QUALICO 2023

12th International Quantitative Linguistics Conference

BOOK OF ABSTRACTS

28-30 June 2023
University of Lausanne
Anthropole building, rooms 1031 & 1129

wp.unil.ch/qualico2023/

UNIL | Université de Lausanne
Section des sciences
du langage
et de l'information

IQLA
International Quantitative Linguistics Association

UNIL | Université de Lausanne
Centre de linguistique
et des sciences du langage

QUALICO 2023
Book of abstracts

Table of contents

Conference program summary

Day 1 (Wednesday 28.6)

	Room A (ANT-1031)	Room B (ANT-1129)	In front of Room A	Unithèque
08:30-09:30			Registration	
09:30-10:00	Opening			
10:00-11:00	Invited lecture 1			
11:00-11:30			Break	
11:30-12:30	A1 - Linguistic variation	B1 - Sign languages		
12:30-14:00				Lunch
14:00-15:30	A2 - Quantitative indices 1	B2 - East-Asian languages		
15:30-16:00			Break	
16:00-17:30	A3 - Dependencies 1	B3 - Stylometry		
17:30			Welcome drink	

Day 2 (Thursday 29.6)

	Room A (ANT-1031)	Room B (ANT-1129)	In front of Room A	Unithèque
09:00-10:30	A4 - Dependencies 2	B4 - Diachrony		
10:30-11:00			Break	
11:00-12:30	A5 - Dependencies 3	B5 - Lexical semantics		
12:30-14:00				Lunch
14:00-15:00	Invited lecture 2			
15:00-16:00	Poster session			
16:00-16:30			Break + posters	
16:30-17:00	A6 - Quantitative indices 2			

After the last session: *social event*

Day 3 (Friday 30.6)

	Room A (ANT-1031)	Room B (ANT-1129)	In front of Room A	Unithèque
09:00-10:30	A7 - Language complexity	B7 - Slavic languages		
10:30-11:00			Break	
11:00-12:30	A8 - Statistical laws	B8 - Neural networks		
12:30-14:00				Lunch
14:00-15:00	A9 - Text classification	B9 - Social media		
15:00-15-30	Closing			
16:00-17:00		IQLA business meeting		

Detailed conference program

Day 1 (Wednesday 28.6)

	Room A (ANT-1031)	Room B (ANT-1129)	In front of Room A	Unithèque
08:30-09:30			*Registration*	
09:30-10:00	*Opening*			
10:00-11:00	*Invited lecture 1. Chair: Aris Xanthos*			
	Tanja Samardžić: Subword tokenization as a method for discovering and comparing linguistic structures (p.11)			
11:00-11:30			*Break*	
11:30-12:30	*Session A1 - Linguistic variation. Chair: Hermann Moisl*	*Session B1 - Sign languages. Chair: Guillaume Guex*		
	Theodore Manning, Eugenia Lukin, Ross Klein and Patrick Juola: Construction & Analysis of a Map-Based Corpus for Tracking Linguistic Variation & Demographic Characteristic Identification (p.13)	**Jan Andres and Jiri Langer**: Persistence of Czech sign language (p.17)		
	Yaqin Wang and Jingqi Yan: Investigating the Linguistic Variation of Lyrics Genre through Quantitative Lens (p.15)	**Jiri Langer and Jan Andres**: Significance of sign parameters based on the quantitative linguistic analysis (p.18)		
12:30-14:00				*Lunch*
14:00-15:30	*Session A2 - Quantitative indices 1. Chair: François Bavaud*	*Session B2 - East-Asian languages. Chair: Adam Pawłowski*		
	Neus Català i Roig, Jaume Baixeries i Juvillà, Lucas Lacasa and Antonio Hernández-Fernández: Semanticity, a new concept in Quantitative Linguistics: an analysis of Catalan (p.19)	**Xinying Chen and Ziyan Wei**: How to define a word in Japanese? Word segmentation in Japanese from the Zipf's law perspective (p.23)		
	Lars Johnsen: Term distance as a relevance measure (p.21)	**Biyan Yu and Lu Fan**: Colligation Diversity in Chinese Grammaticalization: An Entropy-based Approach (p.25)		

	Stefan Th. Gries: A dispersion measure that is by design orthogonal to frequency and its predictive power for lexical decision times (p.22)	**Hua Wang**: A Quantitative Study of Noun Phrase Length in English and Chineseg (p.27)		
15:30-16:00			*Break*	
16:00-17:30	*Session A3 - Dependencies 1. Chair: Sheila Embleton*	*Session B3 - Stylometry. Chair: George Mikros*		
	Michaela Nogolová, Ján Mačutek and Radek Čech: Distributional properties of linear dependency segments (p.29)	**Patrick Juola and Alejandro J. Napolitano Jawerbaum**: A Comparative Analysis of Authorship Attribution in a Creole and Non-Creole Language (p.34)		
	Sonia Petrini and Ramon Ferrer-i-Cancho: The distribution of syntactic dependency distances (p.30)	**Adam Pawłowski and Tomasz Walkowiak**: Can stylometry reveal more than a human reader in a text? A study based on Romain Gary and Emile Ajar's case. (p.36)		
	Lu Fan and Biyan Yu: Probability Distribution of Dependency Distance in Translational language Based on a Treebank Transformed from a Bidirectional Parallel and Comparable Corpus (p.32)	**Jacques Savoy**: French Plays of the 17th Century: A Stylometric Analysis (p.38)		
17:30			*Welcome drink*	

Day 2 (Thursday 29.6)

	Room A (ANT-1031)	Room B (ANT-1129)	In front of Room A	Unithèque
09:00-10:30	*Session A4 - Dependencies 2. Chair: Ján Mačutek*	*Session B4 - Diachrony. Chair: Coline Métrailler*		
	Aixiu An, Yingqin Hu and Anne Abeillé: A gradient model of LDD acceptability (p.40)	**Tereza Klemensová and Michal Místecký**: Long Time No Joe: Piotrowski-Law Development of Personal Names in the Diachronic Perspective (p.46)		
	Felix Bildhauer, Thilo Weber and Franziska Münzberg: Syntactic boundaries or word-count distance? Co-reference configurations and the choice between finite and non-finite adnominal clauses in German (p.42)	**Quentin Feltgen**: A Zipf-Mandelbrot Approach to Diachronic Productivity (p.47)		
	Haruko Sanada: The length and order of grammatical elements in the Japanese clause (p.44)	**Eric S. Wheeler and Sheila Embleton**: Visualizing Character Profile Shifts in English Texts Over The Centuries (p.49)		
10:30-11:00			*Break*	
11:00-12:30	*Session A5 - Dependencies 3. Chair: Emmerich Kelih*	*Session B5 - Lexical semantics. Chair: Arjuna Tuzzi*		
	Michaela Hanuskova, Michaela Nogolová and Miroslav Kubát: Development of mean dependency distance in Czech L2 texts across proficiency levels A1 to C1 (p.51)	**Alizée Lombard, Anastasia Ulicheva, Maria Korochkina and Kathleen Rastle**: The regularity of polysemy patterns in the mind: Computational and experimental data (p.56)		
	Saeko Komori, Masatoshi Sugiura, Ramon Ferrer-i-Cancho, Lluís Alemany-Puig and Wenping Li: Syntactic development and optimality of dependency distances for Japanese as a second language (p.52)	**Kaleigh Woolford**: Modelling semantic differentiation between near-synonyms with word2vec and t-SNE (p.58)		

	Yingqi Jing, Joakim Nivre and Michael Dunn: Multilevel phylogenetic model shows no evidence for dependency locality in Indo-European (p.54)	**Hermann Moisl**: Homomorphism, Voronoi tesselation, and lexical meaning (p.60)		
12:30-14:00				*Lunch*
14:00-15:00	*Invited lecture 2. Chair: Sheila Embleton*			
	George Mikros: Detection of AI-Generated Texts and Quantitative Analysis of Large Language Model Outputs (p.12)			
15:00-16:00	*Poster session. Chair: François Bavaud*			
	Corinne Rossari, Cyrielle Montrichard and Claudia Ricci: Disambiguating adverbs within a quantitative approach. Identification and annotation of polysemy (p.93)			
	Giuseppe Samo: Syntactic strategies for null and pronominal subjects: a quantitative study (p.95)			
	Alessandro Meneghini, Valentina Rizzoli and George Markopoulos: Quantitative analysis of interviews in cooperation contexts: a stylometric profiling of relevant psychological processes (p.97)			
	Petr Pořízka: CapekDraCor database and some aspects of quantitative linguistic analysis of the Čapek brothers' plays (p.99)			
	Takuto Nakayama: Are All Languages Equally Complex?: Information Theory-based Method to Measure the Overall Complexity of a Language (p.100)			
	Yosuke Takubo, Masayuki Asahara and Makoto Yamazaki: Analyzing Japanese texts with evaluation of randomness in binary expression (p.102)			

	Tatsuhiko Matsushita: Text Covering Efficiency and Word Tier Analysis for the proposal of vocabulary learning order and the analysis of text genres (p.104)			
	Woonhyung Chung: Trump's Simple Language: His Idiolect or Global Trend? Exploring Lexical Sophistication in U.S. Presidential Discourses (p.106)			
	Barend Beekhuizen and Kaleigh Woolford: Community-specific Context Typicality as a determinant of lexical variation (p.108)			
	Tatsuhiko Matsushita: Part-of-speech proportion as an index of formality and informality: The case of Japanese (p.110)			
	Martin Hilpert, David Correia Saavedra and Jennifer Rains: Quantifying meaning differences between English clippings and their source words (p.112)			
	Justine Salvadori, Rossella Varvara and Richard Huyghe: Incidence- and abundance-based measures to assess rivalry in word formation (p.114)			
	Yan Liang: Probabilistic Regularity in Translation: A Quantitative Description of Dependency Treebank of Academic Abstracts (p.116)			
16:00-16:30			*Break + posters*	
16:30-17:00	*Session A6 - Quantitative indices 2. Chair: Guillaume Guex*			
	Stefan Th. Gries: Two+-dimensional uncertainty estimates for frequency, dispersion, and association measures (p.61)			

Day 3 (Friday 30.6)

	Room A (ANT-1031)	Room B (ANT-1129)	In front of Room A	Unithèque
09:00-10:30	*Session A7 - Language complexity. Chair: Aris Xanthos*	*Session B7 - Slavic Languages. Chair: Arjuna Tuzzi*		
	Petra Steiner: Morphological Complexity in Lexical Networks (p.62)	**Chenliang Zhou and Junyi Xu**: Uncovering the Relationships Among Slavic Languages: A Lexical Diversity Analysis (p.68)		
	Maud Reveilhac and Gerold Schneider: Measuring language complexity about European politics using different data sources and methods (p.64)	**Ján Mačutek, Emmerich Kelih and Michaela Koščová**: A quantitative approach to noun declension in Slavic language (p.70)		
	Zheyuan Dai and Jianwei Yan: Discourse Markers' Role in Syntactic Complexity of Sentence Structure: A Distance-driven Quantitative Case Study Based on TED Talks (p.66)	**Miroslav Kubát, Radek Čech and Xinying Chen**: Distribution of syntactic functions in different styles and genres (p.71)		
10:30-11:00			*Break*	
11:00-12:30	*Session A8 - Statistical laws. Chair: Emmerich Kelih*	*Session B8 - Neural networks. Chair: Guillaume Guex*		
	Iván G. Torre, Łukasz Dębowski and Antonio Hernández-Fernández: Menzerath-Altmann's law versus Menzerath's law as a criterion of complexity in communication (p.73)	**Olivier Rüst, Marco Baroni and Sabine Stoll**: Getting creative: A Neural Network approach to predicting child utterances in 12 typologically diverse languages (p.79)		
	Jiří Milička: Modelling Menzerath's Law with Gaussian Copula (p.75)	**Julia Lukasiewicz-Pater, Ximena Gutierrez-Vasquez and Christian Bentz**: Entropic analyses of the Voynich Manuscript using a diverse cross-linguistic corpus and neural networks (p.81)		

	Łukasz Dębowski and Iván González Torre: Principled Analytic Corrections of Zipf's Law (p.77)	**Magali Guaresi, Sofiane Haris and Laurent Vanni**: Text Analysis Using Convolutional Neural Networks with Multi-Head Attention (p.83)		
12:30-14:00				*Lunch*
14:00-15:00	*Session A9 - Text classification. Chair: Coline Métrailler*	*Session B9 - Social media. Chair: Radek Cech*		
	Matilde Trevisani and Arjuna Tuzzi: Capturing Distinctiveness: Transparent Procedures to Escape a Pervasive Black-Box Propensity (p.85)	**Wilkinson Daniel Wong Gonzales**: Bayesian and frequentist approaches to explaining (and predicting) morphosyntactic variation in East Asia using social media data (p.89)		
	Lars Johnsen, Adam Pawłowski and Tomasz Walkowiak: Linguistic image of selected decimal classification categories in large bibliographies. Comparative analysis of representative languages of Central Europe and Scandinavia (p.87)	**Prakhar Gupta, Elisa Pellegrino, Leyla Benkais and Aris Xanthos**: Assessing gender impact on paralinguistic accommodation in French WhatsApp conversations (p.91)		
15:00-15:30	*Closing session. Chair: Ján Mačutek*			
16:00-17:00		*IQLA business meeting*		

Tanja Samardžić
University of Zurich

Subword tokenization as a method for discovering and comparing linguistic structures

Abstract

Subword tokenization is unsupervised surface segmentation of words, applied as a preprocessing step when text is given as input to neural networks. For example, the word *coworking*, can be split into *co work ing* and each part is assigned a vector representation (embedding). All pretrained large language models apply some kind of subword tokenization, but the decisions on how this step should be performed remain largely arbitrary and with little reference to the structure of words.

