Uniwersytet Wrocławski

IQLA
International Quantitative Linguistics Association

QUALICO 2018

BOOK OF ABSTRACTS

QUALICO
2018

QUALICO 2018

Information and Language: Coding, Extraction and Applications

Book of Abstracts

Wrocław 2018

QUALICO 2018. Information in Language: Coding, Extraction and Applications. Book of Abstracts

Redakcja
Adam Pawłowski, Anna Cisło

Współpraca
Anna Łach

DTP
Wojciech Sierżęga

Organizatorzy dziękują za wsparcie finansowe i organizacyjne
Rektorowi Uniwersytetu Wrocławskiego,
Dziekanowi Wydziału Filologicznego UWr,
Dyrektorowi Instytutu Informacji Naukowej i Bibliotekoznawstwa UWr,
Dyrektorowi Instytutu Matematycznego UWr.

Organizatorzy dziękują za opiekę naukową
International Quantitative Linguistics Association.

Złożono krojem pisma Brygada 1918. Zdigitalizowana czcionka Brygada, zaprojektowana ok. 1928 r. przez polskiego typografa Adama Półtawskiego na 10-lecie niepodległości Polski. Została odkryta przez Janusza Tryzno, zrekonstruowana przez zespół projektantów w składzie Mateusz Machalski, Borys Kosmynka, Przemysław Hoffer i udostępniona z okazji stulecia odzyskania niepodległości przez Polskę.

The typeface used in this publication is Brygada 1918. The digitised Brygada typeface was originally devised around 1928 by the Polish typographer Adam Półtawski to celebrate the 10th anniversary of Polish independence. It was discovered by Janusz Tryzno, reconstructed by a team of designers consisting of Mateusz Machalski, Borys Kosmynka, Przemysław Hoffer and made available to commemorate the 100th anniversary of the restoration of Poland's sovereignty.

ISBN 978-83-950966-0-0

Uniwersytet Wrocławski
Instytut Informacji Naukowej i Bibliotekoznawstwa
pl. Uniwersytecki 9/13
50-137 Wrocław

Keynote lectures

prof. Łukasz Dębowski
Institute of Computer Science, Polish Academy of Sciences, Warsaw, Poland

The Puzzling Entropy Rate of Human Languages

Abstract

In this talk, we will look back into research in the entropy rates of human languages. Entropy rate is the limiting amount of unpredictability of a random process, measured in bits per unit of the process. Shannon (1951) provided the first estimate of the entropy of texts in English, the famous 1 bit per letter. His method based on guessing by human subjects was followed by many researchers and improved by Cover and King (1978). In contrast, Ziv and Lempel (1977), Brown et al. (1992), and Gao et al. (2008) proposed computational methods of entropy rate estimation based on universal data compression, statistical language models, and match lengths, respectively. Computational methods of entropy estimation were applied to large corpora (Takahira et al., 2016) and many languages (Bentz et al., 2017). This success story takes some twist, however. Hilberg (1990) supposed that Shannon's estimate of the entropy rate of English is an artifact caused by a slow power-law convergence of the estimates, whereas the actual entropy rate of human languages could equal zero. Some versions of this hypothesis were considered by Ebeling and Nicolis (1991) and Crutchfield and Feldman (2003). Also Dębowski (2015) observed experimentally that texts in human languages obey a power-law logarithmic growth of maximal repetition, which implies that conditional Rényi entropy rates are zero, as proved for stationary processes by Dębowski (2017). This property does not generalize to the entropy rate defined by Shannon. Indeed, Takahira et al. (2016), investigating very large corpora for several languages, announced that the entropy estimates follow a power-law convergence but the limiting Shannon en-

tropy rate is close to Shannon's original estimate. Thus, constructing mathematical models of processes with a positive Shannon entropy rate and a zero Rényi entropy rate is an interesting open problem with possible applications to linguistics.

Keywords:

Shannon entropy rate, Rényi entropy rate, human languages

About the author

Łukasz Dębowski received the M.Sc. degree in theoretical physics from the Warsaw University, Warsaw, Poland, in 1994, the Ph.D. degree in computer science from the Polish Academy of Sciences, Warsaw, Poland, in 2005, and the habilitation degree in computer science from the Polish Academy of Sciences, Warsaw, Poland, in 2015. He visited the Institute of Formal and Applied Linguistics at the Charles University in 2001, the Santa Fe Institute in 2002, and the School of Computer Science and Engineering at the University of New South Wales in 2006. Moreover, he held a post-doctoral research position with the Centrum Wiskunde & Informatica from 2008 till 2009 and a visiting professor position with the Department of Advanced Information Technology at the Kyushu University in 2015. He is currently an Associate Professor with the Institute of Computer Science of the Polish Academy of Sciences. His research interests include information theory and statistical modelling of natural language.

□ □ □ □

prof. Nicola Ferro
Department of Information Engineering of the University of Padua, Italy

From Systems to Components: Breaking-down Performance and Discovering Interactions

Abstract

Information Access Systems are pipelines typically constituted by several components developed by neighbouring disciplines, e.g. information retrieval, natural language processing, computational linguistics. When it comes to evaluate such components, you are often faced with two choices: either evaluate them in isolation for some

specific feature, e.g. precision in over/under-stemming for a stemmer, or evaluate them in full pipelines, e.g. the effectiveness of a whole IR systems when using a stemmer. In both cases, you can neither determine the contribution and importance of the single components for the overall performance nor properly study and assess the interaction among components.

We will discuss a new methodology based on Grid-of-Points and General Linear Models which allows us to break-down overall system performance into those of the constituting compenents and to study their interaction. We will show how to apply this methodology firstly to the case of English retrieval and then to multilingual retrieval. We will then discuss how this methodology could be exploited to better understand how components from different disciplines work together, e.g. Word Sense Disambiguation with IR pipelinse. Finally, we will present a visual analytics tool to study to explore and better understand how components work together.

About the author

Nicola Ferro (http://www.dei.unipd.it/~ferro/) is associate professor in computer science at the University of Padua, Italy. His research interests include information retrieval, its experimental evaluation, multilingual information access and digital libraries. He is the coordinator of the CLEF evaluation initiative, which involves more than 200 research groups world-wide in large-scale IR evaluation activities. He was the coordinator of the EU Seventh Framework Programme Network of Excellence PROMISE on information retrieval evaluation. He is associate editor of ACM TOIS and was general chair of ECIR 2016.

□ □ □ □

Long Papers

Sergey Andreev

Adnominal Noun Valency in Modern Russian

The study is devoted to the search of regularities of using noun attributes (adnominals) in Russian fiction. Adnominals are specific sentence elements which are characterized by the tightest syntactic links among all sentence members, possessing a broad spectre of semantic features. The usage of adnominals and the choice of their types by an author is in most cases quite arbitrary which makes them a vivid feature of individual style. At the same time adnominals are an important element of syntactic structure of the sentence forming a kind of secondary predication.

Some previous studies demonstrated the existence of regularity of the distribution of concrete adnominals in Russian literature (Andreev, Popescu & Altmann, 2017). In this study attention is paid to the distribution of nouns, characterized by different adnominal valencies. Adnominal valency is established by the number of adnominals which modify a given noun. The minimal valency is zero, the maximum valency in our data-base is 6.

The data-source of the study includes 18 texts of about 2 thousand words each, written by the most popular Russia modern detective writers (both male and female). The list of adnominals includes different morphological types: adjectival and participial constructions, subordinate sentences, single words of different morphological classes – adjectives, participles, pronouns (demonstrative, possessive, qualifying, negative, indefinite, interrogative, relative), nouns in the genitive, dative and instrumental cases as well as nouns in the nominative case in certain types of apposition, prepositional nouns (of all cases), infinitives and adverbs.

All these types of adnominals were counted in the texts and then analysed in terms of static and dynamic approaches (Naumann, Popescu & Altmann, 2012), revealing proportions of nouns with different va-

lencies and, secondly, alternations of their distribution through the text.

For testing Skinner's hypothesis all the distances between nouns with the same attributive valency were found out. The distances were established by the number of nominal units between nouns with the similar valency. To model the distances the exponential function was used (Boschtan & Best, 2010; Andreev, Popescu & Altmann, 2017: 35). The computation of distances showed very high values of determination coefficients. Only in one text the determination coefficient was 0.8 whereas in all the other texts this coefficient exceeded 0.9, demonstrating very good fitting. This may be considered as another proof that the distribution of adnominals in the text is controlled by a certain law (Andreev, Popescu & Altmann, 2017). Such regularities may exist in different forms in various languages, genres and genders.

According to the number of attributive links in each sentence motifs were singled out which were used for the classification of the authors' styles (cluster analysis).

References:

Andreev, S., Popescu, I.-I. & Altmann, G. (2017). Skinner's hypothesis applied to Russian adnominals. Glottometrics, 36: 32–69.

Boschtan, A. & Best, K.-H. (2010). Diversification of simple attributes in German. Glottotheory, 3(2): 5–9.

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

□ □ □ □

Jan Andres, Dan Faltýnek and Lukáš Zámečník

What Is a Linguistic Law?

The criteria of laws that are set in science disciplines are not satisfactorily met using statistical methods (Andres et al. 2012; 2014) any time in the case of linguistic laws. Therefore, it seems more natural to speak about statistical trends, rather than about laws (Andres, 2014).

Nevertheless, we are not willing to give up trying to examine the hypothesis that the mentioned regularities (tendencies) only hide a deeper level of regularities (principles) whose manifestation are the mentioned tendencies. This way has been taken in earlier reflections

on synergetic linguistics (Köhler, 1986), and now the problem is being investigated in connection with the critical behaviour in phase transitions, in the context of various other disciplines (the problem of universality, Batterman, 2013).

We will introduce one variant of these principles, namely conservation laws often used in sense of economization principles. We will also show problematic issues that have hindered their strict introduction to the linguistic context. We will show how the formulation of this type of principles fits the functional explanation model of synergetic linguistics, and how can be this model reformulated into a topological form (Huneman, 2010; Zámečník, 2014).

References:

Andres, J. (2014). The Moran-Hutchinson formula in terms of Menzerath-Altmann's law and Zipf-Mandelbrot's law. In: Altmann, G., Čech, R.; Mačutek, J. & Uhlirova, L. (eds). Empirical Approaches to Text and Language Analysis. (Studies in Qunatitative Linguistics, 17). Lüdenscheid: RAM-Verlag: 29–44.

Andres, J., Benešová, M., Chvosteková, M., Fišerová, E. (2014). Optimization of parameters in the Menzerath–Altmann law II. Acta Univ. Palacki. Olomuc., Fac. rer. nat., Mathematica, 53(2): 5–28.

Andres, J., Kubáček, L., Machalová, J., Tučková, M. (2012). Optimization of parameters in the Menzerath–Altmann law I. Acta Univ. Palacki. Olomuc., Fac. rer. nat., Mathematica, 51(1): 5–27.

Batterman, R. (2013). The Tyranny of Scales. In: Betterman, R. (ed.). The Oxford Handbook of Philosophy of Physics. Oxford: Oxford University Press: 255–286.

Huneman, P. (2010). Topological explanation and robustness in biological sciences. Synthese, 177: 213–245.

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

Zámečník, L. (2014). The Nature of Explanation in Synergetic Linguistics. Glottotheory, 5(1): 101–120.

□ □ □ □

Jan Andres, Jiří Langer, Dan Faltýnek and Lukas Zámečník

Some Aspects of a Sign Language Quantitative Analysis

We would like to point out some aspects of a sign language quantitative analysis. For this aim, we will present our preliminary results concerning the quantitative analysis of the Czech sign language. Some methodological problems which rose in the processing of sign language quantitative analysis will be also discussed. The exploration based on 30 five minute long dialogues has been focused on the basic text quantitative metrics (entropy, TTR, repeat rate, token length frequency spectrum, average token length, etc.). The second part of our investigation is oriented to the text laws. In particular the Menzerath-Altmann law and the Zipf's law are taken into account with this respect. For the application of the Menzerath-Altmann law, we discuss the appropriate segmentation of the sign language text. Our segmentation concerns sentences, clauses, words and signs. On these levels we have been testing a possible validity of the Menzerath-Altmann law in the sign language text. The main problem of such a segmentation is to identify the sign structure (analogously to the phonetic system).

□ □ □ □

Bernardino Casas, Antoni Hernández-Fernández, Neus Català, Ramon Ferrer-i-Cancho and Jaume Baixeries

Polysemy: A New Bias in Child Language Acquisition?

Children are exposed to millions of words tokens through an accumulation of small interactions grounded in context, immersed in a sea of words that they learn early after their birth. During this process, some words are learned first instead of others and many biases have been hypothesized in the literature in order to explain why some words are learned earlier. Zipf pioneered the study of universal biases that are found in adult language. In this work, we study the possibility of a new bias in vocabulary acquisition in children.

Here we analyse the variation in mean polysemy in children over time with the help of a massive database that contains the transcriptions of conversations between children and adults (CHILDES Database [3]) and using words as our unit of study.

We use two measures of polysemy: the total number of meanings of a word according to the WordNet lexical database, and the number of WordNet meanings of a word that have appeared in an annotated corpus (the SemCor corpus).

Our results show that mean polysemy in children increases over time in two phases, i.e. a fast growth till the 31st month followed by a slower tendency towards adult speech. In contrast, this evolution is not found in adults interacting with children. Various statistical tests indicate that children have a preference for low polysemous words in their early stages of vocabulary acquisition. Interestingly, the evolutionary pattern described above weakens when controlling for syntactic category (noun, verb, adjective or adverb) but it does not disappear completely.

Therefore, we suggest two possible explanations (not necessarily mutually exclusive) for the increase in word polysemy over time in children:

1. Standalone bias. Children have a preference for low polysemous words. This is supported by Zipf's view of polysemy as a cost for the listener: low polysemous words reduce the disambiguation effort for the listener.

2. Side-effect of other biases. When children learn a language, they begin using more nouns than words from other syntactic categories, and then, they increase the percentage of verbs that they use in their conversations over time. Since the mean polysemy of verbs is significantly higher than that of nouns, the mean polysemy increases because the proportion of verbs increases.

□ □ □ □

Radek Čech, Pavel Kosek and Olga Navrátilová

Why Are There Variations of the Word Order Position of Enclitics in Old Czech? The Impact of Length and Frequency

The word order position of enclitics (e.g. mi 'to me', sě [REFL.], ho 'him', ti 'you') has a status of stable enclitics in Modern Czech – they occur after the first phrase of a clause. However, until the beginning of the 20th century, their position varied and they could be used in different positions in a clause. Kosek et al. (2018) described three the most frequent positions of the enclitics in Old Czech and derived stochastic rules for the word order of the enclitics:

R1: if an enclitic appears in a clause, use it after the initial phrase of the clause;
R2: if rule (R1) is not applied, use the enclitic in a postposition of a verb;
R3: if rule (R2) is not applied, use the enclitic in a pre-position of a verb.

In fact, these rules are generalizations of the description of language data and they do not provide any explanation of the phenomenon.

In the study, we set up some hypotheses which try to explain the variation of word order position of the enclitics and test them. First, we assume that the length of the initial phrase of a clause plays a crucial role in the word order position of the enclitics due to acoustic and physiological factors influencing language behaviour. Namely, the longer the first phrase of the clause, the higher probability that a pause is realized (because a speaker needs to take a breath) and, consequently, the lower probability that an enclitic occurs after the first phrase (enclitics almost never occur after the pause because they do not carry stress). To sum up, we hypothesize that the longer the first phrase of the clause, the lower probability of the occurrence of the enclitic after this phrase (H1). Second, the length of the clause is considered too. According to the Menzerath-Altmann law we assume that the longer the clause, the shorter its phrases (in average). Taking into account the hypothesis H1, we hypothesize that the longer the clause, the higher probability of occurrence of the enclitic after first phrase (H2).

Finally, the relationship between rules R2 and R3 is considered. Specifically, we assume that the longer the verb, the higher probability that the enclitic follows the rule R3 instead of R2 because of acous-

tic and physiological factors mentioned above. Thus, we hypothesize that the longer the verb, the higher probability of occurrence of the enclitic in a preposition of a verb (H3). For the hypotheses testing, the historical Czech Bibles are used as a language material (Kosek et al., 2018; Kyas, 1997).

References:

Kosek, P., Čech, R., Navrátilová, O. & Mačutek, J. (2018). On the Development of Old Czech (En)clitics. Glottometrics, 40, 51-62.

Kyas, V. (1997). Česká Bible v dějinách národního písemnictví. Praha: Vyšehrad.

