talk. *Discourse Studies*, *7*(6), 701–732.

Quantitative Analysis of Syntactic Dependency in Czech

Radek Čech [1], Ján Mačutek[2], Michaela Koĺcová[2] and Markéta Lopatková[3]
[1]University of Ostrava, Czech Republic; [2]Comenius University, Slovak Republic; [3]Charles University Prague, Czech Republic

A hierarchical structure of a sentence can be expressed by a dependency grammar formalism (Mel'čuk, 1998; Hudson, 2007). This formalism describes the structure of a sentence in a form of a tree graph; nodes of the graph represent words, while links between nodes represent syntactic relationships between words. Within the approach, there is a syntactic function assigned to each word in a sentence, e.g., predicate, subject, object, attribute; this kind of syntactic function is referred to as analytical function here (Bejček et al., 2013).

The aim of our talk is to present results of quantitative analysis of dependency characteristics of particular analytical functions. For each word in a syntactically annotated corpus (Bejček et al. 2013), a dependency frame is derived first. The dependency frame consists of all analytical functions assigned to its directly dependent words. For instance, from the sentence

Children	love	green	apples
subject	predicate	attribute	object

it is possible to derive two dependency frames. Particularly, the predicate love has the frame [subject; object] because the words Children and apples are directly dependent on the word love according to the dependency grammar formalism; analogically, the object apples has the frame [attribute]. Further, a list of unique dependency frames (with frequency characteristics) is set up for each analytical function and for each basic word form (i.e. the lemma of a word). Based on an expectation that syntactic relationships between analytical functions are ruled by some general mechanisms (cf. Köhler, 2012), we set up hypotheses as follows: (1) there is a regular distribution of dependency frames in general in a language; (2) there is a regular distribution of dependency frames of each analytical functions; differences among distributions of particular analytical functions are caused by their specific syntactic properties; differences are manifested by different distributional models or different parameters of thesame model; (3) the more frequent the analytical function is, the more dependency frames it has; (4) the more particular lemmas occur within the analytical function, the more dependency frames the analytical function has.

To test the hypotheses, a Czech syntactically annotated corpus – the Prague Dependency Treebank 3.0 is used (Bejček et al., 2013).

The results can be interpreted as a generalization of the approach presented by Čech et al. (2010) which is focused on dependency properties of predicates.

Keywords: Syntactic dependency; dependency frame; probability distribution

REFERENCES

Bejček, E., et al. (2013). *Prague Dependency Treebank 3.0*. Prague.

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

Hudson, R. (2007). *Language Networks. The New Word Grammar*. Oxford: Oxford University

Press.
Köhler, R. (2012). *Quantitative syntax analysis* [Quantitative Linguistics (QL), 65]. Berlin / New York: de Gruyter.
Mel'čuk, I. (1988). *Dependency Syntax: Theory and Practice*. Albany, NY: Sunny Press.

Do *L*-Motifs and *F*-Motifs Co-evolve with Word Length?

Heng Chen
Zhejiang University, China

"Word length studies are almost exclusively devoted to the problem of the frequency distribution of word length in text–regardless of the syntagmatic dimension" (Köhler, 2006). What is more, the properties of motifs are similar to those of their basic language units'. Milička's investigation indicates that the distribution of *L*-motifs is inherited from the word length distribution (though not quite sure) (Milička, 2015). As previous studies of word length motifs are almost synchronic ones, in this study, we intend to explore if word length motifs as well as word frequency motifs (since there is a synergetic relationship between word length and word frequency) co-evolve with word length from a diachronic and dynamic point of view.

Our first hypothesis is that word length motifs co-evolve with word length (in written Chinese). In a previous study (Chen & Liu, 2014), we investigate how word length evolves based on the analysis of texts from ancient Chinese within a time span of 1000 years. The results show that the parameter a in Zipf- Alekseev's function (fitting to the static word length distribution statistics) increases with time.

Some investigations show that the text's *L*-motifs rank-frequency relation can be successfully fitted by the Right truncated modified Zipf-Alekseev's function (Milička, 2015; Köhler, 2008). Therefore, to explore if word length motifs (*L*-motifs) co-evolve with word length, we only need also to fit the Zipf-Alekseev's function to the frequency distributions of the *L*-motif types and observe the regular patter of parameter a changes. Further, a correlation analysis can be used to see if the two a value series of fitting modified Zipf- Alekseev's function to word length motifs and word length distributions respectively are highly relevant.

Moreover, the previous study (Chen & Liu, 2014) also shows that the synergetic relation between word length and word frequency (to be specific, the relationship between word length classes and type-token ratio of each word length class) also evolves over time, which can be seen from the decrease of the absolute value of the negative parameter *b* in their relationship function $y = ax^b$. This means that as word length distribution evolves, the word frequencies of different word length classes evolve too. Therefore, our second hypothesis is that word frequency motifs (*F*-motifs) also co-evolve with word length from a diachronic and dynamic point of view.

Indeed, besides word length and word frequency, motif length and motif frequency can also be modelled by the same distributions, e.g. power laws (Mačutek & Mikros, 2015). Therefore, in this study, we will also investigate if motif length and motif frequency evolves over time.

According to (Chen & Liu, 2014), the motivation of word length distribution as well as its relationship with word frequency evolve over time is the increase of word length, to be specific, the disyllabic trend in spoken Chinese. In this study, we reason that this may also be the reason why *L*-motifs and *F*-motifs evolve (if the evolution can be truly validated). To anticipate, this study may give us a much more in-depth understanding of word length motifs as well as word frequency motifs.

REFERENCES

Chen, H., & Liu, H. (2014). A diachronic study of Chinese word length distribution. *Glottometrics*, *29*: 81–94.

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

Köhler, R. (2008). Word length in text. A study in the syntagmatic dimension. In S. Mislovičová (Eds.), *Jazyk a jazykoveda v pohybe* (pp. 416–421). Bratislava: Veda.

Maˇcutek, J. & Mikros, G. K. (2015). Menzerath-Altmann Law for Word Length Motifs. In G. K. Mikros & J. Mačutek (Eds.), *Sequences in Language and Text* (pp. 125–132). Berlin/Boston: De Gruyter Mouton.

Milička, J. (2015). Is the distribution of L-Motifs inherited from the word length distribution? In G. K. Mikros & J. Mačutek (Eds.), *Sequences in Language and Text* (pp. 133–146). Berlin/Boston: De Gruyter Mouton.

A Diachronic Study of Chinese Words Based on Google *N*-gram Data

Xinying Chen[1] and Heng Chen [2]
[1]Xi'an Jiaotong University; [2]Zhejiang University, China

For a long time, diachronic studies of languages, probably widely as known as historical linguistic studies or language evolution studies, are mainly focusing on two aspects: constructing language evolution models and demonstrating different hypothesis by using some small language samples. The insurmountable obstacle of collecting and analyzing authentic diachronic data made the absence of quantitative investigation and hypothesis verification studies based on big data. The situation only has been changed recently due to the advancement of technologies such as OCR, computer memory, text mining, etc. Now, it is possible but still difficult to do a diachronic study by analyzing authentic language data. Despite the difficulties of analyzing and comparing the language data from different times, some pioneer researches (Michel et al., 2011; Chen et al., 2015) have shown the enormous potential of such studies on discovering language change patterns and testifying linguistic hypothesis. Our study here was focusing on Chinese. More specially, we conducted a diachronic studies of Chinese words based on the Google 1-gram data of Chinese in 1710-2010. The goals of our study were:

- To test the reliability of Google *N*-gram data for doing linguistic research,
- To observe the lexicon changes of Chinese during this time.

The Chinese Google 1-gram data is segmented-words frequency data provided by Google Company's Google books project. Although there are some flaws in this data set for doing this study, for instance, some Japanese words and Roman letters were reserved in the data set, we still think that Google 1-gram data is one of the best choices among all the available diachronic data considering the data accuracy and scale. We reconstructed six Chinese words frequency list, 1710-

1760, 1760-1810, 1810-1860, 1869-1910, 1910-1960, and 1960-2010, based on the original Google 1-gram data. Google *N*-gram provided Chinese 1-gram data from 1510-2010, but we did not take any data before 1710 because the data is relatively sparse and we think the segmentation accuracy is declining along with the time backtrack. With 50 years as a threshold, we created six words frequency lists by cleaning up the redundant and invalid data in the original lists, such as Japanese words, Roman letters, Arabic numerals and so on. After recreating the six words frequency lists, we fitted the Zipf's Law (Zipf, 1935, 1949), $y = a * x^{-b}$, to the lists, analyzed the word length distribution of them.

Our results showed that all the six frequency lists fit the Zipf's Law well (the minimum R^2 is 0.90595) despite the huge data scale (the largest list includes more than 21 billion tokens). Therefore, we believe that the texts that Google *N*-gram data created from is very natural and Google *N*-gram data should be suitable for doing some linguistic studies. According to the numbers of types, the vocabulary of Chinese is increasing over time without surprise. In addition, we looked through the word length distribution of Chinese words since the word length directly refers to the numbers of syllabus of Chinese words and some Chinese linguists claimed that there is a multi-syllabification trend of Chinese words during Chinese evolution (Chen et al., 2015).

Our results verified that the average word length of Chinese words are increasing over time and newly appeared words after 1810 are mainly disyllabic and multi-syllabic words. However, it requires a further study to observe.

REFERENCES

Chen, H., Liang, J., & Liu, H. (2015). How Does Word Length Evolve in Written Chinese? *PloS one*, *10*(9), e0138567.

Michel, J. B., Shen, Y. K., Aiden, A. P., Veres, A., Gray, M. K., Pickett, J. P., ... & Aiden, E. L. (2011). Quantitative analysis of culture using millions of digitized books. *Science*, *331*(6014), 176–182.

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

Zipf, G. K. (1949). *Human Behaviour and the Principle of Least Effort*. Cambridge: Addison-Wesley.

Authorship Attribution and Text Clustering for Contemporary Italian Novels

Michele A. Cortelazzo[1], Paolo Nadalutti[2], Stefano Ondelli[2] and Arjuna Tuzzi[1]
[1]Università di Padova; [2]Università di Trieste, Italy

When dealing with authorship attribution (AA), famous cases of disputed authorship naturally come up to one's mind: who wrote William Shakespeare's works? Is Corneille really hiding behind Molières signature in certain successful comedies? Why does Joanne Rowling write under the pen name of Robert Galbraith? And, in Italy, what is Elena Ferrante's real identity?

Style is determined by both words and syntactic structures that a writer decides to use – either consciously or unconsciously - when drafting his or her text and AA methods are called upon to reveal the "author's hand". However, the relevant literature includes hundreds of different proposals (Rudman, 1998; Koppel et al., 2008; Stamatatos, 2009) and no particular approach seems to be

preferable in absolute terms: the choice of one method or another is heavily dependant upon the text type and the objectives of the analysis. Furthermore, at present AA may be considered an – albeit partially – unchartered strain of research because no standard parameters and protocols are available to compare and contrast results achieved according to different procedures (Juola, 2015).

In quantitative approaches, AA is often dealt with as a question regarding the measure of the similarity of (or distance between) two texts, as in the particular case of text clustering. In previous works we have endeavoured to contribute to the debate on AA testing, improve Labbé's intertextual distance (Cortelazzo et al., 2012, 2013) and propose new graphic representation modes to compare the results of different measuring methods (Tuzzi, 2010).

However, we have so far tested different methods on texts whose authors were known. This new analysis deals with a corpus of contemporary novels written in Italian and introduces the following novelties: (1) limited time span; (2) increased number of novels by the same author; (3) focus on cases of disputed authorship (e.g. Giorgio Faletti, Elena Ferrante); (4) last but not least, focus on improving comparison protocols with the introduction of innovative methods to assess results exceeding traditional dichotomous measures (accuracy, precision, recall).

Keywords: Second language acquisition, learner corpora

REFERENCES

Bettoni, C., Di Biase, B. (2015). *Grammatical development in second languages: Exploring the boundaires of Processability Theory* (Eurosla Monograph Series, 3). European Second Language Association & The Authors (Creative Commons Attribution 2.0).

Statistical Regularity of Referential Cohesion in English Texts – A Quantitative Linguistic Study

Yaochen Deng
Dalian Maritime University, China

Cohesion is one of the important properties that contribute to the organization of discourse. Based on a small corpus of academic English, which has been manually annotated with anaphora information, the present study investigates the referential cohesion in English texts from the quantitative linguistic perspective. Results indicate that anaphora in English discourse display a statistical regularity in the frequency distributions and the distances between all the anaphora in a text. The frequency distribution abides by the Frumkina's Law, which best fits the negative binomial distribution. The distances between anaphora represent a monotonously decreasing sequence, which can be captured by the Zipf-Alekseev function and mathematically modelled by $y = ax^{-b-c\ln x}$. The paper concludes with the implications of the results in the interpretation of discourse cohesion as well as potentials for automatic summarization of texts.

Keywords: Anaphora, quantitative linguistics, statistical regularity, text cohesion

An Expanded Quantitative Study of Linguistic vs Geographic Distance Using Romanian Dialect Data

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

In a previous study, we had a quantified case of a partial correlation between linguistic distance (representing dialect differences) and geographic distance, using data from the Crişana region of Romania. In this region, the geography provides more than one way of measuring distance: travel distance, and travel time between locations are not the same as the direct distance. However, that study was limited for practical purposes to only 8 geographic sites.

In this expansion of the study, we look at distances between all pairs of the 120 locations in the available dialect data. Furthermore, there are various ways of subdividing the linguistic data into subsets, such as phonetic, syntactic, etc. or even more specific selections. We examine how well different linguistic aspects correlate with one or another of the geographic distances.

While it is clear that geography cannot account for all of the dialect variation, especially in a modern world where telecommunications of all sorts override the traditional effects of physical distance, and where other factors such as culture and social structure must also play a role, the effects of geography can be measured quantitatively, and a base established against which the other factors operate.