A popular algorithm for performing subword tokenization is Byte-Pair Encoding (BPE), a general-purpose compression algorithm, which, applied to text, improves machine translation and other end-user tasks. Despite its usefulness in language processing, this method is commonly judged as not linguistically relevant, since its output is hard to align with any morphological analysis. The misalignment between BPE and linguistic analysis is puzzling: to compress language data efficiently, BPE needs to find subword patterns that reduce text redundancy. These patterns might not correspond to usual morphological analyses, but they are structural elements.

In this talk, I will show that a systematic analysis of subword units identified by BPE across a set of around 50 typologically diverse languages reveals linguistically relevant patterns. The types of units that have the strongest impact on compression are an indicator of morphological typology: for languages with richer inflectional morphology there is a preference for highly productive units, while for languages with less inflectional morphology, idiosyncratic units are more prominent. The features of BPE subword units can thus distinguish automatically between different morphological types of languages using only raw text.

By monitoring the outcome of compression steps, we can track the cross-linguistic differences in what kinds of redundancy are gradually removed in different languages., which opens a new possibility for describing and comparing languages. For instance, the output of BPE allows us to study the relative length of subword units, revisiting the famous Menzerath–Altmann law on a wide scale. The results of one such analysis show that the length of subword units identified by BPE tends to be rather evenly distributed: as the length of words increases, the length of subword units decreases *evenly and* not only on average. Cross-linguistic variation in the degree of evenness in subword units turns out to be a good criterion for deciding what kinds of languages should be taken into consideration for cross-lingual transfer of pretrained language models, making quantitative linguistic analysis highly relevant to contemporary multilingual natural language processing.

Detection of AI-Generated Texts and Quantitative Analysis of Large Language Model Outputs

George Mikros

Hamad Bin Khalifa University, Qatar

gmikros@gmail.com

Abstract

In recent years, there has been a seismic shift in the landscape of Natural Language Understanding (NLU) and Language Generation (LG) tasks, precipitated by the advent of Large Language Models (LLMs). These models, notably OpenAI's GPT-4 and Anthropic's Claude, have been recognized for their ability to produce high-quality, coherent, and context-specific textual content (Brown et al., 2020). The sophistication of these models is such that their written outputs frequently mirror human-produced text to the extent that eludes detection by most current AI-writing detectors.

In this lecture, we intend to present a quantitative analysis of the textual outputs generated by these two leading-edge LLMs, focusing on discerning linguistic features that distinguish them from human text production. We will scrutinize an extensive array of stylometric and linguistic characteristics and investigate the interrelations among these features utilizing a broad range of statistical methodologies and visualization techniques. Moreover, we will explore the latest advancements in detecting AI-generated writing. However, we argue that, given the current state of technology, it's not feasible to achieve this goal, especially in real-world educational scenarios.

Our wider objective in this talk is to develop a deeper understanding of the stochastic nature of AI-generated writing and to distinguish it from human text production. By doing so, we aim to shed light on the nuanced distinctions between machine-generated and human-generated writing, thereby offering new insights into the evolving field of AI-assisted text production.

Construction & Analysis of a Map-Based Corpus for Tracking Linguistic Variation & Demographic Characteristic Identification

Theodore Manning[1], Eugenia Lukin[2], Ross Klein[3], Patrick Juola[4]

City University of New York[1], Duke University[2], University of Pittsburgh[3], Duquesne University[4]

Project Map Lemon is a linguistic, map-based corpus in its infancy, which is currently in its second iteration. Map Lemon was created to obtain a baseline corpus for linguistic variation among English-speaking North Americans. The corpus currently houses upwards of 21,000 words across 185 participants, 10+ linguistic backgrounds, and 40+ US states and Canadian provinces. It presents a unique method for linguistic data collection, as the HCRC Map Task Corpus (1993) once attempted a similar task, however not for the written medium. Map Lemon additionally houses responses from 91 transgender and non-binary individuals, making it a fantastic resource for analyzing naturally elicited Queer writing. Up until the Map Lemon project, no corpus in the realm of stylometry had included genders outside of cisgender binary Male and Female (Mandell, 2019). Analysis using this corpus has revealed the potential to stylometrically disambiguate gender and sex, as well as region. Further research is currently being conducted to solidify these results. In addition, methods to prevent binarism of non-binary respondents during analysis are currently being explored.

Map Lemon data was gathered electronically via participants on Prolific writing responses in a Google form. They were given a small monetary reward for participation. Participants were asked to be as detailed as possible. In Experiment I, participants were asked to guide, in writing, the fictional Chad LemonLover to a lemonade stand utilizing a hand-drawn map and whatever cardinal directions or landmarks they desired. In Experiment II, participants were asked for their recipe for making lemonade. Demographic information was then collected, including birthplace, gender identity, assigned sex, linguistic background, etc.

Results from conducting stylometric analysis using K-Nearest Neighbor and part of speech tagging in the Java Graphical Authorship Attribution Program indicate that transgender respondents write most similarly to their gender identity rather than sex assigned at birth. Additionally, using the same analysis methods, a naturally occurring unknown nationality in our responses was most similar to Canadian writers (when compared to American and Canadian authors from the same corpus). The respondent later confirmed they are Canadian, showing that Map Lemon can be used to disambiguate region.

We feel Map Lemon demonstrates a unique data collection method for documenting natural written linguistic variation digitally, as well significant results for the field of transgender linguistics. Data collection for this corpus will continue, pending funding.

References

Mandell, L. (2019). *Gender and Cultural Analytics: Finding or Making Stereotypes?* (M. K. Gold & L. F. Klein, Eds.). University of Minnesota Press. https://doi.org/10.5749/j.ctvg251hk

University of Edinburgh. (1993). *HCRC Map Task Corpus*. Linguistic Data Consortium. https://doi.org/10.35111/9GE9-6C05

Title: Investigating the Linguistic Variation of Lyrics Genre through Quantitative Lens

Authors: Yaqin Wang [a], Jingqi Yan [b*]

a Center for Linguistics and Applied Linguistics, Guangdong University of Foreign Studies
b School of Foreign Languages, Zhejiang Gongshang University

* Corresponding author: Jingqi Yan
Address: No.18, Xuezheng Str., Zhejiang Gongshang University, Xiasha University Town, Hangzhou, China. 310018
Email: yjqveronica@live.com

Abstract
The lyrics of popular songs are considered a medium-dependent performed language, presenting strong emotional expressions and distinctive structural, metrical and lexico-grammatical features (Werner 2021). Despite the pervasiveness of popular songs and the uniqueness of lyrics, the genre of the song lyrics does not figure very prominently in quantitative-linguistic studies. Existing literature did not find agreements regarding the genre of lyrics, as some found it similar to the conversational spoken genre (Li & Brand 2009) while others took it as a special genre standing in between the speech-writing continuum (Kreyer and Mukherjee 2007). Moreover, previous studies on lyrics using corpus-based or computational approaches fail to include text size-independent indicators or features on various linguistic levels. In an attempt to better assess the genre of song lyrics and their salient features, the present study conducted a series of principle component analyses on the lyrics corpus based on features from three dimensions. Specifically, part of speech tags as syntactic features, text size independent quantitative stylometric measures from QUITA (Kubát et al., 2014), and psychometric indicators from LIWC 2022 (Boyd et al. 2022) were chosen. Based on clustering results from the pilot study, two music genres from Grammy Awards, i.e., rap and pop, together with three traditional genres from the reference BNC corpus, i.e., informative genre, imaginative genre, and spoken genre, were then introduced in the current research, totaling 400,942 tokens. For the two music genres, the PCA analysis on three levels of text features found a consistent intersection with spoken language but isolation from the informative genre and imaginative genre. The quantitative stylometric features had the best clustering performance, indicating the effectiveness in distinguishing genre differences (e.g., Mandravickaite and Krilavicius, 2018; Kubát and Milicka, 2013). In terms of differences within music genres, pop music seems to be most distinct from other traditional genres, whereas rap music shares more similarities with the spoken genre at every level. This finding was further consolidated when ten important features from three levels were then combined together for a follow-up PCA analysis. To conclude, the present study does not convincingly support the "speech-writing continuum" (Kreyer and Mukherjee 2007) for the lyrics genre. Rather, the findings see lyrics as a special genre with "pseudo-dialogical" (Murphey 1989) and multi-modality,

sharing characteristics of spoken language to a small extent. First employed in lyrics studies, quantitative stylometric indicators give a more revealing insight into detecting genres and styles of lyrics. Meanwhile, combined features from different linguistic levels are proven to best distinguish lyrics from other genres.

Keywords: lyrics; genre; stylometric features; psychometric features; part-of-speech tags

References

Boyd, R. L., Ashokkumar, A., Seraj, S., & Pennebaker, J. W. (2022). *The development and psychometric properties of LIWC-22*. Austin, TX: University of Texas at Austin. https://www.liwc.app

Kreyer, Rolf, and Joybrato Mukherjee. "The Style of Pop Song Lyrics: A Corpus-Linguistic Pilot Study." *Anglia - Zeitschrift Für Englische Philologie* 125, no. 1 (January 2007).

Kubát, M., & Milička, J. (2013). Vocabulary richness measure in genres. *Journal of Quantitative Linguistics*, *20*(4), 339-349.

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

Li, X., & Brand, M. (2009). Effectiveness of music on vocabulary acquisition, language usage, and meaning for mainland Chinese ESL learners. *Contributions to music education*, 73-84.

Mandravickaite, J., & Krilavicius, T. (2018). Quantitative analysis of textual genres: comparison of English and Lithuanian. In *Proceedings of the International Conference on Information Technologies* (pp. 61-67).

Murphey, T. (1989). The when, where, and who of pop lyrics: The listener's prerogative. *Popular Music*, *8*(2), 185-193.

Werner, Valentin. "Text-Linguistic Analysis of Performed Language: Revisiting and Re-Modeling Koch and Oesterreicher." *Linguistics* 59, no. 3 (May 26, 2021): 541–75. https://doi.org/10.1515/ling-2021-0036.

Authors
Prof. Jan Andres
Palacký University Olomouc, Faculty of Science, Czech Republic
jan.andres@upol.cz

Assoc. Prof. Jiri Langer
Palacký University Olomouc, Faculty of Education, Czech Republic
jiri.langer@upol.cz

Title
Persistence of Czech sign language

Keywords
Persistence; sign language; autocorrelation; Hurst exponents; global and local approaches; fractal dimension.

Abstract
By a persistence, we mean a long-time memory (autocorrelation) in language structures. For the desired results, we perform an approximative calculation of the strictly related Hausdorff dimension to a fractal model determined by time series as the reciprocal value of an arithmetic mean of the numerically calculated and statistically verified Hurst exponents on given scaling levels. After providing relevant details from quantitative linguistics, our apparatus is applied to a concrete sign language text. Our experiment so relies on the fractal analysis of a Czech sign language text consisting of 74 sentences, 247 clauses and 893 signs, i.e. on three scaling levels.

References
Andres, J.; Langer, J.; Matlach, V. (2020). Fractal-based analysis of sign language. Communications in Nonlinear Science and Numerical Simulation, 84 (2020), 1–14. https://doi.org/10.1016/j.cnsns.2020.105214

Hřebíček, L. (1998). Hurst's indicators and text. Some properties of word–frequency series. In: Altmann, J., Koch, W. A. (eds.), Systems. New Paradigma for the Human Sciences. Berlin: W. de Gruyter, pp. 572–588.

Hurst, H. E.; Black, R. F.; Simaika, Y. M. (eds.) (1965). Long–term storage: an experimental study. London: Constable.

Tanaka-Ishii, K. (2022). Statistical Universals of Language. Mathematical Chance vs. Human Choice. Springer, Cham.

Authors
Assoc. Prof. Jiri Langer
Palacký University Olomouc, Faculty of Education, Czech Republic
jiri.langer@upol.cz

Prof. Jan Andres
Palacký University Olomouc, Faculty of Science, Czech Republic
jan.andres@upol.cz

Title
Significance of sign parameters based on the quantitative linguistic analysis

Keywords
Sign language; sign parameters; phonological significance; statistical analysis; dendrograms.

Abstract
Signs in sign languages of the deaf are in a language structure analogous to words in spoken languages. In contrast to linearly assembled phonemes during word production, signs in their manual component are arranged simultaneously, when producing phonemes. These are concentrated in clusters of concrete parameters of the sign (hand shape, place of articulation, movement, palm orientation, finger orientation, hands arrangement). For a deeper understanding of the characteristics of sign languages, it is certainly appropriate to find out the phonological significance of individual parameters. Using the quantitative linguistic analysis of several utterances of native speakers in the Czech sign language, dendrograms were compiled using the standard techniques, relied on the statistical analysis, from which it is possible to infer the phonological significance of individual clusters based on the chronology of their breakdown. We will demonstrate that the observed conclusions correspond to the empirical experiences of the members of the sign language community.