□ □ □ □

Xinying Chen

The Power Law Distribution in Universal Dependencies

For a long time, establishing and testing sophisticated cross-language hypotheses for syntax studies face an obstacle of lacking suitable data. One reason is, of course, that fully analysed or annotated data is much less available than plain texts. Another significant problem is lacking a unified criterion for these data sets, which weakens the conclusions of cross-language studies. Newly emerging Universal Dependencies Project provides a neat solution for mentioned problems (Chen & Gerdes, 2017; Nivre & Zeljko Agic, 2017). This on-going project currently released more than 100 dependency treebanks in over 60 languages (UD 2.1). It endorses a more fine-grained cross-language hypothesis. Our goals of the present study are:

- To investigate the distribution pattern of dependencies in a cross-language manner,
- Discuss the correlations between parameters of this distribution function.

Universal Dependencies (UD) is a framework for cross-linguistically consistent grammatical annotation and an open community effort with over 200 contributors producing multi-language treebanks. According to the size of treebanks and the UD annotation schema (the data used for this study is UD 2.0, which includes 70 treebanks from 50 languages), we eliminate the sparse data (treebanks that have a

size smaller than 10,000 tokens) and select only the syntactic dependencies as our observations, then calculate the relative frequencies of each dependency type for every language and fit them to the power law function

$$y = a*x\text{-}b \qquad \text{(Zipf 1935 \& 1949).}$$

Our results (based on OriginPro, 2017) showed that all 42 different languages (with suitable data sizes) fit to the power law distribution rather well, with R2 between 0.67742 to 0.98915. Therefore, we believe that dependencies, which represent syntactic word pairs, behave in a way similar to words. The uneven distribution of dependencies is probably also due to the least effort requirement of language communication. Further investigation reveals:

• The strong negative correlation between the parameters a and b (p-value 0, Pearson correlation

• -0.97616).

• The strong positive correlation between the parameters a and R2 (p-value 0.0000000094412, Pearson correlation 0.75197)

• The strong negative correlation between the parameters b and R2 (p-value 0.0000000000991114, Pearson correlation -0.80798)

The results indicate that the power law distribution of dependencies may due to a single factor since the three main parameters of the function are strongly dependent on each other. The study shows how UD data with a sheer size can contribute to quantitative syntax studies and have a great potential for further discussions.

□ □ □ □

Yuan-Lu Chen and Hsuan-Ying Liu

The More Frequent a Character Word Is the Simpler Form it Has

Human language system has a way to shape its bits to optimize the efficiency. Specifically, Zipf (1949) argues that human language has the pattern of "the more frequent, the shorter", and additionally Piantadosi, Tily, and Gibson (2011) have found that average information content is a better predictor than frequency to predict the length of a word.

Previous studies on alphabet-based languages show that frequent

words tend to be shorter (Zipf, 1949; Sigurd, Eeg-Olofsson & Van Weijer, 2004; Piantadosi, Tily & Gibson, 2011). In addition to frequency, it is also found that the average information content of words is a better predictor than the frequency for the length of the word: specifically, words with higher average information content tend to be longer, while words with low average information content tend to be shorter (Piantadosi et al., 2011). Information content can be understood as how unexpected it is for a word to occur in a certain context. For example, 'America" in "the United States of America" has zero information content because it is totally predicted by the context (i.e. "the United States of"), so it contributes no new information.

In this line of study, the length of bits in human languages is defined by the alphabet length. We expand the study to character-based languages, specifically Mandarin Chinese. We propose that total stoke number is a plausible measurement to gauge the complexity of a word in a character-based system. By conducting corpus studies on Vierthaler (2016), we examine the correlation between the total stroke number of a Chinese word and the frequency of the word, and the correlation between the total stroke number of a Chinese word and the average information content of the word. It is found that the total stroke number is highly correlated negatively with frequency (Traditional Chinese: $r= -0.27$, $n= 7291k$, $p<0.001$; Simplified Chinese: $r= -0.32$, $n= 7291k$, $p<0.001$), and positively correlated with average information content (Traditional Chinese: $r= 0.181$, $n= 7291k$, $p<0.001$; Simplified Chinese: $r= 0.186$, $n= 7291k$, $p<0.001$). Additionally, we compare the traditional Chinese characters and simplified Chinese characters. We have found that the complexity of the word is more strongly correlated with simplified Chinese characters than traditional Chinese ones, suggesting that simplified Chinese is closer to the golden pattern of "the more frequent, the shorter". In terms of communication efficiency, simplified Chinese is better than traditional Chinese.

In conclusion, we have found that Chinese, a character-based language, has the same pattern of optimizing its efficiency as alphabet-based languages.

References:

Piantadosi, S.T., Tily, H. & Gibson, E. (2011). Word lengths are optimized for efficient communication. PNAS, 108(9): 3526–3529.

Sigurd, B., Eeg-Olofsson, M. & Van Weijer, J. (2004). Word length, sentence

length and frequency – Zipf revisited. Studia Linguistica, 58(1): 37–52.
Vierthaler, P. (2016). Late imperial Chinese texts: The corpus for fiction and history: Polarity and stylistic gradience in late imperial Chinese literature. Harvard Dataverse.
Zipf, G. K. (1949). Human behavior and the principle of least effort. Cambridge, MA: Addison-Wesley Press.

□ □ □ □

Maciej Eder, Rafal L. Górski and Joanna Byszuk

Zipf's Law and Subsets of Lexis

The so-called Zipf's law, describing inverse relation in rank-frequency distribution of word frequencies, has long been observed to hold for all languages and types of texts. According to the law, the top of the frequency list is a dominion of function words, with some verbs, but little to no examples of open class parts of speech, which typically land in the lower and lowest parts of the list. On theoretical grounds, one should find a Zipfian distribution even in a subset of a corpus; this is unnecessarily true, however, when a non-random selection of a corpus is excerpted, e.g. one grammatical category.

While a number of studies (e.g. Baroni, 2009; Bentz et al., 2014; Ha et al., 2006; Popescu et al., 2010) were dedicated to examining relation between morphology and frequency distributions, they were focused on learning more about determining syntheticity of a language, rather than the influence of grammar on (non)-Zipfian distributions. Little attention was given to a question, whether Zipf's law holds for grammatical categories. And, if it does not, what might be the reason.

Baroni (2009: 10) observed that "distributions [of ngrams] are even more skewed than those of words" with "few very high frequency types, and long tails of very rare words". Therefore, we also considered n-grams of both words and grammatical categories.

We examined subsets of the lexis and PoS-tags drawn from the balanced version of National Corpus of Polish (ca 300 million segments) to see if they follow Zipf's law, but instead of analysing the distribution of all words encountered in the corpus, we considered particular subsets. We conducted a number of experiments, analysing:

1. words belonging to particular parts of speech.
2. words assigned to particular grammatical categories, e.g. genitive.
3. PoS-tags; note that there are over 1000 unique tags in the NCP-tagset.
4. n-grams (2-6) of lemmata and POS-tags.

Our basic assumption was that classes consisting of a large number of elements would typically follow Zipf's law, however we assumed that smaller subsets would behave differently. Our question, then, was to what extent distribution of frequencies in such subsets would vary from Zipfian and whether that would be explainable by existing linguistic theories.

□ □ □ □

Maciej Eder and Rafal L. Górski

Visualizing the Dynamics of Linguistic Changes

The basic tool used to examine the chronology of linguistic changes is a fairly simple and rather effective method of trend search: the examined features are analysed by mapping the frequency of the described phenomenon on a timeline [@ellegard_auxiliary_1953]. This timeline-centric visualization has become a standard in several studies and corpus tools. Certainly, the most spectacular example is the corpus of several dozens of millions of documents (mainly in English) accompanied by the service Google Books Ngram Viewer http://books.google.com/ngrams, which, according to its authors, enables to examine changes taking place not only in the language, but also in culture [@michel_quantitative_2011].

A significant drawback of simple graphic representation of the trend is a tacit assumption that the researcher knows in advance which elements of the language are subject to change, whereas one might be interested in trend search without any _a priori_ selection of the analysed linguistic changes to be traced. To assess this and similar issues, an iterative procedure of automatic text classification can be applied [@eder_historical_2016], supplemented by various visualizations. Its underlying idea is fairly simple: it excerpts randomly a number of text samples before and after a given year, say 1835, and performs a supervised classification. Then it iterates over the timeline, testing the years

1836, 1837, 1838, 1839, ... for their discriminating power. Any acceleration of linguistic change will be reflected by higher accuracy scores.

The above procedure has been applied to the Corpus of Historical American English (COHA), containing ca. 400 million tokens and covering the years 1810–2009 [@davies_corpus_2010]. In Fig. 1, the classification accuracy rates were shown, while Fig. 2 represents the features exhibiting the biggest variance (or the overall impact on the results).

An access to particular linguistic features – both lexical and grammatical, via POS-tag _n_-grams, will be shown in further visualisations. Also, the method discussed in this study will be compared to other ways of visualizing stylistic drifts over time [@eder_visualization_2017].

□ □ □ □

Yu Fang

Seeing Various Adventures through a Mirror: Stylistic Variation in Literary Translations

Literary translation is about style translation (Boase-Beier, 2006). For a long time, researchers agree on that "the translator's task is simply to reproduce as closely as possible the style of the original" (Baker 2000: 244). Recently, more and more researchers realize that translators are very likely to leave their fingerprints on translations (Huang & Chu, 2014; Jawad, 2014), thus stylistic variations may occur between an original texts and its translations. At the same time, translators are governed by "a number of cultural and professional norms" (Jawad, 2014: 52), for example, by the principle of equivalence (Fung 1998; Luong 2015; Miao 2000), which may constraint the emergence of stylistic variations. This study, combining literary translation and corpus stylistics, aims to explore whether stylistic variations exist in literary translations. Alice's Adventure in Wonderland and its five translations published in different periods of time are selected as our data, and nine sub-corpora from FLOB and nine from LCMC are used as English reference corpora and Chinese reference corpora respectively. Styles of the original text and its translations are determined

by hierarchical cluster analysis based on normalized frequencies of POSs per 1,000 words. The results show that stylistic variations exist in all of the five translations: the style of the original text is most similar to that of mystery fictions and secondarily similar to that of science fictions; while the style of the five translations are all similar to that of general fictions. However, translators also follow the principle of equivalence, which can be seen from the close link to science fictions for both the original text and its translations. The results also indicate differences among styles of the five translations: Zhao's style is secondarily similar to that of science fiction, while the style of other translations are secondarily similar to that of love story. Then reasons for similarities and differences among the five translations are explored by log-likelihood ratio tests: shared POS preference explain their similarities, while distinct using habits for POSs explain different forms of stylistic variations.

□ □ □ □

Gertraud Fenk-Oczlon

Towards a Synergetic Approach to Word Order

The basic ordering of subject (S), verb (V), and object (O) across languages remains a matter of debate. Why are subject first word orders most frequent and what determines SOV/SVO variation? Here, I suggest a systemic or synergetic approach to word order that may advance our understanding of word order evolution and word order variation. The central axiom of synergetic linguistics (Köhler, 1986) is that language systems possess self- regulating and self-organizing mechanisms. In Systemic Typology (Fenk-Oczlon & Fenk, 1999) it is likewise argued that each language goes through self-organizing processes optimizing the interactions between its subsystems (phonological, morphological, and syntactical), and the interaction with its 'natural' environment, e.g. the cognitive system and articulatory system.

The paper starts with an outline of the constant flow of linguistic information hypothesis (Fenk & Fenk, 1980) which is used for explaining the cross-linguistic prevalence of the subject initial word orders SOV

and SVO. Then I will present the results of empirical studies (e.g. Fenk & Fenk, 1999 & 2005) showing systematic interactions between word order and other metric and non-metric linguistic variables (Tab.1).

Table 1. Associations between SVO/ SOV word order and other linguistic features (adapted from Fenk-Oczlon & Fenk, 2005)

SVO
low number of syllables per word
high number of phonemes per syllable
low number of syllables per clause
high number of words per clause
low number of morphological cases
isolating or fusional morphology prepositions
SOV
high number of syllables per word
low number of phonemes per syllable
high number of syllables per clause
low number of words per clause
high number of morphological cases
agglutinative morphology postposition

The interactions found between word order and other linguistic variables may contribute to explaining SOV/SVO variation. For example: It has been argued that SOV word order is preferred in agglutinative languages, because of their tendency to have very long verb forms containing much grammatical information (Fenk-Oczlon, 1983). Placing long and less predictable units late, conforms to the constant flow of linguistic information hypothesis and to Ferrer-i-Cancho's (2017a) principle of predictability maximization. Concerning the principle of dependency length minimization (Ferrer-i-Cancho, 2017b; Liu et. al., 2017), the correlations between word length in number of syllables and SOV/SVO word order, suggest the usefulness of integrating word length in dependency minimization models, as already emphasized by Ferrer-i-Cancho (2017b).

References:

Fenk, A. & Fenk, G. (1980). Konstanz im Kurzzeitgedächtnis – Konstanz im sprachlichen Informationsfluß? Zeitschrift für experimentelle und angewandte Psychologie, 27: 400–414.

Fenk-Oczlon, G. (1983b). Ist die SVO-Wortfolge die „natürlichste"? Papiere zur Linguistik, 29: 23–32.

Fenk-Oczlon, G. & Fenk, A. (2005). Crosslinguistic correlations between size of syllables, number of cases, and adposition order. In: Fenk-Oczlon, G. & Winkler, C. (eds). Sprache und Natürlichkeit. Gedenkband für Willi Mayerthaler. Tübingen: Narr: 75–86.
Ferrer-i-Cancho, R. (2017a) The placement of the head that maximizes predictability. An information theoretic approach. Glottometrics, 39, 2017: 38–71.
Ferrer-i-Cancho, R. (2017b). Towards a theory of word order. Comment on "Dependency distance: a new perspective on syntactic patterns in natural language" by Haitao Liu. et al. Physics of Life Reviews, 21: 218–220.
Köhler, R. (1986). Zur linguistischen Synergetik. Struktur und Dynamik der Lexik. Bochum: Brockmeyer.
Liu, H., Xu, C. & Liang, J. (2017). Dependency distance: a new perspective on syntactic patterns in natural languages. Physics of Life Reviews, 21: 171–193

□ □ □ □

Ramon Ferrer-i-Cancho, Carlos Gómez-Rodríguez and Juan Luis Esteban

The Origins of the Scarcity of Crossing Dependencies in Languages

The structure of a sentence can be represented as a network where vertices are words and edges indicate syntactic dependencies (Mel'čuk, 1988). In this network, the space is defined by the linear order of words in a sentence. The number of crossings of real sentences is related to the computational complexity of language. Interestingly, crossing syntactic dependencies (edges that cross when drawn above a sentence) have been observed to be infrequent in human languages (Hays, 1964; Lecerf, 1960). Only recently, the number of crossings has been shown to be significantly small with respect to different kinds of random baselines (Ferrer-i-Cancho, Gómez-Rodríguez & Esteban, 2018). This leads to the question as to whether the scarcity of crossings in languages arises from an independent and specific constraint on crossings. We provide statistical evidence suggesting that this is not the case, as the proportion of dependency crossings of sentences from a wide range of languages can be accurately estimated by a simple predictor based on a null hypothesis on the local probability that two dependencies cross given their lengths. The relative error of this

predictor never exceeds 5% on average whereas the error of a baseline predictor assuming a random ordering of the words of a sentence is at least 6 times greater (Gómez-Rodríguez & Ferrer-i-Cancho, 2017). Our results suggest that the scarcity of crossings in natural languages is neither originated by hidden knowledge of language nor by the undesirability of crossings per se, but as a mere side effect of the principle of dependency length minimization, namely the minimization of the Euclidean distance between linked words. We will review the statistical support for such a principle and the mathematical theory of crossings that we have developed to reach the conclusions above.

References:

Ferrer-i-Cancho R., Gómez-Rodríguez C. & Esteban J.L. (2018). Are crossing dependencies really scarce? Physica A: Statistical Mechanics and its Applications, 493: 311–329.

Gómez-Rodríguez, C. & Ferrer-i-Cancho, R. (2017). Scarcity of crossing dependencies: a direct outcome of a specific constraint? Physical Review E, 96: 062304.

Hays, D.G. (1964). Dependency theory: A formalism and some observations. Language, 40: 511–525.

Lecerf, Y. (1960). Programme des conflits, modèle des conflits. Bulletin bimestriel de l'ATALA, 1(4): 11–18, (5): 17–36.

Mel'čuk, I. (1988). Dependency Syntax: Theory and Practice. Albany, N.Y.: State University of New York Press.

□ □ □ □

Greta Franzini, Mike Kestemont, Gabriela Rotari, Melina Jander, Jeremi Ochab, Emily Franzini, Joanna Byszuk and Jan Rybicki

Attributing Authorship in the Noisy Digitised Correspondence of Jacob and Wilhelm Grimm

This paper presents the impact of different digitisation strategies in computational text analysis. More specifically, in a body of uncorrected HTR'd (Handwritten Text Recognition) and OCR'd (Optical Character Recognition) correspondence of Jacob and Wilhelm Grimm, it reports on the effect the imperfect digitisation has on the analyses

necessary to computationally identify the different writing styles of the two brothers.

