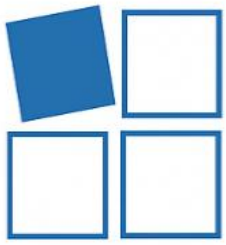

QUALICO 2025

Quantitative Linguistics Conference 2025, Brno, Czechia
June 26–28

Book of Abstracts

IQLA
International Quantitative Linguistics Association

Contents

A dynamical system approach to sign language

Jan Andres[1], Martina Benesova[2], Eva Fiserova[1], Jiri Langer[3]

[1]Palacký University Olomouc, Faculty of Science, Olomouc, Czech Republic. [2]Palacký University Olomouc, Faculty of Arts, Olomouc, Czech Republic. [3]Palacký University Olomouc, Faculty of Education, Olomouc, Czech Republic

Sign language texts have been investigated as time series by means of the Lyapunov-like exponents technique and in terms of topological entropy. Their positivity indicates a sensitive dependence on initial conditions (a bad predictability), resp. a complexity bahaviour (a deterministic chaos). Our quantitative analysis will be oriented into three directions:

(i) a unique sequence of the lengths of signs measured a) in the number of pseudo-syllables and b) in time (seconds),
(ii) two different (i.e. higher and lower) hierarchy levels of the same text,
(iii) an exploration of the sign levels of two different texts.

Some remarks will be supplied concerning the comparison of the obtained results with those for the spoken languages

Random Forest Binary Classification for Word Origin Attribution based on Word Embeddings. A case study in Modern Hindi

Jacek Bąkowski

Institute of Polish Language, Polish Academy of Sciences, Kraków, Poland

Synonymy is commonly occurring and widespread, but also an intriguing linguistic phenomenon. From a functional point of view, apart from stylistic considerations, such as to avoid repetitions, there is no semiotic justification for the existence of absolute synonyms as they don't increase the expressive potential of the language. Even if synonyms do name the same thing, they do in different ways and present different perspectives on a situation. They may also carry different cultural stereotypes.

Synonyms often come from different historical layers that make up a language. Such is also the case of Hindi: a long linguistic contact with Persian, spanning a period of about seven centuries has resulted in multitude of Perso-Arabic loanwords which were assimilated and blended into the tissue of Hindi along with their Sanskrit counterpart. Thus, the question arises whether, several centuries after these borrowings appeared, is it still possible in Modern Hindi to discern their origin and usage patterns based on their embeddings and regardless of their near identical meaning?

Here a Random Forest model was trained on word embeddings of synonym pairs in Hindi of both Sanskrit and Perso-Arabic origin, with the target value being the origin of the word. Quite unexpectedly, it turned out that the model has proven to be an effective tool for classifying synonyms and assigning them their correct origins.

This sheds a new light on synonymy, showing that there are discernible features in synonym usage but also the power of context, through which such classification was possible to achieve.

Dimensionality Reduction Methods for Authorship Attribution

Antonio Calcagnì, Livio Finos, Andrea Sciandra, Arjuna Tuzzi

University of Padova, Padova, Italy

This study explores dimensionality reduction methods for authorship attribution that differ from the selection of the most frequent words (MFWs). We present various approaches, including some techniques not previously used in authorship attribution. The first proposed method is based on Correspondence Analysis, which recently has demonstrated superior results in comparison to MFWs and Large Language Models in the contexts of AI detection and contemporary novels. The second method is based on Principal Component Pursuit, a technique previously used in video surveillance systems, which allows a DTM to be decomposed into two additive matrices: a sparse matrix (foreground) and a low-rank matrix (background). The idea is that by eliminating the 'noise' (sparsity) in the data, we can obtain a structure of predictors that can improve classification performance. Finally, non-Euclidean methods were used to compute intertextual distances, in particular the Riemann distance applied to the covariance matrix of function words frequencies. Multidimensional scaling can be applied to the resulting intertextual distances to obtain a reduced set of coordinates in order to better separate the authors of the texts. The efficacy of these approaches is then evaluated by comparing the results using metrics derived from the confusion matrices. The comparison of methods also involves the interpretability of results, especially in a machine learning framework. These analyses are performed on different corpora (e.g. human vs AI-generated texts, novels) and the preliminary results provided a clear picture of the best performing methods.

Comparative Analysis of Thesauri for Functional Expression Classification Using Machine Learning

Bocheng Chen

Graduate Institute for Advanced Studies, SOKENDAI, Tokyo, Japan

This study evaluates the effectiveness of two Japanese thesauri, the *Kadokawa Ruigo Shin Jiten* and the *Bunrui Goi Hyō(WLSP:Word List by Semantic Principles)*, in differentiating functional expressions "*ni*" and "*ni yotte*" in passive sentences using machine learning models. While thesauri have long been utilized in Japanese linguistic research, their comparative effectiveness for specific tasks has not been well explored. This study employs decision trees, random forests, and support vector machines (SVM) to quantitatively compare their performance.

A dataset of 600 examples (300 for "ni" and 300 for "ni yotte") was extracted from the *Balanced Corpus of Contemporary Written Japanese (BCCWJ)* and annotated with classification codes from both thesauri. Experimental results showed that models using the *Kadokawa Ruigo Shin Jiten* consistently outperformed those using the *Bunrui Goi Hyō* in terms of accuracy, stability, and feature extraction efficiency. The SVM model achieved the highest performance, with 84.44% accuracy on test data using the *Kadokawa Ruigo Shin Jiten*.

Error pattern analysis revealed that misclassifications in the *Kadokawa Ruigo Shin Jiten* were more concentrated, while those in the *Bunrui Goi Hyō* were widely distributed, suggesting that a simpler three-level structure enhances learning. The study concludes that appropriately abstracted thesauri structures improve machine learning performance for functional expression differentiation tasks. Future research should explore larger datasets, additional expression pairs, and advanced models such as deep learning to further validate these findings.

Menzerath’s law: Non-replicable significant inverse relationship between number of phonemes per word and phoneme inventory size

Gertraud Fenk-Oczlon

University of Klagenfurt, Klagenfurt, Austria

This paper reports a study (Fenk-Oczlon & Pilz, 2021) that failed to replicate earlier quantitative typological findings by Nettle (1995, 1998), Wichmann (2011), and Moran and Blasi (2014). All these authors identified a significant inverse relationship between word length, defined as the number of phonemes, and phoneme inventory size. However, the 2021 study, which utilized a different dataset and methodology, did not find significant negative correlations between phoneme inventory size and the number of phonemes per word. Instead, it found only a small, non-significant negative correlation. Interestingly, the study did reveal a significant negative correlation between phoneme inventory size and word length when word length was defined as the number of syllables."

This unexpected result is discussed in light of Menzerath's law and the principle that "the more syllables in a word, the fewer phonemes in a syllable." Our 2021 study not only confirms Menzerath’s (1954) principle, originally demonstrated in German words, across 61 languages but also identifies a highly significant relationship between the number of phonemes per syllable and phoneme inventory size. Taken together, these results suggest that languages with longer words tend to have simpler syllable structures, and simpler syllable structures are associated with smaller phoneme inventories. Thus, the inverse relationship between word length and phoneme inventory size is more pronounced when word length is measured by the number of syllables rather than by the number of phonemes, as previous research has shown.

REFERENCES

Fenk-Oczlon, G. & Pilz, J. (2021). Linguistic complexity: Relationships between phoneme inventory size, syllable complexity, word and clause length, and population size. *Front. Commun.* 6:626032.