References
Andres, J., Benešová, M., Langer, J. (2019). Towards a fractal analysis of the sign language. J. Quantitative Linguist., 1–18. https://doi.org/10.1080/09296174.2019.1656149
Bellugi, U., Fischer, S. (1972). A comparison of sign language and spoken language. Cognition 1(2-3), 173–200.
Langer, J.; Andres, J.; Benešová, M.; Faltýnek, D. (2020). Quantitative Linguistic Analysis of Czech Sign Language. Olomouc: Palacký University. https://doi.org/10.5507/pdf.20.24457277
Stokoe, W.C., (2005). Sign language structure: an outline of the visual communication systems of the American deaf. The Journal of Deaf Studies and Deaf Education 10, 3–37.

Semanticity, a new concept in Quantitative Linguistics: an analysis of Catalan

Neus Català i Roig[a], Jaume Baixeries i Juvillà[b], Lucas Lacasa[c], Antoni Hernández-Fernández[d,e]

[a] *TALP Research Center, Computer Science Departament, Universitat Politècnica de Catalunya, Campus Nord, c/Jordi Girona 1-3, Barcelona, 08034, Catalonia, Spain.*
[b] *LQMC Research Group, Computer Science Departament, Universitat Politècnica de Catalunya, Campus Nord, Barcelona, 08034, Catalonia, Spain*
[c] *Institute for Cross-Disciplinary Physics and Complex Systems (IFISC, CSIC-UIB), Campus Universitat de les Illes Balears, Palma de Mallorca, 07122, Balearic Islands, Spain*
[d] *Societat Catalana de Tecnologia, Secció de Ciències i Tecnologia, Institut d'Estudis Catalans, C/Carme 47, Barcelona, 08001, Catalonia, Spain*
[e] *Institut de Ciències de l'Educació, Universitat Politècnica de Catalunya, Av. Doctor Marañón 44-50, Edifici P, Planta 3, Campus Sud, Barcelona, 08028, Catalonia, Spain*

Abstract

G. K. Zipf formulated two statistical laws on the relationship between the frequency of a word with its number of meanings: the law of meaning distribution, relating the meanings of a word with its frequency rank, and the meaning-frequency law, relating the frequency of a word with its number of meanings. In a previous work we verify these linguistic laws in Catalan in written language and speech, finding two regimes in Zipf's rank-frequency law in large multi-author corpora [1]. However, one of the problems of quantitative studies of word meanings is how to compute the number of meanings of each word, as well as the relationship between meaning and syntax, a crucial problem of Linguistics. In this work we propose a new linguistic concept, ***semanticity*** (s_i), which relates the number of meanings of a word (μ) to the number of neighbouring words at distance i in the linguistic chain (λ_i) as $s_i = \frac{\mu}{\lambda_i}$. We also propose its normalized version (Figure 1).

If we understand the linguistic system as a complex network of words, λ_i is a measure of the centrality of the nodes, so that in general it is to be expected, as a hypothesis, that the so-called function words have a low semanticity (as they are high frequency words and link with many different words) and in contrast the content words have a high semanticity, as they

Preprint submitted to Qualico 2023 *December 19, 2022*

have more potential meanings than connections.

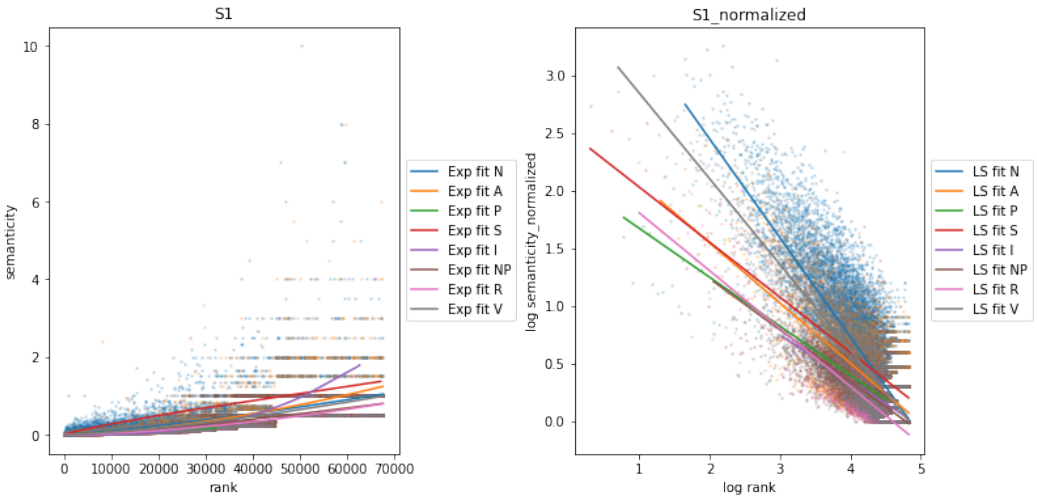


Figure 1: Semanticity versus rank with fitting functions in CTILC Corpus.

After setting the appropriate theoretical framework, we have explored the semanticity of Catalan words and discussed our findings with respect to other approaches in Linguistics and previous works [1, 2]. The findings for Catalan are encouraging for extending the study protocols applied here to other languages and for comparative studies that shed light on the inherent relationship between syntax and semantics in communication systems.

Keywords: Semanticity, Corpus linguistics, Zipf's laws of meaning, Computational linguistics, Catalan, Semantic Network

---

Funding: This work has been funded by the project PRO2023-S03 HERNANDEZ from Secció de Ciències i Tecnologia, Institut d'Estudis Catalans (https://www.iec.cat/).

References

[1] N. Català, J. Baixeries, R. Ferrer-i Cancho, L. Padró, A. Hernández-Fernández, Zipf's laws of meaning in catalan, PLOS ONE 16 (12) (2021) 1–21. doi:10.1371/journal.pone.0260849.

[2] A. Hernández-Fernández, I. G. Torre, J.-M. Garrido, L. Lacasa, Linguistic laws in speech: the case of catalan and spanish, Entropy 21 (12) (2019) 1153.

Distance as a relevance measure

Lars Johnsen
National Libary of Norway

The distance between words and phrases plays an important part in the analysis of text. Distance can be used as a relevance measure for a particular word when doing associative (collocation style) measures between words (e.g. Johnsen (2021)). Distance also enters indirectly when words are counted at certain positions, for example in word embedding models, where relevance is measured in terms of frequency within a window of a particular size (e.g. Mikolov et.al. (2013)). Using treebanks, distance between syntactic structures has come to light (e.g. Chen and Gerdes (2022)), in particular, distances between heads in a dependency structure are of interest. Such information can be used to reason backwards from unannotated text to annotation or analysis.

In this presentation I want to look closer at distance as a relevance measure in and of itself, in particular for word to word relevance. At the same time look at how the measure can be used for texts. Within a given text, words can be measured from any position in the text, and typically, high frequency words tend to be grouped in the middle, measured either from the beginning or the end, while content words position themselves closer or farther into the two halves (a process may of course be repeated in the smaller parts).
I will study this from two perspectives, one is the graph theoretical, and the other statistical.

As for graphs, words are related in a graph structure where they from nodes and each edge is marked with the (average) distance between them. This graph has certain properties or lack of, which may tell us something about the constellation of words. For example, if three words are transitively related, to what extent will they obey the triangle equality, and how will centrality play out?

Statistically, the problem is to find the properties of distance that are to be counted towards relevance. As alluded to above, one such measure is the difference between actual distance and expected distance. For instance, irrelevant words tend to cluster towards the middle of a search window. This happens for high frequency words like conjunctions and prepositions, while some high frequency words that are grammatically connected to the target word typically occur closer (e.g. Johnsen (2021)). However, words that are relevant may occur in the middle, and we suggest to use standard deviation as one of the measures, which directly combines counts and distance. The actual position relative to middle will be discussed.

References:

Chen, Xinying and Gerdes, Kim “Dependency Distances and Their Frequencies in Indo-European Languages”, in Journal of Quantitative Linguistics 1, vol 29, 2022.

Johnsen, Lars G. “Term Distance, Frequency and Collocations.” In Current Issues in Linguistic Theory, edited by Adam Pawłowski, Jan Mačutek, Sheila Embleton, and George Mikros, 356:22–36. Amsterdam: John Benjamins Publishing Company, 2021. https://doi.org/10.1075/cilt.356.02joh.

Mikolov, T., Yih, W. & Zweig, G. Linguistic regularities in continuous space word representations. In NAACL HLT, pp. 746–751, 2013b.

A dispersion measure that is by design orthogonal to frequency and its predictive power for lexical decision times

Stefan Th. Gries
UC Santa Barbara & JLU Giessen

For decades, frequencies of occurrence and co-occurrence have been among the most widely-reported corpus statistics. While the potentially important role of dispersion was recognized already in the 1970s (e.g., Carroll 1970 or Juilland et al. 1970), for some reasons dispersion measures unfortunately never made it into the corpus-linguistic mainstream. By now, a variety of different measures have been proposed (see Gries 2008 for an overview) and evaluated (see Biber et al. 2016, Burch et al. 2017) but one big downside remains for nearly all dispersion measures that have been proposed: While dispersion is cognitively and psycholinguistically not the same as frequency – the former corresponding to 'timing of exposure', the latter to 'amount of exposure' – the vast majority of dispersion measures is very highly correlated with frequency: In our own data, R^2-values for the correlation between frequency and dispersion were never below 0.6, but most often actually >0.9. Thus, one cannot help but wonder whether such measures, which virtually equate frequency and dispersion, do justice to dispersion as a measure that of course is supposed to complement frequency counts, yet also remain conceptually distinguishable and independently interpretable from frequency.

This paper addresses this problem and (i) develops a dispersion measure that is by definition uncorrelated with frequency and then (ii) tests its joint predictive power (with frequency) against a dozen other dispersion measures (with frequency) when it comes to predicting several 10,000 lexical decision times. Specifically, as for (i), we outline how the measure is computed and exemplify its calculation for, first, words of identical frequency but very different ranges in a corpus and then, second, for all word types in several widely-used corpora. The latter exemplification then also allows us to illuminate how the new dispersion measure exhibits much less of a correlation with frequencies and, correspondingly, makes an independent conceptual contribution to our understanding of word usage.

As for (ii), the empirical validation, we use random forests to determine how well the new dispersion measure performs when compared to many existing measures. Comparisons of PRE (proportional reduction of error) statistics show that the new measure (coupled with frequency) outperforms all other measures (also coupled with frequency) in its predictive power for decision times (in all these random forests, word length is also controlled for). We conclude with recommendations regarding how to apply this measure in the future and how to implement it best.

References

- Biber, D.; Reppen, R.; Schnur, E.; & Ghanem, R. 2016. On the (non)utility of Juilland's *D* to measure lexical dispersion in large corpora. *International Journal of Corpus Linguistics*, 21/4, 439–64.
- Burch, B.; Egbert, J.; & Biber, D. 2017. Measuring and interpreting lexical dispersion in corpus linguistics. *Journal of Research Design and Statistics in Linguistics and Communication Science*, 3/2, 189–216.
- Carroll, J. B. 1970. An alternative to Juilland's usage coefficient for lexical frequencies and a proposal for a standard frequency index. *Computer Studies in the Humanities and Verbal Behaviour*, 3/2, 61–65.
- Gries, St. Th. 2008. Dispersions and adjusted frequencies in corpora. *International Journal of Corpus Linguistics*, 13/4, 403–37.
- Juilland, A. G.; Brodin, D. R.; & Davidovitch, C. 1970. *Frequency dictionary of French words*. The Hague: Mouton de Gruyter.

How to define a word in Japanese?
Word segmentation in Japanese from the Zipf's law perspective

Xinying Chen, Ziyan Wei

Xi'an Jiaotong University, P.R. China

Keywords: Japanese, word segmentation, Zipf's Law

Abstract

Zipf's law refers to the fact that the rank-frequency distribution of words in different languages have a similar inverse relation. Theoretically, this law should also apply to Japanese words. However, the definition of word in Japanese is an unsolved problem with ongoing debates. In this study, we aim to address the problem from the perspective of Zipf's law. Different from the research of Yamazaki (2021) which investigated Zipfian distributions of common words in different Japanese texts, our goal is to compare two different word segmentation schema from the perspective of Zipf distribution. We analyze the rank-frequency distribution of lemmas in two Universal Dependencies (UD) treebanks that have the same texts but different word segmentations. Then, we compare the results to see which word segmentation scheme is more aligned with Zipf's law. We therefore can discover which word definition is closer to other languages and is consequently more suitable for cross-language studies.