It is perhaps worth noting that in contrast to many quantitative linguistic studies, which observe a quantified pattern and search for an explanation of the pattern, in this case we have a ready-made explanation and are searching for how much quantified observation there is to support it.

REFERENCES

Embleton, S., Uritescu, D., & Wheeler, E. S. (2015). Exploring Linguistic vs Geographic Distance Quantitatively. *Presentation to VIII. Congress of the International Society for Dialectology and Geolinguistics* (SIDG) 14 - 18 September 2015, Famagusta, North Cyprus.

Statistical Trends Manifested in Protein Analysis (Case Study, QUITA)

Dan Faltýnek, Vladimír Matlach and Lukáš Zámečník
Palacký University Olomouc, Czech Republic

Statistical methods available to quantitative linguistics led to finding/corroboration of mathematized linguistic principles, as e.g. the Menzerath-Altmann law is (MAL). In the fields of bioinformatics and biolinguistics quantitative approaches inspired by quantitative linguistics were applied solely in the DNA analysis. Additionally, QUITA represents potential which exceeds such limitations and enables exploration of the protein structure and functional array. The paper launches the hypothesis on the protein structure and functional array which shifts proteomics towards systemic dynamic description in which economizing principles play the leading role.

The study shows that unlike quantitative linguistics, proteomics is clearer evidence of the role which economizing principles play in the general systemic description. Analyses performed in

protein classes due to QUITA software prove that economizing principles manifested as MAL are employed on the levels (of the so called secondary protein structure) of the protein structural as well as functional arrays. The paper is evidence of extraordinarily auspicious potential which is represented by the quantitative linguistics approach in biology (proteomics).

Are Some Languages Spoken More Quickly Than Others? A Quantitative Typological Study

Gertraud Fenk-Oczlon
Alpen-Adria-Universität Klagenfurt, Austria

Introduction: The question whether languages differ in their intrinsic tempo is attracting great interest. Is it only an impression that "speakers of some languages seem to rattle away at high speed like machine-guns, while other languages sound rather slow and plodding"? (Roach, 1998, 150) Or is there really something in the language structure itself that makes some languages sound faster? And if so: How to measure this "basic" or "intrinsic" tempo of languages, abstracting from the enormous inter-individual and situation dependent variation in speech rate, and from its variation depending on age, gender, education, etc. In Fenk-Oczlon and Fenk (2010) we suggested and applied the simple metric "syllables per intonation unit" in order to analyze and compare languages with respect to their intrinsic tempo, a metric that does without any measuring of duration.
Procedure: Native speakers of 51 languages from all continents (19 European, 32 Non-Indo-European) were asked to translate a matched set of 22 simple declarative sentences encoding one proposition in one intonation unit into their mother tongue. Furthermore, they were asked to count the number of syllables in normal speech. The number of phonemes was determined by the authors, assisted by the native speakers and by grammers of the respective languages.
Results: The 51 languages in our sample show a considerable variation in the mean number of syllables per intonation unit, ranging from 4.64 in Thai up to 10.96 in Telugu. The mean number of syllables per clause is 7.02, and the mean number of phonemes per syllable is 2.24. German shows the highest mean syllable complexity (2.79 phonemes per syllable), followed by Dutch (2.78) and Thai (2.75). The languages with the lowest syllable complexity are Hawaiian (1.76), Japanese (1.88) and Roviana (1.92). The results of a cross-linguistic correlation between the number of phonemes per syllable and the number of syllables per sentence give: $r = -0.73$ (sign. $p < 0.01$).
Discussion: We view these differences as intrinsic tempo differences between languages: The smaller the syllables of a language, the higher the number of syllables per intonation unit and the higher the intrinsic tempo of that language – similar to music where in "phrases containing many notes, the notes are usually very fast" (Temperley, 2001). The number of syllables per intonation unit or the language intrinsic tempo seems to be, moreover, an important variable underlying the classification of languages as "stress-, syllable-, and mora-timed". The findings will be related to complexity trade-offs (e.g. Fenk-Oczlon & Fenk, 2014) and, more generally, to systemic typology.

REFERENCES

Fenk-Oczlon G., & Fenk A. (2010). Measuring basic tempo across languages and some implications

for speech rhythm. *Proceedings of the 11 th Annual Conference of the International Speech Communication Association* (INTERSPEECH 2010), Makuhari, Japan, 1537–1540.
Fenk-Oczlon, G. & Fenk, A. (2014). Complexity trade-offs do not prove the equal complexity hypothesis. *Poznań Studies of Contemporary Linguistics*, *50*(2), 145–155.
Roach, P. (1998). Some languages are spoken more quickly than others. In L. Bauer & P. Trudgill (Eds.), *Language Myths* (pp. 150–158). London: Penguin.
Temperley, D. (2001). *The cognition of basic musical structures*. Cambridge MA/London: MIT Press.

Why Verb Initial Languages are Less Optimized with Respect to Dependency Lengths than other Languages?

Ramon Ferrer-i-Cancho
Universitat Politecnica de Catalunya, Spain

It has been stated that head-final languages such as Japanese, Korean, and Turkish (SOV in all cases) show much less minimization with respect to syntactic dependency lengths than more head initial languages such as Italian (SVO), Indonesian (SVO), and Irish (VSO), which are apparently highly optimized (Futrell et al., 2015). Although this relationship between head finality is presented as a new and unexpected discovery, here will argue that this is a prediction of word order models. The outline of the statistical and mathematical arguments is as follows: SOV languages are a priori less optimized with respect to dependency lengths than SVO languages (Ferrer-i-Cancho, 2008). Although dependency length minimization operates in all languages where it has been tested (Ferrer-i-Cancho, 2004, Ferrer-i-Cancho & Liu, 2014; Liu, 2008; Futrell et al., 2015), a central verb placement can be interpreted as a sign that the principle of dependency length minimization is stronger (Ferrer-i-Cancho, 2014). SVO languages tend to develop from SOV languages (Gell-Mann & Ruhlen, 2011), being dependency length minimization the driving force (Ferrer-i-Cancho , 2015). In turn, VSO/VSO languages tend develop from SVO languages (Gell-Mann & Ruhlen, 2011), namely, verb final languages have passed through a dependency length minimization bottleneck (Ferrer-i-Cancho 2015) and thus verb initial languages are expected to be more optimized than verb final languages, which are closer to an initial or early stage of word order evolution (Ferrer-i-Cancho 2014). Finally, we note that SOV, SVO, VSO and VOS cover the overwhelming majority of languages showing a dominant word order (Dryer & Hasplemath, 2013).

REFERENCES

Dryer, M. S., & Haspelmath, M. (2013). *The World Atlas of Language Structures Online*. Leipzig: Max Planck Institute for Evolutionary Anthropology. (Available online at http://wals.info, Accessed on 2015-10-20.)
Ferrer-i-Cancho, R. (2004). Euclidean distance between syntactically linked words. *Physical Review*, *E 70*, 056135.
Ferrer-i-Cancho, R. (2008). Some word order biases from limited brain resources. A mathematical approach. *Advances in Complex Systems*, *11*(3), 394–414.
Ferrer-i-Cancho, R. (2014). Why might SOV be initially preferred and then lost or recovered? A theoretical framework. In E. A. Cartmill, S. Roberts, H. Lyn & H. Cornish (Eds.), *The*

Evolution of Language - Proceedings of the 10th International Conference (EVOLANG10) (pp. 66–73). Evolution of Language Conference (Evolang 2014), Vienna, Austria, April 14 - 17.
Ferrer-i-Cancho, R. (2015). The placement of the head that minimizes online memory: a complex systems approach. *Language Dynamics and Change*, *5*(1), 114–137.
Ferrer-i-Cancho, R., & Liu, H. (2014). The risks of mixing dependency lengths from sequences of different length. *Glottotheory*, *5*(2), 143-155.
Futrell, R., Mahowald, K., & Gibson, E. (2015). Large-scale evidence of dependency length minimization in 37 languages. *Proceedings of the National Academy of Sciences*, *112*(33), 10336– 10341.
Gell-Mann, M., & Ruhlen, M. (2011). The origin and evolution of word order. *Proceedings of the National Academy of Sciences*, *108*(42), 17290–17295.
Liu, H. (2008). Dependency distance as a metric of language comprehension difficulty. *Journal Cognitive Science*, *9*(2), 159– 191.

Assessment of Linguistic Complexity of the Tatar Language Using Corpus Data (an attempt to study)

Alfiia Galieva, Olga Nevzorova and Dzhavdet Suleymanov
The Tatarstan Academy of Sciences, Kazan Federal University, Russian Federation

---

The main objective of the paper is to discuss the construct of linguistic complexity of Tatar and to demonstrate how it may be measured. Currently linguistic complexity is one of debatable notions in linguistics, and there are different ways of understanding complexity depending on linguistic domains and researchers' aims. We proceed from the assumption that linguistic complexity becomes apparent in those parameters that can be measured. From our viewpoint, among the factors that influence the complexity of a language may be viewed the following:

- universality of means for expressing grammatical categories (or its absence);
- variety of grammatical categories of different types;
- regularity of means of expression of a grammatical category (absence of exception to the rules);
- degree of linguistic redundancy;
- length of affixal chains (average length of affixal chains for each part of speech);
- potential interconversion of parts of speech, etc.

Quantitative properties of the language are essential for describing linguistic complexity. The task of calculating the frequency of grammatical categories and distribution of allomorphs in real texts is the first step in the assessment of linguistic complexity. The paper represents an endeavour to assess the morphological complexity of the Tatar language as a first approximation, by analysing the distribution of certain grammatical categories on corpus data.

We sketch out some features to assess the linguistic complexity of the Tatar language that are relevant for natural language processing.

The Tatar language is an agglutinative language with a highly productive complicated inflectional and derivational morphology. The basic way of word formation and inflection is progressive affixal agglutination when a new unit is built by consecutive addition of regular and

clear-cut monosyllabic derivational and inflectional affixes to the stem, therefore the stem remains unchanged. Affixal agglutination provides unified morphological means for forming derivatives within the same grammatical class of words as well as for changing the part-of-speech characteristic of the word and for turning it into another lexical and grammatical class.

We consider the statistical distribution of grammatical categories on corpus data and emphasize the complex phenomenon of Tatar morphology. The paperdemonstrates that Tatar inflectional affixes tend to be universal for different parts of speech, and all the allomorphs are conditioned phonetically, and irregular affixes do not exist. We also establish the frequency of allomorphs and give some explanations concerning asymmetry of distribution of allomorphs. Particular attention is paid to nominal affixes, and we substantiate statistically that Tatar case affixes and the comparative affix are not bound to the unique grammatical class of stem, but may be joined to a broad range of stems, including verbal ones.

In Tatar the grammatical agreement among nominal parts of speech is absent, which reduces the degree of linguistic redundancy.

Quantitative linguistic data reflects the current state of Tatar National corpus (http://corpus.antat.ru).

The work is supported by the Russian Foundation for Basic Research (project # 15-07- 09214).

Keywords: Linguistic complexity, the Tatar morphology, affixes, distribution of allomorphs

REFERENCES

Becerra-Bonache, L., & Jimenes-Lopez, M. D. (2015). A grammatical inference model for measuring language complexity. In *Advances in Computational Intelligence - 13th International Work-Conference on Artificial Neural Networks*, IWANN 2015, Palma de Mallorca, Spain, June 10-12, 2015. Proceedings, Part I, 4–17.

Bisang, W. (2009). On the evolution of complexity: sometimes less is more in East and mainland Southeast Asia. In G. Sampson, D. Gil & P. Trudgill (Eds.), *Language complexity as an evolving variable* [Oxford Studies in the Evolution of Language] (pp. 34–49). Oxford: Oxford University Press.

Chiperen N. (2009). Individual differences in processing complex grammatical structures. In G. Sampson, D. Gil & P. Trudgill (Eds.), *Language complexity as an evolving variable* (Oxford Studies in the Evolution of Language) (pp. 178–191). Oxford: Oxford University Press.

Dahl, Ö. (2004). *The growth and maintenance of linguistic complexity*. Amsterdam: John Benjamins Press.

Gil, D. (2008). How complex are isolating languages? In M. Miestamo, K. Sinnemäki & F. Karlsson (Eds.), *Language Complexity: Typology, Contact, Change* [Studies in Language Companion Series, 94] (pp. 109–131). Amsterdam: John Benjamins Press.

Groot, C. de (2008). Morphological complexity as a parameter of linguistic typology: Hungarian as a contact language. In M. Miestamo, K. Sinnemäki & F. Karlsson (Eds.), *Language Complexity: Typology, Contact, Change* [Studies in Language Companion Series, 94] (pp. 191–215). Amsterdam: John Benhamins Press.

Guzev V. G., & Burykin A. A. (2007). Common Structural features of agglutinative languages [Obshchiye stroevye osobennocti agglutinativnykh yazykov]. In N. N. Kazansky (Ed.), *Acta linguistica Petropolitana*. Proceedings of ILR RAS. V. 3: Part 1., International Quantitative Linguistics Association, 109-117.

Juola, P. (2008). Assessing Linguistic Complexity. In M. Miestamo, K. Sinnemäki & F. Karlsson (Eds.), *Language Complexity: Typology, Contact, Change* [Studies in Language Companion

Series, 94] (pp. 89–108). Amsterdam: John Benhamins Press.
Kusters, W. (2003). *Linguistic complexity. The Influence of Social Change on Verbal Inflection*. Utrecht: LOT, Netherlands Graduate School of Linguistics.
Oflazer K. (2014). Turkish and its Challenges for Language Processing. *Language Resources and Evaluation*, *48*(4), 639–653.
Oflazer, K., & Kuruöz, Ì. (1994). Tagging and Morphological Disambiguation of Turkish Text. In *Proceedings of the fourth conference on Applied natural language processing*. Association for Computational Linguistics, 144–149.
Suleymanov D., Nevzorova O., Gatiatullin A., Gilmullin R., & Khakimov B. (2013). National corpus of the Tatar language "Tugan Tel": Grammatical Annotation and Implementation. *Procedia – Social and Behavioral Sciences*, *95*, 68–74.
Tatar Grammar [Tatar grammatikası] (2002): in 3 volumes, V. 2. – 448 p. (in Tartar). Moscow: Insan, Kazan: Fiker.
Tatar National Corpus. URL: http://corpus.antat.ru.