The HTR'd version is derived from digital images of the handwritten documents (Collection "340 Grimm") and the OCR'd version of the letters from a print critical edition (Rölleke, H., 2001) (hence, beside recognition errors, there are introduced minor editorial changes), which result in two different instances of processing noise, usually encountered by researchers working with literary texts. The selected sample of 72 letters (28 by Wilhelm, 44 by Jacob; each manually transcribed (MAN), HTR'd and OCR'd) varied in readability, cleanliness of digitisation and date of writing.

In general, HTR produced more errors than OCR. In HTR they were numerous enough to significantly change word distributions as measured by lexical richness scores (Shannon entropy and Simpson's index), which suggests the effect can influence word- or lemma-based authorship attribution methods. As a by-product, we also corrected for the dependence of the richness on text length leading to the conclusion that the letters of the two brothers differ significantly in that respect.

Next, we scrutinised the outcome of the authorship attribution task. Support vector machines on character bi-, tri- and tetragrams with leave-one-out cross-validation show equal accuracy and F1 score for MAN and OCR, and lower results for HTR. The misattributions were not restricted by chronology or letter length; they were also consistent between MAN and OCR, and mostly consistent with HTR.

Further, we checked how a model trained on one digitisation method performs on texts from other methods. Out of the nine pairs, MAN->MAN (method both trained and tested on manually transcribed letters) was significantly superior to others (followed by OCR->OCR), while MAN->HTR and especially OCR->HTR were significantly inferior, the latter indicating the adverse effects of accumulated errors and differences between the methods.

Finally, we analysed the attribution success as a function of a text's error rate with the use of probit regression. The stability of the results was enhanced by a bootstrap procedure consisting of sampling lines of texts with known character recognition error rates. A significant dependence was established for misattributions due to HTR error rates, with an additional bias owing to the uneven ratio of samples coming from the two authors.

In summary, our findings show that OCR digitisation serves as a reliable proxy for the more painstaking process of manual digitisation, at least when it comes to authorship attribution. Our results suggest that attribution is viable even when using training and test sets from different digitisation pipelines.

References:

Collection "340 Grimm" purchased from the Hessen State Archive in Marburg, Germany <http://www.unimarburg.de/uniarchiv/grimm> [accessed: 16 November 2015].

Rölleke, H. (2001). Briefwechsel zwischen Jacob und Wilhelm Grimm. Stuttgart: Hirzel Verlag.

□ □ □ □

Wahed Hemati, Alexander Mehler, Tolga Uslu, Daniel Baumartz and Giuseppe Abram

Evaluating and Integrating Databases in the Area of NLP

Since computational power is rapidly increasing, analysing big data is getting more popular. This is exemplified by word embeddings producing huge index files of interrelated items. The second example is given by digital editions of corpora representing data on nested levels of text structuring. The third example relates to annotations of multimodal communication comprising nested and networked data of various (e.g., gestural or linguistic) modes. While the first example relates to graph-based models, the second one requires document models in the tradition of TEI whereas the third one combines both models. A central question is how to store and process such big and diverse data to support NLP and related routines in an efficient manner.

In this paper, we evaluate six Database Management Systems as candidates for answering this question. This is done by regarding database operations in the context of six NLP routines. We show that none of the DBMS consistently works best. Rather, a family of them manifesting different database paradigms is required to cope with the need of processing big and divergent data. To this end, the paper introduces a web-based multi-database management system (MDBMS) as an interface to varieties of such databases.

□ □ □ □

Antoni Hernández-Fernández, Iván González Torre, Lucas Lacasa, Jordi Luque and Bartolo Luque

Do Linguistic Laws Emerge from the Voice?

Linguistic laws have been routinely investigated in quantitative linguistics both in written corpora and in oral corpora but transcribed to their written "equivalent". This means that inferences of statistical patterns of language are biased by the arbitrary choice of segmentation of the acoustic signal, and makes difficult to establish "physical laws" and the comparative studies between the human voice and other communication systems.

Acoustic communication is fully determined by three physical magnitudes extracted from signals: frequency, energy and time. Human voice seems to be operating close to a critical state (Luque, Luque & Lacasa, 2015) and probably due to that fact when we explore directly human speech we find out some statistical universals of language (so-called linguistic laws) (González Torre et al., 2017). Interestingly, it is well known that we use statistical cues to segment the input and probably share with other species some of these mechanisms.

In a previous work, a method was explored to directly study acoustic speech signals (Luque, Luque & Lacasa, 2015), directly sampling .wav files and automatically generating speech events characterized by energy and time duration. The method has been applied to sixteen different languages (Basque, Catalan, Galician, Spanish, Portuguese, English, Japanese, Vietnamese, Mandarin, Korean, Taiwanese, Arabic, French, German, Hindi and Tamil) recovering successfully some well-known laws of human communication for the very first time both for energy and time duration at levels below the phoneme (González Torre et al., 2017):

- Zipf's Law: here establishes that in a sizable linguistic sample the number of different binned elements N(n) which occur exactly n times decays as $N(n) \sim n^{-\alpha}$.
- Brevity Law: is the tendency of more frequent elements to be "shorter" (here discovered both in energy and duration for the first time) as a result of the compression principle (Ferrer-i-Cancho et al., 2013).
- Heaps-Herdan's Law: shows the sublinear growth of the number of

different elements V in a corpus with size L, measured in total number of elements (V~L^b).

- Guttenberg-Richter's Law: the probability distribution of energy during speech is a power law (P(E)~E^-a) and paves the way of understanding speech production in terms of crackling noise and Self Organized Criticality (Luque, Luque & Lacasa, 2015).

Are these laws the result of processes proper to human physiology? Or are they universal for acoustic communication? Future work is necessary to extend this protocol to other acoustical communication systems (i.e. non-human primates or cetaceans). If these scaling laws are indicative of complex communication we discuss here how these methods further pave the way for new comparative studies in animal communication or the analysis of signals of unknown code.

References:

Ferrer-i-Cancho, R., Hernández-Fernández, A., Lusseau, D., Agoramoorthy, G., Hsu, M. J. & Semple, S. (2013), Compression as a Universal Principle of Animal Behavior. Cognitive Science, 37: 1565–1578 [doi:10.1111/cogs.12061].

González Torre, I., Luque, B., Lacasa, L., Luque, J. & Hernández-Fernández, A. (2017). Emergence of linguistic laws in human voice. Scientific Reports, 7: 43862 [DOI:10.1038/srep43862].

Luque, J., Luque, B. & Lacasa, L. (2014). Scaling and universality in the human voice. J. R. Soc. Interface, 12: 20141344 [DOI: 10.1098/rsif.2014.1344].

□ □ □ □

Lars Johnsen

Generalizing TF-IDF to Collections of Documents

In this talk we present a generalization of the tf-idf measure for terms within a document and across documents. The goal is to find a way of measuring terms within a collection, and comparing collections of documents. The formula we will discuss and employ is delta-tf * delta-df, where delta refers to the ratio of the measure in one collection divided by the same measure in another collection. This method can in turn be generalized to collocation analysis, where word vectors associated with particular words formally can be viewed the same way as documents.

□ □ □ □

Emmerich Kelih

Parts of Speech – Theoretical Problems, Empirical Tendencies and Modelling

Parts of speech play an essential role in morphology, syntax, lexical studies and semantics. From a quantitative point of view in particular the frequency of parts of speech is of outstanding interest. Our contribution starts with a brief discussion of the well-known problem of the theoretical vagueness of these linguistic entities and related problems of an appropriate definition. In the second part known empirical observations about the frequency of parts of speech will be presented, in particular a discussed stability of parts of speech in corpora (e.g. Hudson 1994; Liang & Liu 2013) and the raised hypothesis (sometimes called "Ohno's law" cf. Mizutani, 1989) about the correlation of the occurrence of parts of speech in texts (e.g. number of nouns is correlated with the number of adjectives etc.). Additionally to the mentioned observed empirical tendencies the problem of the modelling of parts of speech frequencies has to be discussed (Kelih & Altmann 2018: 36-39). For modelling the frequencies of parts of speech of Slovene texts (cooking recipes, journalistic articles, sonnets, and dialogues) the function $y = 1 + a*\exp(-bx)$ (Popescu, Altmann & Köhler, 2010) seems to be an appropriate one.

References:

Hudson, R. (1994). About 37% of all word-tokens are nouns. Language, 70: 331–339.

Kelih, E. & Altmann, G. (2018). Problems in quantitative linguistics 6. Lüdenscheid: Ram-Verlag.

Liang, J. & Liu, H. (2013). Noun distribution in natural languages. Poznan Studies in Contemporary Linguistics, 49: 509–529.

Mizutani, Sh. (1989). Ohno's lexical law: its data adjustment by linear regression. In: Mizutani, Sh. (ed.), Japanese Quantitative Linguistics. Bochum: Brockmeyer: 1-13.

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

□ □ □ □

Miki Kimura

Authorship Attribution with Topic Modelling: Alice Bradley Sheldon and Her Contemporaries

This paper presents a quantitative authorship attribution analysis of the works by Alice Bradley Sheldon (1915–1987), an American writer of feminist science fiction, who used two pseudonyms – James Tiptree, Jr. and Raccoona Sheldon – to disguise her true identity. Given that as a commercial strategy Alice Sheldon masqueraded as the male James Tiptree, Jr. for almost a decade, many critics have discussed the author's identity and gender. The most well-known literary critique was made by an American science fiction writer, Robert Silverberg. Silverberg (1975) and Kotani (1999) insisted, with reference to the style of Ernest Hemingway, that James Tiptree's stories were written by a man. In this study, I performed a quantitative stylistic analysis of Sheldon's work, comparing it with those of a group of male and female writers: Theodore Sturgeon, Arthur C. Clarke, Ursula K. Le Guin, and Octavia E. Butler. All the writes I used here were science fiction writers whose careers overlapped Sheldon's. The Sheldon corpus compiled for this study contains all of Alice Sheldon's published works under both her pen names (72 works with 865,802 word tokens). The other four corpora contain all of Sturgeon's works (222 works with 1,777,561 word tokens), Clarke's works (104 works with 467,983 word tokens), Le Guin's works (45 works with 589,481 word tokens), and Butler's works (93 works with 867,396 word tokens).

In this study, by employing a non-supervised machine learning method, Topic Modelling, emphasis is primarily placed on the inter-author variation between Sheldon's works and the works by the other five science writers. The emphasis of this research is placed not only on discriminating the works by these six authors but also on comparing the results from this quantitative authorship attribution with research of such literary criticism scholars as Silverberg (1975) and others. Variables (i.e. Topics) which are effective for this kind of discrimination will be identified. The topics found via Topic Modelling, which are considered effective for discrimination as discussed by Murakami et al., (2017), have been chosen as variables for the anal-

ysis. The discriminate variables chosen from the corpus are sensitive as identifiers, the results from Topic Modelling, for example, should show that they can detect inter-author variation between works by Alice Sheldon's and the ones by other science fiction writers. Then, I will inspect whether the works written by Alice Sheldon are similar to the analysed works by the male or female writers. By comparing all of Alice Sheldon's works with those of her contemporaries, we hoped to find clues about Sheldon's allegedly masculine writing style.

□ □ □ □

Eduard Klyshinsky

The Method of Quantitative Evaluation of Syntactical Inversion

The inversion of word order demonstrates such language phenomena as intonation, negation, question, etc. and was investigated for a variety of languages (e.g. English, French, Greek, Hebrew). Changing word orders among languages makes statistical machine translation more challenging (Birch et al, 2009). However, changing word order in a language makes the learning of this language more complicated since one should formulate more rules in his or her mind. The same is true for syntactic analysis. Thus, there is a great need in verifiable numerical methods for quantitative evaluation of regularity of word order.

For this purposes we evaluated the relation between left and right branching of the same syntactical connections, including a part of speech of head and tail word and a type of the connection. If for the given type of connection the probability of finding a tail word in the left position is equal to the probability of the right position, it means that the considered language has completely irregular phenomenon from the point of view of syntax. Contrariwise, if a tail word could be meet in left or right position only, the connection is completely regular. The average among such values helps evaluate the degree of regularity of a language.

However, the completely irregular connection could be extremely rare. That is why we introduced the importance of a connection calculated as product of degree of irregularity by the frequency of the connection in a corpus.

We evaluated the degree of such syntactical irregularity for 33 treebanks from the Universal Dependencies corpus (Nivre et al., 2016). Our evaluation demonstrates that the Estonian, Finnish and Slovak languages have the most free word order while Japan, Hindi and Urdu have the most strict one. The examined languages demonstrate very good correlation inside the language families. We also have found that the highest probability of inversion corresponds to following connections: <NOUN, VERB, nsubj>, <NOUN, VERB, obl>, <VERB, VERB, advcl>, <ADV, VERB, advmod>.

References:

Birch, A., Blunsom, Ph., Osborne, M. (2009). A quantitative analysis of reordering phenomena. In Proc. of StatMT '09 the Fourth Workshop on Statistical Machine Translation: 197–205.

Nivre, J., Marneffe M.-C., de, Ginter F. et al. Universal Dependencies v1: A Multilingual Treebank Collection. In Proc. of LREC-2016: 1659–1666.

□ □ □ □

Miroslav Kubát, Jan Hůla, David Číž, Kateřina Pelegrinová, Xinying Chen and Radek Čech

Context Specificity Analysis of Function and Content Words

This study aims to analyse the differences between function (synsemantic) and content (autosemantic) words from the perspective of the so-called context specificity of lemma (CSL). This new approach is based on neural networks, specifically the word embedding technique, where each lemma is represented by a vector. The size and orientation of a vector express the position of a lemma in a semantic multi-dimensional space. Thus, it is possible to measure similarities among lemmas. In case there are two lemmas which appear in the very same context in the corpus, these vectors would be identical. We decide to apply namely Closest Context Specificity (CCS) to our analysis. CCS measures the average value of the similarities of the 20 closest lemmas of the target lemma. The research is based on the corpus of the Czech journalism consisting of more than 3 billion tokens.

□ □ □ □

Tatiana Litvinova, Pavel Seredin and George Mikros

The Effect of Author Sex and Gender on Parameters of Written Texts with and without Gender Deception

The problem of studying differences between male and female speech has long been investigated by sociolinguists. Computational linguists are getting involved in this problem as well while gender identification is set to become a text classification task. This growing interest in this task is due not only to the scientific but also practical significance of the issue. As there is an increasing amount of Internet communication, it is currently essential to identify demographic characteristics of authors of texts for marketing where there has to be relevant information about target audiences leaving positive/negative reviews of products and services. In criminology "profiles" of anonymous texts containing threats, etc. have to be established. Despite a large number of papers on the issue, it has not been sufficiently addressed yet. There is currently no analysis of the joint influence of the biological sex and psychological gender of authors (in terms of their femininity and masculinity) on the quantitative parameters of their texts. Linguistic characteristics of texts written with intentional modifications of style in order to imitate an individual of the opposite sex (gender deception) have not been looked much into. It is known that there are parameters that can be imitated and those that cannot but there is no agreement on what exactly these parameters are. What makes it even more challenging is the fact that there are no special text corpora available. We have conducted a pilot study to identify the effect of the biological sex and femininity/masculinity (psychological gender) of authors as well as types of texts (with and without gender imitation and style obfuscation) on different groups of quantitative parameters of a Russian written text using a specially designed text corpus "Gender Imitation Corpus". Gender Imitation Corpus is the first Russian corpus for studies of stylistic deception. Each respondent (n=142) was instructed to write 3 texts on the same topic (from a list). The first text is supposed to be written in a way usual for whoever writes it ('normal style'), the second one should be written as if by someone of the opposite gender ('gender imitation'); the third one should be as if

one by another individual of the same gender so that their personal writing style will not be recognized ('style obfuscation'). Most of the texts are 80-150 words long. Multivariate statistical analysis (MANOVA) revealed the effect of biological sex and type of text on linguistics parameters but femininity/masculinity of authors was shown to have no effect on the text parameters chosen for the study. In particular, it was found that irrespective of text type ('normal' style/gender imitation/style obfuscation) men and women differ in their use of function words. Type of text has an effect on parameters of vocabulary richness. These data can be further exploited for the development of author's gender profiling systems which take into account gender deception scenario.