Japanese words are difficult to segment because: 1) there is no space between words in written Japanese, unlike languages such as English or Czech; 2) the concept of 'word' does not exist in traditional Japanese linguistics (cf. Pringle 2016, Murawaki 2019). As a morphologically rich, agglutinating language, Japanese contains many agglutinative elements. These elements can be suffixes, clitics, or some forms expressing tones, etc. They are often difficult to map directly to individual grammatical categories and it is hard to organize them into a closed system. Therefore, Japanese linguists usually analyze texts based on 'Bunsetsu' (a linguistic unit that contains a basic lexeme and all attached agglutinating elements, kind of phrasal unit). The recent discussions regarding how to define and segment Japanese words are motivated by cross-language studies. Putting aside theoretical controversies, National Institute for Japanese Language and Linguistics (NINJAL) created two operational word units, namely SUW (short unit word) and LUW (long unit word). These two word segmentations are applied in UD Japanese treebanks (Tanaka et al. 2016).

The dataset of this research comes from the Japanese GSD corpus, which consists of 8100 Japanese sentences in Wikipedia (Asahara et al. 2018). In UD, there are two GSD treebanks built on the same texts, namely UD-GSD (with SUW segmentation) and UD-GSDLUW (with LUW segmentation). We first divide each treebank into 15 samples with 540 randomly selected sentences (sampling without replacement). For each sample, we fit the model $y = a\,x^{-b}$ to its rank-frequency distribution of lemmas (since Japanese is a morphologically rich language). Then we conduct paired-samples t-tests to compare the empirical parameters (a, b) of the power law function $y = a\,x^{-b}$ and the coefficient of determination R^2.

The results show that the SUW distributions and LUW distributions are significantly different across all three measurements (with $p < 0.05$). The SUW distributions are more Zipfian thanks to their higher R^2

values. Therefore, SUW seems to be more comparable to ‘word’ in other languages. This finding could be used especially in cross-language studies where the compatibility of units is crucial for comparisons.

References

Asahara, M., Kanayama, H., Tanaka, T., Miyao, Y., Uematsu, S., Mori, S., ... & Murawaki, Y. (2018, May). Universal dependencies version 2 for Japanese. In *Proceedings of the eleventh international conference on language resources and evaluation* (lrec 2018). 1824-1831.

Murawaki, Y. (2019). On the definition of Japanese word. *arXiv preprint arXiv*:1906.09719.

Pringle, G. (2016). Thoughts on the Universal Dependencies proposal for Japanese: The problem of the word as a linguistic unit. Accessed: 2023-01-28. *http://www.cjvlang.com/Spicks/udjapanese.html#otheragglutinative*.

Tanaka, T., Miyao, Y., Asahara, M., Uematsu, S., Kanayama, H., Mori, S., & Matsumoto, Y. (2016). Universal Dependencies for Japanese. In *Proceedings of the tenth international conference on language resources and evaluation (LREC’16)*. 1651-1658.

Yamazaki, M. (2021). Distribution and characteristics of commonly used words across different texts in Japanese. In Pawłowski, A., Mačutek, J., Embleton, S. and Mikros,G.(eds.) *Language and Text: Data, models, information and applications.* John Benjamins.121-134.

Colligation Diversity in Chinese Grammaticalization: An Entropy-based Approach

Biyan Yu[1] Lu Fan[2]
(1. Shool of Foreign Studies, Chang'an University, Xi'an, China
2. Shool of Foreign Studies, Xi'an Jiaotong University, Xi'an, China)

Colligation diversity, which relates to the possible parts of speech that may come before or after a word, reflects the grammaticalization degree of a word. With increasing grammaticalization, an item becomes less autonomous (Meillet 1912), and gradually changes from an independent juxtaposition to an affix or even phonological feature of carrier (Lehmann 2002), whereby it tends to be more attached to other signs. Grammaticalized items are thus suggested to have higher constraints regarding the part of speech they may co-occur, and show less colligation diversity. Nevertheless, it has also been suggested that grammaticalization should be associated with generalization of meaning, which means loss of lexical specificity, resulting in loss of some collocational and other restrictions and hence expanded use (Traugott & Trousdale 2013). In this sense, grammatical items should have more colligation diversity. The seemingly paradoxical relationship between colligation diversity and grammaticalization degree deserves further study.

The measurement of colligation diversity has been explored in the grammaticalization literature, but not much. Previous studies measured colligation diversity of a word by calculating the proportion of the most frequent part of speech it could co-occur (Saavedra 2021; Sun & Saavedra 2020). This measurement actually reflected the constraint of colligation rather than diversity, and in some cases it could not effectively distinguish some words.

This paper aims to explore the differences of colligation diversity between content (lexical) words and function (grammatical) words in Chinese by introducing the notion of entropy. Based on Lancaster Chinese Corpus (LCMC), this study selected 110 modern Chinese function words and 110 content words. For each word, the colligation entropy was obtained as an indicator of its colligation diversity, using the window of 1L (i.e. one word to the left of the node) and 1R ((i.e. one word to the right of the node), marked respectively as ColliDiv1L and ColliDiv1R. It was found that on the whole, ColliDiv1L was slightly higher than ColliDiv1R, which reflected that no matter content words or function words, more parts of speech could appear before a word. Besides, both ColliDiv1L and ColliDiv1R of function words were higher than that of content words, revealing that function words could co-occur with more parts of speech than content words. Be that as it may, there was no significant difference, neither in ColliDiv1L nor ColliDiv1R, between the content words and function words in modern Chinese.

The measure of entropy reflects more diversity information since it takes into consideration different colligation types, their frequencies, and their distribution (Liu et al. 2022). A higher entropy corresponds to more colligation diversity, meaning not

only that there are more possible parts of speech for a word to co-occur with, but also that those colligations are much more equally diversified instead of an absolute dominant colligation.

Quantitative approaches of grammaticalization can enrich our ways to think about the transition from more lexical to more grammatical and bring new inspiration to the traditional grammaticalization research.

Key word: colligation diversity; grammaticalization; entropy

References

Lehmann C. Thoughts on Grammaticalization[M]. Arbeitspapiere des Seminars für Sprachwissenschaft der Universität Erfurt 9. 2002:146.

Liu K., Liu Z., Lei L. Simplification in translated Chinese: An entropy-based approach[J]. Lingua, 275. 2022.

Meillet A. L'évolution des formes grammaticales[J]. Scientia (Rivista di Scienza), 1912,12(6):384–400.

Saavedra D. C. Measurements of grammaticalization : developing a quantitative index for the study of grammatical change[M]. Berlin:Boston. 2021.

Sun L., Saavedra D. C. Measuring grammatical status in Chinese through quantitative corpus analysis[J]. Corpora, 2020,15(3):317–342.

Traugott E C. & Trousdale G. Constructionalization and Constructional. Changes. Oxford: OUP. 2013.

A Quantitative Study of Noun Phrase Length in English and Chinese

Hua Wang

Phrase is a fundamental component and an essential unit in language, among which noun phrase is of vital importance and usually loaded with quantities of information. However, literature concerning the quantitative aspects on noun phrase is few and far between. In the present study, based on the Penn Chinese Treebank 9.0 (CTB 9.0) and the British Component of International Corpus of English(ICE-GB), we investigate whether Zipf-Alekseeve function, a function that can be used to describe the length distribution of syllable, word, sentence and some other linguistic units, can be used to describe the length distribution of noun phrase and also focus on the quantitative aspects of noun phrase length of different genres in English and Chinese. Results show that the length distribution of noun phrase in Chinese and English can both be described with Zipf-Alekseeve function, which for the first time verifies and extends the applicability of this function to the level of phrase, and further shows that the background mechanism of all the length distribution of linguistic units might be the same. Beyond that, a mirror-like and linear dependent relationship is found between parameter *a* and *b* in Zipf-Alekseev function; the values of parameter *a* fluctuate among different genres no matter in Chinese or in English, while the values of *b* can reflect the difference between Chinese and English. In addition, genre also has a significant influence on the mean length of noun phrase. To be specific, the more informal and colloquial the texts are, the shorter the mean length of noun phrase is. In addition, the study decides the range of noun phrase length and this range is then verified through more texts.

Keywords: noun phrase; length; Zipf-Alekseeve function; genre

References

Berlage, E. 2014. *Noun Phrase Complexity in English*. Cambridge: Cambridge University Press.

Fan, F. & Deng, Y. 2010. *Quantitative Linguistic Computing with Perl*. Lüdenscheid: RAM-Verlag.

Köhler, R. & Altmann, G. 2000. Probability distributions of syntactic units and properties. *Journal of Quantitative Linguistics*, 7(3): 189-200.

Liang, J. & Liu, H. 2013. Noun distribution in natural languages. *Poznań Studies in Contemporary Linguistics,* 49(4): 509–529.

Popescu, I.-I., Best, K.-H. & Altmann, G. 2014. *Unified Modeling of Length in Language*. Lü- denscheid: RAM-Verlag.

Wang, H. 2012. Length and complexity of NPs in written English. G*lottometrics,* 24:79-87.

Wimmer, G., Köhler, R., Grotjahn. R. & Altmann, G. 1994. Towards a theory of word length distribution. *Journal of Quantitative Linguistics*, 1: 98-106.

Authors affiliations: School of Foreign Languages, Dalian University of Technology, Dalian, China

E-mail: wanghuazju@163.com

Distributional properties of linear dependency segments

Nogolová, M., Mačutek, J., Čech, R.

Keywords: *linear dependency segment, dependency syntax, rank-frequency distribution, length distribution*

Abstract

The aim of the presentation is to analyse distributional properties of a linear dependency segment (LDS; Mačutek et al. 2021). A LDS is a recently proposed linguistic unit that hierarchically lies between word and clause. Its determination reflects both clause word order and syntactic dependency structure. Specifically, it is defined as “the longest possible sequence of words belonging to the same clause in which all linear neighbours (i.e., words adjacent in a sentence) are also syntactic neighbours (i.e., they are connected by an edge in the syntactic dependency tree which represents the sentence)” (Mačutek et al. 2021). The motivation for its introduction arose from the Menzerath-Altmann law (MAL) studies. The MAL reflects the relationship between a language construct and its constituents, specifically, the longer the construct, the shorter the constituent on average. On a syntactic level, different units were considered as constituents of clauses (e. g. words and phrases), however, the results are not conclusive. The analysis of MAL based on LDS (Mačutek et al. 2021) shows that the longer the clause in LDS, the shorter the LDS in words on average. We hypothesize that the rank-frequency distribution and the distribution of LDS length will exhibit properties similar to other linguistic units. Indeed, the rank-frequency distribution of LDS can be modelled by a power law, while the hyper-Poisson distribution can be used as a model for LDS length frequencies.

Acknowledgement

J. Mačutek was supported by research projects VEGA 2/0096/21 and APVV-21-0216.

References

Mačutek, J., Čech, R., Courtin, M. (2021). The Menzerath-Altmann law in syntactic structure revisited. In: *Proceedings of the Second Workshop on Quantitative Syntax*, Association for Computational Linguistics, pp. 65–73.

The distribution of syntactic dependency distances

Sonia Petrini[a] and Ramon Ferrer-i-Cancho[a]

[a] Quantitative, Mathematical and Computational Linguistics Research Group. Departament de Ciències de la Computació, Universitat Politècnica de Catalunya (UPC), Barcelona, Catalonia, Spain.

The syntactic structure of a sentence can be represented as a graph where vertices are words and edges indicate syntactic dependencies between them. In this setting, the distance between two syntactically linked words can be defined as the absolute value of the difference between their positions. These distances obey the well-known principle of Dependency Distance minimization, namely they are shorter than expected by chance (Ferrer-i-Cancho, 2004; Ferrer-i Cancho et al., 2022; Futrell et al., 2015; Liu, 2008), and, consistently, the shape of their probability distribution has been described with either an exponential (Ferrer-i-Cancho, 2004) or a power-law (Liu, 2007) curve in previous research. Here we aim to contribute to a further characterization of the actual distribution of syntactic dependency distances, and unveil its relationship with short-term memory limitations.

We propose a new double-exponential model in which speed of decay in probability is allowed to change after a break-point distance. The model is motivated both empirically and theoretically. On one hand, it builds on the sudden decrease in speed of decay observed in real sentences after a certain distance (Ferrer-i-Cancho, 2004), apparently in contrast with the minimization principle itself. On the other hand, it is consistent with the chunking model for language processing (Christiansen and Chater, 2016), where the change in speed could mirror the transition from the processing of individual words to higher-level structures. Long dependencies are in fact a burden for memory, and grouping linguistic units yields a new graph in which nodes are chunks, and thus distances between related elements are shorter, easing the production/understanding process of a written sentence.

Two hypotheses are tested, first that syntactic dependency distances are distributed following two exponential regimes, second that the break-point between the regimes is stable across languages. The latter would provide further support for the connection between the two-regime model and the structure of memory. Finally, we also give an account of the relation between the best estimated model and the closeness of syntactic dependencies, as measured by a recently introduced optimality score (Ferrer-i Cancho et al., 2022). The analysis is carried out on a parallel corpus of 20 languages, from different families and writing systems. Among the alternative distributions included for model selection, we consider a double-regime model where the first part of the curve is distributed following a power-law, building on what has been found in Chinese and other languages (Haitao, 2016; Liu, 2007).