Testing Hypotheses on English Compounds

Hanna Gnatchuk
Alpen-Adria University, Austria

The present article is devoted to testing 5 hypotheses on English compounds. We focus on the connection between the number of the compounds and the following variables – length, age, polysemy and word class. The material of our research consists of two dictionaries – Longman Exams Dictionary (2007) and The Oxford Dictionary of English etymology (1966). The links have been captured applying Zipf-Alekseev, Lorenz and the exponential function with an additive constant. The hypotheses under consideration have been positively confirmed.

REFERENCES

Longman Exams Dictionary (2007). Pearson Longman.
Onions, C. T. (1966). *The Oxford Dictionary of English Etymology.* Oxford: At the Clarendon Press.

Four Cases of Diachronic Change in Polish, and Piotrowski Law

Rafał L. Górski and Maciej Eder
Polish Academy of Sciences, Poland

Raijmund G. Piotrowski (1974, cf. Altmann, 1983) suggested that language change is described by a formula which was used in other branches of science, chiefly in biology, as a model of growth processes. The aim of this paper is to confront Piotrowski law with empirical data gathered from a

diachronic corpus of Polish. The corpus contains 8,561,904 running words. The oldest text was published in 1532, the most recent one in 1774. In a commonly accepted periodisation, around 1546 commences the Middle Polish period, which lasts until mid 18th century. Thus the corpus covers slightly more than this period. It is rather opportunistic, consisting chiefly of belleslettres, with a handful of sermons, registers of rural community courts as well as learned texts. Of the changes to be examined three are examples of epenthesis: a change of *na* > *naj*, a prefix which is a marker of superlative, and two isolated epentheses: [s] in *wszytko* > *wszystko* 'all'; [d] in *barzo* > *bardzo* 'very'; the fourth example is the fate of elision of [l] in *albo* > *abo* 'or'.

The corpus was divided into time brackets. We experimented with different spans of these brackets – between 20 and 50 years, yet they were all shifted by 10 years, e.g. 1550-1570, 15601580 ... 17401760, 17501770 or 15501600, 15601610 etc. This procedure can be seen as a kind of smoothing, however its motivation is twofold. On the one hand, the timespan should be wide enough so as to guarantee a balanced (or rather varied) sample which is cumulated as a data point. On the other hand, in order to assure feasible analysis the number of data points should be as high as possible. Yet, if each data point represents a longer timespan, than outermost texts within it are much more distant than two texts in neighbouring stretches. The proposed procedure assures both a higher number of datapoints and rather varied group of texts which represent a data point. However, what's most important is that we diminish an effect of a Procrustean bed that is cutting the data into arbitrary slices.

Of these four examples two are described with a Piotrowskian curve: *na* > *naj* and *barzo* > *bardzo*, with a reasonable fit. The evolution of *wszytko* doesn't even slightly resemble the curve in question, although we observe a shift from a rather limited amount of the new forms (17%) up to 0,7% at the end of the examined period. Contrary to that, the elision of [l] in albo is an unsuccessful innovation. The share of the innovative form raised rapidly to 50%, oscillated for some 60 years between 50% and 27%, and again fell to almost 0% in our corpus so as to become the only accepted form in modern Polish. Our paper supports the observation that the law named after its discoverer is a strong tendency with a considerable number of exceptions, rather than a linguistic law in a strict sense.

REFERENCES

Altmann, G. (1983). Das Piotrowski Gesetz und seine Verallgemeinerungen. In K.-H. Best & J. Kohlhase (Eds.), *Exakte Sprachwandelforschung* (pp. 59–90), Göttingen: Herodot.

Klemensiewicz, Z., Lehr Spławiński, T., & Urbańczyk S. (1965). *Gramatyka historyczna języka polskiego*, Warszawa: Państwowe Wydaw; Naukowe.

Piotrowskaja, A. A., & Piotrowski, R. G. (1974). Matematičeskie modeli diachronii i tekstoobrazovanija. In *Statistica reči i avtomatičeskij analiz teksta* (pp. 361–400), Leningrad: Nauka.

Word Length Research: Coming to Rest. Or Not?

Peter Grzybek
University of Graz, Austria

The state and history of word length research seems to witness a paradoxical phase in our days. For many decades, research in this field had been shaped by predominantly local (or "ad hoc")

perspectives, concentrating on more or less isolated observations, resulting in theoretical models which eventually were combined with the claim that the results obtained might be of universal relevance (cf. the works by Wilhelm Fucks' in the 1950s). Such universalist claims increasingly became to be seen as illusions, when Wimmer and Altmann (1996) developed their "Theory of word length" which would soon be integrated in these authors' "Unified derivation of some linguistic laws" (2005) – now, word length was seen as being derived from one general approach, with different kinds of extensions and (local) modifications, and the hope was to identify language-specific, text-type specific, author-specific and possibly other boundary conditions, leading to a whole variety of (discrete) probability distributions. Ironically enough, it was exactly 20 years after the first steps "Towards a Theory of Word Length" Wimmer et al. (1994), that Popescu et al. (2014) put forth their attempt to find a unified model of length distribution of any unit in language, including words, last not least guided by the intention to avoid the search for ever new models whose validity is always merely local. The general model suggested now is a continuous function, the well-known Zipf-Alekseev function $y = Cx^{a+b\ln(x)}$, the varying parameter values of which are now seen to be sufficient to explain for what previously different models were needed – in fact, Altmann (2015) encourages scholars from quantitative linguistics to test this model for word length with as many data as available.

The present contribution will take this demand at face value: both data from previous studies will be re-analyzed and new data from various languages will be presented and analyzed; a particular focus will be laid on the question why the Zipf-Alekseev function in its discrete form (i.e., the Zipf-Alekseev distribution) might have played but a minor role in word length research until today.

REFERENCES

Altmann, G. (2013). Aspects of word length. In R. Köhler & G. Altmann (Eds.), *Issues in Quantitative Linguistics*, 3 (pp. 23–38). Lüdenscheid: RAM.

Altmann, G. (2015). *Problems in Quantitative Linguistics* 5 [Studies in Quantitative Linguistics; 21]. Lüdenscheid: RAM-Verlag.

Best, K.-H. (2005). Wortlängen. In R. Köhler, G. Altmann & R. G. Piotrovski (Eds.), Q*uantitative Linguistik – Quantitative Linguistics. Ein Internationales Handbuch – An International Handbook* (pp. 260–273). Berlin et al.: de Gruyter.

Fucks, W. (1956). Mathematical theory of word formation. In C. Cherry (Ed.), *Information theory* (pp. 154–170). London.

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

Grzybek, P. (2013). Word Length. In J. R. Taylor (Ed.), *The Oxford Handbook of the Word* (pp. 1–25). Oxford: Oxford University Press.

Popescu, I.-I., Best, K.-H., & Altmann, G. (2014). *Unified Modeling of Length in Language* [Studies in Quantitative Linguistics; 16]. Lüdenscheid: RAM-Verlag.

Wimmer, G., Köhler, R., Grotjahn, R., & Altmann, G. (1994). Towards a Theory of Word Length Distribution. *Journal of Quantitative Linguistics*, *1*(1), 98–106.

Wimmer, G. & Altmann, G. (1996). The Theory of Word Length: Some Results and Generalizations. *Glottometrika*, *15*: *Issues in General Linguistic Theory and the Theory of Word Length*, 112–133.

Wimmer, G. & Altmann, G. (2005). Unified derivation of some linguistic laws. In R. Köhler, G. Altmann & R. G. Piotrovski (Eds.), *Quantitative Linguistik – Quantitative Linguistics. Ein Internationales Handbuch – An International Handbook* (pp. 791–807). Berlin et al.: de

Gruyter.

Why the Lognormal Distribution Seems to be a Good Model in Quantitative Film Analysis

Peter Grzybek and Ernst Stadlober
University of Graz, Austria

Recent studies on shot length frequencies in movies have culminated in a vivid and controversial discussion, which led to the publication of a series of articles (Baxter, 2015; DeLong, 2015; Redfern 2015) in the re-named electronic journal Digital Scholarship in the Humanities (previously known as Literary and Linguistic Computing). Given that the frequency distributions of shot length typically display an asymmetrical right-skewed shape, the discussions basically focused on the question if the lognormal distribution can be an adequate model or not. Most participants in this many-voiced choir united to the unison support of this model as an adequate one, although the crucial question if there is some "philosophy" behind it, which might serve as an explanation for the observations made, has remained unanswered till today. Unfortunately, being so tremendously busy with defending the lognormal distribution, the authors of that discussion did not take notice of a systematic study of Russian Soviet movies, which one year before had proved another model to be adequate, which is well-known from quantitative linguistics. In fact, Veronika Koch, in her 2014 dissertation on Quantitative film studies, could show that the Zipf-Alekseev function, which had already previously been introduced into film analysis by Grzybek and Koch (2012), " is generally a suitable theoretical model for modelling of shot length frequencies" (Koch, 2014, p. 134).

In this contribution, we will not confine ourselves to the possibly embarrassing question which of the two models is the empirically better one: from a merely statistical point of view, it is well likely that both models turn out to be equally apt. Rather, we will, on the one hand, try to attempt from a mathematical meta-perspective, to explain why these just two models appear to be productive, by demonstrating their common roots, and on the other hand, we will attempt to offer an explanation for the Zipf-Alekseev function's efficiency, referring to the well-known Weber-Fechner law and the related more recent concept of "just noticeable differences". Our theoretical arguments will be sided by a re-analysis of shot length frequencies in early Laurel and Hardy movies, providing additional empirical arguments to throw overboard the lognormal distribution as a good choice in this context, from a theoretical point of view.

REFERENCES

Baxter, M. (2015). On the distributional regularity of shot lengths in film. *Digital Scholarship in the Humanities*, *30*(1), 119–128.

DeLong, J. (2015). Horseshoes, handgrenades, and model fitting: the lognormal distribution is a pretty good model for shot-length distribution of Hollywood films. *Digital Scholarship in the Humanities*, *30*(1), 129–136.

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

Koch, V. (2014). *Quantitative film studies: regularities and interrelations exemplified by shot*

lengths in Soviet feature films. Ph.D.diss., University of Graz: http://unipub.uni-graz.at/download/pdf/242930.
Redfern, N. (2015). The log-normal distribution is not an appropriate parametric model for shot length distributions of Hollywood film. *Digital Scholarship in the Humanities*, *30*(1), 137–151.

Script Complexity Distribution Based on Construction Theory of Chinese Character Form

Wei Huang
Beijing Languange and Culture University, China

Script complexity is a quantitative concept to describe the graphemic complexity of characters. Studies on script complexity and some other properties are helpful to construct the synergetic theory of a writing system, and to explore the script evolution process of this system. Our study here will focus on a relative special writing system: Chinese characters.

Comparing with other languages, Chinese has a distinctive script system that is not alphabetic. Therefore, the definition of script complexity in most of existing researches is not compatible with Chinese characters. Some linguists have utilized the number of strokes in a Chinese character to define the script complexity. However, some important factors are not considered in this way of measurement from a philological point of view and we think that the script complexity of Chinese characters should be defined based on the enormous acknowledged achievements of Chinese philology.

Many Chinese philology studies have come down to the script complexity feature. For instance, more frequent a Chinese character is used, less strokes it has. However, as lots of other studies, this research also only considered a single factor: the number of strokes. There are some other presentations in formalizing the Chinese characters, such as segment of stroke and dot matrix (lattice). These methods are geared to the needs of computational linguistics and do not fit the requirements of analyzing human recognition process. A reasonable and commonly recognized theory of Chinese scripts is that a character is a hieratical
construct of components. In other words, a Chinese character consists of one or more components, which is consists of one or more strokes. Thus the definition of script complexity of Chinese character should consider the following characteristics as much as possible: the number of strokes, the number of components, the number of component levels, and the positional relationship among components.

This research observed the script complexity distribution in both of the Chinese writing system and the authentic texts. Firstly, we proposed a new operational definition of script complexity based on the construction theory of Chinese character form mentioned in last paragraph. Secondly, the rationality of this definition was tested by using the data of Chinese Proficiency Test (Hanyu Shuiping Kaoshi, HSK) to find out that whether the learning difficulty would increase along with the complexity. Thirdly, the distributions of the script complexity of simplified Chinese characters in different charsets are illustrated and discussed. These charsets include Frequently-used Characters in Modern Chinese (Xiandai Hanyu Changyong Zibiao) with 2500 Level I characters and 1000 Level II characters, Common Characters in Modern Chinese (Xiandai Hanyu Tongyong Zibiao) with 7000 characters, and Common Characters Standard (Tongyong Guifan Hanzi Biao) with 8105 characters, etc. Besides these charsets, we also collected some written texts in different genres and made a comparison study of the distributions of the script complexity of these texts.

The construction theory of Chinese character form is compatible with not only simplified Chinese characters, but also traditional Chinese characters and the Chinese characters used in Japanese and Korean. The work presented in this study would provide a practical method and some preliminary findings for studies in Chinese character evolution as well as the research in relationship between script complexity and its frequency.

Keywords: Script complexity, Chinese characters, distribution of script complexity, construction theory of Chinese character form

REFERENCES

Kandler, A. (2009). Demography and Language Competition. *Human Biology*, *81*(2-3), 181–210.