Moran and Blasi (2014) “Cross-linguistic comparison of complexity measures in phonological systems,” in Measuring Grammatical Complexity, eds F. J. Newmeyer and L. B. Preston (Oxford: Oxford University Press), 217-240.

Nettle, D. (1995). Segmental inventory size, word length, and communicative efficiency. *Linguistics* 33, 359–367.

Nettle, D. (1998). Coevolution of phonology and the lexicon in twelve languages of West Africa. *J. Quant. Linguist.* 5, 240–245.

Wichmann, S., Rama, T., Holman, E. W. (2011). Phonological diversity, word length, and population sizes across languages: the ASJP evidence. *Linguist. Typol.* 15, 177–197.

No hard constraint on word order in the noun phrase. The science of the unseen

Ramon Ferrer-i-Cancho

Universitat Politècnica de Catalunya, Barcelona, Spain

The frequency of the preferred order for a noun phrase formed by demonstrative, numeral, adjective and noun has received significant attention over the last two decades. Some researchers have argued for a hard constraint, e.g., universal grammar or some universal cognitive mechanism, that would explain why not all 24 possible orders are attested (Medeiros et al 2016). We revisit Cysouw's (2010) hypothesis: all orders are a priori possible but some are not attested due to undersampling. We present a statistical framework that (a) allows one to ascertain if certain orders are banned (b) explains why some orders have not yet been observed and (c) predicts when more orders will be observed. Our theory predicts that all or almost all orders are likely to be present on Earth and also predicts that a sample larger than 4600 languages is needed so that all 24 orders are observed with high probability. The fact that research so far has explored a rather small fraction of all languages on Earth offers a more parsimonious explanation to unattested orders than some hard constraint.

REFERENCES

Cysouw, Michael. 2010. Dealing with diversity: Towards an explanation of NP-internal word order frequencies. Linguistic Typology 14(2-3). 253–286.

Ferrer-i-Cancho. 2024. The exponential distribution of the orders of demonstrative, numeral, adjective and noun. https://arxiv.org/abs/2502.06342

Medeiros, David P. & Piattelli-Palmarini, Massimo & Bever, Thomas G. 2016. Many important language universals are not reducible to processing or cognition. Behavioral and Brain Sciences 39.

How do translators deal with repeated reporting verbs in literary texts? A multifactorial comparative Study in English-to-Russian and English-to-Slovak language pair

Łukasz Grabowski[1], Daniel Borysowski[1], Filip Kalaś[2], Lorenzo Mastropierro[3]

[1]University of Opole, Opole, Poland. [2]University of Business and Economics, Bratislava, Slovakia. [3]University of Insubria, Varese, Italy

In this multifactorial study, interfacing stylistics, corpus linguistics and translation, we aim to identify the predictors of repetition or lexical variety in the translation of reporting verbs from English into Russian and from English into Slovak. Using a sample of 20 literary novels (English-to-Russian) and 14 literary novels (English-to-Slovak) from InterCorp v. 15 (Rosen et al. 2022), we fit multiple negative binomial regression with mixed effects to assess the effect that selected predictor variables (e.g. frequency of a ST verb, its number of senses in Princeton WordNet) have on the response variable: the number of TT reporting verb types (lemmas) a ST reporting verb is translated. If the number of types is high it means that translators opted for lexical variety (i.e. used various TT reporting verbs as translation equivalents of a single repeatedly used ST reporting verb in English-original novels). The overall model fit per the lowest AIC and BIC values obtained through backward elimination reveals that semantic category of a ST reporting verb, its frequency and translation date as well as the translator as a random effect have the largest individual contributions to explaining the proportion of variation (75%) in the response variable in the Russian translations. The low variance (0.05) in the random effect means that the impact of individual translators is relatively similar: there is some variability between the translators, but it is relatively small, and no single translator significantly influenced overall results. We later compare these findings with the ones in the English-Slovak language pair.

Quantitative Linguistics and *Large Language Models*: Epistemological Challenges and Applications

Antoni Hernández-Fernández, Bernardino Casas, Jaume Baixeries Juvillà, Neus Català Roig

Universitat Politècnica de Catalunya, Barcelona, Spain

Quantitative linguistics (QL) has long sought to uncover universal patterns governing language phenomena through mathematical and statistical modeling. Classical linguistic laws provide explanatory frameworks grounded in the scientific method, showing that texts as symbolic representations of speech degrade the fit of these laws (Torre et al., 2019). However, the rise of "*Large Language Models*" (LLMs) introduces new epistemological challenges for language analysis and QL. Yet, are LLMs truly *models*? While QL emphasizes explicit theoretical formulation and hypothesis testing, LLMs rely on data-driven optimization and emergent linguistic representations. This contrast raises fundamental questions about explanatory frameworks in linguistics and the balance between theoretical rigor and empirical adaptability. Can emergent patterns from LLM analysis be considered genuine linguistic laws or mere artifacts of statistical training?

Comparative analyses of classical linguistic laws in LLMs reveal foundational parallels while exposing critical divergences in methodology, generalizability, interpretability, and scalability. We discuss the epistemological tension between QL's hypothetico-deductive frameworks and the inductive paradigms of LLMs (or even *Small Language Models*), proposing hybrid methodologies integrating these approaches (Català et al., 2024). This convergence advances both linguistic theory refinement in QL and Artificial Intelligence explainability enhancement.

REFERENCES

Català, N., Casas, B. & Hernández-Fernández, A. (2024). The semanticity of Catalan words: quantitative linguistics in the era of large language models. Madrid: Dykinson, 2024, p. 249-268. http://hdl.handle.net/2117/413308

Torre, I., Luque, B., Lacasa, L., Kello, C & Hernández-Fernández, A. (2019). On the physical origin of linguistic laws and lognormality in speech. *Royal Society Open Science*, *6*(8), 191023. https://royalsocietypublishing.org/doi/10.1098/rsos.191023

Distance-Based Relevance Zones: A Case Study from Norwegian Digitized Collections

Lars Johnsen

Oslo, Norway

This paper extends distance and frequency-based approaches for measuring distributional relevance between words by introducing a three-zone model of contextual relevance. Using data from the Norwegian newspaper digitization project, we demonstrate how the right context of target words can be systematically divided into three distinct zones: near (1-3 words), middle (4-8 words), and distant (>8 words), each capturing different types of linguistic relationships.

Our analysis shows that these zones exhibit distinct statistical patterns in terms of probability ratios (PMI) and positional variance. Using non-logarithmic PMI preserves the direct interpretation of word associations as probability ratios, showing how many times more likely words are to occur together than by chance. The near zone typically captures syntactic relationships with high probability ratios and low positional variance, while the distant zone reveals discourse-level relationships with selective high ratios and moderate variance. The middle zone serves as a transition area, characterized by higher variance and generally lower association strengths.

We combine these probability ratios with position statistics to create zone-specific word embeddings and graphs. The data is captured in tuples of the form (w_1, w_2, pmi, dist, dist_variance), where:

- pmi = $p(w_2|w_1)/p(w_2)$ directly measures association strength
- dist = avg(position(w_2)-position(w_1)) captures average distance
- dist_variance = variance(dist) indicates positional consistency

Initial results from Norwegian data show distinct patterns in how different word classes behave across these zones, offering new perspectives on historical language change and textual organization.

Automatic quantitative text analysis of contemporary French-language literatures from Europe, Africa and Canada.

Ewa Kalinowska[1], Adam Pawłowski[2], Tomasz Walkowiak[3]

[1]Uniwerity of Warsaw, Warsaw, Poland. [2]University of Wrocław, Wrocław, Poland. [3]Wrocław University of Technology, Wrocław, Poland

Most quantitative studies of literature use stylometry as a frame of reference and often focus on authorship attribution. Similar assumptions also guide research on text taxonomy. In this case, the focus is not on linguistic derivatives of human genotype, but on other text characteristics that arise from social, or environmental conditions.