A two-regime model – where the first regime follows either an exponential or a power-law decay – is found to be the most likely one in all 20 languages we considered, independently of sentence length and annotation style (Petrini and Ferrer-i Cancho, 2022). Moreover, the break-point is fairly stable across

languages and averages values of 4-5 words, suggesting that the amount of words that can be simultaneously processed abstracts from the specific language to a high degree. This work unveils the presence of a universal pattern in the languages included in the study, which could be linked to the functioning of short-term memory via the well-established chunking mechanism for language processing and production (Christiansen and Chater, 2016). Furthermore, it finds the common ground of previous results on the topic, providing an account of the distribution of syntactic dependency distances with respect to sentence length, text content, and annotation style.

References

Christiansen, M. H. and Chater, N. (2016). The now-or-never bottleneck: A fundamental constraint on language. *Behavioral and brain sciences*, 39.

Ferrer-i-Cancho, R. (2004). Euclidean distance between syntactically linked words. *Physical Review E*, 70:056135.

Ferrer-i Cancho, R., Gómez-Rodríguez, C., Esteban, J. L., and Alemany-Puig, L. (2022). Optimality of syntactic dependency distances. *Physical Review E*, 105(1):014308.

Futrell, R., Mahowald, K., and Gibson, E. (2015). Large-scale evidence of dependency length minimization in 37 languages. *Proceedings of the National Academy of Sciences*, 112:10336 – 10341.

Haitao, L. Q. L. (2016). Does dependency distance distribute regularly. *Journal of Zhejiang University (Humanities and Social Science)*, 2(4):63–76.

Liu, H. (2007). Probability distribution of dependency distance. *Glottometrics*, 15:1–12.

Liu, H. (2008). Dependency distance as a metric of language comprehension difficulty. *Journal of Cognitive Science*, 9:159–191.

Petrini, S. and Ferrer-i Cancho, R. (2022). The distribution of syntactic dependency distances. *https://arxiv.org/abs/2211.14620*.

Probability Distribution of Dependency Distance in Translational language Based on a Treebank Transformed from a Bidirectional Parallel and Comparable Corpus

Lu Fan (School of Foreign Studies, Xi'an Jiaotong University, Xi'an, China)
Biyan Yu (School of Foreign Studies, Chang'an University, Xi'an, China)

Abstract

As a constrained language, translational language attracts considerable attention from linguistic researchers. It arises from the process that coded elements are rendered into other codes, referred as the "third code", "third language", "hybrid language"and "constrained language". In an attempt to contribute to this expanding field of research, this study examines the distinctive features of translational language by adopting the quantitative linguistic approach to investigate the probability distribution of the individual dependency distance (DD) and that of typical dependency type in translational language. The probability distributions were tested in a bidirectional parallel and comparable corpus where both intralanguage and interlanguage comparisons were conducted. The results show that: (1) macroscopically, the distributions of both DD and the dependency type *nsubj* follow Right truncated modified Zipf-Alekseev distribution, one of the power-law distribution, in either native or translational language, indicating that this feature of translational language is part of language universal rather than translation universal; (2) however, as a language variation, translational language presents disparity in microscopic parameters *a* and *b* in distributions comparing source and native language. These findings suggest that, on the one hand, translational language exhibits the phenomenon of minimizing dependency distance as much as native language, reflecting a universal trend toward reducing cognitive load and following "the least effort principle" during human language production; and on the other hand, due to the simultaneous activation of the source language and the target language systems in the brain during the translating process, translational language is subject to the gravitational pull from both the source and the target language systems, demonstrating "compromise features". Additionally, translational language exhibits unique features that do not vary with translation direction, for instance, the larger parameter *a* and smaller parameter *b* in translational language while fitting the distribution of the DD of dependency type *nsubj*, which can possibly be regarded as universal features of translation. This study provides new insights into the research of translational language and illustrates the power of syntactic quantitative methods in translation studies.

Keywords: Probability Distribution; Dependency Distance; Translational Language

References

Baker, M. (1996). Corpus-based translation studies: the challenges that lie ahead. In H. Somers (Ed.), *Terminology, LSP and Translation: studies in language engineering: in honour of Juan C. Sager* (Vol. 18, pp. 175-186). Amsterdam and Philadelphia: John Benjamins.

Chen, X., & Gerdes, K. (2020). Dependency Distances and Their Frequencies in Indo-European Language. *Journal of Quantitative Linguistics*(1), 1-19.

Duff, A. (1981). *The Third Language: recurrent problems of translation into English: it ain't what you do, it's the way you do it*. Oxford: Pergamon Press.

Fan, L., & Jiang, Y. (2019). Can dependency distance and direction be used to differentiate translational language from native language? *Lingua, 224*, 51-59. doi:10.1016/j.lingua.2019.03.004

Fan, L., & Jiang, Y. (2020). A syntactic dependency network approach to the study of translational language. *Digital Scholarship in the Humanities, 36*(3), 595-606. doi:10.1093/llc/fqaa030

Ferrer-i-Cancho, R. (2015). The placement of the head that minimizes online memory: a complex

systems approach. *Language Dynamics and Change, 5*(1), 114-137. doi:10.1163/22105832-00501007

Futrell, R., Mahowald, K., & Gibson, E. (2015). Large-scale evidence of dependency length minimization in 37 languages. *Proceedings of the National Academy of Sciences, 112*(33), 10336–10341. doi:10.1073/pnas.1502134112

Heringer, H. J., Strecker, B., & Wimmer, R. (1980). *Syntax Fragen-Lösungen Alternativen*. München: Wilhelm Fink Verlag.

House, J. (2008). Beyond intervention: universals in translation? *trans-kom, 1*(1), 6-19.

Hřebíček, L. (1996). *Word associations and text*. Trier: Wissenschaftlicher Verlag Trier.

Hudson, R. (1995). *Measuring Syntactic Difficulty*. http://dickhudson.com/wp-content/uploads/2013/07/Difficulty.pdf.

Hudson, R. (2010). *An Introduction to Word Grammar*. Cambridge: Cambridge University Press.

Jiang, J., & Liu, H. (2015). The effects of sentence length on dependency distance, dependency direction and the implications-based on a parallel English-Chinese dependency treebank. *Language Sciences, 50*, 93-104. doi:10.1016/j.langsci.2015.04.002

Jiang, J., & Liu, H. (2018). *Quantitative Analysis of Dependency Structures*. Berlin/Boston: De Gruyter.

Kruger, H., & Rooy, B. V. (2016). Constrained language: A multidimensional analysis of translated English and a non-native indigenised variety of English. *English World-Wide, 37*(1), 26-57.

Liu, H. (2009). Probability Distribution of Dependencies Based on a Chinese Dependency Treebank. *Journal of Quantitative Linguistics, 16*(3), 256-273.

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. doi:10.1016/j.plrev.2017.03.002

Narisong, Jiang, J., & Liu, H. (2014). Word Length Distribution in Mongolian. *Journal of Quantitative Linguistics, 21*(2), 123-152. doi:10.1080/09296174.2014.882191

Nivre, J. (2006). *Inductive Dependency Parsing*. Dordrecht: Springer.

Ouyang, J., & Jiang, J. (2017). Can the probability distribution of dependency distance measure language proficiency of second language learners? *Journal of Quantitative Linguistics, 25*(4), 1-19. doi:10.1080/09296174.2017.1373991

Pande, H., & Dhami, H. S. (2012). Model generation for word length frequencies in texts with the application of Zipf's order approach. *Journal of Quantitative Linguistics, 19*(4), 249–261.

Popescu, I. I., Best, K. H., & Altmann, G. (2014). *Unified modeling of length in language (Studies in quantitative linguistics 16)*. Lüdenscheid: RAM-Verlag.

Pustet, R., & Altmann, G. (2005). Morpheme length distribution in Lakota. *Journal of Quantitative Linguistics, 12*(1), 53-63.

Tesnière, L. (1959). *Éléments de syntaxe structurale*. Paris: Klincksieck.

Wang, Y., & Liu, H. (2017). The effects of genre on dependency distance and dependency direction. *Language Sciences, 59*, 135-147.

Wang, Y., & Yan, J. (2018). A Quantitative Analysis on a Literary Genre Essay's Syntactic Features. In J. Jiang & H. Liu (Eds.), *Quantitative analysis of dependency structures* (pp. 295-314). Berlin/Boston: de Gruyter.

Zipf, G. K. (1936). *The psychobiology of language*. London: Routledge.

Zipf, G. K. (1949). *Human behavior and the principle of least effort: an introduction to human ecology*. Cambridge, MA: Addison-Wesley Press.

A Comparative Analysis of Authorship Attribution in a Creole and Non-Creole Language

Patrick Juola
Duquesne University, USA
juola@mathcs.duq.edu

Alejandro J. Napolitano Jawerbaum
Duquesne University, USA
napolitanojawea@duq.edu

Keywords
Stylometry, authorship attribution, creole linguistics, individual variation.

Abstract

Stylometry (Ainsworth and Juola 2018), the computational study of writing style, has proven itself to be a practical method of answering questions of authorship in a wide variety of languages. However, previous research has focused on non-creole languages. Creole languages are "new languages that develop out of a need for communication among people who do not share a common language" (Siegel 2008). In contrast to older languages, they are often considered to be unusual, in having less complexity and stylistic variation (Siegel 2004). However, this view is somewhat controversial, in part due to a lack of formal quantitative studies (Robinson, 2008).

The theory of authorship analysis (Coulthard, 2004) states that authors can be distinguished by "cumulatively unique rule-governed choices" among the available options in the language. With fewer stylistic options, are the opportunities for choice reduced? And is this reduced choice reflected in the performance of standard stylometric methods?

*Juola and Napolitano Jawerbaum (2022) presented an analysis of 100 parliamentary speeches in Seychellois Creole (*seselwa *or* kreol*, a French-based creole spoken in the Seychelles islands) by ten different speakers and showed that five commonly used stylistic analyses were able to correctly identify the speaker with performance far above chance. We extend this analysis with a comparable corpus of standard (Québécois) French. As with the previous study, we collected ten samples each by ten different speakers in the Québec National Assembly. These samples ranged in size from 737 to 7747 words (compare to the Seychellois corpus' 814-8079 words), and are in a nearly identical genre. With ten authors each, the problem of authorship attribution is formally identical between the two corpora, thus providing us with an opportunity to determine if authorship attribution is more or less difficult in a creole than in its related acrolect language.*

These documents were analyzed for authorship using the JGAAP software package and five typical feature sets. With ten authors, random guessing would be expected to achieve 10 successes (10%), with 16 or more successes indicating ``significant'' ($p < 0.05$) results. The table below shows results for both seselwa and French.

Feature Set Studied	***Seselwa Docs Correct***	***French Docs Correct***
50 most frequent words	*57*	*57*
Character 10-grams	*62*	*72*

Character 4-grams	*74*	*80*
Word 2-grams	*79*	*72*
(Histogram of) word lengths	*33*	*31*

In both languages, all five feature sets outperformed chance, but there is no clear pattern showing that the authorship attribution task is more difficult in either language. Of the five trials, French outscored seselwa twice, underscored twice, and tied once. While it is clear that word lengths performed relatively poorly compared to the other feature sets, none of the interlingual differences appear significant, an observation confirmed by a z-test for population proportions.

We therefore conclude that, potential typological differences between creole and non-creole languages aside, stylometric analysis is no more difficult (or easy, for that matter) in creole languages. We further conclude, pending further replication and exploration in other language sets, that there is, in fact, no less complexity and stylistic variation in creole languages.

Bibliography

Ainsworth, Janet, and Patrick Juola. "Who wrote this: Modern forensic authorship analysis as a model for valid forensic science." Wash. UL Rev. 96 (2018): 1159.

Coulthard, Malcolm. "Author identification, idiolect, and linguistic uniqueness." Applied linguistics 25, no. 4 (2004): 431-447.

Juola, Patrick and Napolitano Jawerbaum, Alexander J. "Stylometric Authorship Attribution in Seychellois Creole." DH_BUDAPEST_2022, Budapest, Hungary. (2022.)

Robinson, Stuart. "Why pidgin and creole linguistics needs the statistician: Vocabulary size in a Tok Pisin corpus." Journal of Pidgin and Creole Languages 23, no. 1 (2008): 141-146.

Siegel, Jeff. "Morphological elaboration." Journal of Pidgin and Creole Languages 19, no. 2 (2004):333–362.

Siegel, Jeff. The emergence of pidgin and creole languages. Oxford University Press, 2008.

Can stylometry reveal more than a human reader in a text? A study based on Romain Gary and Emile Ajar's case.

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

Keywords:
stylometry, Z-Score, TF-IDF, Pointwise Mutual Information, taxonomy, Romain Gary, Emile Ajar

The essence of stylometry, as conceived by its founder Wincenty Lutosławski, is the belief that an author, regardless of his age, the style of his writing, and the subject matter, is unable to change certain characteristics of his texts. In his explanations Lutosławski used the analogy of style to handwriting, reminding that signature was used to identify customers in legal or banking systems. Modern forensic science would claim that certain biometric and psychometric traits remain unchanged in mature individuals throughout their lives. Of course, the most glaring examples of this invariability are fingerprints, the spectrum of voice and DNA. However, this is not relevant for culture, since humans can create extremely different outputs regardless of their genotype. In literature, using various tools of stylistics, specific vocabulary, syntactic patterns, etc. an author can create different "selves" that are not part of his identity.