□ □ □ □

Jan Mačutek, George Mikros, Andrij Rovenchak and Valentin Vydrin

Menzerath–Altmann Law across Several Levels of Language Unit Hierarchy

The Menzerath-Altmann law (e.g. Cramer, 2005a) says that the mean size of constituents is a function of the size of the construct (e.g., word length influences the mean length of syllables of which the word consists). It was shown to be valid in many languages and between many language units (e.g., words and syllables, clauses and words, sentences and clauses, etc.). The relation between the sizes can be expressed as the function

$$y(x)=a\, x^{b}\, e^{(-cx)},$$

where $y(x)$ is the mean size of constituents if the size of the construct is x; a, b, c are parameters.

Most often, a special case of the formula with $c=0$ is sufficient.

While the validity of the law is widely accepted, an interpretation of its parameters remains an open problem. This question was presented already by Cramer (2005b), and recently by Mačutek et al. (2018). In order to gain some insight into this area, it seems reasonable to model the law across several language units in the same text so that an impact of the language unit can be investigated and typical parameter

values can be observed.

We will present results (parameter values for the Menzerath-Altmann law) for the neighbours in the language unit hierarchy (syllable – word – clause – sentence, and syllable – word – word length motif) in two Indo-European languages (Modern Greek and Ukrainian) and in one language from the Niger-Congo family (Bamana).

References:

Cramer, I.M. (2005a). Das Menzerathische Gesetz. In: Köhler, R., Altmann, G., Piotrowski, R.G. (eds).
Quantitative Linguistics. An International Handbook. Berlin, New York: de Gruyter: 659–688.
Cramer, I.M. (2005b). The parameters of the Menzerath-Altmann law. Journal of Quantitative Linguistics, 12(1): 41–52.
Mačutek, J., Chromý, J., Koščová, M. (2018). Menzerath-Altmann law and prothetic /v/ in spoken Czech. Journal of Quantitative Linguistics [DOI: 10.1080/09296174.2018.1424493].

Acknowledgement:

Supported by grant VEGA 2/0054/18 (JM).

□ □ □ □

Piotr Malak, Elżbieta Herden and Adam Pawłowski

Reading Bibliographies: Methods of Semi-Automatic Categorization of Short Texts

There are many unsupervised algorithms of text classification of proven accuracy. The canonical way traverses from texts pre-processing to terms weighting, excluding stop words and finally text normalisation. Then relevant features are selected and models are built for automatic classification. A new approach provided in recent years by NLP researchers consists in deep learning based on rather complex and foggy algorithms, which, however, prove to be relatively efficient and accurate. One of such algorithms is word2vec, the other one is called fastText. We compare the efficiency and pertinence of classification of short text performed with the help of traditional and modern classifi-

cation approaches. In particular we compare MLP, SGDClassifier and SVMLinear versus word2vec and fastText techniques. The research was carried out on a corpus of ca 500,000 bibliographical records extracted from the database of the National Library of Poland. While short texts to be classified are limited to document titles, keywords and description fields of the Dublin Core format will be used as reference for verification. It should be emphasised that short texts are challenging for NLP and ML methods. That is why techniques less vulnerable to text density and its inner complexity may be promising solutions.

□ □ □ □

Vladimir Matlach, Diego Krivochen, Lukáš Zámečník and Jiří Milička

Randomness Classification

Nowadays, data scientists face a phenomenon which we might call the question of randomness. This phenomenon impacts not only hard data scientists but nearly anyone in the modern world, counting in statisticians, genome and DNA analytics, online-banking users, computer users in general, and even linguists. The simplest way to understand this problem is looking at a string, such as this one (where T-A-C-G are nucleotides):

...AGCTGCGGCTGCATGCTCCGCAAAGCTTCGATCG...

or this one:

...10101011101000011111011001001110100010101011001...

The first example is a DNA substring which stores instructions for protein assembly. It is a pre-defined process with very specific combinatory rules, and finding proteins is a matter of searching for anything that at least looks as if it was made by a grammar. The second example is an important string as well. In contrast to DNA, in which linear order conveys structural meaning, we wish to characterize it as random as it can be: how can we measure the randomness of a string?

This work introduces a new method of studying randomness of a given string, quantifying its properties and allowing to cluster (or group) analysed strings together based on their characteristics. This is rele-

vant for the study of both formal and natural languages, because we can address questions about the nature of the computational system that underlies human language by analysing the quantitative properties of its output. What we propose is a simple method that can distinguish random strings from non-random strings with relative success. The method itself consists of a simple algorithm studying string complexity and its combinatorics resulting to a vector characterizing the randomness. Moreover, we will show that strings which have been obtained by means of different formal grammars (Finite-State to Context-Sensitive) cluster together in a way that betrays the recursive procedure that has been employed to generate them. Pseudo-random texts (e.g. Lorem Ipsum) and 'monkey-typing' can also be readily identified and also distinguished from true random texts. The method proposed here complements existing algorithmic procedures in cryptography (NIST 2008; Hamano & Yamamoto 2010) as well as autocorrelation and density measures (Patel et al. 2015, for Lindenmayer systems), and methodically outperforms many of the diehard tests based on brute force mechanisms (Marsaglia, 1995). The resulting vectors can be used as input for machine learning or data mining algorithms, such as SVM or other classifiers.

□ □ □ □

Jiří Milička and Alžběta Růžičková

Demand and Supply in the Communication Process: The Case of Lexical Richness and Phonological Features

Production and perception are two sides of the same coin – the shape of a language and the properties of a text are products of human abilities in both of them. The quantitative linguistic theories – e.g. the Zipfian principle of least effort (Zipf, 1949), the Altmannian theory proposal (Altmann, 1978) and the Köhlerian synergetic control circle (Köhler, 1986) – take into account that the language is a trade-off between the demands and capabilities of the text receiver and the capabilities of the producer. In other words, the producer tries not only to minimize their effort but also to maximize the success of the communication

and therefore accommodates their texts to satisfy the needs of the receiver. The main objective of the study is to explore how texts meet the demands of readers. While the corpora are mostly utilized to study the production of texts, there is no inherent reason why they could not be used to study perception. We compiled a corpus of Czech blogs to explore the relation between lexical richness and phonological features of texts and their success. Moving average type-token relation (MATTR) and moving average entropy (MAH) were used as lexical richness metrics, and several phonological features (PF) connected with the euphony, such as vowel/phoneme ratio, open syllable ratio, consonant cluster distribution etc., were taken into account. The number of views and relative number of likes were employed as text success metrics.

The following competing hypotheses emerged: 1. there is no relation between the lexical richness (or PF) and the text success at all;

the lexically less rich texts are more popular than the lexically richer ones, as they are more readable;

1. for the text producer it is difficult to attain the lexical richness (PF values) level that is ideal for the receiver, thus the positive correlation between lexical richness (or PF) and text success can be observed;

2. the text producing abilities evolved so that the lexical richness (and PF) level which is ideal for the producer is the same as the ideal level for the text receiver.

It has turned out that the most frequent lexical richness metrics values correspond with the values of the texts whose average number of views is the highest one, which is in accordance with the theories that expect some degree of self-organization in language. Contrary, lexically rich texts have on average a higher relative number of likes and with the descending lexical richness metrics values the average relative number of likes is also descending. The text success is almost independent of the phonological features of the texts.

References:

Altmann, G. (1978). Towards a Theory of Language. Glottometrika, 1: 1–26.

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

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

□ □ □ □

Michal Místecký

Calculating the Victory Chances: A Stylometric Insight into the 2018 Czech Presidential Election

The contribution focuses on performing a set of stylometric analyses on a corpus of texts connected to the Czech presidential election, which took place in January 2018. A similar analysis has already been carried out for Czech, Italian, and American presidents (Čech, 2014; Kubát & Čech, 2016; Rimkeit-Vit & Gnatchuk, 2016; Zörnig & Altmann, 2016). The samples will include presidential candidates' web pages articles and interviews, and will be studied via traditional style-measuring indexes (e.g., vocabulary richness indicators, information-measuring entropy, average tokens length, activity, verb distances, and thematic concentration – cf. Čech et al., 2014; Kubát et al., 2014) in order to investigate the extent of interpretable style features (extent of lexis, degree of information and intellectuality, narrative / descriptive character, syntactic complicatedness, topic-focus / topic fragmentation). The counts will undergo multifarious follow-up calculations: the differences between the candidates will be tested on statistical significance, and cluster analyses will be carried out, so that the closeness of the potential presidents is accounted for. A special attention will be paid to the texts of Jiří Drahoš, who was the opponent of Miloš Zeman, the incumbent president, in the second round of the election; here, a collocation analysis will be elaborated, so as to assess the main attributes he added to the core notions of the campaign (such as "state", "nation", or "president").

The goal of the contribution is to uncover various tactics used by individual candidates to attract potential voters, and evaluate the level of their success in the election. Although it will not be a point of the presentation to pay heed to the reasons of Drahoš's failure, it will try to prove the utility of the used calculations for political marketing, or at least for the aftermath analysis of the results of the voting struggle.

□ □ □ □

Hermann Moisl

Topological Mapping for Visualisation of High-Dimensional Historical Linguistic Data

Discovery of the chronological or geographical distribution of collections of historical texts can be more reliable when based on multivariate rather than on univariate data because, assuming that the variables describe different aspects of the texts in question, multivariate data necessarily provides a more complete description. Where the multivariate data is high-dimensional, however, its complexity can defy analysis using traditional philological methods. A variety of mathematical and statistical methods is available for help in such cases; the present discussion proposes topological mapping of high-dimensional data abstracted from historical text corpora into low-dimensional space as a way of visualising structure which may be latent in the data but invisible to direct inspection.

The discussion is in four main parts.

- The first part introduces some mathematical concepts used in the discussion: vectors, vector spaces, and manifolds;
- The second part identifies two problems with high-dimensional multivariate data: interpretability and nonlinearity;
- The third part shows how topological mapping solves these problems by projecting high-dimensional nonlinear data into a low-dimensional linear vector space;
- The fourth part exemplifies the application of topological mapping to high-dimensional nonlinear data abstracted from a collection of English-language texts from different historical periods, showing how the map is able to identify the known relative chronology of the texts.

□ □ □ □

Jinghui Ouyang and Jingyang Jiang

Minimization and Probability Distribution of Dependency Distance in the Process of Second Language Acquisition

Dependency distance minimization (DDM), resulting from the constraint of working memory capacity and the effect of "the principle of least effort" on syntactic structure is found to be a universal quantitative property of many natural languages. However, there are no such studies on second language learners' language system. To investigate whether second language learners develop the interlanguage system under the same constraint, we selected 367 participants of Chinese English learners of nine consecutive grades, and built one second language dependency treebank and two corresponding random treebanks (RL1: with one word as root and randomly selected another word as its governor; RL2: random words as root and governor, but without crossing edges) and fitted different probability distribution models to dependency distances. It was found that:

(1) The mean dependency distance (MDD) of English writings by Chinese students increases significantly across nine grades, suggesting MDD can measure language proficiency of second language learners. However, the MDD of higher-level learners remains stable and doesn't reach the level of English native speakers in our contrastive dependency treebank extracted from Wall Street Corpus.

(2) The MDDs of learners' interlanguage at different learning stages are significantly lower than their corresponding random languages. The relative stability of MDD at the advanced learning stage and the lower MDDs as compared with RL1 and RL2 indicates the tendency of DDM.

(3) The distribution of dependency distances of RL1 of second language learners' writings cannot fit the Zipf-Alekseev distribution, but that of RL2 of the same writings can. Besides, the parameters in the Zipf-Alekseev distribution of RL2 have no correlation with second language learners' language proficiency.

Our previous study (Ouyang & Jiang, 2017) found that the distribution of dependency distances of Chinese English learners' interlanguage can fit the Zipf-Alekseev distribution and well reflect learners' language proficiency at different learning stages. Projectivity is the

mechanism behind that makes the probability of dependency distances of both natural language and RL2 fit the Zipf-Alekseev distribution. The phenomenon that the parameters of natural languages can reflect second language learners' language proficiency, while the parameters of RL2 cannot, can only be explained by syntax. The current study corroborates that DDM is a language universal not only in first language, but also in second language. This helps clarify the relationship between human cognition and language.

□ □ □ □

Katharina Prochazka and Gero Vogl

Language Contact in Austria-Hungary – Differences between the Two Parts and Consequences for Mathematical Modelling

From its outset, the Habsburg Empire was a multinational and multilingual state, with language contact occurring in many areas. By the Austro-Hungarian Compromise of 1867, the state was divided into two parts which were mostly independent, i.e. the dual monarchy of Austria-Hungary.

In both parts of Austria-Hungary, language shift processes (people giving up use of one language in favour of another) took place. People switched from the minority to the majority language German or Hungarian. This can be seen e.g. from census data from 1880 onwards. Although the census data do not necessarily reflect actual linguistic competence (Brix, 1982; Lieberson, 1966; Vries, 2006), they cover the entire population and thus allow quantitative research on the temporal and spatial development of language shift in Austria-Hungary. In a second step, these data can then be mathematically modelled (Prochazka & Vogl, 2017) to get an insight into the underlying processes that influence language shift. However, mathematical models of language shift depend on both the available data and the circumstances of language shift.

This presentation compares two language contact situations in the different parts of Austria-Hungary:

(1) Slovenian-German in Carinthia (Austrian half)

(2) German-Hungarian in Baranya/Tolna (Hungarian half)

We will examine the different circumstances in these two situations and the implications of these differences for a quantitative mathematical description. In particular, we will look at two key factors: settlement and population structure, which is important for a spatial description of language shift and different language politics, which accelerate or slow down language shift. How can information on these factors be exploited for building an accurate model?

References:

Brix, E. (1982). Die Umgangssprachen in Altösterreich zwischen Agitation und Assimilation: die Sprachenstatistik in den zisleithanischen Volkszählungen 1880 bis 1910. Wien: Böhlau.

Lieberson, S. (1966). Language questions in censuses. Sociological Inquiry, 36(2): 262–279.

Prochazka, K. & Vogl, G. (2017). Quantifying the driving factors for language shift in a bilingual region. PNAS, 114(17): 4365–4369.

Vries, J., de (2006). Language Censuses. In: Ammon, U. et al. (eds). Sociolinguistics. An International Handbook of the Science of Language and Society. Vol. 3.2. Handbücher zur Sprach- und Kommunikationswissenschaft/Handbooks of Linguistics and Communication Science (HSK). Berlin: Walter de Gruyter: 1104–1116.

□ □ □ □

Marija Radojičić, Biljana Lazić, Sebastijan Kaplar, Ranka Stanković, Ivan Obradović and Ján Mačutek

Quantitative Analysis of Syllable Properties in Some Slavic Languages

Initial results from the Slovak-Serbian bilateral project "Quantitative analysis of syllables in Slavic languages (Russian, Slovak, and Serbian)" will be presented. The analysis of syllables properties will be based on two parallel texts, namely, translations of several chapters from Russian novels "The Master and Margarita" by Mikhail Bulgakov, and "How the Steel Was Tempered" by Nikolai Ostrovsky. For the purpose of this project, a software tool for a syllabification of texts was

developed. The syllabification is based on two principles: (1) maximal onset, (2) sonority hierarchy (e.g. Zec, 1995).

We will focus on modelling of two properties: (1) the rank-frequency distribution of syllables (syllable frequencies were studied in the past mostly from a psycholinguistic point of view, see several contributions in Cairns and Raimy, 2011; the frequencies are presented in percentages, if at all, and therefore are impossible to model), (2) syllable length (measured on the number of graphemes). Models for both syllables and canonical syllable types (CV-structure, where C stands for a consonant and V for a vowel) will be presented. The relation between syllable length and syllable frequency will be investigated as well.

Data obtained within the project will also be used to further develop a synergetic model for syllables which was briefly sketched by Kelih (2012: 139-150) and to test several hypotheses mentioned in Strauss et al. (2008).

References:

Cairns, C.E., Raimy, E. (eds). (2011). Handbook of the Syllable. Leiden, Boston: Brill.

Kelih, E. (2012). Die Silbe in slawischen Sprachen. Von der Optimalitätstheorie zu einer funktionalen Interpretation. München et al.: Otto Sagner.

Strauss, U., Fan, F. & Altmann, G. (2008). Problems in Quantitative Linguistics, 1. Lüdenscheid: RAM-Verlag.

Zec, D. (1995). Sonority constraints on syllable structure. Phonology, 12(1): 85–129.

Acknowledgement:

Supported by grants SK-SRB-2016-0021 (MR, BL, RS, JM) and VEGA 2/0054/18 (JM).

□ □ □ □

Haruko Sanada

N-grams of Valency Types and Their Significant Order in the Clause

Aim of the study: The present study is one of a series of empirical studies on Japanese valency. It investigates a statistically significant order of valency types (complements and adjuncts) in the clause by employ-

ing a frequency of n-gram data of valency types. The literature shows that there are broad rules for the order of valency types in the clause though Japanese is an agglutinative SOV language and that complements are omittable. Our former studies (Sanada, 2017) found that there are common patterns of neighbouring valency types.