The study of (post)colonial languages presents a unique opportunity to scientifically investigate the influence of a wide range of factors on a text. These languages, used across different continents, maintain their core morphological, lexical, and syntactic structures but are influenced by local cultures or education systems in various ways.

We believe that quantitative methods of text analysis should allow for the creation of taxonomies that highlight the influence of social, religious, historical, and environmental factors on literary texts written in the same language but within different cultures.

The research material consists of several hundred contemporary novels (from the 20th and 21st centuries) written in French by European authors (from France, Belgium, and Switzerland), as well as Canadian and African authors.

In this research, we will apply methods based on Hierarchical Clustering (both agglomerative, bottom-up approaches and divisive, top-down approaches), LDA (Latent Dirichlet Allocation, topic modeling), and Deep Learning techniques, such as word embeddings (e.g., Word2Vec, GloVe) and sentence embeddings (e.g., BERT, Sentence-BERT). We will compare taxonomies and word maps based on the entire vocabulary as well as selective subsets of lexemes (Pattern-Based Approach). Subsets created for selective taxonomies will be generated semi-manually, using synonym dictionaries and wordnets.

Gustav Herdan and mobile Quantitative Linguistics: Brno/Brünn – Vienna – Bristol

Emmerich Kelih

Institute for Slavonic studies, Vienna, Austria

On the occasion of the Qualico 2025 conference in Brno (Czech Republic), the idea was born to give a short lecture on the role and work of Gustav Herdan (1897-1958). He was born in Brno, studied in Prague and Vienna and was forced to move to the UK in 1938 for political reasons. Having studied both linguistics and statistics, his main contribution could be seen in his seminal works, among which are Language as Choice and Chance 1956, The Calculus of Linguistic Observations 1962, Quantitative Linguistics 1964 and The Advanced Theory of Language as Choice and Chance 1966. From today's point of view, Herdan's contribution to Quantitative Linguistics seems to be almost forgotten, and therefore we would like to present some "highlights" of his conception of Quantitative Linguistics (Type-Token-Relation, Vocabulary Size, Zipf's Law), but also to show how far he stuck to “orthodox” structuralist concepts.

REFERENCES

Best, Karl-Heinz; Altmann, Gabriel (2007): Gustav Herdan (1897-1968). In: Glottometrics (15), S. 92–96.

Confirming the Minimization Effect in MDD and MHD among Second Language Learners

Saeko Komori[1], Masatoshi Sugiura[2]

[1]Chubu University, Kasugai, Japan. [2]Nagoya University, Nagoya, Japan

This study aims to confirm whether the distance minimization effect occurs in second language learners by measuring MDD and MHD and comparing them with those of native speakers. Komori et al. (under review) analyzed longitudinal data from first- and second-year Chinese learners of Japanese. They found that while MDD increased in the first year, its rate of increase declined and flattened in the second year, suggesting that dependency distance minimization (DDM) occurred after the first year, with learners' MDD approaching that of native speakers. However, MHD increased steadily across both years, remaining well below native speakers' MHD even at the end of the second year. This study further investigates this hypothesis.

RQ1: Is the DDM effect consistent in MDD across different second-year learner data?

RQ2: Does the MDD of higher third-year learners exceed the native speakers' MDD threshold?

RQ3: Does the minimization effect occur in MHD as learners progress from the second to the third year?

In this study, we used corpus data from Komori (2019) with 38 second-year learners, 32 third-year learners, and 35 native speakers. Participants wrote opinion essays on the same topic.

Results showed the DDM effect in MDD for the second year learners (1.78) and the third year learners (1.95), which came close to the threshold of native speakers' (1.99). The third-year learners' MHD, however, was much lower (1.95) than the native speakers' (2.77). The distance minimization effect in MHD was not observed even in the third-year, which requires further investigation with more advanced learners.

Pseudosyllable in sign language

Jiri Langer[1], Jan Andres[2], Martina Benesova[3], Eva Fiserova[2]

[1]Palacký University Olomouc, Faculty of Education, Olomouc, Czech Republic. [2]Palacký University Olomouc, Faculty of Science, Olomouc, Czech Republic. [3]Palacký University Olomouc, Faculty of Arts, Olomouc, Czech Republic

Although sign languages of the Deaf are natural languages, they differ fundamentally from spoken languages due to their distinct modality (visual-motoric). One key difference lies in the simultaneous production of phonemic components within individual signs, which are grouped into parameters such as hand shape, place of articulation, movement, palm orientation, finger orientation, and hand arrangement. This simultaneity poses challenges in the efforts to segment signs into constituent units analogous to syllables in spoken languages.

While several theories have been proposed to identify and define syllables in sign languages, these approaches fall short of capturing the functional equivalence of syllables in spoken languages (e.g., under one specific definition of the syllable, most signs are typically categorized as monosyllabic, limiting its usefulness for quantification purposes.). However, cluster analysis utilized in our experiments has identified a novel unit termed the pseudosyllable, which demonstrates adherence to the Menzerath-Altmann law, offering new insights into the hierarchical structure of sign language.

Communicative efficiency: laws and orders

Dr. Natalia Levshina

Radboud University, Nijmegen, The Netherlands

Communicative efficiency has been a prominent theme in linguistics and cognitive science. There is plenty of evidence showing that language users try to communicate efficiently, saving time and effort while making sure that they transfer the intended message successfully. In my talk I will present three case studies that illustrate the main principles of efficient communication and methodological challenges that we face. In the first one, I will discuss Zipf's law of abbreviation and engage with recent controversy about the functional explanations of the law, as well as about the role of frequency and informativity. In the second case study, I will use corpus data from typologically diverse languages to resolve a paradox: why are SOV languages predominant, despite having longer dependency distances? I will demonstrate that full SOV clauses are infrequent in verb-final languages, so the word order is less problematic than one could assume. This highlights the importance of gradient, token-based approaches in cross-linguistic research on communicative efficiency. Finally, I will explore correlations between thirteen linguistic, cultural and demographic variables related to the cues to Subject and Object (the famous 'who did what to whom' question) and argue that we should explore efficiency at the level of a culture or community, considering cross-cultural variation in communicative needs.

Government proposals for 2022 Elections in Brazil: Topic Modelling and Political Spectrum

Rodrigo de Lima Lopes

Universidade Estadual de Campinas, Campinas, Brazil

This research analyses the 2022 Brazilian presidential candidates' government proposals using Reinert's (1983, 1990) hierarchical topic modelling methodology, implemented via a package in R. This corpus is part of the BRPoliCorpus, a corpus of Brazilian Political Documents. It applies a mixed-methods approach, scraping proposals from the Supreme Electoral Court (TSE-Tribunal Superior Eleitoral) website, pre-processing texts, and segmenting them into context units. The text corpus comprises 136,463 tokens across diverse political parties. Key methodological steps included: 1) segmenting texts into 30-word units with a tolerance for preserving sentence coherence; 2) assigning unique IDs for each segment based on party affiliation; 3) employing Reinert's method to identify lexicon patterns within a Document Term Matrix (DTM), optimising clusters through chi-squared tests and partition refinement and 4) validating the optimal number of topics qualitatively. Results identified statistical lexical patterns in four topics: labour, development, infrastructure, and government. Lexical choices might align with candidates' political spectra, as left-wing parties highlight labourers' rights, aligning development with sustainable growth and infrastructure with social welfare. Right-wing parties advocate for deregulation, economic liberalisation, and privatisation, portraying development as resulting from industrial expansion.