The example of Romain Gary proves that readers can be easily led astray by creating new fictional identities. The new authorial personality produced by this prominent writer proved to readers and critics – even very studious and competent ones – not only attractive, but above all dissimilar to other, well-known, writers. Gary published 4 novels under the pseudonym Ajar, which were very successful in the francophone world. Proof of this is the Goncourt Prize, awarded in 1975 for the novel *La vie devant soi*, as well as its 1977 film adaptation, which won the French César and the Oscar for non-English language film.

So, for stylometry, the situation here is close to ideal: one and the same person is writing but appears in the media as two different authors. Not only a massive audience of readers and cinephiles, but also professional critics believed that Ajar was an authentic persona – a young debuting writer – and not the deliberate creation of an experienced and seasoned author of more than twenty novels. Gary went so far as to substitute the person of his cousin, Paul Pavlovitch, as the "real" Ajar to the media. No one pointed out that in addition to their kinship, both pseudonyms (Gary's real name was Kacev) are phonetic transcriptions of words referring to fire and embers in Polish or Russian – it can be simplistically said that they are variants of the same name. Could this be a coincidence? Nevertheless, the revelation of the true identity of the author of *La vie devant soi* and *Gros-câlin* came as a great surprise.

The working hypothesis that clearly arises in this context is the following: the analysis of the corpus of texts signed Gary and Ajar, carried out by advanced methods of stylometry that do not refer to the most visible stylistic or content layer, should prove that both sets of texts are authored by one and the same person. Thus, referring to the title, we assume that although stylometry has no sensitivity to aesthetic features, it can recognize the authorship of texts better than a human being by comparing author's stylistic fingerprint in the same manner as one compares, for example, the DNA of seemingly different individuals.

The subject of analysis will be samples taken from all of Gary's texts written in French, as well as a comparative collection made up of samples of selected novels by Michel Tournier,

Louis Aragon and Raymond Queneau. These authors were suggested by the French press as potential authors of texts signed Ajar. In the study we will apply methods based on statistical definition of style. Consequently lexemes will be represented numerically as vectors. These vectors will be then weighted and on this basis a measure of stylistic similarity of the analyzed samples will be determined. The following methods of determining stylistic features will be used: the most frequent words in the corpus, selected function words, punctuation, parts of speech and their n-grams. The corpus will be explored using the following weighting methods: Z-Score, TF-IDF, Pointwise Mutual Information (PMI), and simplified PMI. Out of the known similarity/distance measures, we will apply cosine similarity, Euclidean distance, Burrows delta, Eder distance, and Jacquard similarity.

References

Labbé Dominique (2008), *Romain Gary et Emile Ajar*. hal-00279663

Pawłowski Adam (1998), *Séries temporelles en linguistique. Avec application à l'attribution de textes: Romain Gary et Émile Ajar*. Paris, Genève: Champion-Slatkine.

Savoy Jacques (2020), *Machine Learning Methods for Stylometry. Authorship Attribution and Author Profiling*. Cham: Springer Nature Switzerland.

Tirvengadum Vina (1996), *Linguistic Fingerprints and Literary Fraud*. Digital Studies/le Champ Numérique, (6). DOI: http://doi.org/10.16995/dscn.187

French Plays of the 17th Century: A Stylometric Analysis

Jacques Savoy

Computer Science Dept., University of Neuchatel
rue Emile Argand 11, 2000 Neuchatel (Switzerland)
Jacques.Savoy@unine.ch

ABSTRACT

PROBLEM. French plays written during the 17th century present an interesting authorship attribution problem, due in part by the lack of rights protecting the true author. Moreover, some plays could have been authored by two (or more) writers such as with *Psyché* (1671) written by Corneille, Molière and Quinault. A similar practice appears in UK (e.g., *The Two Noble Kinsmen* (1634) authored by Shakespeare and Fletcher). When a work is published under a single name, the large majority of French scholars assume that the author's name appearing in the front page is the true author. We aim to verify this so-called signature hypothesis based on an authorship attribution procedure with 20 plays written by nine possible authors (J.G. de Campistron (1656–1723), Champméslé (1642–1701), Chevalier (16..–1673), Hauteroche (1617–1707), Montfleury (1608–1667), P. Quinault (1635–1688), J. Racine (1639–1699), Tristan (1601–1655), and T. Corneille (1625–1709)). This corpus contains works corresponding to the same text genre (comedy in verse) and each text contains more than 10,000 tokens (see Table 1).

METHODS. Using Labbé's intertextual distance (Labbé, 2007), one can estimate the stylistic distance between two plays based on their lemmas. When a distance is lower than 0.2, this outcome is strong evidence that both texts have been written by the same writer. This procedure was verified with an Italian corpus (150 novels, 40 authors) (Savoy, 2018) and a French corpus composed on 200 novels written by 30 authors (Kocher & Savoy, 2019).

FINDINGS. Analyzing the authorship of these 20 plays (open-class problem), we discovered strong stylistic similarities between 18 of those plays, an outcome in contradiction with the signature hypothesis. Internal evidence cannot favor one name over the others. However, one can argue that small intertextual distance between two plays could be, in part, explained by two texts dealing with very similar topics (e.g., two lovers facing the opposition of the father, or a story with characters having false identities). With external consideration, one might suggest that one (or two) author(s) could be behind all those similar plays. To confirm these findings, the NSC method (Jockers, 2013) have been applied.

LIMITS. We assumed that each play was written by a single author, and thus ignoring possible co-authorship. We have worked play by play and we didn't consider that all plays published under a given name were authored by the same single (unknown) writer (e.g., like the Ferrate's question). In addition, due to the authorship debate between Molière (1622—1673) and P. Corneille (1606—1684), these two names have been discarded. Finally, a larger corpus must be built to confirm our findings.

Author	Title	Year	Tokens
Campistron	Le jaloux désabusé	1709	14383
Champméslé	Le parisien	1684	18091
Champméslé	Ragotin	1684	15596
Chevalier	L'intrigue des carrosses	1662	10154
Chevalier	Le pédagogue amoureux	1665	16977
Hauteroche	L'amant qui ne flatte point	1668	20908
Hauteroche	Crispin musicien	1671	22350
Montfleury	La femme juge et partie	1669	17464
Montfleury	Le comédien poète	1673	21771
Quinault	Les rivales	1653	18680
Quinault	La mère coquette	1665	19452
Racine	Les plaideurs	1668	10063
Tristan	Amaryllis	1652	16124
Tristan	Le parasite	1654	18701
T. Corneille	Dom Bertran de Cigarral	1651	20911
T. Corneille	L'amour à la mode	1651	20819
T. Corneille	Le galant doublé	1659	21152
T. Corneille	Le baron d'Albikrac	1667	20558
T. Corneille	La comtesse d'Orgueil	1670	21124
T. Corneille	Le festin de Pierre	1677	20068

Table 1: Selected plays present in our analysis

References

Burrows, J.F. (2002). Delta: A measure of stylistic difference and a guide to likely authorship. *Literary and Linguistic Computing*, 17(3), 267-287.

Jockers, M.L. (2013). Testing authorship in the personal writings of Joseph Smith using NSC classification. *Digital Scholarship in the Humanities*, 28(3), 371-381.

Labbé, D. (2007). Experiments on authorship attribution by intertextual distance in English. *Journal of Quantitative Linguistics*, 14(1), 33-80.

Kocher M., & Savoy J. (2019). Evaluation of text representation schemes and distance measures for authorship linking. *Digital Scholarship in the Humanities*, 34(1), 189-207.

Savoy J. (2018). Is Starnone really the author behind Ferrante? *Digital Scholarship in the Humanities*, 33(4), 902-918.

Savoy J. (2023). Stylometric analysis of characters in Shakespeare's plays. *Digital Scholarship in the Humanities,* to appear.

A gradient model of LDD acceptability

Aixiu An[1], **Yingqin Hu**[2], **Anne Abeillé**[2]

[1]Massachusetts Institute of Technology, [2]Université de Paris

Conditions on non-local dependencies (LDD) ("island constraints") (Ross, 1967) play a major role in syntactic theories and are meant to capture the intuitive difference between (1-a) and (1-b). Such conditions have started to get tested with human behavioral experiments (Sprouse & Hornstein 2013), along with neural network models (Wilcox et al 2018).

Three questions are of relevance for linguistic theories: i) are these constraints categorical? ii) do they apply similarly across languages? iii) do they apply similarly across different constructions within a language?

(1) a. ??[Which activities] did certain aspects of - endanger the employees' health?

b. [Which activities] did many employees fear certain aspects of - for their health?

We present a gradient acceptability model to evaluate different theories regarding island constraints, by collecting published experimental data testing different constructions (wh-questions, relative clauses, it-clefts), for two island types (subject and adjunct) across English and French. We rely on a penalized linear model of weighted violable binary constraints (An, 2020), in order to see: (i) which constraints play a role in each language (ii) what is the relative weight of each constraint. In this model, the well-formedness of a structure is the weighted sum of constraint violations.

$$\text{well-formedness} = w_0 + \sum_{i=1}^{m} w_i C_i(s_k)$$

The English dataset includes 35 conditions from 11 acceptability experiments (Abeillé et al. (2020), Gibson et al (2021), Chaves and Dery (2019), Phillips (2006)). The French dataset includes 31 conditionsfrom 13 experiments Abeillé et al. (2020), Pozniak (2018) and Paape (2017)). We used mean acceptability rating (with scale norming) of each condition as the response variable. All conditions were annotated with various constraints proposed in the literature (1 if they are violated, 0 otherwise):

- Typological/Processing constraints : *extr-no-subj*: only extract a subject, *dist-filler* (normalized distance between filler-gap) (Gibson et al. 2000's DLT: minimize distance between filler and gap), *embed* (preference for matrix gap)

- Syntactic constraints: *extr-adj-hd* (Adjunct island: no extraction out of adjunct), *extr-nom-hd* (Complex NP island: no extraction out of NP), *extr-subj-hd* (Subject island: no extraction out of subject) (Chomsky 1973; Huang 1982).

- Semantic constraints, *inter-rel* (no interfering element (same animacy) between filler and gap), *inter-syn* (no interfering NP between NP filler and gap), *no-pro-filler* (preference for nominal fillers (Pesetsky

2000) or complementizers).

- Discourse based constraints: *focal-domain* (no focal extraction out of a non-focal constituent) (Goldberg 2006)(Abeillé et al., 2020).

We train the models with (1) by adding elastic net regularizations using the package *glmnet* in R, and evaluated them with leave-one-out cross-validation. Figure1 shows the best model for each language. For English the subject island constraint plays a major role, followed by semantic (relative interference) and discourse (fconstraints; the adjunct island constraint does not play a role. For French, only the discourse, processing (*distance*) and semantic (*no-pro-filler*) constraints play a role.

These results cast doubt on a purely syntactic approach to LDD, and call for an interplay of semantic, pragmatic, and processing factors Liu et al. (2022).

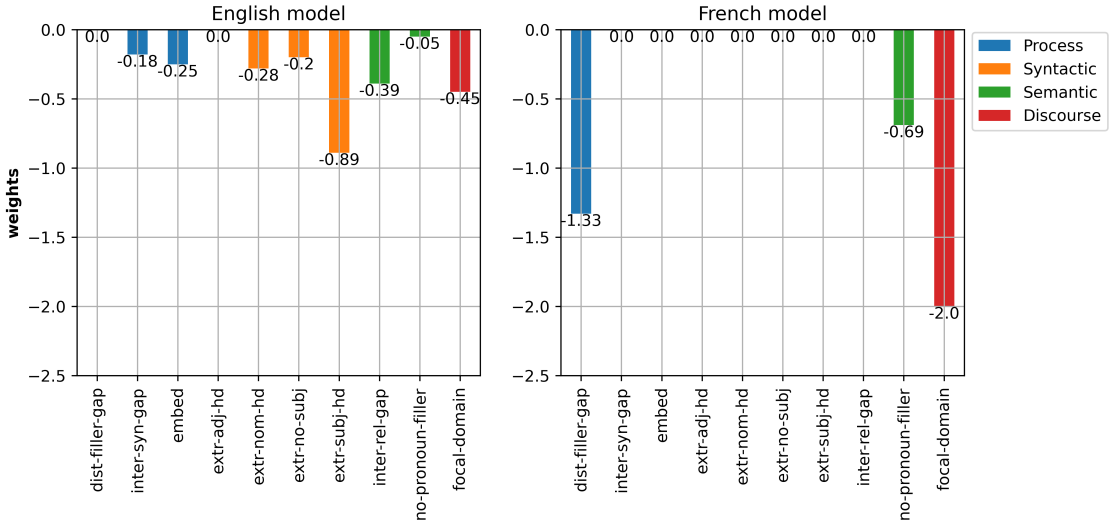


Figure 1: The best model for English (**left**), with the leave-one-out cross-validation error 0.69 and for French (**right**), leave-one-out cross-validation error is 1.43.

References

Abeillé, A., Hemforth, B., Winckel, E., and Gibson, E. (2020). Extraction from subjects: Differences in acceptability depend on the discourse function of the construction. *Cognition*, 204:104293.