Data: We employed the Japanese valency database (Ogino et al., 2003), which is the same as the one employed in our last studies. 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.

Hypotheses and methods of analyses: The clause with the verb meet consists of complements – subject (or topic), object, and predicate, which is located at the end of the clause. We set the following hypotheses: (1) The other valency types, namely time, place, and occasion, should be allocated as adjuncts in no particular order between the subject and object; (2) the subject and object must form a kind of 'bracket structure' of the valency in the clause.

For each valency type, we investigated the frequencies of bi-grams of the valency types forward and backward. We performed '2-sample tests for equality of proportions without continuity correction data' for pairs of the forward and backward bi-grams to confirm a significant order of neighbouring valency types. For significant pairs, Cohen's d was also obtained as the effective size.

Results and conclusions: We concluded the following points as significant preference of the orders of valency types:

1. The subject appears at the first position, and the object appears in the position close to the end of the clause.

2. Time and place appear before the object, and time comes before the place.

3. Occasion takes a position before the subject. It also takes a position after the object, and it can appear between the subject and object. Therefore, occasion gives an impression that the Japanese sentence is a free order language.

The subject and object form a kind of 'bracket structure' of the valency in the clause. However, occasion can also appear outside of the 'bracket structure', and the subject and object are omittable in the clause. Therefore, their role should be called as 'anchors' in the clause.

References:

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

Sanada, H. (2017, submitted). Negentropy of dependency types and parts of speech in the clause. In: Jiang, J. & Liu H. (eds). Quantitative analysis of dependency structures. Berlin: Mouton de Gruyter.

□ □ □ □

Tatiana Sherstinova

Some Statistics of Russian Everyday Speech in Sociolinguistic and Pragmatic Aspects

The report presents recent quantitative studies carried out on the data of the ORD corpus of Russian everyday speech known as "One Day of Speech" corpus, which contains real-life recordings made in everyday settings. Now, the ORD collection exceeds 1250 h of recordings, presenting speech of 130 respondents and hundreds of their interlocutors. 2850 macro episodes of everyday spoken communication have been already annotated, and the speech transcripts add up to 1 m tokens. The unique recordings of the ORD corpus allow to make multilevel research of everyday speech in linguistic, sociolinguistic, and pragmatic aspects. The obtained results showed that practically on each linguistic level one there are features exhibiting a very high similarity between different sociolects. At the same time, a list of linguistic features peculiar to certain social groups was revealed as well. Different word frequency lists have been compiled, and the statistics concerning the distribution of speech acts in everyday conversations was obtained.

□ □ □ □

Petra Steiner

Some Quantitative-Linguistic Hypotheses on Case

In linguistics, case is a complex and multi-layered term. Starting from Fillmore's (1968) seminal paper, its terminological meaning shifted more and more to the scope of syntax and semantics. This paper is part of a project which relates the syntactic and semantic aspects of case (valency) to morphological properties (morphological case). In Indo-European languages, valency is a linguistic property of verbs and other parts of speech which is mostly defined in a qualitative way. Therefore, we set quantitative definitions for the semantic and syntactic levels of case. Then we relate these quantitative properties to the level of morphological case by deriving the connections between verb valency and nominal inflectional cases: 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. This paper first 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.

Concerning the syntactic-semantic side, we derive the relations between semantic and syntactic properties of verbs, for example the relation between the obligatoriness of arguments and their ellipsis or positions in constructs. We relate the complexity of syntactic arguments to the well-known Menzerath's law and build on two laws of Behaghel (1932a & 1932b) for the derivation of the quantitative property of argument position in constructs.

The relation between morphological case and syntactic case can be derived from the general principle of functional equivalence (Köhler, 2005: 765). Furthermore, the position of syntactic arguments in linguistic constructs has effects on the frequency of their inflectional markers. A synergetic model for all these linguistics units and properties is derived and corroborated for lexical and textual data. These investigations are providing missing links of former work on Modern Germanic languages (Steiner, 2009 & 2015).

References:

Behaghel, O. (1932a). Deutsche Syntax: Eine geschichtliche Darstellung. Band IV Wortstellung.
Periodenbau. Heidelberg: Winters (Germanistische Bibliothek - Sammlung Germanistischer Elementar und Handbücher – Grammatiken, 10).
Behaghel, O. (1932b). Von deutscher Wortstellung. Zeitschrift für Deutschkunde, 44, 1930: 81–89.
Fillmore, C.J. (1968). The Case for Case. In: Bach, E. & Harms R.T. (eds). Universals in linguistic theory. New York: Holt, Rinehart and Winston: 1–88.
Köhler, R. (2005). Synergetic linguistics. In: Köhler, R., Altmann, G. & Piotrowski R.G. (eds). Quantitative Linguistik/Quantitative Linguistics: Ein internationales Handbuch/An International Handbook. Berlin-New York: de Gruyter. (Handbücher zur Sprach- und Kommunikationswissenschaft / Handbooks of Linguistics and Communicative Science, 27): 760–774.
Steiner, P. (2009). Diversification in Icelandic Inflectional Paradigms. Köhler, R. (ed.), Issues in
Quantitative Linguistics. Lüdenscheid: RAM-Verlag. (Studies in Quantitative Linguistics, 5): 126-154.
Steiner, P.C. (2015). Diversification in the Noun Inflection of Old English. In: Tuzzi, A., Benešová, M. & Ján Mačutek, J. (eds). Recent Contributions to Quantitative Linguistics. Berlin-Boston: de Gruyter Mouton (Quantitative Linguistics, 70): 181–201.

□ □ □ □

Mao Sugiyama

What the First President Said and How Journalists Understood His Words: Comparing the Usage of Key Words from Yeltsin's Presidential Addresses

The aim of this study is to investigate the gap in understanding between the speaker and the reporters, in other words, what the first Russian president Boris Yeltsin delivered to the citizen in the federal assembly in his Russian presidential addresses and how journalists understood what Yeltsin mentioned.

Гаврилова (Gavrilova 2012) researched some features of Yeltsin's rhetoric. According to her analysis, Yeltsin tried to spread an under-

standing of the meaning of 'president' in his different statements, such as an inaugural speech, interviews and biography to research metaphorical and rhetorical expressions.

This study focuses on Yeltsin's style of political speech in his Russian Presidential Addresses to the Federal Assembly, 1994–1999. From this study, Yeltsin's political speech style, as the first president of Russia, was obvious. In addition, this study uses the broadsheet Независимая газета (Nizavishimaya gazeta, Independent Newspaper), in which journalists reported about Yeltsin's addresses. In 1991, Yeltsin changed the media law. However, media changed from an organ party to an independent media, and as Iijima (2009) points out, this situation caused a fund shortage in some newspaper companies. Because of this situation, some struggling newspaper companies had to rely on oligarchs for financial support. On the other hand, the chief editor of "Independent Newspaper" managed to get over this financial crisis.

The result of the correspondence analysis shows the key words in Yeltsin's presidential addresses to the Federal Assembly such as: безопасность 'safety', свобода 'freedom', and кризис 'crisis'. The contexts of these words: кризис 'crisis', described the economic situation of Russia; and, безопасность 'safety' and свобода 'freedom', described Yeltsin's responsibility as a leader of the state and, all three, as elements of a democratic country. On the other hand, journalists reported a different view of Yeltsin's statements. When journalists used the word безопасность 'safety', they referred to the safety assurance system as the organ. Journalists made up the word демократура (demokratura), which was observed in 1994 in the Independent Newspaper. This term is composed of the word democracy and dictatorship. Journalists evaluated Yeltsin's policy with a critical eye and they were concerned about the increased authority of Yeltsin as a leader of the Russian Federation.

□ □ □ □

Arjuna Tuzzi

The Importance of Being Earnest (and Average)

When the aim of a study is comparing and contrasting texts of the same genre and achieving a good arrangement for a text clustering, we often resort to lexical-based approaches and appropriate measures of similarity/distance (Burrows, 2002; Juola, 2008; Rudman, 1998, Stamatatos, 2009; Labbé & Labbé, 2001; Tuzzi, 2010) between texts, e.g. cosine similarity, Burrows's Delta, Labbé's intertextual distance, etc. Given the properties and the formula of a distance, we obtain a square matrix that includes n×n cells and n(n-1)/2 positive non-zero non-redundant values that can be exploited for an automatic classification of the n available texts. This distance matrix might be read from an alternative perspective, i.e. as a ranking system: for each text we can sort all the other n-1 texts from the closest to the furthest. The distribution of these ranks among texts represents an interesting object of research (Alvo & Yu, 2014) when we consider the whole corpus and also when we observe groups of texts that share some properties (e.g. they belong to the same author).

A preliminary experiment involved a large corpus of contemporary Italian novels and showed that we can identify some novels that systematically top positions in all rankings and prove to be close to most of the available texts; on the contrary, we have novels that do not show strong similarities in any list and systematically lie in the furthest positions. This study compared results achieved through different measures and formulated some hypothesis to understand when in text clustering it is worth either to distinguish "average" and "eccentric" novels or disregard them in in-depth investigations.

References:

Alvo, M. & Yu, P.L. (2014). Statistical Methods for Ranking Data. Springer.

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

Juola, P. (2008). Authorship Attribution. (Foundations and Trends in Information Retrieval, vol. 1(3)): 233–334.

Labbé, C. & Labbé, D. (2001). Inter-textual distance and authorship attribution. Corneille and Molière. Journal of Quantitative Linguistics, 8(3): 213–31.

Rudman, J. (1998). The state of authorship attribution studies: Some problems and solutions. Computers and the Humanities (31): 351–365.
Stamatatos, E. (2009). A survey of modern authorship attribution methods. Journal of the American Society for Information Science and Technology, (60): 538–556.
Tuzzi, A. (2010). What to put in the bag? Comparing and contrasting procedures for text clustering. Italian Journal of Applied Statistics/Statistica Applicata, 22(1): 77–94.

□ □ □ □

Gero Vogl and Katharina Prochazka

Decline and Persistence of a Language Minority

In the Hungarian part of the so-called Austrian-Hungarian dual monarchy censuses were conducted from 1881 on which included registration of the mother tongue. Speakers of German language were distributed all over Hungary, but we have confined the present investigation to the districts Tolna and Baranya, sometimes named "Schwäbische Türkei" (Swabian Turkey). The name derives from the fact that until the end of the seventeenth century this region was under Turkish administration and more or less depopulated when the Turks left. Afterwards very quickly the region was re-populated by Germans. In great parts of these two districts German was the dominant language until WW II. We have studied the language shift from German to Hungarian in this region, 1881–1930, digitizing the language data for each settlement from the ten-year censuses.

We have observed two characteristic phenomena of language shift: The tendency to giving up German and speaking Hungarian) (1) was higher in large communities than in small ones, (2) was lower in communities with German speaking majority. We quantified these phenomena and found that both effects were less distinct in communities with less than about 1000 inhabitants, but clearly recognizable in larger ones.

Fig. 1 (left). Increasing magyarisation with increasing size of community.

Fig. 2 (right). Increasing persistence of German language with increasing fraction of German spoken in 1881. For explanations see text.

(1) Fig. 1 shows the ratio of the fractions of German speakers in 1930, fG(30), and in 1881, fG(81), as a function of the total population of the respective community in 1881, n(81) for communities with more than 1200 inhabitants.

The relation can be described by an exponential starting at zero loss of German language in smaller communities and indicating a tendency of complete language loss for extremely large communities:

$$fG(30) / fG(81) = \exp(-m*n(81))$$

Law of decline m may be regarded as a "magyarisation factor" per inhabitant of the community. For the Swabian Turkey it is 0,00007 per inhabitant. As could be expected, there is considerable scattering in the data but we think that the general trend prevails being also visible, though of course less strongly, in the data of the 1900, 1910 and 1920 censuses.

(2) Fig.2 depicts the ratio of the fractions of German speakers in 1930, fG(30), and in 1881, fG(81), as a function the original fraction of German language in 1881, fG(81), again for communities with more than 1200 inhabitants. We hypothetically describe the relation with a logistic curve starting with fG(30) / fG(81) about 0,2 in communities with a small percentage of German speakers and reaching a value of about 0,86 in purely German speaking communities.

$$fG(30) / fG(81) = 1/(\exp(-(fG(81)-fT)/k)+1)$$

Persistence law

We call the factor k in the exponential of the logistic curve a persistence factor of the German language. Again there is a continuous development from 1881 through 1900, 1910 and 1920 to 1930. With progressing time fT shifts to higher fractions fG(81) implying that at long times even communities with originally large majority of German speakers will lose their language. fT may be regarded as a "turning point". Whether the laws found for decline and persistence of the German language minority in Hungary can be applied to other language minorities remains to be tested.

□ □ □ □

Rania Voskaki and George Mikros

A Modern Greek Readability Tool: Development of Evaluation Methods

The present research reports on an online Modern Greek Readability tool (MOGRead), developed by the Centre for the Greek Language. It constitutes part of an ongoing postdoctoral research project that aims to evaluate MOGRead in terms of effectiveness and efficiency, in order to propose novel methods that will enhance the tool's accuracy.

MOGRead was developed in order to meet the needs of teachers of Modern Greek as a second/foreign language (L2). Provided that the existing Modern Greek corpora are few and rarely updated, the need of an effective readability tool is imperative, especially for less used and less taught languages, such as Modern Greek.

Under this scope, the objective is to provide reliable results for any text as it concerns the level of adequacy (A1 to C2) according to Common European Framework for languages. In order to evaluate MOGRead's reliability, we have created two distinct corpora of plain texts: (a) a representative set of texts, according to their language level A1 to C2, available online on the Portal for teaching Modern Greek as L2 , and (b) a verification corpus that we have annotated and classified to the equivalent language level. In order to tag the verification corpus we opted for the TreeTagger tool (Schmid, 1994). As for the analysis of both corpora by quantitative methods, we used QUITA (Kubát et al., 2014).

The applied evaluation methods will permit the development of novel methods for the readability analysis of texts, taking into account the readability that differs among readers (DuBay, 2004) and especially among learners of a second/foreign language; in our study Modern Greek learners' readability. Consequently, the objective of this study is to underline the findings that we came up with after the implementation of evaluation methods on both corpora, to present the outcomes, and to set the basis for the procedure that has to be followed in terms of larger-scale reliable results.

References:

DuBay, W.H. (2004). The Principles of Readability. Costa Mesa, California: Impact Information.

Kubát, M., Matlach, V. & Čech, R. (2014). QUITA - Quantitative Index text Analyzer Lüdensheid: RAM.

Schmid, H. (1994). Probabilistic Part-of-Speech Tagging Using Decision Trees., In: Proceedings of International Conference on New Methods in Language Processing, Manchester, UK.

□ □ □ □

Relja Vulanović

Grammar Efficiency and the One-Meaning–One-Form Principle

A formula for measuring the degree of violation of Anttila's (1972) one-meaning–one-form principle (from now on, the 'Principle') has been proposed in (Vulanović & Ruff, to appear). Mathematically speaking, the formula measures how far a relation between two sets (the set of meanings and the set of forms) is from a bijection. The measure is constructed so that it is equal to 1 when the relation is a bijection, that is, when the Principle is satisfied. Otherwise, it is greater than 1 proportionally to the extent of violation of the Principle. These properties make the measure suitable to be included as a factor in a new formula for evaluating grammar efficiency. Whereas it is obvious that grammar efficiency should be inversely proportional to the degree of violation of the Principle, it needs to be discussed how to modify the previous grammar-efficiency formula from (Vulanović, 2003 & 2007). This is done in the present paper.

The most important factor in the old formula is the so-called parsing ratio. One of the roles of the parsing ratio is nothing else but to measure the extent of violation of the Principle, only, this is done indirectly, via parsing. Now that a direct measure from (Vulanović & Ruff, to appear) is available, it is used to replace this component of the parsing ratio. Then the remaining roles of the parsing ratio can be taken over by a formula which requires fewer and easier calculations. The

resulting new grammar-efficiency formula is therefore simpler than the previous one. This is illustrated by calculating grammar efficiency for various flexible part-of-speech systems that have been considered in (Hengeveld et al., 2004) and (Hengeveld & Van Lier, 2010a & 2010b).

References:

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

Hengeveld, K., Rijkhoff, J., and Siewierska, A. (2004). Parts-of-speech systems and word order. Journal of Linguistics, 40: 527–570.

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

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

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

Vulanović, R. (2007). On measuring language complexity as relative to the conveyed linguistic information. SKY Journal of Linguistics, 20: 399–427.

Vulanović, R. & Ruff, O. (to appear). Measuring the degree of violation of the one-meaning–one-form principle. QUALICO 2016.