REFERENCES

Reinert, M. (1983). Une méthode de classification descendante hiérarchique: Application à l'analyse lexicale par contexte. Les Cahiers de l'analyse Des Données, 8(2), 187–198. Reinert, M. (1990). Une méthode de classification des énoncés d'un corpus présentée à l'aide d'une application. Les Cahiers de l'analyse Des Données, 15(1), 21–36.

Rolling topics in Positivist and Young Poland novels: a comparative study

Wojciech Łukasik

Uniwersytet Jagielloński, Kraków, Poland

In this paper, I intend to use topic modelling to research two epochs in Polish literature: Positivism and Young Poland. I will start with observations related to the typology of topics obtained from a corpus of prose texts, and then proceed to describe the topics generated by a corpus of 190 Positivist and Young Poland texts. I will compare the topics with ones obtained from a Young Poland-only corpus, pointing to similar topics being generated from both corpora and discuss the possible implications of these data for a more general classification of topics generated by corpora of literary texts. Next, I will present examples where high probabilities of particular topics co-occur with particular characteristics of texts related to theme and style. I will discuss the differences between the two groups of texts in the corpus in a search for quantifiable markers of Positivist and Young Poland style. The final part of my paper will include several "rolling topics" analyses where the texts will be divided into even parts to be analysed separately. This allows one to illustrate the changing probabilities of topics throughout one given text in a plot, which is then used to locate segments where a given topic is highly probable, and to verify the topic's connection to thematic or stylistic characteristics of a text basing on concrete examples. These analyses will be used to discuss the role of topics in texts, and the differences between Positivist and Young Poland texts, in more detail.

On a mathematical formulation of a theory of language

Ján Mačutek, Gejza Wimmer, Michaela Koščová

Mathematical Institute, Slovak Academy of Sciences, Bratislava, Slovakia

Gabriel Altmann often wrote about language laws as statements deducible from language theory. From a purely mathematical point of view (and with a huge simplification), language laws are special cases of one general mathematical formula, or, seen from the other side, a linguistic theory is a mathematical formula generalizing language laws. However, constructing such a theory is not an easy task (we remind the still ongoing quest of physicists for a "theory of everything").

In linguistics, almost all language laws (and observed regularities for which mathematical models exist) are special cases of a general approach suggested by Wimmer and Altmann, published in the Handbook of Quantitative Linguistics in 2005. There is, however, one exception: the Piotrowski law, which models a diachronic development of a language property. This law is usually modelled by the logistic function, which is not a special case of the general formula mentioned above.

In our contribution, we suggest a mathematical approach to modelling which unites the formula by Wimmer and Altmann with that of Piotrowski. We will also show that our model provides a good fit for the diachronic development of word length in Chinese.

The Wheel of Time: The Computational Stylometry Comparison of Robert Jordan and Brandon Sanderson

Nikolaos Maniotis, Athanassios Karassimos

School of English, Aristotle University of Thessaloniki, Thessaloniki, Greece

This study investigates a closed set of candidates, focusing on a comparative stylometric analysis between the late Robert Jordan, author of The Wheel of Time series, and his successor, Brandon Sanderson, who completed the final three books. Three key research questions are addressed in this study. The first explores whether Sanderson, as the continuator of Jordan's legacy, sought to emulate his predecessor's writing style or allowed his own stylistic tendencies to surface while building on Jordan's material. The second investigates claims found in online forums regarding the authorship of the final book's epilogue and explores whether it was written by Jordan, Sanderson, or a combination of both. Lastly, the third research question examines other supplementary linguistic characteristics. Using the 'stylo' package in R to investigate word and character frequencies on 2-gram and 3-gram levels, the final three Wheel of Time books were compared with the twelve books of the same saga authored by Jordan and four of Sanderson's earlier works. Initial principal components analysis and hierarchical cluster analysis experiments found that the books in question were authored by Sanderson, indicating that he did not act as a ghostwriter but retained his authorial signature. The results for the last book's epilogue were mixed, with most tests attributing it to Sanderson, while some suggested Jordan as the author. This may have been due to editorial intervention or the corpus size of the epilogue. Empirical testing, such as the approach taken in this study, is crucial for refining methods and improving analysis reliability.

The BALAGHA Score: A Quantitative Linguistics Approach to Arabic Rhetorical Analysis

Mandar Marathe

SOAS University of London, London, United Kingdom

The "**B**alāgha **A**ssessment for **L**iterature in **A**rabic with di**G**ital **H**umanities **A**pproaches" (BALAGHA) Score is an interdisciplinary, mixed-methods framework for the quantitative assessment of Arabic Rhetoric (*al-balāgha*), derived from well-established Quantitative Linguistics methodologies.

Widely employed across Arabic literature – including in poetry, prose, and the Qur'ān – Arabic Rhetoric utilises approximately 100 rhetorical devices such as metaphor, simile and allegory to enhance textual cohesion, aesthetic impact, and persuasive force. However, the literary criticism of Arabic Rhetoric remains qualitative and subjective, leading to large inconsistencies in critical assessments. There are currently no quantitative frameworks for evaluating Arabic Rhetoric, limiting empirical comparisons across different texts and authors.

The BALAGHA Score draws upon previous work in lexical density and relative frequency analysis (Ure, 1971; Halliday, 1985, 2014; Al-Wahy, 2019). It uses rhetorical device density as a surrogate marker of the rhetorical qualities of Arabic discourse. The BALAGHA Score is calculated as the total number of rhetorical devices in a sample, divided by the total number of morphemes in the sample (as a measure of sample size), scaled by a multiplication factor of 100. This easy-to-calculate metric enables cross-corpus, diachronic, inter-author, and cross-genre analyses of rhetorical expression, facilitating empirical comparisons across historical periods, authors, and genres.

Establishing a new quantitative foundation for Arabic rhetorical assessment, the BALAGHA Score is a novel application of Quantitative Linguistics in Arabic literary studies. It provides objective and reproducible insights into Arabic rhetorical use. Furthermore, its methodological framework is adaptable to other languages, contingent upon appropriate language-specific modifications for rhetorical feature extraction.

Evaluating the Effectiveness of Different Topic Modeling Techniques on a Technical Corpus: A Comparative Study

Alessandro Meneghini, Arjuna Tuzzi

University, Padova, Italy

The rapid growth of digital text data in policy-making has led to an increasing reliance on topic modeling techniques in the social sciences to uncover themes through a distant reading perspective. However, the proliferation of different approaches to topic modeling has created a challenge for researchers and decision-makers in choosing the most effective method. This challenge is further complicated by the emergence of Word Embeddings-based techniques, which offer promising results but introduce new complexities. This study compares the efficiency of five topic models: Latent Dirichlet Allocation (LDA), Non-Negative Matrix Factorization (NMF), and two Word Embeddings-based approaches (BERTopic with HDBscan and BERTopic with K-Means). The models are evaluated on a corpus of abstracts from European Cohesion Policy projects in Italy and Greece (2014–2020), characterized by a technical language and a high percentage of hapax. Using metrics such as Topic Coherence, Topic Diversity, and Intertopic Distance, the results indicate that BERTopic with K-Means and HDBscan outperform other models, confirming findings from studies on common news datasets. However, the performance metrics vary significantly depending on the subcorpus and number of topics targeted. Furthermore, the study highlights that the clustering technique used in Word Embeddings-based models has a significant impact on these performance metrics. This suggests that the automatic use of such methods should be approached with caution, emphasizing the importance of considering the specific clustering approach for optimal results.