An, A. (2020). *Theoretical, empirical and computational approaches to agreement with coordination structures*. PhD thesis, Université Paris Cité.

Chaves, R. P. and Dery, J. E. (2019). Frequency effects in subject islands. *Journal of linguistics*, 55(3):475–521.

Liu, Y., Winckel, E., Abeillé, A., Hemforth, B., and Gibson, E. (2022). Structural, functional, and processing perspectives on linguistic island effects. *Annual Review of Linguistics*, 8:495–525.

Phillips, C. (2006). The real-time status of island phenomena. *Language*, pages 795–823.

Ross, J. R. (1967). *Constraints on variables in syntax*. PhD thesis, Massachusetts Institute of Technology.

Syntactic boundaries or word-count distance? Co-reference configurations and the choice between finite and non-finite adnominal clauses in German

Felix Bildhauer*, Thilo Weber*, Franziska Münzberg†

January 27, 2023

Keywords— German syntax, co-reference, control, finiteness

In German, certain nouns (such as *Vorteil* ('advantage')) can take a subordinate clause that can be realised in finite form (as a *dass*-clause ('that'-clause)) (1a) or as a *zu*-infinitive ('to'-infinitive) (1b). The latter is considered the more restricted variant in that it is possible only where its implicit subject (often represented by PRO) is "controlled by" (co-referent with) an expression in the surrounding context or has arbitrary/generic reference.

(1) Solche Ausstellungen$_i$ hätten den **Vorteil**,
such exhibitions have.PAST.SUBJ the advantage
'Such exhibitions would have the advantage'
a. dass sie$_i$ Unternehmen zusammenbringen
that they companies bring.together.3RD.PL.PRES
'that they bring companies together'
b. Unternehmen zusammenzubringen
companies together.bring.INF
'of bringing companies together'
'Such exhibitions would have the advantage of bringing companies together.'

We investigate the distribution of the two variants based on samples drawn from the German Reference Corpus (Kupietz et al. 2010) and the German web corpus DECOW16B (Schäfer & Bildhauer 2012), thus covering both conceptually written registers (as typically found in newspaper texts) and less formal registers (as typically found in internet forums). Among other things, we examine Brandt's (2019, 289) suggestion that the quality of the control configuration plays a major role in the syntactic realisation of the subordinate clause (the better the quality, the more likely the clause is to be realised as an infinitive rather than a *that*-clause). Inter alia, this predicts that constructions in which the subject of the subordinate clause has a co-referent expression within the head-noun-NP are more likely to contain an infinitive than are constructions with a co-referent expression that occurs outside the head-noun-NP or even outside the clause of the head-noun-NP.

In a first step, cases were identified in which *that*-clauses and *to*-infinitives indeed appear to be interchangeable ('choice contexts' in the sense of Rosenbach 2013; inter-rater agreement κ_{Cohen} = 0.8). The resulting set of attestations (n=6,268), of which 651 are finite) was analysed in a mixed-effects logistic regression model considering not only the position of the co-referent expression but

*Leibniz-Institut für Deutsche Sprache, Mannheim, {bildhauer,weber}@ids-mannheim.de
†beredt, info@beredt.de

also a number of other factors hypothesised to govern the choice between the two variants, such as the modality, voice, and syntactic complexity of the subordinate clause.

Figure 1 shows the coefficient estimates. One of the results is that, as hypothesised, constructions with an NP-internal co-referent expression are most likely to contain an infinitive, followed by constructions with an NP-external (but still clause-internal) co-referent expression, in turn followed by constructions with a clause-external co-referent expression. However, as it turns out, the syntactic position of the co-referent expression is highly correlated with its distance (measured in number of words) to the subordinate clause subject, which raises the question of whether it is necessary at all to refer to syntactic boundaries (NP, clause). We therefore compare the original statistical model to an alternative one that operates with distance rather than syntactic boundaries. We show that the alternative model explains the variation equally well and we propose that an explanation in terms of distance can also plausibly be derived from accessibility theory (Arnold 2010).

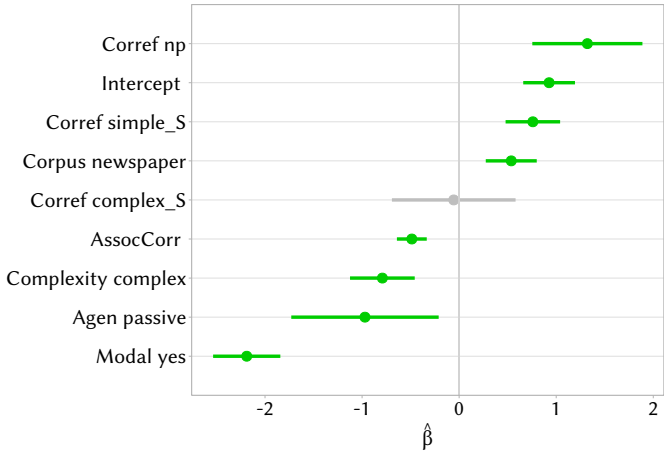


Figure 1: Point estimates and 95 % confidence intervals for a mixed-effects logistic regression model. Positive coefficients increase the probability of the 'to'-infinitive. Marginal R^2 (fixed effects only) = 0.25, conditional R^2 (entire model) = 0.36 (Nakagawa & Schielzeth 2013). Estimated group level error (*lemma*) $\hat{\sigma}_\alpha$ = 0.72

References

Arnold, Lawrence. (2010). How speakers refer: the role of accessibility. *Language and Linguistics Compass* 4(4). Publisher: De Gruyter, 187–203.

Brandt, Patrick. (2019). Alternation von *zu-* und *dass*-Komplementen: Kontrolle, Korpus und Grammatik. In Eric Fuß, Marek Konopka & Angelika Wöllstein (eds.), *Grammatik im Korpus: Korpuslinguistisch-statistische Analysen morphosyntaktischer Variationsphänomene*, 211–297. Tübingen: Francke Attempto.

Kupietz, Marc, Cyril Belica, Holger Keibel & Andreas Witt. (2010). The German reference corpus DeReKo: A primordial sample for linguistic research. In Nicoletta Calzolari, Khalid Choukri, Bente Maegaard, Joseph Mariani, Jan Odijk, Stelios Piperidis, Mike Rosner & Daniel Tapias (eds.), *Proceedings of the seventh international conference on language resources and evaluation (LREC '10)*, 1848–1854. Valletta, Malta: European Language Resources Association (ELRA).

Nakagawa, Shinichi & Holger Schielzeth. (2013). A general and simple method for obtaining R^2 from generalized linear mixed-effects models. *Methods in Ecology and Evolution* 4. 133–142.

Rosenbach, Anette. (2013). Combining elicitation data with corpus data. In Manfred Krug & Julia Schlüter (eds.), *Research methods in language variation and change*, 278–294. Cambridge, MA: Cambridge University Press.

Schäfer, Roland & Felix Bildhauer. (2012). Building large corpora from the web using a new efficient tool chain. In Nicoletta Calzolari, Khalid Choukri, Thierry Declerck, Mehmet Ugur Dogan, Bente Maegaard, Joseph Mariani, Jan Odijk & Stelios Piperidis (eds.), *Proceedings of the Eighth International Conference on Language Resources and Evaluation (LREC'12)*, 486–493. Istanbul: ELRA.

Author
Haruko Sanada

email
hsanada@ris.ac.jp

Correspondence address
Rissho University, Faculty of Economics
4-2-16, Osaki, Shinagawaku,
1418602 Tokyo, Japan

Title
The length and order of grammatical elements in the Japanese clause

Abstract
(416 words excluding keywords, authors affiliations and references)

Aim of the study:
The present study is one of a series of empirical studies of the Japanese clause. Our last study (Sanada 2021) investigated the preferred orders of "grammatical functions" (complements and adjuncts, hereafter "grammatical elements") in the clause, finding significant order preferences in the clause. The present study empirically verified a relationship between the length and order of grammatical elements, as only two papers (Miyajima 1964, Saeki 1975) briefly mentioned a relationship whereby grammatical elements are shorter the closer they appear to the end of the clause.

Data:
We employed the Japanese valency database (Ogino et al. 2003), from which 240 sentences were extracted, including 243 clauses containing the verb 'meet'. The clauses contain in total 766 grammatical elements (subject, object, time, place, occasion) and predicates, which are always located clause-finally. We number them by position starting from 1 at the start of the clause. The length of grammatical elements is measured as both the number of morphemes and the number of morae.

Hypotheses and methods of analyses:
We posited the following hypotheses: (1) The length of a grammatical element is shorter the closer the element is to the end of the clause; and (2) the average length of each type of grammatical function varies significantly because grammatical elements have preferred positions in the clause. We identified 971 pairs of grammatical elements and predicates, e.g., the average length of subjects and the average length of objects if the subject and object appeared in the same clause, as Japanese complements are omittable. We performed an unpaired two-sample *t*-test for unequal variances (Welch's *t*-test) on the pairs. For significant pairs, Cohen's *d* was also obtained as a measure of effect

size.

Results and conclusions:

We concluded that grammatical elements are shorter the closer they are to the end of the clause as follows:

1. Overall, the average length and positions of grammatical elements show a negative relationship.
2. Average lengths of occasion and predicate are significantly shorter than those of other grammatical elements.
3. The lengths of places appearing before versus after an object and before versus after an occasion are significantly different, and a place that appears before an object is shorter.

We conclude that hypothesis (1) holds and (2) partly holds. The results for place follow the conclusion of above two papers, while subjects and objects do not follow the length property. We thus consider there to be a grammatical property of subjects and objects that has a higher priority than the length property.

References:

Miyajima, Tatsuo. (1964). Joshi jodoshi no yoho (Usage of postpositions and auxiliary verbs). In: National Language Research Institute. (ed.) Gendai Zasshi 90shu no Yogo Yoji: Dai3bunsatsu: Bunseki (Vocabulary and Chinese Characters in Ninety Magazines of Today, vol. 3: Analysis of Results), pp. 69-239. Tokyo: Shuei Shuppan.

Ogino, Takano; Kobayashi, Masahiro; Isahara, Hitoshi. (2003). *Nihongo Doshi no Ketsugoka* (Verb valency in Japanese). Tokyo: Sanseido.

Saeki, Tetsuo. (1975). *Gendai Nihongo no Gojun* (Word order in modern Japanese). Tokyo: Kasama Shoin.

Sanada, Haruko. (2021). *N*-grams of valency types and their significant order in the clause. In: Pawłowski, Adam, Jan Mačutek, Sheila Embleton and George Mikros (eds.), *Language and Text. Data, Models, Information and Applications*, pp. 69-91. Amsterdam, The Netherlands: John Benjamins.

Keywords:

Sentence structure, linguistic length, word order, complement, adjunct, position in the clause, Japanese, Synergetic Linguistics.

Long Time No Joe: Piotrowski-Law Development of Personal Names in the Diachronic Perspective

Tereza Klemensová – Michal Místecký

Abstract: The study attempts to apply Piotrowski Law on the development of selected Czech personal first names in the 1920–2016 period. The Piotrowski-Law function is able to model changes in language, making use of three parameters – *a* (the beginning of the change), *b* (its sharpness), and *c* (the overall final frequency of the phenomenon); it thus allows to study the names from a diachronic perspective (cf. Altmann et al., 1983; Místecký et al., 2018). The frequencies of the names for particular years will be based on the data from the Ministry of the Interior; we will work with the absolute numbers of name bearers and analyse the five highest-scoring names for both sexes (as of 2016 – *Jiří*, *Jan*, *Petr*, *Josef*, and *Pavel* for men; *Jana*, *Marie*, *Eva*, *Hana*, and *Anna* for women; Malačka, 2011). The goal of the contribution is to test the approach against the onomastic material and to determine historical and contemporary trends among top Czech anthroponyms.

Keywords: Czechia; diachrony; modelling; onomastics; personal names; Piotrowski Law

References:

Altmann, Gabriel – von Buttlar, Haro – Rott, Walter – Strauß, Udo (1983). A law of change in language. In: Brainerd, B. (ed.). *Historical Linguistics*. Bochum: Brockmeyer, 104–115.

Malačka, Ondřej (2011). *KdeJsme.cz* [online; accessed 20 January 2023]. Available at: <https://www.kdejsme.cz/>.

Místecký, Michal – Andreev, Sergey – Altmann, Gabriel (2018). Piotrowski Law in Sequences of Activity and Attributiveness: A Four-Language Survey. *Glottometrics*, 42, 21–38.