Schulze, C., Stauffer, D., & Wichmann, S. (2008). Birth, Survival and Death of Languages by Monte Carlo Simulation. *Communications in Computational Physics*, *3*(2), 271–294.

Application of Quantitative Linguistic Properties to Translatorship Recognition

Zhan Ju-hong and Jiang Yue
Xi'an Jiaotong University, China

This study proposes a method of using quantitative linguistic properties to contrast different translation styles and to recognize the translators of unknown translation texts. The Chinese translations of Pride and Prejudice by two different translators were cleaned, segmented, tagged and used as two corpora for analysis. Both the two corpora were divided into halves, the first halves to be used as training translation texts (TTT1 and TTT2) and the second as experimental translation texts (ETT1 and ETT2) whose translators assumed unknown. Fourteen linguistic properties were explored and compared for significant difference between two TTTs, eventually discovering five contrastive and differentiative linguistic properties typical of two different translators. Data of the five linguistic properties of two ETTs were also acquired and cross-compared with those of TTTs for statistical significance of differences by using Significance Test of Difference. It was found that the five linguistic properties show no statistical significant difference between TTT1 & ETT1, and between TTT2 & ETT2; while the five properties show highly significant difference between TTT1 & ETT2, TTT2 & ETT1, and ETT2 & TTT2. Thus, we might draw a conclusion that ETT1 and TTT1 belong to the same translator; while ETT2 and TTT2 belong to another. The paper has successfully determined the contrastive linguistic properties between two translators, and based on these quantitative linguistic properties, ideally recognized the translators of experimental texts. This paper provides a new method for the research of different translation styles, translators' styles and translatorship recognition or even identification, which is intended to improve the accuracy, objectivity and explainability of the traditional studies of translation and translator styles.

Keywords: Translatorship recognition, quantitative linguistic properties, significance test of difference

Are Most People On Twitter Introverts? A Distributional Analysis of Personality Types on Twitter

Patrick Juola[1], Sean Vinsick[1] and Michael V. Ryan[2]
[1]Duquesne University; [2]EthosIO

Authorship profiling, and in particular, social media analysis for authorship, is an emerging problem that has received much research attention in recent years (Quercia et al., 2011; Rangel et al., 2015; Juola et al., 2013; Noecker et al., 2013). There is a strong business case for developing this capacity, but it also supports important public policy concerns, such as identifying/profiling cyberbullies or other on-line criminals.

We focus here on a specific social media platform: Twitter. Twitter is a microblogging site that allows users to post short (140 character) status messages, called "tweets." Due to the brevity of these messages (and due to message conventions produced by this constraint), they can be challenging to analyze using standard NLP technology. Most research has focused specific types of Twitter users (cite) and the social aspects (such as number of followers or number followed) (Quercia et al., 2011) .

For this research, we attempted to collect a representative sample of US-based Twitter users and applied a commercially-available personality analyzer (EthosIO) to see what the personality is, not of a uncharacteristic type (e.g. "highly read," see Quercia et al., 2011), but of a typical user.

The sample was collected from the Twitter "public stream", a small random sample of all publish status updates. From this stream, we performed two analyses. For our first analysis, we extracted a small set of user names found that identified their location as the "US" and that posted in English. This provided us with a set of 103 user names, each of which were analyzed using the EthosIO personality analyzer to determine each user's most probable personality type using the Myers-Briggs personality taxonomy (MBTI).

The distributions of MBTI personality types in the US population are well known („How Frequent...?", 2013). Of the four axes, three are roughly evenly distributed (50/50), while approximately 75% of the US are type S (as opposed to 25% as type N). Our results differed strongly from these:

- 82/103 were type I, while only 21/103 were type E
- 52/103 were type S, 51/103 were type N
- 87/103 were type F, 16/103 were type T
- 52/103 were type P, 51/103 were type J

The single most common type was INFP, with 28 usernames (27.18%) being identified as that type. We saw no instances at all of ENTJ or ESTP, and only a single instance each of ESFP, ENTP, ENFJ, and INTP. By contrast, the INFP personality type occurs in about 4.4% of the US population. A larger-scale analysis based on self-reporting in the Twitter stream produces similar discrepancies. Detailed results of both experiments will be presented.

These results show that the Twitter-using population differs strongly in distribution of personalities from the general population. For example, introverts strongly dominate extroverts in both samples, in marked contrast to the general public. This is an important piece of information both for those looking to market using Twitter as well as to linguists looking to study Twitter-specific behavior.

REFERENCES

"How Frequent Is My Type?" Available at http://www.myersbriggs.org/my-mbti-personality-type/my-mbti-results/how-frequent-is-my-type.htm accessed 30 Nov 2015.

Juola, P. et al. (2013) Keyboard Behavior Based Authentication for Security. *IT Professional*, *15*, 8–11.

Noecker, J., Ryan, M., & Juola, P. (2013). Psychological Profiling through Textual Analysis. *Literary and Linguistic Computing*, *28*, 382–387.

Quercia, D., Kosinski, M., Stillwell, D., & Crowcroft, J. (2011). Our Twitter profiles, our selves: Predicting personality with Twitter. In *Privacy, Security, Risk and Trust (PASSAT) and 2011 IEEE Third Inernational Conference on Social Computing* (*SocialCom*) (pp. 180–185), Boston, MA: IEEE.

Rangel, F., Celli, F., & Rosso, P. (2015). Overview of 3rd Author Profiling Task at PAN 2015. In L. Cappellato, N. Ferro, G. Jones and E. San Juan (Eds.), *CLEF 2015 Labs and Workshops, Notebook Papers*, 8 – 11 September, Toulouse, France.

How Plausible is the Hypothesis that the Population Size is Related to the Phoneme Inventory Size? Comment's from a Quantitative Linguistics Point of View

Emmerich Kelih
University of Vienna, Austria

The recently controversial discussion (Bybee, 2011; Atkinson, 2011) about a supposed interrelation between the population size and the phoneme inventory size is taken as a starting point for this presentation. Whereas the related discussion is almost (Sproat, 2011; Cysouw et al., 2012) devoted to extra-linguistic factors (community size and degree of language contact, climate, shared amount of conveyed information etc.) we would like to focus on some basic epistemic and methodological question of this hypothesis. Furthermore, in this discussion the theoretic status of the phoneme inventory size remains open, since interrelations between the phoneme inventory size and other linguistic characteristics and features has to be taken into consideration (Kelih, 2015). The focus of the presentations therefore lies on interrelations between the phoneme inventory size and phonological, morphological, syntactical, lexical and semantic features. Finally, based on corpus data and the Swadesh-list for Slavic languages the empirical impact (in addition to Nettle, 1995; 1998) of the phoneme inventory size will be shown in detail. As a result it can be stated that the phoneme inventory size mainly influences phonological level and to a lesser extent other hierarchically higher levels.

REFERENCES

Atkinson, Q. D. (2011). Phonemic diversity supports a serial founder effect model of language expansion from Africa. *Science*, *332*(6027), 346–349.

Bybee, J. (2011). How plausible is the hypothesis that population size and dispersal are related to phoneme inventory size? Introducing and commenting on a debate. *Linguistic Typology*, *15*(2), 147–153.

Cysouw, M., Dediu, D., & Moran, S. (2012). Comment on "Phonemic Diversity Supports a Serial

Founder Effect Model of Language Expansion from Africa". *Science*, *335*(6069), 656–657.
Kelih, E. (2015). *Phonologische Diversität – Wechselbeziehungen zwischen Phonologie, Morphologie und Syntax*. Frankfurt am Main: Peter Lang.
Nettle, D. (1995). Segmental inventory size, word length, and communicative efficiency. *Linguistics*, *33*(2), 359–367.
Nettle, D. (1998). Coevolution of Phonology and the Lexicon in Twelve Languages of West Africa. *Journal of Quantitative Linguistics*, *5*(3), 240–245.
Sproat, R. (2011). Phonemic diversity and the out-of-Africa theory. *Linguistic Typology*, *15*(2), 199–206.

Quantitative Evaluation of Morphological Ambiguity of European Languages

Eduard Klyshinsky[1] and Varvara Logacheva[2]
[1]Higher School of Economics, Moscow, Russia; [2]University of Sheffield, UK

The evaluation of the language complexity needs both theoretical background and numerical methods. The former provides an explanation of different language phenomena (Dahl, 2004). On the other hand, numerical evaluation allows finding new patterns and rules of the language development. However, some quantitative evaluation methods are disputable. For example, the Kolmogorov complexity theory defines a string's complexity as the complexity of model which can generate this string. Kolmogorov complexity coefficient is usually approximated with the archiving coefficient – i.e. the ratio between the size of initial text and its archived version. This method is sensitive to a wide range of features of a considered language as well as a given language, and there is no way to understand how much all the factors contributed to the final result. Thus, there is a great need in more verifiable numerical methods, as the one described in (Schepens et al. 2013; Lindh-Knuutila, 2014).

In this paper, we introduce a new taxonomy for morphological ambiguity based on six types of ambiguity. The first type consists of unambiguous words. The second type includes word forms that correspond to more than one set of grammatical parameters, e.g. the German '*wohnen*' can be infinitive or plural 3rd person. Words of the third type can belong to more than one part of speech, e.g. the English 'close' can be a verb, a noun, or an adjective. The fourth type consists of word forms that belong to different words, e.g. the Russian 'вина' ('*vina*' – '*guilt*') can be either noun 'вина' ('*vina*' – '*guilt*') in singular, nominative case or noun 'вино' ('*vino*' – '*wine*') in singular, genitive case. The fifth type of word forms have ambiguities in both part of speech and initial form, e.g. the French '*est*' can be the noun '*est*' ('*east*') or the verb '*être*' ('*to be*'). The sixth type are out-of-vocabulary words.

We analyzed monolingual news wire texts for seven European languages: English, French, Spanish, Italian, German, Russian, and Polish. We also used the parallel News Commentary corpus for French, German, and Spanish. We calculated the distribution of words among introduced types of ambiguity and found out that languages belonging to the same language family demonstrate higher correlation than languages from different families. The computed results confirm the intuition that the English language has an extremely high percentage of POS-ambiguous words, while Slavic languages demonstrate their inflectional nature.

In our experiments, we found out that varying such experimental parameters as number of grammatical features, size of the corpus or vocabulary do not lead to higher similarity rate of different distributions. Thus, we can state that the distribution of words of a text among type of

ambiguity classes reflects some inner properties of a language. Languages related to the same family demonstrate more similarity than other languages; therefore, the introduced distribution can be used as a numerical evaluation of the languages similarity.

Investigating the Chronological Variation of Lyrics of Popular Songs Through Lexical Indices

Yuichiro Kobayashi, Misaki Amagasa and Takafumi Suzuki
Toyo University

Popular songs can be regarded as a fine representation of modern society and culture. In particular, the lyrics of popular songs are the most important aspect for understanding the sense of values and linguistic sensitivity in a given generation and community (Suzuki & Hosoya, 2014). Although some sociological and linguistic studies have examined the language used in lyrics, most of them focused on specific artists. However, it is necessary to conduct the comprehensive description of a wider variety of popular lyrics in multiple generations for observing the macroscopic shifts and changes in cultural trends.

The purpose of the present study is to investigate the chronological variation of popular Japanese songs using stylometric techniques. This study compares the use of lexical indices found in Japanese hit songs over the past three decades. From the viewpoint of computational stylistics, the analysis of lyrics can be a case study of short texts whose running and distinctive words are quite limited. It can also contribute to the development of a methodology for investigating the chronological variation of linguistic styles.

This study draws on the lyrics of 773 songs, which appeared on the Oricon annual top 20 single hit chart between 1977 and 2012. The text data of these lyrics were obtained from Uta Net (http://www.uta-net.com) and Uta Map (http://www.utamap.com), two databases for popular Japanese songs. The frequencies of five different types of lexical indices, namely (a) number of words, (b) parts-of-speech, (c) word types, (d) character types, and (e) vocabulary level were calculated using jReadability (http://jreadability.net/). The size of the data was about three hundred thousand words.

Multiple regression analysis was conducted to explore the chronological change in the frequencies of lexical indices. Representing words occurring in lyrics as their lexical indices, stylistic differences among texts can be focused on more clearly than thematic differences (Tabata, 2004). In this study, 26 types of lexical indices were used asexplanatory variables for multiple regression analysis and the release year of each lyric was used as a criterion variable. Moreover, Akaike information criterion (AIC) was checked in order to identify the lexical indices that distinguish the lyrics between different generations as well as to avoid multicollinearity issues.

The results of the present study suggest that the frequencies of word types and character types greatly changed around 1990. At that time, frequencies of foreign vocabulary and *katakana* characters fell sharply and those of Sino-Japanese words and Chinese characters rose in the lyrics of popular songs. The frequent use of loan words written in foreign vocabulary or katakana characters in Japanese lyrics was considered fashionable during the 1980s, but the trend has become dramatically outmoded due to changes in linguistic sensitivity in Japan after the 1990s. Furthermore, the frequencies of subordinating verbs, such as adverbs and adnominal adjectives, gradually increased year after year, whereas those of nouns decreased. The findings suggest that a turning point in cultural trends in Japan corresponds with the burst of the bubble economy around

1990.

Dynamic Development of Vocabulary Richness of Text

Miroslav Kubát and Radek Čech
University of Ostrava, Czech Republic

Any text is a sequence of various units such as phonemes, words, sentences, etc. These units can be used for a measurement of many properties of text, e.g. vocabulary richness, thematic concentration, entropy, descriptivity. These text features are usually expressed by one resulting value, mostly in the interval < 0; 1 > (cf. Popescu et al., 2009; Tuzzi et al., 2010; Čech et al., 2014).