Investigating the Impact of Semantic Similarity on Stylometric Attribution Using Controlled Artificial Texts and Delta Distances

George Mikros[1], Radek Čech[2], Petra Mutlová[2]

[1]Hamad Bin Khalifa University, Doha, Qatar. [2]Masaryk University, Brno, Czech Republic

This study investigates how semantic similarity affects stylometric methods, focusing on Delta distances for authorship attribution. It was inspired by Czech authors' analysis of medieval Latin texts, where stylometric methods for authorship attribution failed. In particular, these methods proved to be extremely sensitive to the initial settings, meaning that small changes in the parameters led to fundamentally different results. We assume that this is caused by the high similarity of the texts.

To analyze the impact of similarity on methods, controlled artificial text corpora were created to simulate varying levels of semantic overlap and assess clustering robustness. Using OpenAI o1-mini, 60 texts (300 words each) were generated for three virtual authors (A, B, and C). Each text consisted of two blocks: a shared block common across all authors (inter-overlap) and an author-specific block unique to each author (intra-overlap). Multiple corpora were systematically designed with varying word overlap percentages, ensuring controlled semantic and stylistic overlap for analysis.

K-means clustering and hierarchical clustering analysis were used to examine intra-author and inter-author separability. Both intra-author and inter-author overlaps we manipulated (90%, 50%, 10% intra-author overlaps, 10%, 50%, 90 inter-author overlaps) to observe how it influences author differentiation.

The findings confirm that increasing semantic similarity diminishes the reliability of Delta distances in distinguishing authors. By employing controlled experimental conditions, this study underscores the importance of mitigating semantic overlap in stylometric analyses and emphasizes the need for robust feature selection in authorship attribution methodologies, particularly in contexts of high semantic similarity.

Stylometric Analysis for Machine Translation Evaluation: Investigating Delta Distance as a Measure of Stylistic Fidelity

George Mikros[1], Vilelmini Sosoni[2], Vasilis Manousakis[3], Kelly Polychroniou[4]

[1]Hamad Bin Khalifa University, Doha, Qatar. [2]Ionian University, Corfu, Greece. [3]University of Patras, Patras, Greece. [4]Boston University, Boston, USA

This study explores the potential of stylometric analysis as an alternative measure of machine translation (MT) quality, focusing on the Greek-to-English translation of nine short stories written by the author and translator Dr. Vasilis Manousakis. Traditional MT evaluation metrics—such as BLEU, BLEURT, COMET, and TER—primarily assess lexical and semantic adequacy, but they may not fully capture stylistic fidelity. To address this gap, we apply Burrows' Delta, a well-established authorship attribution method, to quantify the stylometric distance between human translations (HTs) and MT outputs produced by ModernMT platform. Our hypothesis is that higher-quality MT outputs should exhibit lower Delta distances to HTs, indicating greater stylistic similarity.

Using the R Stylo package, we computed Delta distances between HTs and MT outputs and then merged these values with existing evaluation scores. Correlation analysis reveals that Delta distances negatively correlate with BLEURT (-0.45), BLEU (-0.42), and COMET (-0.29), indicating that more fluent and accurate translations tend to preserve stylistic features better. Conversely, Delta distances positively correlate with TER (0.32), reinforcing the notion that poor translations exhibit greater stylistic divergence. These findings suggest that stylometric measures, particularly Delta, can complement existing MT evaluation frameworks by providing insights into stylistic fidelity.

This research contributes to the growing field of computational stylistics in MT evaluation, offering a novel approach to assessing translation quality beyond lexical and semantic accuracy.

Comparison of Human-Written and LLM-Generated Texts: A Complete QL Package

Jiří Milička

Charles University, Prague, Czech Republic

This study systematically examines various quantitative linguistic relationships and looks at how human-written texts differ from comparable texts generated by large language models.

The main idea of this paper is to exhaustively review as many QL relationships as possible and determine where and why they differ (Zipf's Laws, Herdan's Law, distributions of segment lengths on various levels, Menzerath-Altmann Law, frequency distributions of morphological categories and syntactic attributes, properties of syntactic graphs, lexical diversity...). Thanks to this systematic approach, we will be able to explore whether the differences can be explained by the Köhlerian synergetic control cycle.

The study is based on the PseudoBrown and PseudoKoditex corpora, which are collections of texts produced by LLMs to be comparable with the original Brown (English) and Koditex (Czech) corpora. The corpora contain several tens of thousands of texts generated by both current and historical models from both Western and Chinese companies (Anthropic, OpenAI, Meta, DeepSeek). At the time of the conference, these corpora will be publicly available and thus accessible for further research.

Neural iImplementation of lexical meaning using topographic representation of environmental input topology

Hermann Moisl

Newcastle University, Newcastle upon Tyne, United Kingdom

This paper addresses the general problem of how to implement lexical meaning in a natural or artificial physical neural system. It proposes that this can be done by preserving the similarity structure of environmental input stimuli in the activation patterning of the neural system, and that this structure preservation can be mathematically modelled by an artificial neural network architecture that homeomorphically maps a Voronoi input manifold topology to a system state topology represented as a Delaunay graph. The motivation is both scientific and technological. The scientific motivation is the age-old question of how linguistic meaning arises in the mind and how this relates to the structure and dynamics of the physical brain. The technological one is that of artificial intelligence, that is, how to build an artificial language system which inorporates meaning; this has recently become topical in that questions are being asked and claims made about whether or not large language models are conscious. The discussion is in three main parts: the first part defines 'meaning' and 'topographic representation' as understood by the discussion, the second describes the neural architecture and the associated mathematical topics of Voronoi topology and Delaunay graph, and the third exemplifies their application to implementation of lexical meaning.

On Linguistic Complexity: Form-Meaning Pairing Through Subword Token Polysemy

Takuto Nakayama

Keio University, Tokyo, Japan

Recently, whether "all languages are equally complex" has intrigued researchers. However, there is no consensus on how linguistic complexity should be measured. In particular, semantic aspects are often overlooked in complexity calculations. To address this gap, this study aims to compute linguistic complexity by focusing on form-meaning pairing.

The pilot study is designed as follows. First, subword embeddings are obtained from a corpus using BERT. Second, kernel density estimation (KDE) is applied to the embeddings of each subword to estimate how frequently a subword is used with a specific meaning. Third, the Shannon entropy of each subword is calculated based on the KDE results, and the average entropy is taken as the entropy of the language. The pilot study uses randomly selected Wikipedia articles as the dataset. The procedure is conducted on 100 articles at a time and repeated 10 times. For a precise analysis, subwords with a frequency greater than 1,000 are considered.

The pilot study examines seven languages (Basque, English, Finnish, French, Hungarian, Indonesian, and Turkish), covering multiple language families (Altaic, Austronesian, Basque, Indo-European, and Uralic). The result shows that Basque has the highest entropy (0.1807), while English has the lowest (0.0364). This suggests that the seven languages have approximately 1.0255 to 1.1334 meanings per subword. Therefore, from the perspective of form-meaning pairing, polysemy across languages appears to fall within a relatively narrow range. For further research, the dataset should be expanded, in terms of the number of languages and the volume of data within each language.