□ □ □ □

Yaqin Wang

Analysis on English Text Genre Classification Based on Dependency Types

The present study aims to explore whether dependency type can be used as a distinctive text vector for classifying English genres. Dependency type across four genres, i.e., spoken, informative, imaginative, and written to spoken, were used as text vectors. Text mining methods, namely, principal component analysis (PCA), hierarchical clustering, and random forest were employed to investigate the clustering effect. PCA and agglomerative hierarchical clustering were first conducted and results show that dependency types can be used as an effective parameter in distinguish text genres, especially between

spoken genre and written genre. Random forest, which determines the importance of numerous variables, was then employed. The result shows that not all dependency types are important in classifying genres. Ten important types were also tested again to see whether they improve performance of the clustering result. The study may be useful for future applications in text genre categorization and dependency studies and offer new thoughts in the applications of NLP community.

□ □ □ □

Ming Xu and Yue Jiang

Stylometric Comparison between L1 Translations and L2 Translations of The True Story of Ah Q

Lun Xun's works were translated into English numerously from two translational directions: into (L1 translations) and out of the translator's mother tongue (L2 translations). The present article, based on the corpus composed of two L1 translations (by William A. Lell & Julia Lovell) and two L2 translations (by Wang Jizhen & Yang Xianyi and Gladys B. Tayler) of The True Story of Ah Q, one of Chinese literary classical novels written by Lu Xun, compares the stylometric features of the four translations from the perspectives of vocabulary richness, thematic concentration, activity and text similarity to find the differences and similarities in between. The combination of the four methods, in which the first three, independent of text length, are simple in operation and linguistic interpretation (Kubát, M. & Čech, R., 2016) and the last one can directly and clearly reflect the similarities of the translations as a whole, allows to investigate the difference and similarity of the four versions in a powerfully linguistically comprehensive view. The result indicates that:

1) Wang's and Lell's translations share a much higher similarity in vocabulary richness, thematic concentration, activity and text similarity;

2) The differences between two L1 translations are bigger than those between two L2 translations;

3) Lovell's translation, with the feature of richest vocabulary, lowest thematic concentration and activity, is strikingly different from the other three.

□ □ □ □

Yingying Xu

A Corpus-based Empirical Study on Inter-Textual Word Family Growth

Based on a corpus of 271,275 words, this paper examines the vocabulary size and inter-textual vocabulary growth patterns of the corpus in terms of word types, lemmas and word families, and tests the fitness of Brunet's, Tuldava's, Herdan's and Köhler-Martináková's models to the word family growth of the corpus. The results show that:

1) the vocabulary size of the corpus decreases greatly after lemmatization (about 6,300 words), and further reduces after turning lemmas into word families (about 3,600 words);

2) the inter-textual vocabulary growth patterns of the corpus can be better described by the word family growth curve; and

3) the Herdan's model proves to be good for the description of the inter-textual word family growth for the corpus.

□ □ □ □

Makoto Yamazaki

Distribution and Characteristics of Co-occurrence words across different texts in Japanese

In general, texts consist of a few function words which have very high frequencies of occurrence and many content words having relatively low frequencies. This is the effect of vocabulary balance which Zipf (1949: 22–27) pointed out. It is also the result of textual coherence as Halliday and Hasan (1976) claimed. As a consequence of these textual characteristics, we suppose that the distribution of co-occurrence words across different texts would follow the same rules. To confirm this hypothesis, we had experiments using large-scale corpus. We randomly chose same size text samples from the Balanced Corpus of Contemporary Written Japanese (BCCWJ) and counted the number of occurrences across the texts for each word. For example, when using ten text samples, there are ten kinds of co-occurrences, namely

1 to 10. We chose ten text samples for 100 times and calculated the average frequency of co-occurrences. For each ten text samples, we assigned different text length from 100 to 1,000 in 100 increments. And the number of text samples chosen ranged from 10 to 100 in 10 increments. Thus we have a total of 10,000 distribution data. The corpus data we used was the book samples in BCCWJ. BCCWJ has 20,668 book samples (Maekawa et al., 2014: 348). We didn't use the whole text because the length of text sample differs from sample to sample. So we extract the first N words from each text sample.

Experimental results show as follows. (1) Between the degree of co-occurrence and the number of co-occurrence words in that degree, a distribution like Zipf's law was recognized. (2) But the curve was not linear at the end. In ascending order of degree of co-occurrence, the curve turns to increase 2 or 3 degrees before the full co-occurrence (when using N text samples, full co-occurrence is N). (3)As the text length increases, the number of co-occurrence word also increases linearly. As for the full co-occurrence, the number of co-occurrence word steadily increase, and there seems to be no saturation point. (4) About the part of speech, the rate of function words increases and the rate of noun decreases as the degree of co-occurrence increases.

References:

Halliday, M.A.K. & Hasan, R. (1976). Cohesion in English. London: Longman.

Maekawa, K. et al.(2014). Balanced corpus of contemporary written Japanese. Language Resources and Evaluation, 48: 345–371 [DOI 10.1007/s10579-013-9261-0].

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

Acknowledgement:

This work was supported by JSPS KAKENHI Grant Number 15H03212 and NINJAL collaborative research project "A Multifaceted Study of Spoken Language Using a Large-scale Corpus of Everyday Japanese Conversation"(PI: Hanae Koiso).

□ □ □ □

Ling Yan

Distribution of Evaluative Language in English News Opinion

Sentiment analysis at present mainly extracts adjective expressions to identify sentiment orientation while fails to deal with other evaluative expressions, such as verbs, nouns, and adverbs. People instinctively feel that adjective expressions are closely related to attitude and evaluation. Other words, however, also include evaluative meaning, e.g. prevent, deny, cheer up, etc.

The study seeks to answer two questions: (1) What is the proportion of words with evaluative meaning in their own word categories? For example, how many nouns with evaluative meaning are there in all nouns of all subjective sentences? We call this proportion PE. (2) Are the differences of PEs significant? If there is a significant difference between two PEs, we should pay more attention to the higher proportion of word category. If not, we should give equal attention to all word categories when we do sentiment analysis.

The research design is as follows: (1) Build a news opinion corpus. Opinion section is a rich mine full of evaluative resources. (2) Identify subjective sentences. (3) Tag part-of-speech automatically and annotate the words with evaluative meaning manually. (4) Use SPSS to calculate the proportions and differences.

The findings are: (1) The order of PEs, from high to low, is adjective, verb, noun, adverb. (2) There is no significant difference between PE of an adjective and PE of a verb. However, there are significant differences among others.

Suggestions are: Other word categories, except adjectives, do not frequently carry obvious evaluative meaning, however, they bear more powerful evaluative meaning. A bias topic itself voices a clear attitude to all readers, even though the whole sentence seems very objective without any adjective expressions. It seems that the sentence level and the text level, besides the word level, should be taken into integrated consideration if we want to do a better sentiment analysis.

□ □ □ □

Biyan Yu and Yue Jiang

Probability Distribution of Syntactic Divergences of Determiner His-(Adjective)-Noun Structure in English-to-Chinese Translation

Studies of translation divergences reveal that lexical divergences in English-to-Chinese translation some regularities which can be modelled by a probability distribution. Besides lexical divergences, we hypothesize that there might be a great number of syntactic divergences since Chinese and English belong to two typologically divergent languages. In this study, we investigated whether the Chinese translation of English determiner his-(adjective)-noun structure (DNS (his) hereafter) diversifies in syntactic construction and whether the distribution of diversified Chinese translations conforms to the regularity of diversification process introduced by Altmann (2005). It is found that the twenty-two Chinese syntactic constructions corresponding to one single English DNS (his) do form a decreasing rank-frequency distribution, which is in comfortable agreement with the usual Zipf-Alekseev function. This diversification process may be ascribed to both linguistic and cultural factors such as contextual factors, translator's subjectivity, functional equivalents, and the tendency for minimal effort in language coding and decoding during translating process.

□ □ □ □

Posters

Joanna Byszuk

Analysis of Cross-Lingual Semantic Change in Professional Discourse with Quantitative Methods

Professional discourse is a popular object of studies, both with quantitative and qualitative means, most often focusing on features of discourse and on spoken text. The study underlying this paper aims at discovering and attempting to explain stylistic differences in professional discourse, especially in the lexical choices made by professionals of different nationalities writing in a foreign language.

The paper presents a possible application of methods used in stylometry and distributional semantics, such as supervised machine learning classification algorithms and word vector models, in the study of linguistic differences between the texts of similar topics but written by authors with various cultural background and mother tongues.

The study was conducted on a corpus of blog posts written by programmers from Polish-, English-, German- and Spanish-speaking countries, both in their native languages and in English, collected and cleaned by the author of the paper.

In the empirical part of the analysis used were two subsets of the corpus:

1) all texts written in English, marked for mother tongue of the author – used for comparing against one another and examining language variation,

2) texts in other languages – used for discovering linguistic features of discourse in selected language speaking communities and testing cross-lingual methods.

The aim of the paper was three-fold:

1) to discover mother-tongue dependent stylistic variation between texts of similar topic and use;

2) to reveal cross-lingual semantic shifts, e.g. differences in the use of terminology;

3) to evaluate possible limitations of used methods, e.g. constraints in creating cross-lingual word representations and in comparing both mono- and multi-lingual models.

The paper discusses preliminary observations concerning both the results and usefulness of above-mentioned methods, with hope to begin a discussion on other possibly useful methodologies for tracing language variation, namely lexical choices and syntax, and thus examining cross-lingual influences.

□ □ □ □

Lu Fan and Yue Jiang

Can Dependency Distance and Direction Be Used to Differentiate Translational Language from Original Language?

Translational language, due to its distinctive features, is called "the third code" (Frawley, 1984) or "the third language" (Duff, 1981). However, whether "the third language" truly exists or not has been an issue of controversy in linguistics studies. In this study, in the framework of dependency grammar (Liu, 2009), we used mean dependency distances (MDD) and dependency direction to analyse translated texts and original texts in an attempt to differentiate translational language and original language from a typological perspective. We built a dependency-annotated treebank consisting of English translations of Report on the Work of the Government delivered on annual China National People's Congresses by Chinese premiers and English originals of the State of the Union by American presidents, representing translated language and original language respectively. To rule out the possible influence of sentence length and text size on MDD and dependency direction, we controlled the sentence length within 5 to 40 words, and made sure each text has similar size. Hence, three treebanks were built in total. It was found that (1) MDD of the translated texts is significantly different from that of original texts, with MDD of the former much longer than that of the original texts; (2) texts of the two types show a similar pattern in dependency direction; 3) contrary to our hypothesis, the translated texts are more head-initial than head-final

compared with the original texts. Difference in MDD is probably influenced by the source language and difference in dependency direction is possibly due to the tendency of normalization in translating process, one of translation universals proposed by Mona Baker (Baker, 1993). These findings suggest that

1) dependency distance and direction can be used to differentiate translational language from original language;

2) typologically speaking, translational language, a third language which distinctive from original language does exist; and

3) quantitative linguistic methods can be applied to translation studies.

□ □ □ □

Łukasz Gągała

Hydra: Integrated Tagger-Lemmatiser with Deep Learning and Parallel Computing

Rapid growth of electronic resources in digital humanities we observe today requires also a constant endeavour of tools development for members of the research community. In the field of historical computer linguistics we invariably face two difficulties related to reliable (pre-)processing, more precisely tagging and lemmatisation of premodern texts. The first problem is the nature of premodern spelling, both in Latin and in the vernacular languages of Europe, exposing a high degree of variance (between regions and even between particular scribes, as far as scribal culture is considered) that inevitably impedes the automatic text processing on scale. The second concern is the question of easily deployable algorithmic approach that would be able to deal with that abundance of historical dialects and scribal traditions exceeding the contemporary linguistic variety of Europe.

To address that matter we propose an integrated tagger-lemmatiser based on recurrent neural networks with parallel computation. Deep Learning as a flourishing domain of research has proofed that deep neural networks can effectively explore complexity of language (e.g. by tagging and lemmatising). However a character-based approach that

would be very desired for non-standard spelling of premodern texts, both in Latin and in the vernaculars, requires a heavy model of neural networks. This in turn results in a very long computation time making the enterprise highly unpractical and unattractive. To improve a general neural-net approach for tagging and lemmatising we introduce a parallel computing technique (gossip protocol) to the training that aggregates multiple GPUs and shortens the total computation time. Additionally, we implement a novel recurrent-neural-net model (SRU) being much faster than state-of-the-art solutions (LSTM, RNN).

The gain of time (by parallelisation and the SRU-algorithm) is used to improve upon tagging and lemmatising multi-label tokens and multi-token labels (Figure 1 and 2) that characterise premodern vernacular languages without consistent norms of orthography and punctuation. Moreover, due to an extended context window our approach is meant to catch words dependencies in languages with a more free word order (like Latin and Old Church Slavonic/Old Russian). Finally, our deep-learning approach can be beneficiary for preprocessing of any language, as far as an annotated training corpus is provided (even for a pipeline process by building up a corpus, where new token forms may be mapped into already existing labels).

□ □ □ □

Hongjian Han and Yue Jiang

A Corpus-based Comparison of the General Features of Human Translation and Online Machine Translation

Based on a parallel-corpus of *Pride and Prejudice* (Austin, 1970) consisting of two human Chinese translations, one online-machine Chinese translation and the original English text, this study compared the 16 quantitative and stylistic features of the human translations with that of the online translation. Results of the study indicate that the online translation is far behind human translations at lexicon and syntactic levels and in sentiment expression in terms of quality. It is abundant in features of "explicitation" (Baker, 1993) such as using pronouns, auxiliaries and conjunctions – but it lacks "simplicity" in stand-

ard type/token ratio and lexical density, and fails to show the tendency of syntactic "explicitation" as previous descriptive translation studies has described. These may be attributed to its weakness in identification and transformation of parts of speech and identification of the contextual meanings, semantic sentiments and tense of the original sentences. The two human translations show different lexical and syntactic features and differed in degree in "explicitation", "simplification" and syntactic "explicitation", which are probably due to the different translation purposes and subjectivities of the translators.

□ □ □ □

Xinlei Jiang and Yue Jiang

Effect of Dependency Distance on Disfluencies in Interpreting

Quantifying assessment of syntactic difficulty helps better predict the challenges presented by a given source text. Dependency distance is widely acknowledged as a metric of syntactic complexity (Liu, 2008), bridging textual factor and cognitive load in language processing, while disfluencies, as shown in psycholinguistic literature, are the predictors of cognitive effort due to language perception or production (Shreve, Lacruz, & Angelone, 2011).

Given abundant corpus-based text analysis, there has been hardly any report on empirical experiments that explore the effect of dependency distance on interpreting process. The present study is intended to investigate the relation between maximum dependency distance of source language and disfluencies in interpreting. Compared with mean dependency distance, which is usually used as a stable measurement in treebank analysis, maximum dependency distance is more feasible in operationalizing syntactic difficulty of each sentence in interpreting experiments. Disfluencies, as indicator of interpreting quality and manifestation of an interpreter's cognitive load, include silent pause, filled pause, repetition and self-monitoring, with location, duration and frequency recorded.

20 pairs of sentences were used in this study, with sentence length ranging from 10 to 20 words and maximum dependency distance

ranging from 5 to 16. Within each pair, textual factors such as lexical difficulty, sentence length, syntactic structure and information load are control variables while maximum dependency distance differs markedly. Two trials are formed by extracting sentences from every pair alternately, with 20 items in each trial. 30 subjects, from universities in Xi'an, are required to do sight interpreting of 50 sentences, including 30 fillers and a trial, presented in an individually-randomized order. In the recruitment of participants, gender is balanced, and working memory capacity as well as language proficiency are controlled at a similar level. Besides, post-test questionnaires are performed to collect qualitative data to answer the call for triangulation. Recordings are transcribed manually and then coded according to the above disfluencies scheme aided by audio tools. Processed by statistical software, ANOVAs are performed to analyse the quantitative data in order to uncover the relation between maximum dependency distance and disfluencies in interpreting.

Results show that (1) in general pattern, (when going beyond a fixed bound,) the longer the maximum dependency distance is, the more frequently disfluencies in interpreting occur or the longer disfluencies last; (2) based on between-pair analysis, the direct relations between the maximum dependency distance and disfluencies in interpreting manifest themselves distinctively to some extent, probably because of differences in the number of chunking, syntactic structure, and sentence length among the 20 sentence pairs.

□ □ □ □

Susanne Kabatnik

(eine) Frage stellen – zada(wa)ć pytanie: Functions of German and Polish Support Verb Constructions in Texts – a Corpus-based Case Study

As the largest multilingual online source of information worldwide, Wikipedia is known for its comprehensibility (Van Dijk, 2010: 86ff.). This comprehensibility is realized through different ways of explicit and implicit linguistic devices to which various formats of infor-

mation linking and resumption belong (Schwarz-Friesel & Consten, 2014). Empirical studies on Support Verb Constructions (SVC; Storrer, 2007 & 2013) indicate its function in creating coherence, as in the following example of attribution and pronominal resumption:

(1) Alle Fragen, die eben gestellt wurden, und die alle berechtigt sind, werden in Metzlers Text beantwortet.