To extend this approach, we propose to analyse also a dynamic development of the given text property. It brings the more detailed view on the text characteristics; for instance, two texts can have the same resulting value of the index representing the given property (e.g. type-token ratio), however, they can significantly differ with regard to the development of the property in the text.

Specifically, the vocabulary richness of text is analysed in our study. As a method, the moving average type-token ratio (MATTR) is used due to its text length independence (Köhler & Galle, 1993; Covington & McFall, 2010; Kubát & Milička, 2013). To analyse the development of MATTR, the text must be first segmented into parts. Three different ways of text segmentation are chosen and analysed texts are divided into the following parts: (a) sequences of arbitrarily large number of tokens - so called "moving windows", (b) paragraphs, and (c) chapters. For each part of text, the value of the MATTR is computed, then, the values are put in a graph and subsequent values are connected by a line. The length of the line between subsequent values is computed (cf. Hřebíček, 2000) and, further, used for the analysis of the development of MATTR. Finally, the obtained differences among texts are tested by Wilcoxon-Mann-Whitney test (cf. Čech, 2015). The corpus consists of Czech texts of various genres.

Keywords: MATTR, sequence of text, language unit, vocabulary richness, text development

REFERENCES

Čech, R. (2015). Development of thematic concentration of text (in Karel Capek's books of travel). *Czech and Slovak Linguistic Review*, *1*(2015), 8–21.

Čech, R., Popescu, I.-I., & Altmann, G. (2014). *Metody kvantitativní analýzy (nejen) básnických textů*. Olomouc: Univerzita Palackého v Olomouci.

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.

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

Köhler, R., & Galle, M. (1993). Dynamic Aspects of Text Characteristics. In L. Hřebíček & G. Altmann (Eds.), *Quantitative Text Analysis* (pp. 46–53). Trier: WVT.

Kubát, M., & Milička, J. (2013). Vocabulary Richness Measure in Genres. *Journal of Quantitative Linguistics*, *20*(4), 339-349.

Popescu, I.-I., Altmann, G., Grzybek, P., Jayaram, B. D., Köhler, R., Krupa, V., Macutek, J., Pustet, R., Uhlirova, L., & Vidya, M. N. (2009). *Word frequency studies*. Berlin/New York: Mouton de Gruyter.

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

Quantitative Properties of the Dialogue in Brothers Karamazov

Haotian Li
Zhejiang University, China

The world is filled with a myriad of information, and language is the carrier of information. In Applied linguists' view, through the language, we can observe the real world. Content analysis, as a branch of applied linguistics, is a research method to analyze the information carried in language and observing the real world, mainly used in the field of sociology, psychology. Literary works as a means to reflect the real world by writer, can also use this method.

Fyodor Mikhailovich Dostoevsky is world famous Russian writer, and his last work – the *Brothers Karamazov* is his master piece. Dostoevsky is good at describing the tragic fate of the characters through a lot of contradictory concepts, such as life and death, smart and stupid, the opposition between truth and falsehood, perfect and the imperfect, to make the hero faces the major decision of life, and revealing the protagonist's real personality under the critical state. This technique is used throughout many of his major works. The *Brothers Karamazov's* contradictory performance is represented in "the existence of god", "patricide motivation and action".

Based on the theory of content-analysis, and through two angles of conceptual analysis and relationship analysis, mathematic method is used to study the dialogue in *Brothers Karamazov* including the contradictory concept, the relationship between concepts. This paper tries to study the extent to which the contradictory language contributes to the tragedy ending of the story and how it affects the reader's interetation, through the analyses of dialogue between characters.

Models of Noisy Channels that Speech Gets Over

Jiří Milička and Karolína Vyskočilová
Charles University, Prague

Language, as a method of communication, is a robust system that is able to overcome various noisy channels (in the Shannonian sense). The contemporary quantitative linguistic research takes into account the theory of information, understanding that various language features function as a redundancy, which is necessary to get over the noise. Thus using an appropriate model of typical noisy channels is necessary to model both language and communication. This paper provides empirical testing of some existing models and proposes several new ones. The paper consists of four parts:

1. The first part explores some well-known models (Shannon, 1948; Gilbert, 1960) of the noisy channel and introduces a few new ones (1st author 2015).
2. The models are compared with some typical real noise which interferes with communication between people and their parameters are fitted to the empirical data that were obtained

during building a spoken corpus (2nd author 2013).

3. The ability of a recipient to filter out various kinds of noise is predicted according to the theory of information.
4. The theoretical predictions are compared with perception tests.

REFERENCES

Gilbert, E. N. (1969). Capacity of a Burst-Noise Channel. *Bell System Technical Journal*, *39*(5), 1253–1265.

Milička, J. (2015). *Teorie komunikace jakožto explanatorní princip přirozené víceúrovňové segmentace textů* (The Theory of Communication as an Explanatory Principle for the Natural Multilevel Text Segmentation). PhD thesis, Prague: Charles University.

Shannon, C. E (1984). A Mathematical Theory of Communication. Bell System Technical Journal, 27(3), 379–423.

Vyskočilová, K. (2014). *Tvorba specializovaného korpusu banátské češtiny a jazyková analýza vybraných jevů* (Building Specialized Corpus of Banat Czech and Linguistic Analysis of Selected Features). Master thesis, Prague: Charles University.

The Mathematics of Meaning

Hermann Moisl
Newcastle University, U.K.

Mathematics and statistics have been successfully applied to modelling of the formal aspects of language ranging from phonetics through phonology, morphology, and lexis to syntax. Their application to modelling of natural language meaning has, however, been limited to elementary concepts in set and function theory as part of of truth-conditional semantics. The present discussion argues that the scope for mathematical modelling of natural language meaning is far greater than that, and proposes a model based on the mathematical theory of nonlinear dynamical systems.

The discussion is in four main parts:

The first part locates the proposed model in the context of current approaches to natural language semantics in cognitive science, philosophy of language, and linguistics. In terms of the concept of supervenience in the philosophy of science, the model is intended as an implementation-level account underlying cognitive-level models of linguistic meaning.

The second part briefly describes the essentials of nonlinear dynamical systems theory and motivates its application to modelling of natural language meaning. The motivation derives from recent neuroscientific results in which the brain is seen as a physical dynamical system with nonlinear neural dynamics and extensive feedback, whose operation is best captured by the mathematical theory of dynamical systems.

The third part describes a model of sentence meaning based on a specific class of mathematical dynamical systems: artificial neural networks (ANN). ANNs were designed to model neural morphology and dynamics, and are therefore an appropriate choice for the present proposal; the subclass of ANNs used here feature nonlinear activation functions, feedback, and point attractor dynamics. The most notable feature of the proposed model is that it eliminates recursive phrase structure as a basis for the construction of sentence meaning, which is fundamental to current

cognitive approaches to natural language semantics, and replaces it with trajectories in high-dimensional phase space.

The fourth part presents results derived from a computer simulation of the proposed artificial neural network model.

Context Predictability Methods in Twitter Social Network Analysis

Tataina Nikulina and Elena Yagunova
St.-Petersburg State University, Russian Federation

Today the social networks play undoubtedly important role: as a source of information, as a mean of mobilization and as an environment for the discussion and reflection. Nowadays investigations of social attitudes take new turns using language technologies (text analysis, data mining, opinion mining, information extraction and so on). This investigation is based on Twitter as this network can give us the most coherent text subcorpora. The microblogging network was chosen as it is a platform of users' rapid reaction. The language of Twitter is short (messages have the fixed length – 140 symbols) and the user is not able to edit a message (further: tweet). All these features create the virtual space with natural living language. Therefore there are difficulties in proceeding because of this specific language. Main challenge is connected with a huge amount of grammatical mistakes, abbreviations, authorial neologisms, slang. Another problem is to define jokes, sarcasm, trolling etc. The focus of the following research is concerned with identification of Twitter language specifics, collocations and drawing an analogy with spoken language. We tried to define if Twitter language is surprisal or predictable. We used contextual predictability (entropy measure, surprisal measure and some measures of connectedeness, and other characteristics) and have come to the preliminary results, ex., high number of collocations were found (bigrams) but still huge number of unpredictable ones happened. Moreover we found around 2 hundreds high- frequency neologisms that also require analysis.

Table 1. Example of collocations.

bigram	*probability*	*entropy*	*translation*
половины россиян	0.25	11.81	(half of Russians)
ветке метро	0.375	11.43	(metro line)
призрак оперы	0.33	16.43	(the Phantom of the Opera)
сыграют матч	0.33	12.99	(play the match)
чемпионка россии	0.33	7.87	(championess of Russia)
майкл джексон	0.4375	16.20	(Michael Jackson)
дэвид кэмерон	0.4	17.01	(David Cameron)
мид рф	0.40	9.28	(abbr. the Ministry for Foreign Affairs of Russian Federation)
снимут фильм	0.5	12.64	(shoot a film)
заседаний госдумы	0.5	12.4	(sederunt of State Duma)

bigram	*probability*	*entropy*	*translation*
дальнего востока	0.69	13.52	(Far East)
нанести удар	0.67	13.69	(Strike a blow)
народной республики	0.7	10.88	(people's republic)

One of our hipotesis is the changing tweets language at the time of the hard sociopolitical events. We have analyzed specific data collected in February-April 2014 – the time of hard socio-political events in Ukraine and Olympic games in Sochi, the most data was concerned to these events. This fact assigns users' activity and rapid spread and frequent usage of collocations (and also neologisms).The research is in progress. We are collecting data for large corpus to make a Twitter frequency dictionary to unify and simplify further proceeding with probabilistic methods. We hope that next steps of research will allow us to solve lots of tasks connecting with preproceeding using entropy, surprisal, contextual predictability, and other characteristics.

Keywords: Context predictability methods, social networks, text analysis, collocations, tweets language

---

Harmony in Diversity: Language Data Codes in English-Chinese Poetry Translation – Quantitative Study on Shakespearean Sonnets Translation

Xiaxing Pan[1], Xinying Chen[2] and Bingli Liu[1]
[1]National Huaqiao University; [2]Xi'an Jiaotong University, China

---

The translation of poetry is a very complex process. The paradoxical nature of untranslatability and translatability of poetry has been discussed by many famous scholars, such as Robert Frost, who purported "poetry is what gets lost in translation", and Susan Bassnet, who advocated "poetry is what we gain in translation" (Hai, 2005). However, the both drastic opinions addressed one common theme that poetry translation is no more a repetition of the original works than a reproduction. Therefore, there are both similarities and discrepancies between the translated works and the original pieces (Huang, 1998; Bian, 1989; Zheng, 2001; Hai, 2005; etc.).

Furthermore, which are these similarities and discrepancies specifically? How can we find and test them in a clear and objective manner? For answering these questions, we quantitatively compared 20 English poetries (randomly selected from Shakespearean sonnets, which are No. 5, 8, 11, 17, 28, 36, 52, 60, 65, 69, 91, 93, 104, 109, 110, 124, 133, 138, 142, and 151) with their Chinese translated versions (from four different translators, which are Zongdai Liang, An Tu, Zhengkun Gu, and Minglun Cao).

We observed several quantitative linguistic features from the perspectives of vocabulary, word frequency distribution, and POS distribution of these poetry texts, such as the a-*index*, *writer's view*, *index b of the Zipf's law* (Condon, 1928; Zipf, 1935; Popescu & Altmann, 2006, 2007; Martináková et al., 2008; Tuzzi et al., 2010a, 2010b; Popescu et al., 2012), etc. After running ANOVA test and comparing these features, we utilized Alternative Least Square Scaling (ALSCAL) analysis and Cluster analysis (Biber, 1993; Ji, 2013) by using these features as parameters to capture the similarity or difference between these poetry texts.

Our preliminary results showed that:

1) There are not many differences in terms of the vocabulary size between the translated poetries and original pieces according to their *a-index*.

2) According to the *index b of the Zipf's law* and *writer's view*, there are significant differences between the original pieces and the translated works, while there are no statistical significant differences between the translated works of four authors. It shows that the word frequency distribution and POS distribution features of translated works are clearly different from that of the original texts. The differences are probably due to the general language difference of Chinese and English.

3) The results of ALSCAL analysis and clustering analysis illustrated that by using the POS distribution feature as the parameter we can distinguish the original pieces from the translated works, as well as distinguish different groups of translators: professional translators and professional poets.

Keywords: Shakespearean sonnets, poetry translation, quantitative linguistics

REFERENCES

Biber, D. (1993). The Multi-Dimensional Approach to Linguistic Analyses of Genre Variation: An Overview of Methodology and Findings. *Computers and the Humanities*, *26*, 331–345.

Bian, Zh. (1989). Merits and demerits of the poetry translation since 'May 4th'. *Translations*, *4*, 182–88.

Condon, E.U. (1928). Statistics of Vocabulary. *Science*, *67*, 300.

Hai, A. (2005). The translation of poetry by the translator-cum-poet. *Chinese Translator Journals*, *6*, 27–30.

Huang, W. (1988). Influences from American and British on May 4th new poetry. J*ournal of Beijing University*, *5*, 25–38.

Ji. M. (2013). *Exploratory Statistical Techniques for the Study of Literary Translation*. Lüdenscheid: RAM-Verlag.

Martináková, Z., Mačutek, J., Popescu, I., & Altmann, G. (2008). Some Problems of Musical Texts. *Glottometrics*, *16*, 80–110.

Popescu, I., & Altmann, G. (2006). Some Aspects of Word Frequencies. *Glottometrics*, *13*, 23–46.

Popescu, I., & Altmann, G. (2007). Writer's View of Text Generation. *Glottometrics*, *15*, 71–81.

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

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

Tuzzi, A., Popescu, I., & Altmann, G. (2010b). The Golden Section in Texts. In A. Wilson (Eds.), *Emperical Text and Culture Research 4: Dedicated to Quantitative Emperical Studies of Culture* (pp. 30–41). Lüdenscheid: RAM-Verlag.