Language models as cognitively realistic mathematical models of grammar

Andrea Nini

University of Manchester, Manchester, Great Britain

The way grammar has been typically modelled mathematically in Linguistics is via the use of formal languages. For example, a Phrase Structure Grammar would conceive a natural language grammar as a set of terminal words that can be combined using a list of rules. Although this model of grammar can work effectively at a certain level of abstraction, this same model is ineffective at capturing various linguistic phenomena that characterise real language use (e.g. Goldberg, 2003; Hilpert, 2014; Sinclair, 2004). In response to the needs of working on real language, usage-based frameworks have proposed the idea that, in reality, grammar is an end of a continuum that starts with the lexicon. This model of grammar or, more appropriately, of lexicogrammar, does not have a mathematical formal framework in the same way that formal language theory is to Phrase Structure Grammars. In this talk, I will propose the idea that a language model is the correct mathematical formal framework to model this way of conceiving grammar. I will do so by borrowing from Cognitive Psychology findings about language processing, which point to the mechanisms of grammar processing to be similar to a probability distribution over sequences (Ullman, 2015). I will then show some indirect empirical corroborations of this idea from an authorship verification study that uses a method inspired by this model called LambdaG (Nini et al., 2025). Although computationally simple, this method outperforms various sophisticated neural network baselines, possibly indicating that the method is exploiting better approximations to the reality of linguistic individuality in grammar processing.

REFERENCES

Goldberg, A.E., 2003. Constructions: A new theoretical approach to language. Trends in Cognitive Science 7, 219–224.

Hilpert, M., 2014. Construction Grammar and its Application to English. Edinburgh University Press, Edinburgh.

Nini, A., Halvani, O., Graner, L., Gherardi, V., Ishihara, S., 2025. Grammar as a behavioral biometric: Using cognitively motivated grammar models for authorship verification. https://doi.org/10.48550/arXiv.2403.08462

Sinclair, J., 2004. Trust the Text: Language, Corpus and Discourse. Routledge, London.

Ullman, M.T., 2015. The Declarative/Procedural Model: A Neurobiological Model of Language Learning, Knowledge, and Use, in: Hickok, G., Small, S.L. (Eds.), Neurobiology of Language. Elsevier Science & Technology, San Diego.

The Menzerath-Altmann law: Good news and bad news

Michaela Nogolová[1], Ján Mačutek[2], Radek Čech[3]

[1]University of Ostrava, Ostrava, Czech Republic. [2]Mathematical Institute, Slovak Academy of Sciences, Bratislava, Slovakia. [3]Masaryk University, Brno, Czech Republic

The Menzerath-Altmann law says that shorter language units (constructs) consist, on average, of shorter parts (constituents). The law has been corroborated in many languages and for several language units. However, systematic results for units across all levels, from phonemes to (at least) sentences, are still lacking.

The good news is that, for the first time, we will present such results (for the Czech language). Relations between construct length and the mean constituent length will be modelled by a simple power law for these linguistic units: sentence – independent clause – clause – phrase – subphrase – chunk – word – phoneme.

The bad news is that the Menzerath-Altmann law does not hold for the relation between words length and the mean syllable length in some languages. Languages with simple syllable structure (e.g. ones that allow only V and CV syllables) do not "obey" the law. In such languages, the only way to shorten the mean syllable lengths would be to use words with many vowel clusters, which would be difficult to pronounce as well as to process. Such words would also contradict the so-called "horror aequi" principle, according to which there is a tendency in language to avoid sequences of adjacent (near-)identical units or structures. However, the validity of the Menzerath-Altmann law can probably be "saved" if we re-interpret it so that it does not speak only of length, but of a more general "cost" in the sense of the Zipf's least effort principle.

Computational Thematics and Lexical Variation in Brazilian Novels: The Phenomenon of Regionalism

Adriana Pagano[1], André Coneglian[1], Maciej Eder[2]

[1]Universidade Federal de Minas Gerais, Belo Horizonte, Brazil. [2]Institute of Polish Language (Polish Academy of Sciences), Krakow, Poland

In this paper, we undertake the question of regionalism in Brazilian literary novels from the turn of the 19th century and the early 20th century. To this end, we train a topic model to discover the latent thematic structure of the entire corpus. Then we use supervised classification, combined with SHAP values analysis, in order to see which textual features (in our case: proportions of particular topics) are responsible for the distinction between Regionalism and Urbanism. This is a demanding research problem, because we face a number of confounding factors here. Firstly, particular authors exhibit their own stylistic voice. Secondly, the corpus covers a certain time span, which inevitably introduces a temporal signal. While it is relatively simple to automatically group the target novels into, say, 19th-century vs. 20th-century ones, it is non-trivial to isolate the temporal signal alone. The same holds true for other signals, including, at the first place, the trace of Regionalism.

In order to extract the most distinctive features, we used SHAP values. In the case of the Brazilian novel corpus, we deal with a binary classification problem, therefore we compute SHAP values based on logistic regression models. The topics that turned out to be distinctive, capture the vocabulary related to nature and animals (água 'water', rio 'river', gado 'cattle', cobra 'snake'), farm (facão 'machete', fazenda 'farm') weather conditions (seca 'drought'), morphoclimatic zones (caatinga 'scrub'). However, some less intuitive topics were also highlighted by our SHAP model.

LexaMorf: A Tool for Lexico-Morphological Quantitative Text Analysis

Petr Porizka

Palacký University, Olomouc, Czech Republic

LexaMorf is a new tool for linguistic annotation and quantitative analysis of morphological categories and lexical diversity/dispersion. It is modularly built in Python and R, including a graphical user interface. One of the aims was to create a user-friendly tool, with the possibility of adding more features in the future and combining functions that would require several different tools. The lexical module allows batch processing of multiple files and their comparison. It provides a series of quantitative parameters, both traditional (TTR, Root TTR, U, C, K, I, Vm, etc.) and newer measures (MATTR, MSTTR) that do not depend on the different sizes of texts to be compared and neutralising this factor; it also creates a frequency dictionary of texts and calculates different frequencies (AF, RF/i.p.m., ARF, ARF_Ratio) with subsequent visualisation of the dispersion. The second, morphological module, allows the linguistic annotation (lemmatisation and tagging) of texts in different languages according to the chosen tagger/parser (currently: MorphoDiTa and UDPipe) and the output can be used both for building of an annotated corpus and for the quantitative analysis of different morphological parameters such as POS, grammatical number, case, verb tense, mode, person, etc. Thanks to the implementation of UDPipe, it is a versatile tool that allows the quantitative exploration of grammatical parameters in almost 80 languages (for some of them even diachronically, e.g. modules for Old French, Ancient Greek, Old East Slavic, Old Church Slavic, etc.), thus opening up new possibilities for the quantitative analysis of linguistic data.

Stylometric Analysis of Dramas by the Čapek Brothers

Petr Porizka

Palacký University, Olomouc, Czech Republic

Karel and Josef Čapek were Czech writers active in the first half of the 20th century. Both also wrote plays: some separately, and three together. But they had different approaches to the language of their plays: while Karel used contemporary language and a more colloquial style, Josef used a bookish to archaic style. Because of these differences, the data will be divided into three groups: the dramas of (1) Karel, (2) Josef, and (3) their joint plays. These three sub-collections will be analyzed using stylometric methods (the Stylo package in R). Other basic quantitative characteristics will also be provided, exploring the different structural layers (dialogues of all characters, only of the protagonists, situational/authorial comments, stage directions) and comparing the linguistic aspects in the works of the two authors. While Josef tends towards linguistic exclusivity, where all characters speak in a standardised code, even in a bookish style, his brother Karel uses a more authentic and colloquial style of dialogue and the principle of linguistic contrast. He chooses different stylistic variants according to the functional stratification in terms of social affiliation, i.e. a colloquial, non-standard variety of language for the lower classes (servants, etc.), and literary Czech for the socially higher-ranking or more educated characters. The position of the jointly written dramas in the dendrograms (and other outcomes of the stylistic analysis) will help to answer the question of which of the two brothers had more authorial influence in the joint dramas.