(1′) All questions which just have been asked and which are all legitimate are answered in Meddlers text.

In this paper, I select and analyse support verb constructions in the context of discourse coherence. Following the research paradigm of grammar and text linguistics, I compare and contrast these selected constructions with their Polish equivalents. I employ corpus linguistics (Perkin, Keitel & Kupietz, 2012; Taborek, 2017) in combination with traditional syntactic tests (Pittner & Berman, 2012; Wöllstein, 2014). The data of this research project comes from Wikipedia-corpora, which the Institute of German Language (German: Institut für Deutsche Sprache) provides in Polish and German. Firstly, I describe the possibilities of information enrichment and its re-establishment. I then focus on the placement of SVCs and the consequences for the structuring of information. I conclude with a discussion on the findings and the meaning for the German and Polish research traditions.

References:

Storrer, A. (2007). Corpus-based investigations on German support verb constructions. In: Fellbaum, C. (ed.). Idioms and Collocations. Corpus-based Linguistic and Lexicographic Studies. London: Continuum.

Storrer, A. (2013). Variation im deutschen Wortschatz am Beispiel der Streckverbgefüge. In: Deutsche Akademie für Sprache und Dichtung: Reichtum und Armut der deutschen Sprache. Berlin u.a.: De Gruyter: 171–209.

Schwarz-Friesel, M. & Consten, M. (2014). Einführung in die Textlinguistik. Darmstadt: WBG (Wiss. Buchges.).

Taborek, J. (2017). Funktionsverbgefüge in bilingualen deutsch-polnischen Wörterbüchern. Korpusbasierte Analyse – syntagmatische Muster – Äquivalenz. In: Jesenšek, V. & Enčeva, M. (eds). Wörterbuchstrukturen zwischen Theorie und Praxis. Herbert Ernst Wiegand zum 80. Geburtstag gewidmet. (= Lexikographica.Series Maior). Berlin: de Gruyter [submitted].

Van Dijk, Z. (2010). Wikipedia. Wie Sie zur freien Enzyklopädie beitragen. München: Open Source Press.

Wöllstein, A. (2014). Topologisches Satzmodell. Zweite aktualisierte Auflage. Heidelberg: Winter.

□ □ □ □

Ivan Kondyurin

The Problem of Automatic Ontology Supplementation by Means of Semantic Analysis of Definitions (in Case of Nouns Denoting Locations in Russian)

This paper investigates the possibility to add new elements into the linguistic ontology automatically using the results of the semantic analysis of definitions from different sources. It also concerns the problems that emerge in the process of automatisation and are caused by inconsistency of definitions and abundance of ellipses and suggests a possible algorithm for automatic supplementation.

When it comes to natural language processing, knowledge about the real world can be of considerable importance. One of the effective ways to take this knowledge into account is to use a linguistic ontology. In terms of information science, ontology can be defined as a formal, simplified representation of a certain domain, which contains concepts and relationships between them.

In the case of Russian, ontological semantics is already exploited in several projects, including ABBYY Compreno and AIIRE, where the ontology is integrated into the language processor. Herein the language processing and semantic analysis is considered regarding the AIIRE system, since its ontology is public and available online.

Complete and accurate analysis of any sentence in this system is impossible without each word being properly processed at morphological, syntactic and semantic levels. In order to achieve this and thus to enable all the semantic constraints function properly, one should constantly add new links and even concepts to the ontology. Moreover, since the language is volatile through its very nature, new words appear regularly, which is particularly important in respect of proper nouns and nomens. On account of vast volume and variability of lexicon along with sophisticated structure of concepts the new elements can scarcely be added manually or imported from existing structured databases, except other ontologies.

One of the most challenging and important problems regarding ontological semantics is therefore the issue of automatic supplementation. The goal of the present research is to discover whether semantic

analysis of pre-extracted definitions is capable of producing correct concepts and relations which can be loaded into the ontology, and if so, to define the way it should be performed and the preferable information source. The effectiveness of supplementation using each of the two sources, namely definitions from Wikipedia articles and entries of Wiktionary, is also compared. It is worth noting that data in this case is considered pre-extracted, for Wikipedia mining is still the issue of further investigation.

A parallel corpus of definitions from both sources is created for the semantic group of nouns denoting locations (including toponyms and nomens), and an attempt to elaborate the algorithm for automatic extraction of elements for ontology supplementation is carried out. All the emerging problems caused by inconsistency of definitions and abundance of ellipses are described and investigated.

The conclusion is made about the possibility of extrapolating this algorithm to larger semantic groups and its reasonability for NLP.

□ □ □ □

Tatiana Litvinova and Olga Litvinova

A Study of Texts of an Extremist Forum "Kavkazchat" Using Linguistic Inquiry and Word Count (LIWC)

Internet has turned into a powerful tool of manipulating the minds and behaviour of young people. This gives rise to autonomy of countless youth extremist movements. Internet resources are also massive tools of promotion and propaganda for terrorist activities. It is thus of growing importance to analyse extremist and terrorist online language using modern linguistic tools to understand psychology of threats. The material for the current study are texts of a Russian extremist forum 'Kavkazchat' accessed via The Dark Web Portal (Qin et al. 2005). The texts were analysed by means of the Linguistic Inquiry and Word Count (LIWC) software (Pennebaker, Francis & Booth, 2001) with the help of the main dictionary of the software (Kailer & Chung, 2011) as well as those designed by the authors. This software allows large bodies of texts to be analysed along a whole range of linguistic and psychological categories.

LIWC is commonly used to analyse extremist and terrorist texts, mostly in English, for example language of the ISIS, Al-Qaeda, etc. (Chung & Pennebaker 2011; Pennebaker & Chung, 2005; Pennebaker & Chung, 2008). The software is easy to use to analyse such texts as hidden motives and interests of terrorist groups could be examined in their dynamics as well. Unlike traditional content-analysis software, it is also a unique tool of identification of latent psychological characteristics of the authors. For the first time Russian texts of an extremist forum have been analysed (according to the topics) using LIWC along a wide range of linguistic and psychological categories. The obtained results were compared with the data for extremist texts in English.

□ □ □ □

Gregory Martynenko, Alexey Melnik and Tatiana Sherstinova

Russian Short Stories of the Twentieth Century: The Computer Anthology and Quantitative Analysis

The Computer Anthology of Russian short stories of the twentieth century is currently being created at St. Petersburg State University. Theoretical foundations for its creation are based on the systemic ideas by Yury N. Tynyanov, who was a famous writer, literary critic, scholar and a known member of the Russian Formalist school. We supplied the Tynianov's approach by statistical method and raised requirements for ensuring the homogeneity of populations under study. This means that genre homogeneity of research data should be fully maintained. To achieve this homogeneity, we decided to limit the analysis to a genre of short story. Our decision is justified by the fact that short story, belonging to "small" prosaic forms, is a very common genre, allowing to involve into research a large number of texts and writers. Besides, that is short story that perform an exploratory function in the process of literary development, it is a genre of rapid response, vividly reacting to the demands and challenges of time and sometimes even foreseeing future times. The report will present main approaches to quantitative analysis of these literary texts, as well as to taxonomy and clustering of writers according to their literary styles.

□ □ □ □

Evgenia Mescheryakova

Referential Hierarchies: From Ordinal Scales to Dissimilarity Matrices… and Back Again

Referential hierarchy, first introduced by Silverstein (1976) and Moravcsik (1978), has been widely used in linguistic typology, as it allows to analyse various cross-linguistic regularities, e.g. differential argument marking strategies (Haude & Witzlack-Makarevich, 2016; Witzlack-Makarevich & Seržant, 2017) and verb-argument agreement (Woolford, 1999), etc. However, its universality is arguable (Bickel et al. 2015), and its internal structure, in theory, varies from one language to another. The hierarchy is traditionally represented as an ordinal scale (Dixon, 1979) or a list of logically independent scales, with elements denoting the properties of NPs and their referents such as person, animacy, definiteness, and number (Croft, 2003). An alternative view on linguistic scales was proposed in (Cysouw, 2015): any 'scale of linguistic structure' can be generalized to a dissimilarity matrix.

We adopt this approach, as it was done by Bickel et al. (2015), because it allows us to derive the 'hierarchy' empirically. However, we propose another dissimilarity measure, so that the 'hierarchy' for every language is a matrix, and not just for a language sample.

In this presentation, we show how these 'hierarchies' can be compared quantitatively, how to compare them to the scales offered by previous theoretical works, and how to analyse them together with geo- and phylogenetic data about the languages.

□ □ □ □

Piotr Mirocha

Corpora and Collocations in Discourse Analysis: The Representations of Europe in Croatian and Serbian Newspapers

Linguistic pictures of the world shared by various interest groups were in the focus of critical discourse studies since the foundation of this discipline (Van Dijk, 1998). However, its use of discourse samples

gathered and selected manually by an analyst raised criticism on the account of being biased. Application of language corpora and quantitative tools – such as collocation analysis – may pose a solution to this problem (Baker, 2008).

This presentation is a report from the study of corpora consisting of articles from three Croatian (Novi list, Jutarnji list, Večernji list) and three Serbian (Danas, Politika, Večernje novosti) daily newspapers, encompassing texts issued between October 2012 and October 2017. The text corpora were chosen to represent various ideological profiles present in the mainstream quality press, ranging from centre-left to centre-right. The data was downloaded from the online issues of the respective newspapers, then, followingly, tagged with part-of-speech labels and lemmatised.

This enabled an analysis of collocations of the lexeme Europa – both contingent, appearing only in single newspapers or single yearly subcorpora, as well as consistent ones, reemerging in a bigger number of subcorpora. While contingent collocates often reflect passing 'hot topics' in the discourse and their study can explain mechanisms of topical dynamics in discourse, the consistent collocates are more telling about cultural stereotypes fixed in the studied corpora.

The goal of the research was to reconstruct a linguistic representation of Europe respectively in Croatian and Serbian media discourse, as well as to discover their akin and divergent features. A part of this task was to recognise differences between collocates in national newspapers of discrepant political lines. In consequence, similarities and distinctive traits in the resulting collocation sets, as well as differing semantic connotations of the collocates enabled to reconstruct the linguistic representation of Europe in Croatian and Serbian media.

In sum, the presentation demonstrates the potential offered by quantitative and corpus approaches to discourse analysis and the study of linguistic pictures of the world.

References:

Baker, P., Gabrielatos, C., Khorsavinik, M., Krzyżanowski, M., McEnery, T. & Wodak, R. (2008). A useful methodological synergy? Combining critical discourse analysis and corpus linguistics to examine discourses of refugees and asylum seekers in the UK press. Discourse & Society, 19(3): 273–306.

Van Dijk, T. (1998). The Study of Discourse. In: Van Dijk, T. (ed.). Discourse as Structure and Process: Discourse Studies: A Multidisciplinary Introduction. London: Sage.

□ □ □ □

Branislava Šandrih

SMS Sentiment Classification Based on Emoticons, Informal Abbreviations and other Text Features

Short message exchange is one of the most popular communication styles of the present times. In this paper we investigate the influence of emoticons, informal words' short forms and text features on a sentiment of a short message.

We collected a small corpus of ~5,500 SMS messages from one person's cell phone in a four-year time interval. Most of the senders are in their middle twenties and use the "popular" texting language, i.e. short form of words, abbreviations and emoticons. Therefore, this research relies on the informality of the corpus and tries to discover the influence of modern language patterns on messages' sentiment.

We expect that emoticons have the highest influence on the impression about the mood of a message sender. Therefore, we listed 102 different emoticons in a form of regular expressions in Python and extract them as features. Usage of short word forms, slang words and other kind of abbreviations is very common in texting. We listed 135 different abbreviations that are widely used in textual communication among Serbian speaking persons, although many of them are actually in English. Absolute count of each abbreviation occurrence is then used as a single feature per short message instance. For this dataset, 20 text features (also referred as "linguistic") were added to the final feature set, e.g. message length, number of exclamation marks, number of question marks, number of dots, etc. These features were selected after careful human analysis of this certain dataset, as it seemed they could help with differentiating formal and informal tone in messages. Afterwards, 16 more features were added as a ratio of already mentioned feature counts.

We build a linear Support Vector Machine (SVM) classifier that is able to determine positive, negative and neutral sentiments for a short message with average accuracy score of 96% in a 5-fold cross validation setting and average F1-score of 96.8%, 76.6% and 96.6% in favour of positive, negative and neutral class, respectively. This technique best performs on informal text written in Serbian language, since it uses as a subset of features a set of predefined, informal slang abbreviations.

□ □ □ □

Krzysztof Topolski, Adam Pawłowski and Piotr Malak

Statistical Distributions of Vocabulary in Authorial Corpora as a Stylometric Tool

Methods of stylometry are dominated by the use of multidimensional techniques, based on the occurrence of the most frequent units (mainly lexemes). Such units are regarded as independent of the topic of the analysed works and of their authors' intentions. It is emphasized that the units are not consciously used in the text by its author but, at the same time, indirectly express some qualitative features of the authorial style (e.g. the degree of complexity of the text or the type of narrative). For example, some conjunctions are characteristic of subordinate sentences (complex and difficult texts), others of equivalent sentences, whereas in simple sentences prepositions are rather prevalent (simple texts or dialogues). The choice of pronouns is a stylistic option related to the proportion of dialogues and descriptions and to their character (who speaks to whom). Equally effective, but slightly different in assumptions, seems to be an approach based on the quantitative analysis of word n-grams. In both cases an important argument in favour of the most frequent tokens is that they are uninflected and thus independent of lemmatisation. The disadvantage of this approach, however, is the consistent omission of the polysemy phenomenon. We argue that the study of statistical distribution of text elements can also become an interesting option. This kind of distribution takes into account the whole spectrum of vocabulary, not only the most common words, and, at the same time, it is characterized by a high degree of syntheticism. As our previous research has shown, among the most appropriate models for Polish texts there are Zipf-Mandelbrot finite distribution and the generalized inverse Gauss-Poisson distribution.

The subject of this paper will be a quantitative analysis of the texts by 11 Polish authors from the late 19th and early 20th centuries (6 men and 5 women). Original prose texts will be examined. Translations and poetry are not taken into account. Two types of distributions will be fitted to data, and then the texts will be classified on the basis of the parameter values of the obtained models of distribution.

□ □ □ □

Mayumi Usami and Makoto Yamazaki

Quantitative Characteristics of the BTSJ Japanese Natural Conversation Corpus (BTSJ-Corpus) ver. 2018: Focusing on the Differences of the Use of Polite Forms According to Sub-Groups

The present study first introduces the purpose and significance of developing the BTSJ Japanese Natural Conversation Corpus (hereafter BTSJ-Corpus) ver. 2018, which has compiled 333 Japanese Natural Conversations with sounds and transcriptions. Then, the overall quantitative characteristics of the corpus such as the number of total words and TTR are described. The BTSJ-Corpus is one of the largest corpus (928,070 words) which compiles Japanese spontaneous oral conversations in various settings such as those between unacquainted people and between intimate friends of different social sub-groups. BTSJ is the abbreviation of the transcribing rules named 'Basic Transcription System for Japanese' which had developed by considering the characteristics of Japanese language and Japanese interaction such as the phenomena that each sentence-final indicates the different politeness levels and there are many backchannels and overlapped utterances.

After describing the overall quantitative characteristics of the BTSJ-Corpus, we analysed it from the viewpoint of the use of polite-forms depending on the settings and the relationships between the speaker and hearer. It is because in Japanese conversation the choice of speech-level, basically either polite-forms or non-polite forms, has crucial functions for smooth communication and it represents the relationships between the speakers conspicuously. The following four settings were analysed:

Conversations between unacquainted people:

1) between native speakers.

2) between native speaker and non-native speaker.

Conversations between intimate friends:

3) between native speakers. And

4) between native speaker and non-native speaker.

The major results of the use of polite forms are as follows:

1) In both results of native and non-native speakers, the percentage

of the polite-forms are higher in conversations between strangers than those of between intimate friends.

2) In conversations between intimate friends, the percentage of polite-forms used by non-native speaker is higher than that of native speakers.

3) Native speakers use few polite-forms in conversations between intimate friends while non-native speakers use about the same percentage of polite forms in conversations between friends as in conversations between strangers.

These results show that the major difference in the use of polite forms between native and non-native speakers is that the fact that non-native speakers do not switch the use of polite-forms depending on the interlocutor. And there are possibilities that this fact may cause some uncomfortableness especially in conversations between friends.

In the presentation, we explain the coding system applied and discuss what these results tell us about characteristics of Japanese conversation by showing other results as well. We also argue the functions of the use of polite-forms in Japanese more in detail.