Zheng, H. (2001). On dissimilation and optimization of target language. *Chinese Translator Journals*, *3*, 3–7.

Zipf, G. K. (1935). *The Psychobiology of Language*. Boston: Houghton-Miflin.

Statistical Distributions of Lexeme Frequencies in a Text Corpus. A Comparative Analysis

Adam Pawłowski, Maciej Piasecki and Krzysztof Topolski
University of Wrocław, Poland

In the present paper will estimate and compare statistical distributions of vocabulary in a big text corpus of Polish and in its subcorpora. The subcorpora under study will be excerpted from the corpus using two kinds of criteria: formal (grammatical classes) and semantic (inter alia proper names). Additionally, the afore–mentioned subcorpora will be compared (using the same statistical method) with a randomly generated (and equal in size) reference corpus. We shall test a hypothesis which states that statistical distributions in the entire corpus and in its subcorpora are different with regard to type and/or parameter values (in case of similar models).

The advanced hypothesis relies on the generally accepted assumption that language is a coherent and efficient system of communication and knowledge representation, yet its internal structure reveals many inconsistencies on the level of grammar and vocabulary. Contrary to programming or artificial languages, that are constructed according to some well-defined principles, natural language is the effect of a long and complex process of phylogenetic development guided by human adaptive behaviour and depending on various environmental and, at the recent stages of the evolution, cultural factors. In this perspective language is a "system of subsystems" susceptible of analytical (quantitative and qualitative) descriptions of every particular layer (i.e. grammar, vocabulary, subsets of the vocabulary etc.). Consequently, we consider the concepts of "general population" or "general language" for a given ethnolect as useful for theory, however, as lacking convincing empirical evidence.

The statistical analysis will contain, among others, fitting the empirical word tokens and word type distributions to the one of the theoretical distributions from the classes such as lognormal, Chitashvili's extended generalized Zipf's law, class of generalized inverse Gauss-Poisson structure type distribution and mixture distributions. To evaluate goodness of fit we use standard chi-square test, as well as distribution of specialized tests dedicated to this type of analysis. To discover the structure in multivarate data sets we will use hierarchical clustering and classification trees.

The research will be carried out on the material of a balanced corpus of Polish texts created and administrated by the team of the language engineering group of the Wrocław University of Technology (Department of Computational Intelligence). Excerption of data will be carried out using NLP programming tools prepared in the framework of the CLARIN-PL consortium.

REFERENCES

Dahl, Ö. (2004). T*he growth and maintenance of linguistic complexity*. Amsterdam: John Benjamins.

Lindh-Knuutila, T. (2014) *Computational Modeling and Simulation of Language and Meaning: Similarity-Based Approaches. Aalto University publication series* DOCTORAL DISSERTATIONS 49/2014.

Schepens, J., Van der Slik, F., & Van Hout, R. (2013). Learning Complex Features: A Morphological Account of L2 Learnability. *Language Dynamics and Change*, *3*(2), 218–244.

Personality Prediction in Facebook Status Updates Using Multilevel *N*-gram Profiles (MNP) and Word Features

Vassiliki Pouli, Ephtychia Triantafyllou and Georgios Mikros
National and Kapodistrian University of Athens, Greece

The aim of this paper is to develop a methodology for predicting the author's personality using short written texts. A lot of research in this field incorporates texts containing at least 50 or more words. However, most texts in social media and microblogging platforms are shorter (e.g. Facebook status updates contain on average 5-7 words) and further research should also focus in extracting personality information from these tiny text fragments.

This paper attempts to bridge this gap by determining efficient models for personality prediction using the information from Facebook status updates posted on a user profile.

To this end, 7000 user status updates were collected through the MyPersonality.org project[1] (Kosinski et al., 2015) along with users' personality scores user for verification of the results in the final analysis steps. By using Natural Language Processing methods, several grammatical and syntactical features were extracted to verify their importance on the personality traits.

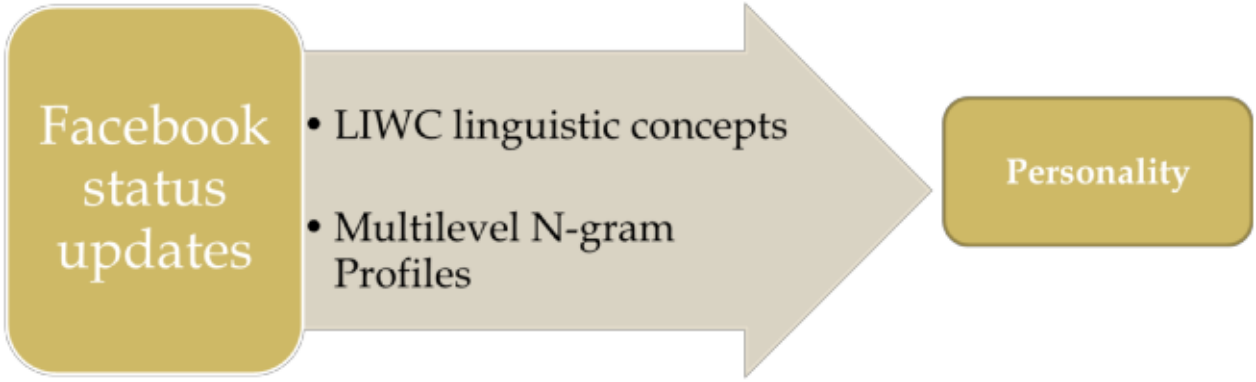

Fig. 1. Personality prediction through text analysis using Facebook status updates.

The analysis IS/WAS comprised of two approaches (Figure 1). The 1st one exploits the word frequencies and quantifies the user status updates on psychological dimensions using LIWC (Francis & Pennebaker, 1993). The 2nd approach (Mikros & Perifanos, 2013) dynamically fragments the text into multi-level *N*-grams of increasing length forming a contiguous sequence of *N* text pieces (Figure 2). Multi-level refers to both levels of words and characters, while increasing length considers fragments of $n=2$ and $n=3$. The *N*-grams offer extra lexical most important stylometric features to pinpoint an author from the written text since are language-independent.

1 http://mypersonality.org/wiki

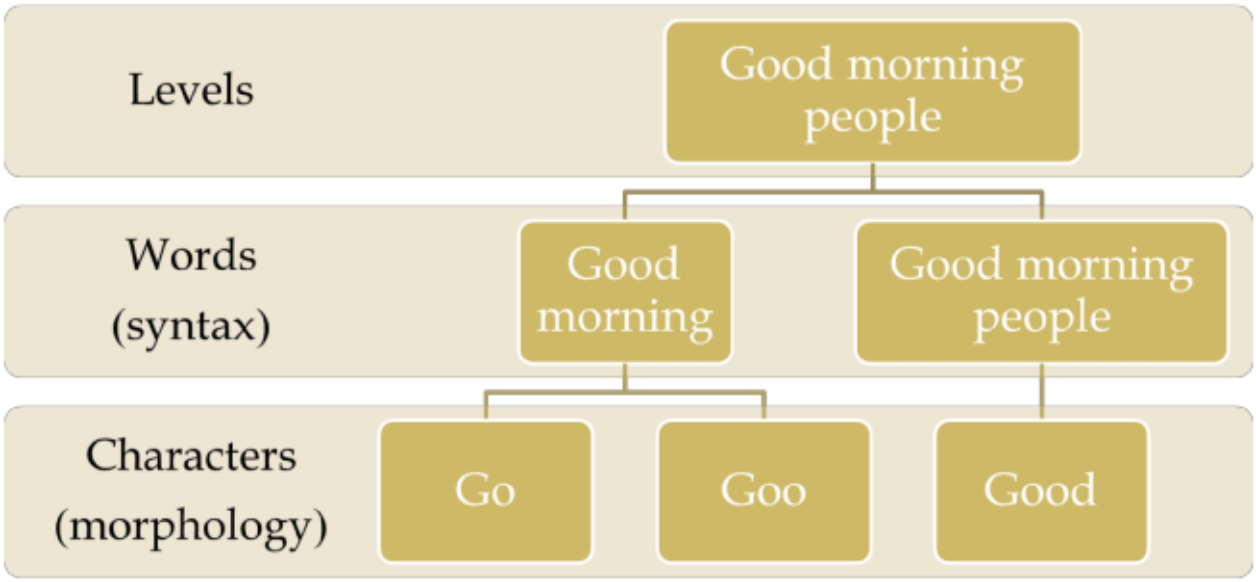

Fig. 2. Multi-level *N*-grams Profiles.

N-grams have been used successfully in the past for paternity identification in the investigations of Koppel and Ordan (2011) and Grieve (2007), among others. However, they have not been applied yet to Facebook status-updates where the text is dramatically shorter. *N*-grams computation was done by means of Perl and *N*-gram Statistic Package (NSP) (Banerjee, 2003) and then frequencies were normalized to avoid asymmetric text length calculations.

In our analysis, LIWC and multi-level *N*-grams of increasing length were used as features in the linear regression machine learning method to predict the user personality. The analysis proved to be a very powerful tool for the personality prediction while providing interpretable and language accepted models. Each personality trait was significantly correlated with particular LIWC linguistic characteristics while *N*-grams were able to determine with 29% accuracy the "Extroversion" and "Conscientiousness" personality categories. Even with extremely short users' status updates text (5-7 words on average) the accuracy in result prediction is quite high compared with other surveys that use larger texts (such as emails, blogs etc.) where the text normally extends to more than 30 words and can reach thousands of words.

The results can be used as a source of information for a better personality understanding whilst also offering more personalized services in social networks.

ACKNOWLEDGMENT

We would like to thank the MyPersonality.org project for providing the data for our research.

REFERENCES

Banerjee, S., & Pedersen, T. (2003). The design, implementation, and use of the Ngram Statistic Package (pp. 370–381). *Proceedings of the Fourth International Conference on Intelligent Text Processing and Computational Linguistics*, Mexico City, Mexico, February 2003.

Francis, M. E., & Pennebaker, J. W. (1993). LIWC: Linguistic Inquiry and Word Count. *Technical Report*. Dallas, TX: Southern Methodist University.

Grieve, J. W. (2007). Quantitative authorship attribution: an evaluation of techniques. *Literary and Linguistic Computing*, *22*(3), 251–270.

Koppel, M., & Ordan, N. (2011). Translationese and its dialects (pp. 1318–1326). *Proceedings of*

the 49th Annual Meeting of the Association for Computational Linguistics: Human Language Technologies (ACL-HLT '11), Portland, Oregon, USA, June 19 – 24, 2011.

Kosinski, M., Matz, S., Gosling, S., Popov, V., & Stillwell, D. (2015) Facebook as a Social Science Research Tool: Opportunities, Challenges, Ethical Considerations and Practical Guidelines. *American Psychologist*, *70*(6), 543–556.

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

Using Physics to Model Language Diffusion: Benefits and Challenges

Katharina Prochazka and Gero Vogl
University of Vienna, Austria

Investigating language change is an often difficult and labour-intensive process further complicated by the fact that many linguistic concepts cannot be quantified. Even so, efforts have been made to apply methods from the natural sciences to linguistics. Mathematical models are especially attractive for the study of language spread because they widen the timescale: If fed with enough data from the past, it should be possible–with all due caution–to predict language distribution in the future. This is particularly enticing for sociolinguistic research on minority languages and their decline through language shift (speakers abandoning use of one language for another).

We discuss the application of physics to the study of language spread: Can language diffusion (the transport of something immaterial) be modelled by techniques used in solid state physics to describe the transport of matter (physical diffusion)?

Using data on language use in Austria, we show that physical models make it possible to study language spread as a quantitative phenomenon to get a complete overview of past and possibly future spatio-temporal development. The results might then provide starting points for intervention measures to influence language spread. Thus, physical models offer a large-scale complement to the traditional sociolinguistic smaller-scale study of language shift.

However, the limitations and challenges in choosing adequate physical models for linguistic applications must also be addressed. In the past language spread (increase and decline in a language's use) has been most often described by differential equations based on Fick's laws of physical diffusion (Kandler, 2009). These equations model the behaviour of a proportion of speakers in the population. Here we propose the combination of these classic diffusion models with so-called agent-based models (Schulze et al. 2008). Agent-based models simulate the actions of individual speakers, leading to a more realisti,c description of local effects influencing language use. Furthermore, we argue that any language shift model should primarily rely on empirical data rather than including parameters that cannot be reliably measured (e.g. prestige of a language). Only when empirical data over a certain time and space scale are available and used for calibration can the strength of physical models be fully exploited to advance the study of language spread.

Keywords: Diffusion, language spread, language shift, agent-based modelling

REFERENCES

Kandler, A. (2009). Demography and Language Competition. *Human Biology*, *81*(2-3), 181–210.
Schulze, C.,Stauffer, D., & Wichmann, S. (2008). Birth, Survival and Death of Languages by Monte Carlo Simulation. *Communications In Computational Physics*, *3*(2), 271–294.

Diffusion Model for Language Shift in Austria

Katharina Prochazka and Gero Vogl
University of Vienna, Austria

Language shift occurs when speakers give up use of one language for another. This shift can lead to a loss in linguistic and cultural diversity and eventually to language extinction. Therefore, language shift needs to be monitored to understand the reasons and target supportive measures. Quantitative monitoring is possible through the use of computer models.

We present a spatio-temporal model for language shift in Carinthia, Austria. Historically, there have been two prevailing languages in Carinthia, Slovenian and German. However, use of Slovenian has been steadily declining (Austrian census 1880 to 2001, Busch, 2001; Priestly, 2003). We model this decline using a simulation technique based on an approach originally developed for ecology (Vogl et al., 2008; Smolik et al., 2010; Richter et al., 2013). In our model, the surveyed geographic area is divided into 1x1 km2 lattice cells and for each cell, the number of speakers of one language is simulated based on language-influencing factors such as the presence of schools, parish language or the closeness to other settlements where the same language is spoken. The model is built and calibrated using data on language use from the Austrian census which is available from 1880. This allows us to satisfactorily describe the evolution of language use in Austria over time and space for which we present results. Additionally, we will discuss the feasibility of predictions using our model and its application to other language shift situations.