AI as a Democratic Actor: Investigating LLM Narratives on Democracy

Maud Reveilhac

University of Zurich, Zurich, Switzerland

Artificial intelligence (AI) models like large language models (LLMs) are now used in political discourse, decision-making support, and public debate mediation. This study examines how LLMs portray AI's role in democracy when responding to different ideological and political perspectives, using profile endorsement such as AI governance advocate, techno-skeptic democrat, and institutionalist democrat. Specifically, it investigates whether LLMs adapt their rhetorical strategies, reinforce biases, or maintain neutrality across diverse viewpoints. LLMs are already influencing public discourse through chatbot interactions, policy reports, and automated journalism. Understanding how they construct their narratives on AI's democratic role is essential for evaluating their potential impact on public perception and policy recommendations. Each profile is prompted with structured political inquiries regarding AI's role in referendums, digital democracy, and citizen decision-making. The study applies quantitative linguistic methods to analyze AI-generated discourse, focusing on variations in lexical complexity, argument structure, and framing patterns across different ideological perspectives. Responses are collected from multiple LLMs to detect variation in ideological framing and rhetorical strategies and processed using computational text analysis techniques. We measure lexical diversity, textual readability, and surprisal to assess linguistic variation and coherence. Additionally, topic modeling and sentiment analysis are employed to identify ideological positioning. This study provides novel insights into how AI systems internally frame their role in governance, whether they reflect or challenge ideological biases. It also addresses theoretical and methodological questions in computational linguistics by assessing the interplay between AI, argumentation, and political text analysis through mathematical and statistical modeling.

Recursive characteristics of the length and order of elements in Japanese by linguistic layer: Relationship between morpheme length and position within clause constituents

Haruko Sanada

Rissho University, Tokyo, Japan

Through a series of studies analyzing the relationship between the length and position of linguistic elements, we determined the relationship between the length and position of clause components in a clause and between the length and position of a clause in a sentence (Sanada 2018; in print). It has been shown that the constituent elements become shorter as they approach the end of a clause or sentence. This paper presents the results of a study of 'the relationship between morpheme length and position within clause constituents'. On average, morphemes are shorter towards the ends of complements, adjuncts, and predicates. This is similar to the position and length of the elements that comprise a clause and the clauses that comprise a sentence. Furthermore, because these characteristics appear recursively, that is, nested, in multiple linguistic layers, we discuss the impact of these recursive features on communication.

REFERENCES

Sanada, Haruko. (2018). Quantitative aspects of the clause: the length, the position and the depth of the clause. *Journal of Quantitative Linguistics*, vol.26 (4), pp.306 - 329.
Sanada, Haruko. (In print). The length and order of grammatical elements in the Japanese clause. In: Pawlowski, A.; Embleton, S.; Mačutek, J.; Xanthos, A. (eds.) *Mathematical Modelling in Linguistics and Text Analysis: Theory and Applications*. Amsterdam, The Netherlands: John Benjamins.

Studying Borel normality of Spontaneous Japanese in binary expression

Yosuke Takubo[1,2], Masayuki Asahara[3,4], Makoto Yamazaki[3]

[1]Niihama College, Niihama, Japan. [2]High Energy Accelerator Research Organization (KEK), Tsukuba, Japan. [3]National Institute for Japanese language and Linguistics (NINJAL), Ichikawa, Japan. [4]The Graduate University for Advanced Studies (SOKENDAI), Hayama, Japan

Random numbers are fundamental for encrypted communication. To ensure secure data transmission and prevent information leakage or wiretapping, it is crucial to quantitatively verify the randomness of bit sequences. Techniques from this field may also be applicable to the statistical analysis of natural language texts, leveraging their inherent randomness.

Analyzing Borel normality is one of such method that evaluates the frequency distribution of n-bit sequences. For completely random numbers, each n-bit sequence appears uniformly. In our previous studies, the Borel normality of written Japanese was examined by transforming text into binary sequences using various character encodings, such as UTF-8, SJIS, and EUC. The results revealed that different registers such as newspapers, magazines, and books, exhibited similar feature of Borel normality, regardless the encoding method.

Building on our previous findings, this study investigates the Borel normality of written Japanese composed solely of Hiragana, Katakana, and Romaji (Romanized Japanese), excluding Kanji, to isolate the randomness attributable to other elements. This approach allows us to analyze the Borel normality of reading in written Japanese. Additionally, the Borel normality of spoken Japanese is examined, using Corpus of Spontaneous Japanese (CSJ), which is divided into five registers such as academic presentation speech and simulated public speaking. Finally, this study further explores its relationship with linguistic metrics such as TTR (Type-Token Ratio) and Zipf's law.

Corpus-driven threshold-based argument categorization and word-sense disambiguation on verbal semantic selection for subjects

Nándor Virág

University of Szeged, Szeged, Hungary

Verbs impose selectional constrains on their arguments, one of them being semantic selection. This can be observed in corpora by the distribution of subjects next to each verb. Thus, a verb's selectional properties can be deduced from the frequency of its subjects' semantic categories. However, questions arise: what are the semantic categories of the subjects, and how can a verb's selectional constrain be determined based on its subjects?

To address these questions, WordNet can be used as base for a hierarchical structure in finding a fitting semantic category by the hypernymy relation between synsets of a verb's subjects: the concept, that is the most specific hypernym to each subject, acts as category tag. Given the variability in language, handling outliers, that appear as subjects even though they don't fit the constrain imposed by the verb, becomes essential. This can be tackled by setting an appropriate threshold (minimum percentage of words considered) to filter out exceptions.

Setting this threshold poses an opportunity: using this method, we can disambiguate distinct word senses in the verbs standing as predicates. For example, the verb *roll* has subjects with differing semantic categories for its distinct meanings, i.e. *John rolled the ball* / *The ball rolled*. The key quantitative linguistic task is twofold:

- defining an optimal threshold for both outlier filtering and WSD,
- determining an acceptable level of entropy within a semantic group of subjects.

By fine-tuning these parameters, we aim to develop a reliable method for semantic classification and enhancing NLP applications in argument structure analysis.

Investigating Context Awareness in Automatic Pragmatic Annotation

Nándor Virág, Tibor Szécsényi

University of Szeged, Szeged, Hungary

Determining pragmatic features poses a greater challenge compared to identifying morphological or syntactic ones. During pragmatic annotation of a corpus, human annotators must consider not only the formal characteristics of individual words but also the broader, often multi-sentential, context of the annotated expression. This process requires drawing inferences based on contextual information and world knowledge to determine pragmatic functions. This study investigates whether large language models, such as BERT, trained for similar tasks rely on the available context in a comparable manner, and examines the differential impact of closer versus more distant context.

We trained a BERT model on a manually annotated corpus containing information about the functions of imperative elements. The reliability of the automatic annotation was evaluated using precision, recall and F1-score. Our findings indicate that BERT can annotate with a reliability comparable to human annotators.

Three experiments were conducted to examine the context awareness of the trained model. The result of the first experiment was that increasing the fixed-size context window improves the reliability of the analysis. The second experiment explored how annotation reliability depends on the position of the target element within a fixed-size context, distinguishing between left and right as well as small and large contexts. Finally, we varied the size of the left and right contexts surrounding the target element to determine the minimal context size required.

The experiments offer scalar, one-, and two-dimensional representations for analyzing the context awareness of the model, providing insights into how it processes contextual information during pragmatic annotation.