Acknowledgement:

This study was conducted as a part of the sub-project 'The study on the Language Use of Japanese Language Learners' (project leader: USAMI Mayumi) in 'Multiple Approaches to Analysing the Communication of Japanese Language Learners' (Project Leader: ISHIGURO Kei), NINJAL.

□ □ □ □

Letao Wang and Yue Jiang

The Dynamic and Complex Adaptive Process of Lexical Rank Frequency Distribution of Translations in the Translator's View Triangle

The study investigates the differences in h-points of lexical rank-frequency distribution and the changes in translator's view triangles between four translators based upon 32 fragments of their literary Chinese-to-English translations. It is found that (1) the changes in the translator's view and the synsematic view angles based upon

h-point are largely related to the influence of the source texts and that the lexical difference between the translations by the translators is mostly found in the autosemantic view angles; (2) L1 translators show a significantly higher notional word richness than L2 translators; (3) a synergic relationship is found between the three angles in the translator's view triangle, thus reflecting retrospectively the dynamic and complex adaptiveness of the translating process. The study is indicative of the applicability and validity of the relevant indicators of lexical rank-frequency distribution to contrastive translation studies and suggests that translator's view triangle can be used to observe the dynamic distribution of lexical rank-frequency among different translations, and esp. the conscious and unconscious control by the translator of his lexical use in translating process.

□ □ □ □

Peng Zhang and Hong Zhu

Production of Inflections in L2 Japanese – a Picture Naming Study

Inflection is regarded as a productive morphological transformation by adding a grammatical affix to a stem. Many previous experimental studies have found that the processing of inflections is not only related to language cognition but also concerned with the distinction between procedural and declarative memory in human minds (Ullman et al., 2001 & 2002). Although results from the Lexical Judgement Task and Priming Paradigm have led to a number of consistent and replicable models, they could only describe the recognition and storage of inflections. However, they cannot appropriately depict a more complex production procedure, which is thought to involve a sequence of distinct processes, including visual perception, lemma retrieval, phonological encoding, motor programming and articulation (Clahsen et al., 2010; Gor & Cook, 2010; Indefrey & Levelt, 2004; Tabak & Baayen, 2010).

The present study aims to investigate the explicit mechanism and two possible influencing factors (the degree of regularity and the whole-word frequency) of L2 learner's production of Japanese inflectional morphology through two naming experiments. 30 advanced

L2 Japanese Learners (L1: Chinese) and 30 Japanese native speakers were tested in each experiment. Experiment 1 used a cross-tense naming task to investigate the naming latency in low processing cost condition. The task required the participants to name the past-tense form when presented with the infinitive form, and vice versa. Experiment 2 used a self-paced picture naming task to investigate the naming latency in high processing cost condition. In the task, participants were first familiarized with the targeted picture names by being taken through the picture book that had the infinitive form printed below each photograph. Afterwards, they carried out an on-line self-paced naming task, in which subjects were requested to use the past /infinitive tense form to name each picture.

The results of Experiment 1 indicated that the production strategy of Japanese L2 learner's inflections was affected by the complexity of declension rules and whole-word frequency. Japanese L2 learners produced inflections with simple declension rules faster than those with complex rules. The easier the declension rules are, the more the decomposition strategy will be used. It can be attributed to the highly regularized transformation system in Japanese inflections. The facilitating frequency effect was also observed in producing all types of inflections, that is, the naming latency of high frequency inflections was shorter than low frequency ones. When the processing cost increased (Experiment 2), the regularity and frequency effects were both inhibited. L2 learners preferred to use whole-word retrieval strategy in producing inflections. Rule-based decomposition strategy was only observed in inflections with relatively simple declension rules. Taken together, the results of the two experiments suggest that it is relative to the morphological classification of inflections based on the phonetic features of segments in classroom instructions.

References:

Clahsen, H., Felser, C., Sato, M. & Silva, R. (2010). Morphological structure in native and nonnative language processing. Language Learning, 60: 21–43.

Pinker, S. & Ullman, M. (2002). The past and future of the past tense. Trends in Cognitive Sciences, 6: 456–463.

□ □ □ □

Haiyan Zhou and Yue Jiang

Frequency Distribution of Personal Pronoun Subjects in English Translations of Tao Te Ching

Previous studies have indicated that English is more explicit than Chinese in the use of personal pronouns, but most of those studies were focused on contemporary novels, economic texts, and political texts in a qualitative way, whereas few concentrated on ancient Chinese classics and quantified explicitation from the perspective of narratology. Chinese has its own distinctive features in terms of its formation and grammatical development within thousands of years (Ying Liu, 2016). Hence, based upon a self-built bilingual parallel corpus, this study is intended to compare the Chinese original of Tao Te Ching (Wang Bi version) with eight English translations in terms of their frequency distribution of personal pronoun subjects in them. The translations include those by Raymond B. Blakney (1955), Witter Bynner (1944), Herman Ould (1946), Arthur Waley (1934), Lin Yutang (1955), Wu Jingxiong (1961), Chen Rongjie (1963) and Liu Dianjue (1963). Furthermore, we also investigated the diversification of narrative perspectives in the aligned paragraphs and sentences on a one-on-one basis. In addition, the causes for the emergence of divergence were explored as well.

Our study found that

1) personal pronoun subjects in the translated texts are much more explicitly used than in the source text, esp. the third person pronoun;

2) Chinese translators (L2 translators) prefer to add the second person pronoun, which is rarely used in novels, probably to establish or build up a rapport with their readers and in the meantime adopt an instructive attitude in the interaction intended, whereas native English-speaking translators (L1 translators) tend to conform to the original;

3) our syntactic comparison reveals that due to the interference of the ancient Chinese language, personal pronoun subjects in the source text are rendered into varied sentence patterns, such as relative pronouns, finite verb phrases, formal subject, nominal clauses, etc.;

4) despite that for the first person "吾(wú)" and "我(wǒ)" there is no statistical difference between translators in both directions of trans-

lation, some L1 translators misinterpreted "吾(wú)" as a plural pronoun, causing a wrong shift of this pronoun.

There are two factors that may help explain the cause of the above differences between the translations. One is direction of translation and the other is related to the influence of the source text, esp. the characteristics of ancient Chinese classics with more oblivious subjects. The results of this study have further evidenced that

1) the rendition of ancient Chinese classical texts shows more explicitation of personal pronouns, a translation universal, than that of modern or contemporary Chinese novels or texts.

2) lexical and syntactic diversification occurs in the translating process of ancient Chinese. For example, a sentence, even a single word in a source language, may be translated into different forms in a target language (Yu Fang & Haitao Liu, 2015; Biyan Yu & Yue Jiang, 2017).

□ □ □ □

The List of Contributors

Author	E-mail	Affiliation
Andreev Sergey	smol.an@mail.ru	Smolensk State University, Russia
Andres Jan	jan.andres@upol.cz	Palacký University Olomouc, Czechia
Baumartz Daniel	baumartz@stud.uni-frankfurt.de	Goethe University Frankfurt
Baixeries Jaume	jbaixer@cs.upc.edu	Universitat Politècnica de Catalunya, Spain
Byszuk Joanna	joanna.byszuk@ijp.pan.pl	Institute of the Polish Language, Polish Academy of Sciences, Poland
Casas Bernardino	bcasas@cs.upc.edu	Universitat Politècnica de Catalunya, Spain
Català Neus	ncatala@cs.upc.edu	Universitat Politècnica de Catalunya, Spain
Čech Radek	cechradek@gmail.com	University of Ostrava, Czechia
Chen Xinying	chenxinying@mail.xjtu.edu.cn	Xi'an Jiaotong University, China
Chen Yuan-Lu	cheny@email.arizona.edu	University of Arizona, Linguistic Department, USA
Číž David	davidciz95@gmail.com	University of Ostrava, Czechia
Dębowski Łukasz	ldebowsk@ipipan.waw.pl	Polish Academy of Sciences, Poland
Eder Maciej	maciejeder@gmail.com	Institute of Polish Language, Polish Academy of Sciences, Poland
Esteban Juan Luis	esteban@cs.upc.edu	Universitat Politècnica de Catalunya, Spain
Faltýnek Dan	dan.faltynek@upol.cz	Palacký University Olomouc, Czechia
Fan Lu	fl_91910@126.com	Xi'an Jiaotong University, China
Fang Yu	fydiana@163.com	Zhejiang University, China
Fenk-Oczlon Gertraud	gertraud.fenk@aau.at	Alpen-Adria-Universität Klagenfurt, Austria
Ferrer-i-Cancho Ramon	rferrericancho@cs.upc.edu	Universitat Politècnica de Catalunya, Spain
Ferro Nicola	ferrodei.unipd.it	University of Padua, Italy
Franzini Emily		Decoded Ltd., UK

Franzini Greta		University of Göttingen, Germany
Gągała Łukasz	lukaszgagala@wp.pl	University of Göttingen, Germany
Gómez-Rodríguez Carlos	cgomezr@udc.es	Universidade da Coruña, Spain
Górski Rafał L.	njgorski@cyf-kr.edu.pl	Institute of Polish Language, Polish Academy of Sciences, Poland
Han Hongjian	hanhongjian99@163.com	Xi'an Jiaotong University, China
Hemati Wahed	hemati@em.uni-frankfurt.de	Goethe Universität, Germany
Herden Elżbieta	elzbieta.herden@uwr.edu.pl	University of Wrocław, Poland
Hernández-Fernández Antoni	antoni.hernandez@upc.edu	Universitat Politècnica de Catalunya, Spain
Hůla Jan	jan.hula21@gmail.com	University of Ostrava, Czechia
Jander Melina		University of Göttingen, Germany
Jiang Jingyang	jy-jiang@zju.edu.cn	Zhejiang University, China
Jiang Xinlei	396348813@qq.com	Xi'an Jiaotong University, China
Jiang Yue	yuejiang58@163.com	Xi'an Jiaotong University, China
Johnsen Lars	yoonsen@gmail.com	National Library of Norway, Norway
Kabatnik Susanne	skabatni@mail.uni-mannheim.de	University of Mannheim, Germany
Kaplar Sebastijan	kaplar@uns.ac.rs	University of Novi Sad, Serbia
Kelih Emmerich	emmerich.kelih@univie.ac.at	University of Vienna, Austria
Kestemont Mike		University of Antwerp, Belgium
Kimura Miki	mk_ling@meiji.ac.jp	Meiji University, Japan
Klyshinsky Eduard	eklyshinsky@hse.ru	National Research University, Higher School of Economics, Moscow, Russia
Kondyurin Ivan	ivan.kondyurin@gmail.com	Saint Petersburg State University, Department of Computational Linguistics, Russia
Kosek Pavel	kosek@phil.muni.cz	Masaryk University, Czechia

Krivochen Diego	d.krivochen@reading.ac.uk	University of Reading, UK
Kubát Miroslav	miroslav.kubat@gmail.com	University of Ostrava, Czechia
Lacasa Lucas	l.lacasa@qmul.ac.uk	Queen Mary University of London, UK
Langer Jiří	jiri.langer@upol.cz	Palacký University Olomouc, Czechia
Lazić Biljana	biljana.lazic@rgf.bg.ac.rs	University of Belgrade, Serbia
Litvinova Olga	olga_litvinova_teacher@mail.ru	Voronezh State Pedagogical University, Corpus Sociolinguistics and Authorship Profiling Researches Laboratory, Russia
Litvinova Tatiana	centr_rus_yaz@mail.ru	Voronezh State Pedagogical University, Russia
Liu Hsuan-Ying	hsuanying.liu@UND.edu	University of North Dakota, USA
Liu Yanchun	liuyanchunone@163.com	Communication University of China, China
Luque Bartolo	bartolome.luque@upm.es	Universidad Politécnica de Madrid, Spain
Luque Jordi	jls@tid.es	Telefónica Research, Spain
Mačutek Ján	jmacutek@yahoo.com	Comenius University in Bratislava, Slovakia
Malak Piotr	piotr.malak@uwr.edu.pl	University of Wrocław, Poland
Martynenko Gregory	g.martynenko@gmail.com	St. Petersburg State University, Russia
Matlach Vladimir	vladimir.matlach@upol.cz	Palacký University Olomouc, Czechia
Mehler Alexander	Mehler@em.uni-frankfurt.de	Goethe-University Frankfurt am Main, Text Technology Group, Germany
Melnik Alexey	st064458@student.spbu.ru	St. Petersburg State University, Russia
Mescheryakova Evgenia	e-meshch@ya.ru	National Research University, Higher School of Economics, Moscow, Russia
Mikros George	gmikros@isll.uoa.gr	National and Kapodistrian University of Athens, Greece

Milička Jiří	milicka@centrum.cz	Department of Comparative Linguistics, Faculty of Arts, Charles University, Prague, Czechia
Mirocha Piotr	pamirocha@gmail.com	Jagiellonian University, Poland
Místecký Michal	mmistecky@seznam.cz	University of Ostrava, Czechia
Moisl Hermann	hermann.moisl@ncl.ac.uk	Newcastle University, UK
Navrátilová Olga	olganav@mail.muni.cz	Masaryk University, Czechia
Obradović Ivan	ivan.obradovic@rgf.bg.ac.rs	University of Belgrade, Serbia
Ochab Jeremi	jeremi.ochab@uj.edu.pl	Institute of Physics, Jagiellonian University, Poland
Ouyang Jinghui	oyjh@zju.edu.cn	Zhejiang University, China
Pawłowski Adam	adam.pawlowski@uni.wroc.pl	University of Wrocław, Poland
Pelegrinová Kateřina	a15048@student.osu.cz	University of Ostrava, Czechia
Prochazka Katharina	katharina.prochazka@univie.ac.at	University of Vienna, Austria
Radojičić Marija	marija.radojicic@uns.ac.rs	University of Belgrade and University of Novi Sad, Serbia
Rotari Gabriela		University of Göttingen, Germany
Rovenchak Andrij	andrij.rovenchak@gmail.com	Ivan Franko National University of Lviv, Ukraine
Růžičková Alžběta	ruzickovaalzbeta@seznam.cz	Institute of Phonetics, Faculty of Arts, Charles University, Prague, Czechia
Rybicki Jan		Institute of English Studies, Jagiellonian University, Poland
Sanada Haruko	h_sanada@nifty.com	Rissho University, Japan
Šandrih Branislava	branislava.sandrih@fil.bg.ac.rs	Faculty of Philology, University of Belgrade, Serbia
Seredin Pavel	paul@phys.vsu.ru	Voronezh State University, Russia
Sherstinova Tatiana	sherstinova@gmail.com	St. Petersburg State University, Russia
Stanković Ranka	ranka@rgf.rs	University of Belgrade, Serbia
Steiner Petra	steiner@ids-mannheim.de	Institut für Deutsche Sprache, Germany
Sugiyama Mao	sugiyama.mao0420@gmail.com	Osaka University, Japan

Topolski Krzysztof	krzysztof.topolski@uwr.edu.pl	University of Wrocław, Poland
Torre Iván González	ivan.gonzalez.torre@upm.es	Universidad Politécnica de Madrid, Spain
Tuzzi Arjuna	arjuna.tuzzi@unipd.it	Università di Padova - Dipartimento FISPPA, Italy
Usami Mayumi	usamima@gmail.com	National Institute for Japanse Language and Linguistics, Japan
Uslu Tolga	uslu@em.uni-frankfurt.de	Goethe University of Frankfurt, Germany
Vogl Gero	gero.vogl@univie.ac.at	University of Vienna, Austria
Voskaki Rania	rvoskaki@hotmail.com	National and Kapodistrian University of Athens, Greece
Vulanović Relja	rvulanov@kent.edu	Kent State University at Stark, USA
Vydrin Valentin	vydrine@gmail.com	LLACAN, France
Wang Letao	546189249@qq.com	Xi'an Jiaotong University, China
Wang Yaqin	wyq322@126.com	Zhejiang University, China
Xu Ming	xuming318@sina.com	Xi'an Jiaotong University, China
Xu Yingying	xuyingying04@126.com	School of Foreign Languages, Dalian Maritime Universit, China
Yamazaki Makoto	yamazaki@ninjal.ac.jp	National Instutute for Japanese Language and Linguistics, Japan
Yan Ling	yanling67@sina.com	School of International Studies, Communication University of China, China
Yu Biyan	yubiyan813@163.com	Xi'an Jiaotong University, China
Zámečník Lukáš	lukas.zamecnik@upol.cz	Palacký University Olomouc, Czechia
Zhang Peng	yaoyuan2046@163.com	Zhongnan University of Economics and Law, China
Zhou Haiyan	zhouhy@sxnu.edu.cn	Xi'an Jiaotong University, China
Zhu Hong	shukou005@hotmail.com	Zhongnan University of Economics and Law, China

ISBN 978-83-950966-0-0