Keywords: Diffusion, language extinction, language shift, agent-based modelling, sociolinguistics

REFERENCES

Austrian census 1880 to 2001, published by k.k. Statistische Central-Commission/Österreichisches Statistisches Zentralamt/Statistik Austria, Vienna.

Busch, B. (2001). Slovenian in Carinthia – a sociolinguistic survey. In G. Extra & D. Gorter (Eds.), *The Other Languages of Europe: Demographic, Sociolinguistic and Educational Perspectives* (pp. 119–137). Clevedon: Multilingual Matters.

Priestly, T. (2003). Maintenance of Slovene in Carinthia (Austria): Grounds for guarded optimism? *Canadian Slavonic Papers*, *45*(1/2), 95–117.

Richter, R., Dullinger, S., Essl, F., Leitner, M., & Vogl, G. (2013). How to account for habitat suitability in weed management programmes? *Biological Invasions*, *15*(3), 657–669.

Smolik, M. G., Dullinger, S., Essl, F., Kleinbauer, I., Leitner, M., Peterseil, J., Stadler, L.-M., & Vogl, G. (2010). Integrating species distribution models and interacting particle systems to predict the spread of an invasive alien plant. *Journal of Biogeographie*, *37*(3), 411–422.

Vogl, G., Smolik, M., Stadler, L.-M., Leitner, M., Essl, F., Dullinger, S., Kleinbauer, I., & Peterseil,

J. (2008). Modelling the spread of ragweed: Effects of habitat, climate change and diffusion. *The European Physical Journal - Special Topics*, *161*, 167–173.

Quantitative Interrelations of Properties of the Complement and the Adjunct

Haruko Sanada
Rissho University, Japan

Aim and problems of the study: The present study is one of a series of empirical studies on Japanese valency. The present study focuses on (1) the proportion of complements and adjuncts in the clause, (2) relationships between complement and adjunct numbers and the position of clauses in the sentence, and (3) relationships between complement and adjunct lengths and the position of clauses in the sentence, and (4) differences between complement and adjunct lengths in the non-embedded clause and those in the embedded clause. The number of adjuncts is not related to the number of complements, but both are related to the position of the clause in the sentence and related to the type of clause, whether it is non-embedded or embedded.
Data: We employed the Japanese valency data-base (Ogino et al., 2003), which is the same as the one employed in our last studies (2012, 2013, 2014a, 2014b, 2015, to appear a, to appear b). For the present study, 240 sentences were extracted from the valency database, including 243 clauses containing the verb "meet". From the 243 clauses we obtained 348 complements and 174 adjuncts. We also obtained a number of morphemes using a morphological analyzer.
Results and conclusions: We investigated the four problems above with statistical tests and conclude as follows: 1. The number of adjuncts is almost stable regardless of the number of complements. The average number of adjuncts is significantly less than the average number of complements. 2. Relationships between the length of clauses containing the verb "meet" (i.e. length of arguments or morphemes) and the position of the clauses in the sentence have a decreasing function. Namely, the 12015/09/09 Abstract clause is longer in the beginning of the sentence, and according to the position of the clause closing to the end, the shorter the clause in the arguments or in the morphemes. This tendency is the same for relationships with sentences containing various verbs. 3. To make the clause shorter according to the position of the clause closing to the end, two options are possible. One is to make the number of arguments smaller, the other is to make arguments themselves shorter in morphemes. If the number of adjuncts is relatively stable, the length of adjuncts in morphemes must be less. Compared to the functions of the length in morphemes, functions of the number of complements or adjuncts fit the curves better. This was interpreted to mean that omitting arguments from the clause and making the number of arguments smaller is more preferable than making arguments themselves shorter in morphemes. 4. Complements in the embedded clause are significantly shorter or smaller in number than ones in the non-embedded clause. Embedded clauses contain significantly less adjuncts than non-embedded clauses. However, the length of adjuncts in embedded clauses and in non-embedded clauses does not significantly differ. We must investigate these problems with other verb data in the future.

Keywords: Valency, sentence structure, length, frequency, complement, adjunct, position in the sentence, Japanese, synergetic linguistics

Stylistic Evolution of US Political Speeches

Jacques Savoy
University of Neuchatel, Switzerland

Over time and space, the language is changing in all of its underlying dimensions, at the phonological, lexical, syntactical or semantics level. However, the rate of change may vary and it is still difficult to estimate it precisely (Juola, 2003; Rule et al., 2015). This communication focuses on the stylistic (Biber & Conrad, 2009) evolution of US governmental speeches. To achieve this, we have created a corpus composed of 57 inaugural speeches and 229 State of the Union addresses from 1789 through 2016. Within these corpora, the context of each intervention is relatively stable (similar audience, form, objectives) allowing a good basis for comparisons. To analyze the frequencies of various style markers, we have used statistical methods and all reported results are statistically significant. Moreover, our main findings will be illustrated with concrete examples.

As a first style marker, we consider the mean sentence length. Over time, we can see a clear decrease in the mean number of words per sentence with, for example, a mean value between 65 words/sentence (inaugural speech delivered by Washington or Adams) to an average of around 23 with Obama. Over the years, the US presidents have adopted a simpler and more direct style to obtain a better understanding of their purposes (and not to go over the "heads of the Congress"). Complex explanations with various specifications and limits require longer sentences that were the norm during the 18 th or 19 th century.

As "Politics is the art of the vocabulary" (B. Constant), the lexicon was more intensively studied. As another style marker, the mean number of letters per word tends to decrease over the years. More recent presidents prefer using more common and shorter words than complex and longer lexical terms. This result indicates that modern presidents adopt a simple vocabulary to be more clearly understood by the citizens themselves (Tulis, 1987). As a related measure, the type-token ratio (TTR) tends to decrease over time, however this reduction is not always statistically significant.

According to recent psychological studies (Pennebaker, 2011), the frequencies of pronouns and their variations can reveal the speaker's emotional state, profile, and even some of his/her psychological traits. With our corpora, the relative frequency of the pronoun we tends to increase over the years. Under the Obama presidency, this value raises to 6.5% (inaugural speeches), compared to 0.2% with Washington (or 0.6% with Lincoln). For other pronouns, we cannot detect any clear tendency over time.

After applying a part-of-speech (POS) tagger, we can analyze the distribution of the different grammatical categories. Over the whole corpus, the relative frequencies of determiners and prepositions tend to decrease over time. On the other hand, the number of punctuation symbols has the tendency to increase over the years. These aspects are related to thereduction of the mean sentence length. We have also observed an increase in the number of adverbs used by the US presidents, with a peak under Raegan's presidency.

As another way to analyze the style and the lexical choices, we have applied the LIWC 2007 semantic classes (Tausczik & Pennebaker, 2010). Based on a general label (e.g, Posemo), the LIWC's system regroups all words or word stems belonging to this positive emotion class (e.g., joy, pretty, support, happ*, etc.).

Based on those semantic groups, we do not detect any significant variations of general positive or negative emotions, except a slight increase in the Sad class (e.g., missing, tears, suffering, suffer). For the class *Tentat* (e.g., some, perhaps, most, seemed), we observe a large variability among presidencies, with high values for Adams (4.2%, 1797, inaugural speech) or Lincoln (3.9%, 1861)

compared to small percentages for others (0.6% for T. Roosevelt in 1905). For a few semantic classes, we can detect significant changes over the years. For example, the class *Social* (e.g., talk, child, citizen, inform*) shows a clear increase. Contemporary presidents use more terms belonging to this category. Less noticeable, but still statistically significant, we can also perceive a growth of terms belonging to the class *Health* (e.g., sick, pains, life) or *Family* (e.g., son, mother, family, relatives).

Overall, the presidential style has clearly changed over the last two centuries (Lim, 2002). The presidents must convince not only the Congress but the American people as well. We observe a decrease in the complexity at the syntactic and lexical level. Although the number of speeches per year was relatively stable until F.D. Roosevelt, their number increases to reach one speech / day with Carter's presidency (Tulis, 1987).

Keywords: Discourse analysis, stylistic evolution, political speeches

REFERENCES

Biber, D., & Conrad, S., (2009). *Register, genre, and style*. Cambridge: Cambridge University Press.

Hart, R. P., Childers, J. P., & Lind, C. J. (2013). *Political tone. How leaders talk & why*. Chicago: Chicago University Press.

Jockers, M. L. (2014). *Text analysis with R for students of literature*. Heidelberg: Spinger-Verlag Press.

Juola, P. (2003). The time course of language change. *Computers and Humanities*, *37*, 77–96.

Lim, E. T. (2002). Five trends in presidential rhetoric: An analysis of rhetoric from George Washington to Bill Clinton. *Presidential Studies Quarterly*, *32*(2), 328–348.

Pennebaker, J. W. (2011). *The secret life of pronouns. What our words say about us*. New York: Bloomsbury Press.

Rule, A., Cointet, J.-P., & Bearman, P. S. (2015). Lexical shifts, substantive changes, and continuity in State of the Union discourse 1790–2014. *Proceedings PNAS*, *112*(35), 10837–10844.

Tausczik, Y. R., & Pennebaker, J. W. (2010). The psychological meaning of words: LIWC and computerized text analysis methods. *Journal of Language and Social Psychology*, *29*(1), 24–54.

Tulis, J. (1987). *The rhetorical presidency*. Princeton: Princeton University Press.

Quantitative Properties of Case Syncretism and Valency of Indo-European Languages

Petra Steiner
Institut für deutsche Sprache, Mannheim

In Indo-European languages morphological case marking is an effect of semantic and syntactic valency. Verb valency and morphological case can be considered as two sides of the coding of predicate-argument structures. On both linguistic levels, principles of economy can be observed.

Case syncretism (see Jakobsen, 1971, p. 67) - the "homonymy of inflection markers in some grammatical domain" (Müller et al., 2004, p. 6) - is a typical property of inflectional Indo-European languages. 'Explanation' of syncretism is often based on underspecification (e.g. Müller et al., 2004, p. 6f.) or some linguistic principles, e.g. the Principle of Contrast (Müller, 2004, p. 197),

according to which inflectional class features trigger the form of inflection (Müller, 2004, p. 211).

However, this assumption has no explanatory force. It rather is itself in need of an explanation. In contrast, this paper describes how the Zipfian forces of unification and diversification apply to inflectional morphemes and paradigms. Linguistic hypotheses about the frequency distributions of inflectional suffixes are derived and corroborated by statistical testing on empirical data. These investigations are providing missing links of former work on Modern Germanic languages (Steiner, 2009; Steiner & Prün, 2007).

Verb valency, the other side of morphological case, is derived from highly-structured processes in text. Based on the assumptions of Thurner et al. (2015) for sample-space collapse of word frequencies in sentences, a model for case patterns in sentences is derived, corroborated for textual data and linked with the hypotheses on morphological case.

REFERENCES

Jakobson, R. (1971). Beitrag zur allgemeinen Kasuslehre – Gesamtbedeutungen der russischen Kasus. In Jakobson, R. (Ed.), *Selected Writings 2: Word and Language* (pp. 23–71). The Hague: Mouton.

Müller, G. (2004). On Decomposing Inflection Class Features: Syncretism in Russian Noun Inflection. In Müller, G., Gunkel, L., Zifonun, G. (Eds.), *Explorations in Nominal Inflection* [Interface Explorations; 10] (pp. 189–227). Berlin/New York: de Gruyter.

Müller, G., Gunkel, L., & Zifonun, G. (2004). Introduction. In G. Müller, L. Gunkel & G. Zifonun (Eds.), *Explorations in Nominal Inflection* [Interface Explorations; 10] (pp. 1–20). Berlin/New York: de Gruyter.

Steiner, P. (2009). Diversification in Icelandic Inflectional Paradigms. In R. Köhler (Ed.), *Issues in Quantitative Linguistics* [Studies in Quantitative Linguistics; 5](pp. 126–154). Lüdenscheid: RAM-Verlag.

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

Thurner, S., Hanel, R., Liu, B., & Corominas-Murta, B. (2015). Understanding Zipf's Law of word frequencies through sample-space collapse in sentence formation. *Journal of Royal Society, Interface 12*: 20150330.

Authorship Attribution of Yasunari Kawabata's Novels: Who Actually Wrote *Otome no minato* and *Hana nikki*

Hao Sun and Mingzhe Jin
Doshisha University, Japan

Yasunari Kawabata (1899–1972) was a famous Japanese novelist who won the Nobel Prize for Literature in 1968. He is known to have suffered from anxiety and required sleeping pills. Tragically, he chose to end his life in 1972. It has long been argued that many of his novels are actually written by ghostwriters. Although the issue of ghostwriters associated with Yasunari Kawabata has long been debated by literary scholars, the question has not yet been resolved due to a

lack of evidence. Tsuneko Nakazato, one of Yasunari Kawabata's disciples, is said to be one of the authors of Yasunari Kawabata's novels for girls, *Otome no minato* and *Hana nikki*. To date, two pieces of evidence have been used to show that Tsuneko Nakazato is also one of the authors. The first is a part of the *Otome no minato* manuscript (Chapter 7) found in 1989, which was written by Tsuneko Nakazato. The second is advice about how to write *Otome no minato* and *Hana nikki* in letters from Yasunari Kawabata to Tsuneko Nakazato. However, these novels are still parts of Yasunari Kawabata's collected writings. This means these novels are officially recognized to be written by Yasunari Kawabata alone. The traditional method of investigating the issue is by using historical records for verification. Letters between Yasunari Kawabata and ghostwriters or the manuscripts of ghostwriters are recognized as compelling evidence. Providing an alternative to traditional methods, new evidence from an authorship attribution perspective is provided in this study. Statistical and computational methods in Japanese authorship attribution have a long history. They were first used in the 1950s to study Biten Yasumoto's authorship of the *Genjimonogatari*. Over the past decade, this field has dramatically developed, taking advantage of research in machine learning and natural language processing, and, in particular, the emergence of Japanese tokenizer/parser tools. Since then, several types of textual measurements and algorithms for Japanese authorship attribution have been proposed. In our research, an integrated classification algorithm is used as authorship attribution method. Character–symbol bigrams, tag bigrams, particle bigrams, phrase patterns are used as textual measurements. Adaptive boosting, high-dimensional discriminant analysis, logic model tree, random forest and support vector machine are used as classifiers. The results provide new evidence to prove that *Otome no minato* and *Hana nikki* are collaborative works of Yasunari Kawabata and Tsuneko Nakazato.