The distribution of dependency distance and hierarchical distance in contemporary written Japanese and its influencing factors

Linxuan Wang[1], Shuiyuan Yu[2]

[1]Renmin University of China, Beijing, China. [2]Beijing Language and Culture University, Beijing, China

To explore the relationship between Dependency Distance (DD) and Hierarchical Distance (HD), this article analyzes the characteristics of both based on the Balanced Corpus of Contemporary Written Japanese.

Although the increase in sentence length causes an increase in the entropy of the DD distribution, the proportion of dependencies where DD is less than 10 shows little fluctuation, and the proportion of DD=1 remains around 60%. It suggests that regardless of sentence length, native Japanese speakers tend to use more adjacent dependencies to reduce Mean DD (MDD) due to cognitive resources limitation. In contrast, as sentence length increases, the value of HD with the highest proportion gradually increases and the proportion of HD=1 decreases since the HD distribution is more greatly affected by the valency of the predicate than the DD distribution.

Besides, the intersection of MDD and MHD between sentence length=5 and 6 also reflects the role of valency. Before the intersection, MDD is greater than B; after that, their relationship is reversed. The maximum valency of the predicate in Japanese is 4 and before the valency reaches saturation (i.e., when sentence length is less than 6), contemporary written Japanese will prioritize increasing linear complexity (i.e., DD). After reaching saturation, hierarchical complexity (i.e., HD) remains higher than linear complexity.

Finally, MDD and MHD have a significant negative correlation after controlling for sentence length, also reflecting the trade-off relationship between DD and HD because the valency of the predicate will promote the growth of MDD while restricting the growth of MHD.

Does Menzerath-Altmann Law hold true for English-as-a-foreign-language learner English?

Mengge Wang, Jingyang Jiang

Zhejiang University, Hangzhou, China

The study explores whether the Menzerath-Altmann Law (MAL) holds true for compositions written by EFL learners. The study was conducted under the framework of dependency grammar, with a self-built corpus of 325 compositions (47, 761 tokens) across five proficiency levels. 23 types of learner errors (syntactic, lexical, and mechanical errors) were annotated according to a combination of error parsing results from ChatGPT (GPT-4) and a professional college English teacher ($F_{0.5} > 70\%$). Referring to these error annotations, learner language was segmented into different linguistic units. The MAL analyses were conducted at both the sentence-clause-phrase level and the clause-phrase-word level. Additionally, based on whether a non-finite verb is a mandatory element of a clause, clauses were segmented in two different ways to examine whether the clause definition influences the MAL fitting. The results indicate that 1) the correlation between sentence length and clause length abides by the MAL irrespective of composition levels, with nearly all goodness-of-fit values exceeding 0.96. 2) The correlation between clause length and phrase length follows the MAL in three composition levels, all yielding satisfactory fitting results. 3) The segmentation of clauses does not influence the MAL fitting. Overall, the findings indicate that, despite the presence of various linguistic errors and the factor of proficiency level on learner linguistic structures, the MAL is still valid in learner language, though the degree of fitting can be influenced by the composition proficiency and the linguistic units being concerned.

Long-Range Dependence in Word Time Series

Paweł Wieczyński[1], Łukasz Dębowski[2]

[1]unaffiliated, Gdańsk, Poland. [2]Institute of Computer Science, Polish Academy of Sciences, Warszawa, Poland

It has often been expressed that texts in natural language exhibit long-range dependence (LRD) (Altmann et al., 2009; Lin and Tegmark, 2017; Mikhaylovskiy and Churilov, 2023) as opposed to an exponential decay of correlations being characteristic of Markov processes (Lin and Tegmark, 2017). The observation of LRD for sufficiently large lags means that generation of texts in natural language cannot be modeled by a Markov process of a relatively small order (Lin and Tegmark, 2017). In particular, the existence of LRD would explain the necessity of using complex memory architectures in successful large language models (Vaswani et al., 2017; Behrouz et al., 2024).

In this paper, we systematically explore a simple measure of dependence to check whether texts in natural language actually exhibit a power-law decay characteristic of LRD. Our data set is the Standardized Project Gutenberg Corpus (Gerlach and Font-Clos, 2020). Similarly as Mikhaylovskiy and Churilov (2023) but differently than Lin and Tegmark (2017), we seek for LRD on the level of words rather than on the level of characters. This yields a more prominent effect. Rather than the mutual information, which is difficult to estimate, we consider the centered cosine similarity between words treated as vectors. As a vector representation of words, we consider the word2vec embeddings (Mikolov et al., 2013). Given these assumptions, we observe a power-law decay characteristic of LRD, confirming the earlier results of Mikhaylovskiy and Churilov (2023), who considered a similar setup.

Distribution and Frequency of Words in Texts: An Analysis of Occurrence Intervals and Positions

Makoto Yamazaki

National Institute for Japanese Language and Linguistics, Tachikawa, Japan

This study quantitatively examines word distribution in texts, focusing on word occurrence intervals and positional distribution—two underexplored but crucial aspects of vocabulary studies.

We analyzed two texts: Shayo (The Setting Sun) by Dazai Osamu (1947), a novel, and Yuki (Snow) by Nakaya Ukichiro (1938), a scientific essay. Both contain 50,000–60,000 words.

Using morphological analysis, we found that in the novel text, symbols had the shortest intervals, followed by particles, nouns, affixes, verbs, adjectives, and conjunctions. Scientific essay showed almost the same results. The Wilcoxon rank sum test showed that this trend of occurrence was statistically significant.

A comparison of the part-of-speech composition ratios of words with narrower intervals (1000 words or less) and words with wider intervals (10,000 words or more) showed that the ratios of verbs and adjectives increased as the interval increased, while the ratios of nouns and particles decreased. This trend was similar for both texts.

In terms of occurrence position, we examined the distribution of the first and last occurrence positions of words. The results showed that as the frequency of occurrence in the text increased, the first occurrence position was closer to the beginning of the text and the last occurrence position was closer to the end of the text. This suggests that, in general, as the frequency of a word increases, the word appears more evenly throughout the text, which can be seen as a manifestation of lexical cohesion as argued by Halliday & Hasan (1976).

Language universals in sentence length: Comparing sentence length distributions of 10 languages

Yikai Zhou[1], Haitao Liu[2]

[1]Zhejiang University School of International Studies, Hangzhou, China. [2]Fudan University College of Foreign Languages and Literature, Shanghai, China

The patterns of human language abstracted by probabilistic models can help improve the proficiency of natural language processing. Numerous studies have proved that the universality of human language can be reflected in the distribution of word lengths, but the case of sentence lengths remains unclear. Sentence length demonstrates the flow of the mind about decisions to arrange pauses in speech and written texts for clear delivery, and it is an intermediate level between words and text structures. We compared the sentence length distributions of news texts in 10 languages and found that the lengths of full sentences (the interval between two consecutive full stops) and minor sentences (the interval between two consecutive either full or minor stops) in all languages conform to the extended positive negative binomial distribution, and the interlingual differences in minor sentence length distribution are smaller than that of full sentence length; the parameters of sentence length distribution can measure the differences between languages, and the clustering results are generally consistent with the linguistic genealogical relationships. Moreover, genre adaptation in terms of sentence length is also unified across languages. We propose that the universal pattern of sentence length distribution in human languages reflects the common cognitive resources of humans; the interlingual differences in the characteristics of sentence length distribution reflect the language-specific rules of structure and unit organization; and genre differentiation in sentence length roots in humans' common cognitive adaptive capability.