Measuring the Degree of Violation of the One-Meaning–One-Form Principle

Relja Vulanović and Oliver Ruff
Kent State University at Stark, Ohio, USA

According to Miestamo (2008), the linguistic principle 'one-meaning–one-form' (from now on, the 'Principle') has been known under this name since Anttila's (1972) book. Speaking of the Principle in his book, Anttila states (p. 181): "A maximally efficient system avoids polysemy (forms with many [related] meanings, especially if these occur in the same semantic sphere) and homophony, two (unrelated) meanings getting the same form. [...] Avoidance of homophony and polysemy provide clear evidence of this mental force 'one meaning, one form'." The book provides many examples of the manifestations of the Principle. Nevertheless, examples of the Principle being violated also abound. It is therefore of interest to be able to measure how much a linguistic system departs from the Principle. This kind of measure can become a component of the measure of grammar complexity or grammar efficiency (Miestamo et al., 2008; Vulanović, 2003). For instance, the Principle is proposed in (Miestamo, 2008) as one of the criteria for the absolute (theory-oriented, objective, as opposed to user-oriented) approach to language complexity.

We approach the question of how to measure the degree of violation of the Principle in a formal, mathematical way. We consider general relations between two sets in contrast to the situations when a bijection can be established between them. In order to be used as a factor of grammar complexity, the measure of how much the Principle is violated should be equal to 1 when the relation is a bijection; otherwise, it should be greater than 1 proportionally to the extent of how much the relation is far from a bijection. We develop formulas satisfying this property and apply

them to two linguistic examples of morpho-syntactic nature. The first example is a relatively simple example related to subject and object marking. In the second example, we consider more complicated linguistic structures, viz. parts-of-speech systems in the sense of Hengeveld (1992). The examples confirm the viability of the proposed formulas.

REFERENCES

Anttila, R. (1972). *An Introduction to Historical and Comparative Linguistics*. New York: Macmillan.

Hengeveld, K. (1992). Parts of speech. In M. Fortescue, P. Harder & L. Kristoffersen (Eds.), *Layered Structure and Reference in Functional Perspective* (pp. 29–55). Amsterdam/Philadelphia: Benjamins.

Miestamo, M. (2008). Grammatical complexity in a cross-linguistic perspective. In M. Miestamo, K. Sinnemäki & F. Karlsson (Eds.), *Language Complexity: Typology, Contact, Change* [Studies in Language Companion Series; 94] (pp. 23–41). Amsterdam: Benjamins.

Miestamo, M., Sinnemäki, K., Karlsson, F. (2008). *Language Complexity: Typology, Contact, Change* [Studies in Language Companion Series; 94]. Amsterdam: Benjamins.

Vulanović, R. (2003). Grammar efficiency and complexity. *Grammars*, *6*, 127–144.

Polyfunctionality Studies in German, Dutch, English and Chinese

Lu Wang
Dalian Maritime University, China

Each word belonged to one and only one part of speech in the moment of its creation. Then, some words are expanded to new parts of speech by way of adopting affixes, e.g. *die Zahl* (n.), *zählen* (v.), and *zahlreich* (adj.) in German, or without any change ("conversion"), e.g. 左右 (n. v. adv.) in Chinese and *run* (n. v.) in English. The latter words, which have more than one part of speech, are called **polyfunctional words**. The number of parts of speech of a word is called **polyfunctionality**.

Actually, parts of speech as any nominal entities form two kinds of distribution: not only the rank distribution (where the variable is the rank associated with the corresponding number of words), but also the spectrum i.e. polyfunctionality distribution (obtained by counting the polyfunctional words with the given number of parts of speech).

Most of the previous works on parts of speech studied the rank distribution. The polyfunctionality distribution is left almost untouched. The present paper focuses on this topic. We consider that polyfunctionality should be lawfully distributed. We attempt to investigate polyfunctionality distribution to test whether it can be captured by distribution models and whether there is a unified model to capture the distributions in different languages. Four languages are adopted: German, Dutch, English (from celex dictionary) and Chinese (from *Modern Chinese Dictionary*). In the 4 dictionaries, German has 1.29% polyfunctional words; Dutch has 1.42%, Chinese has 7.31% and English has the most 11.49%. An individual word in German can carry at most 4 parts of speech; a Dutch word can have 5; a Chinese or an English word can include up to 6 parts of speech. Given that polyfunctionality is the kind of spectrum aspect of linguistic property, we tested many proper models. Finally, the Waring distribution is found to capture all the data and perform excellent goodness-of-fit results.

Further, the polyfunctional words form part-of-speech patterns. As shown in the above examples, the part-of-speech pattern of the word 左右 is "n. v. adv.", which polyfunctionality is 3. The pattern of the word *run* is "n. v.", which polyfunctionality is 2. In this way, part-of-speech patterns also form a polyfunctionality distribution. Again, we expect to find a model to capture such distributions. The fitting results show that the data of all the 4 languages abide by the binomial distribution.

The results show that, despite of the linguistic typological difference, polyfunctional words widely exist in different languages; polyfunctionality distributions can be captured by waring distribution; the polyfunctionality of part-of-speech patterns also distribute regularly and the distribution abide by the binomial distribution.

Coherence and Quantitative Measures of Texts

Makoto Yamazaki
National Institute for Japanese Language and Linguistics, Japan

---

This study investigates the relationship between the coherence and quantitative measures of a text. To capture the essence of coherence, we created random texts and examined how their quantitative measures vary by the type of randomness.

First, we chose several samples of the same size from the Balanced Corpus of Contemporary Written Japanese (BCCWJ). Then, we merged two texts and compared the type/token ratio (TTR) of the merged text with that of the original texts. Almost all the merged texts showed greater TTR values than those of the original texts. This suggests that less coherent texts show greater TTR values.

Second, we made three types of randomized texts using BCCWJ and compared their TTR values and other quantitative measures. We also used the TTR values of the original texts to set a baseline for comparison. The baseline (Type 0) and the randomized three types of texts (Type 1, 2 and 3) were as follows:

Type 0: *n*-part simply divided text (baseline).
Type 1: *n*-part divided text in which vocabulary is randomized.
Type 2: *n*-part text composed of words in which serial numbers share the same residual.
Type 3: *n*-part text composed of consecutive *n*-size words which appear at regular intervals.

We examined 252 such texts, and the following results were obtained:

1. Type 1 and type 2 show smaller TTR values than those of the baseline.
2. In type 3, TTR value decreases as consecutive words gets longer.
3. In type 1 and type 2, the number of hapax legomena decreases as the number of divisions gets bigger.
4. In type 3, the number of hapax legomena is stable even if the consecutive words get longer.

The Rank-Frequency Distribution of Part-of-speech Motif and Dependency Relation Motif in the Deaf Learners' Compositions

Jingqi Yan
Zhejiang University, China

Motif, originally called segment or sequence, refers to the longest sequence of monotonously increasing numbers which represent a certain quantitative property of a linguistic unit (Köhler, 2006).The linguistic motif is a unit for the sequential analyses in quantitative studies. This paper makes an exploratory study of Part-of-speech (POS) motif and dependency relation motif on the basis of the treebank of deaf students' writing in three learning stages. Three aspects of motif distribution are analyzed. The fitting to the POS motif data with Zipf-Mandelbrot distribution yields a good result. Yet the fitting is not good for the distribution in the dependency relation motif. The dependency relation motif distribution fits well by the Popescu-Altmann-Köhler function instead. A further observation of the top-ranking motifs discovers that the most frequent motifs are formed among the most frequently-used word classes and dependency relations. More modifying relations are added in the texts of higher learning stages. Moreover, the hapax-token ratio is counted for the respective two motif types and the result shows an extremely heavy proportion of hapax, which may indicate the discreteness of sentence structure. Motifs are more compact at the higher language level. Through the three analyses, some changes of sentence structure and the development of syntax are witnessed along the language learning course.

Keywords: Second language acquisition, learner corpora

A VSM-based Algorithm for Comparing Online Translation and Human Translation of English Passives into Chinese

Yingxian Zhang and Yue Jiang
Xi'an Jiaotong University, China

In this study, by combining the vector space model (VSM) with principal component analysis (PCA), a new algorithm is proposed to compare the similarity between human translated texts and an online machine translated text of English passives into Chinese. 63 features are retrieved from a bilingual parallel corpus (one-to-five) consisting of the English original text of Pride and Prejudice, four Chinese translations of it by humans and one translation of it by an online machine translator. Besides, an "Average human translation (AHT)" was simulated by calculating the mean value of each feature. The 63 features include macro- and micro-ones ranging from lexical, syntactic and semantic levels, 18 linguistic quantitative features, 7 ways of rendition, 18 kinds of Chinese rendition syntactically corresponding to English passives and 20 words of high frequency of occurrence. After the retrieval, firstly, all the texts are vectorized by using the above 63 features. Then PCA is applied to dimension deduction and all the features are finally condensed to 3 principal components to form a new 3-dimension vector space. Afterwards, the vector distance between all these translated texts are measured by calculating the weighted mean of the included angle cosine distance and Euclidean distance.

The results of the calculation show that in vector distance, all the four human translated texts are close to each other but far from the online translation, thus implying a low similarity between human translation and online translation. One of the four human translated texts (WKY) is the farthest from online translation in vector distance compared with the other three human translated texts whereas another human translation text (ZJH) is the closest to the online translation, indicating that the human translations vary from one another in their distance to the online translation. Compared with other human translations, SZL text is the closest to AHT, implying that it might be the optimal version among the four human translations in terms of vector distance.

Conclusively, an algorithm combining VSM with PCA is applied to compare the textual similarity between online translation and human translations of English passives into Chinese. Based on our study, we may suggest that this proposed algorithm can be applied to compare different translations of some special syntactic structures, and can also be generalized to assess the textual similarity of whole texts for a more objective comparison of different translated texts.

Keywords: Online machine translation, vector space model, principal component analysis, passives, vector distance

---

Distribution of Combined Structures of Parts of Speech and Syntactic Functions

Hongxin Zhang and Haitao Liu
Zhejiang University, China

This paper examines the "*dependent*+*head*=*function*" structure in a dependency treebank of Chinese news broadcasts. The equation means that a dependent modifies a head (whether head initial or head final) and fulfills a certain function. For instance, "*n*+*v*=*obj*" represents a structure where a noun (dependent) serves as the object of the verb (head). From the corpus, we choose 5 relatively separate pieces of news for investigation, and their lengths range from 573 to 737 words.

The paper examines 6 sub-structures: 1) dependents only, 2) functions only, 3) *dependent*+[ ]=*function*, regardless of the type of the head, 4) [ ]+*head*=*function*, 5) *dependent*+*head*=[ ], and finally 6) the complete form, *dependent*+*head*=*function*. It finds that all the data abide by the right truncated modified Zipf-Alekseev distribution pattern with R^2 ranging from 0.94 to 0.9974. The research findings indicate that all these 6 types of structure are results of diversification processes and justify the practice of combining certain properties and examining their distribution patterns.

This study extends traditional syntactic distributions beyond only one property (e.g. parts of speech, functions). The examination of the newly combined features as an integral structure is an interesting endeavor, which is expected to work in multiple situations.

Text characterization using syntactic motifs

R. Köhler
Trier University

R-motifs, formed from categorical instead of quantitative sequences, are used to characterise texts with respect to parts of speech as a syntactic feature. The attempt to classify the end-of-year speeches of the Italian presidents by means of not more than some parameters of the thus formed motifs fails. Instead, it could be shown that Ord's criteria and two other characteristics of the empirical frequency distributions of the motifs reveal a continuous development over time which follows the Piotrowski-Altmann Law.

Keywords: Motif, S-motif, text characterisation, syntactic properties, parts of speech, Piotrowski-Altmann Law

---

Syntactic complexity of dependency structures

S. Naumann and A. Beliankou
University of Trier

The verbal predicate-argument structure is a long standing topic for the research community. We are going to describe some quantitative properties of the argument structure focusing on the syntactic complexity both for dependency and phrase-structure style grammars.
Syntactic complexity is one of the key concepts used in the synergetic model of the syntactic subsystem that was developed by R. Köhler (2012). The model suggests that the data should follow the hyper-Pascal distribution and first empirical tests using phrase-structure annotated corpora seemed to support this claim. But further investigations using additional material did not confirm the model.
One possible reason for these inconclusive results could be the fact that phrase structure descriptions based on different grammar formalisms can vary a lot and that the used corpora were annotated using different schemes. We expect the dependency annotated corpora not to have this drawback. Based on the previous work by Naumann (2014) we pursue a comparison of the model performance on both annotation formalisms.

Keywords: syntactic complexity, synergetic linguistics, tree-banks, corpus linguistics, linguistic formalism

REFERENCES

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

Naumann, Sven. (2014). *Syntactic complexity in quantitative linguistics.* In: Selected Papers of the 9th International Conference on Quantitative Linguistics, Olomouc, 2014