Tereza Klemensová
Department of Czech Language, Faculty of Arts, University of Ostrava
Reální 3, 701 03 Ostrava
tereza.klemensova@osu.cz

Michal Místecký
Department of Czech Language, Faculty of Arts, University of Ostrava
Reální 3, 701 03 Ostrava
michal.mistecky@osu.cz

A Zipf-Mandelbrot Approach to Diachronic Productivity

Quentin FELTGEN – Ghent University (Belgium)

Keywords: Zipf-Mandelbrot distribution, productivity, Construction Grammar, diachrony, sampling

Productivity is a key feature of schematic constructions (that is, constructions that offer a free, fillable slot), insofar as it quantifies the openness of their schema, the versatility of their use, and the scope of their meaning (Hilpert 2013). Several empirical measures of productivity have been offered, ranging from those defined by Baayen (2009), to measures of diversity inspired from Ecology (Sundquist 2020). However, the study of productivity in a diachronic perspective (Barðdal & Gildea 2015, Enghels 2021), has proved especially challenging. The traditional measures suffer from a lack of comparability across decades, notably due to the different number of tokens (Gaeta & Ricca 2003). To alleviate this, the tracking of new types (Baayen & Renouf 1996, Cowie & Dalton-Puffer 2002, Bergs 2020) and projecting the fillers in a constructed semantic space (Perek 2014, 2018, Desagulier 2022), have been explored.

In this contribution, I focus on the relationship between number of types and number of tokens as embodied by the Zipf-Mandelbrot structure of the fillers spectrum (Baayen 2001, Zeldes 2012, Koplenig 2018, Tunnicliffe & Hunter 2022), mostly exploited in diachrony to extrapolate the number of types to larger sample sizes (Hartmann 2018). Here, I use Evert's method of fitting data (Evert 2004, Evert & Baroni 2005). I first evaluate, over a range of parameters, which sample size is required for the model to come with a reliable estimation of the parameters' values, and show that sample sizes ranging from 60 to 120 (depending on the parameters values) are sufficient. Next, I study the impact of fitting error on the estimation of the number of types, to assess how dramatic it may be in a productivity study. Despite large fitting error for very low sampling sizes, there is virtually no consequence on the average number of types, notably thanks to the known correlation between the two parameters (Iszák 2006).

To study specific constructions, I rely on the COHA corpus (Davies 2010). I then assume that the distribution is homogenous over the 20 decades of the corpus. To best recover its parameters, I compare taking the average of the parameters over each decade or pooling all the data together, and show that the former is significantly better. From this Zipf-Mandelbrot distribution, I then generate, for each decade, a distribution over the number of types given the number of tokens. This allows to detect where the number of types is significantly high or low, and more generally, whether the productivity increases or decreases over time.

To assess the robustness of this method, I apply it to 10 English constructions, for the most part already studied in the literature to provide a comparison: *many a N* (Hilpert & Perek 2015), *in the midst/middle/heart/center of* (Desagulier 2022), *V the hell out of* (Perek 2016), *so ADJ a* (Rudnicka 2021), N_i *after* N_i (Sommerer & Baumann 2021), *far too ADJ*, *let alone Vinf*, *never to V*, *a hint of N*, *on the edge of N*. The main lesson of this overview is that productivity tends to be homogenous over the frequency rise of the construction (that is, the type frequency rises with the token frequency in a way which is consistent with a global Zipf-Mandelbrot law), while the productivity declines when the frequency decreases, therefore introducing a fundamental asymmetry between the two processes.

References

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

Baayen, R. H. (2009). Corpus linguistics in morphology: Morphological productivity. In *Corpus linguistics. An international handbook* (pp. 900–919).

Baayen, R. H., & Renouf, A. (1996). Chronicling the Times: Productive lexical innovations in an English newspaper. *Language*, *72*(1), 69–96.

Barðdal, J., & Gildea, S. (2015). Diachronic Construction Grammar: Epistemological context, basic assumptions and historical implications. *Diachronic Construction Grammar*, 1–49.

Berg, K. (2020). Changes in the productivity of word-formation patterns: Some methodological remarks. *Linguistics*, *58*(4), 1117–1150.

Cowie, C., & Dalton-Puffer, C. (2002). Diachronic word-formation and studying changes in productivity over time: Theoretical and methodological considerations. In J. E. Díaz Vera (Ed.), *A changing world of words* (pp. 410–437). Brill.

Davies, M. (2010) *The Corpus of Historical American English (COHA).* Available online at https://www.english-corpora.org/coha/.

Desagulier, G. (2022). Changes in the midst of a construction network: A diachronic construction grammar approach to complex prepositions denoting internal location. *Cognitive Linguistics*, *33*(2), 339–386.

Enghels, R. (2018). Towards a constructional approach to discourse-level phenomena: The case of the Spanish interpersonal epistemic stance construction. *Folia Linguistica*, *52*(1), 107–138.

Evert, S. (2004). A simple LNRE model for random character sequences. *Proceedings of JADT*, *2004*.

Evert, S., & Baroni, M. (2005). Testing the extrapolation quality of word frequency models. *Proceedings of Corpus Linguistics*, *2006*.

Gaeta, L., & Ricca, D. (2006). Productivity in Italian word formation: A variable-corpus approach. *Linguistics*, *44*(1), 57–89.

Hartmann, S. (2018). Derivational morphology in flux: A case study of word-formation change in German. *Cognitive Linguistics*, *29*(1), 77–119.

Hilpert, M. (2013). *Constructional change in English: Developments in allomorphy, word formation, and syntax*. Cambridge University Press.

Hilpert, M., & Perek, F. (2015). Meaning change in a petri dish: Constructions, semantic vector spaces, and motion charts. *Linguistics Vanguard*, *1*(1), 339–350.

Izsák, J. (2006). Some practical aspects of fitting and testing the Zipf-Mandelbrot model: A short essay. *Scientometrics*, *67*(1), 107–120.

Koplenig, A. (2018). Using the parameters of the Zipf–Mandelbrot law to measure diachronic lexical, syntactical and stylistic changes–a large-scale corpus analysis. *Corpus Linguistics and Linguistic Theory*, *14*(1), 1-34.

Perek, F. (2014). *Vector spaces for historical linguistics: Using distributional semantics to study syntactic productivity in diachrony*.

Perek, F. (2016). Using distributional semantics to study syntactic productivity in diachrony: A case study. *Linguistics*, *54*(1), 149–188.

Perek, F. (2018). Recent change in the productivity and schematicity of the way-construction: A distributional semantic analysis. *Corpus Linguistics and Linguistic Theory*, *14*(1), 65–97.

Rudnicka, K. (2021). So-adj-a construction as a case of obsolescence in progress. In S. Kranich & T. Breban (Eds.), *Lost in Change: Causes and processes in the loss of grammatical elements and constructions* (Vol. 218, pp. 51–73). John Benjamins Publishing Company.

Sommerer, L., & Baumann, A. (2021). Of absent mothers, strong sisters and peculiar daughters: The constructional network of English NPN constructions. *Cognitive Linguistics*, *32*(1), 97-131.

Sundquist, J. D. (2020). Productivity, richness, and diversity of light verb constructions in the history of American English. *Journal of Historical Linguistics*, *10*(3), 349–388.

Tunnicliffe, M., & Hunter, G. (2022). Random sampling of the Zipf–Mandelbrot distribution as a representation of vocabulary growth. *Physica A: Statistical Mechanics and its Applications*, *608*, 128259.

Zeldes, A. (2012). *Productivity in argument selection*. De Gruyter Mouton.

Visualizing Character Profile Shifts in English Texts Over The Centuries

Eric S. Wheeler (York University, Toronto, Canada) and Sheila Embleton (York University, Toronto, Canada and Laurentian University, Sudbury, Canada)

In past work, we have measured texts using a very basic alphabetic profile (called a "character profile") to show that different languages can be readily distinguished from one another using multidimensional scaling (MDS) pictures (Wheeler 2020). Tools we had already developed for studying dialect variation (see e.g. Embleton, Uritescu and Wheeler 2015) were adapted to visually display the results, and the results were both striking and easy to see.

In this work, we follow up on a casual observation from the first study that English texts from different eras were also distinguished in the MDS picture. Perhaps, we speculated, one could show a language progressing through time with this very simple measurement, visually displayed on the MDS picture.

Using English texts, downloaded mostly at the Project Gutenberg library, from Anglo-Saxon writings through Middle English, Elizabeth and Jacobean, Renaissance and Reformation, Augustan, and Victorian authors, we find a range of text character profiles that, indeed, confirm a temporal component. But, the resulting pictures also raise considerations of authorship, and genre, and possibly editorial and transcriptional interference. So, the final result is not a simple path across the MDS picture, but it is a most suggestive one, all the more readily seen in its visual presentation.

Since the quantity of letters in a text is hardly a conscious style choice of any author, we must treat this result as an "inherent" property of language itself, and look for systematic explanations for it. Others have studied such aspects using more elaborate measures (e.g. Andreev 2007, Brainerd 1972 especially chapter 6, Fangxiang 2007 and many others) but here we have a simple measure showing change in language. In particular, the existence of differences in genre from the same author indicates something of interest. We discuss some of the possibilities without claiming we have the final conclusion.

References

Project Gutenberg. https://www.gutenberg.org/

Andreev, Sergej N. 2007. A diachronic study of the style of Longfellow. in Grzybek & Köhler. pp 1-12.

Brainerd, Barron. 1972. Weighing Evidence in Language and Literature. A statistical approach. University of Toronto Press.

Embleton, Sheila, Dorin Uritescu and Eric S. Wheeler. 2015. The Advantages of Quantitative Studies for Dialectology. In Arjuna Tuzzi, Martina Benesova, Jan Macutek (eds). Recent Contributions to Quantitative Linguistics. Quantitative Linguistics. 70. De Gruyter Mouton. pp 51-61.

Fangxiang, Fan. 2007. A corpus based quantitative study on the change of TTR. word length, and sentence length of the English language. in Grzybek and Köhler. pp 123-130.

Grzybek, Peter and Reinhard Köhler (eds). 2007. Exact methods in the Study of Language and Text. Quantitative Linguistics 62. Mouton de Gruyter.

Wheeler, Eric S. 2020. Language Identification by Simple Character Profiles. in Emmerich Kelih and Reinhard Köhler. ed. Words and Numbers; In Memory of Peter Grzybek (1957-2019). RAM-Verlag. pp 44-52

Development of mean dependency distance in Czech L2 texts across proficiency levels A1 to C1

M. Hanušková, M. Nogolová, M. Kubát
University of Ostrava

Keywords: MDD, dependency syntax, Czech, second language

Syntactic analysis is a crucial part of second language acquisition studies from the beginning. During the last 30 years, many measures have been invented for depicting the syntax development of second language (L2) writers. Most of them are focused on the average length of clauses, t-units, or sentences, using different units (e. g. words, phrases). Recently, scholars have focused on finding measures that can reflect the dependency structures of sentences (cf. Ouyang & Jiang 2018; Ouyang et al. 2022). Following this trend, we apply the mean dependency distance (MDD) for analysing the syntactic development of Czech L2 texts. MDD is calculated as the average distance (measured in the number of words) between the governors and the dependent words in the sentence (cf. Liu 2008; Liu et al. 2017). In this analysis, we use 5,721 texts from the CzeSL-SGT learner corpus (Šebesta et al. 2014) that cover A1–C1 levels of language proficiency (according to the CERF). Furthermore, the results are compared with the reference corpus (REF-CZ) consisting of texts written by Czech native speakers (SKRIPT2012; Šebesta et al. 2013). We also examine the cross-linguistic influence of the learners' first language. Specifically, we compare the results of texts written by Slavic and non-Slavic speakers at the same proficiency levels. The differences between proficiency levels are statistically tested, as well as the differences between the results of the Slavic and non-Slavic groups. The results indicate that the higher the level of language proficiency, the higher the MDD. There are statistically significant differences between all pairs except B2 and C1. Furthermore, Slavic L1 groups have a higher MDD in most cases and statistically significant differences at lower levels of language proficiency.

References

Liu, H. (2008). Dependency distance as a metric of language comprehension difficulty. Journal of Cognitive Science, 9(2), 159-191.

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.

Šebesta, K., Bedřichová, Z., Šormová, K., Štindlová, B., Hrdlička, M., Hrdličková, T., Hana, J., Petkevič, V., Jelínek, T., Škodová, S., Poláčková, M., Janeš, P., Lundáková, K., Skoumalová, H., Sládek, Š., Pierscieniak, P., Toufarová, D., Richter, M., Straka, M. & Rosen, A. (2014): CzeSL-SGT: CzeSL-SGT – a corpus of non-native speakers' Czech with automatic annotation, version 2 from 28 Sep 2014. Ústav Českého národního korpusu FF UK, Praha.

Šebesta, K., Goláňová, H., Jelínek, T., Jelínková, B., Křen, M., Letafková, J., Procházka, P., Skoumalová, H.: SKRIPT2012: akviziční korpus psané češtiny, přepisy písemných prací žáků základních a středních škol v ČR. Ústav Českého národního korpusu FF UK, Praha 2013.

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

Ouyang, J., Jiang, J., & Liu, H. (2022). Dependency distance measures in assessing L2 writing proficiency. Assessing Writing, 51, 100603.

Syntactic development and optimality of dependency distances for Japanese as a second language

Saeko Komori[1], Masatoshi Sugiura[2], Ramon Ferrer-i-Cancho[3], Lluís Alemany Puig[3], and Wenping Li[4]

1 Chubu University, Japan.
2 Nagoya University, Japan.
3 Universitat Politècnica de Catalunya - Barcelonatech (UPC), Catalonia, Spain.
4 Shanghai University of Finance and Economics, China.

Keywords: Second language learning, dependency syntax, hierarchical distance, dependency distance

The purpose of this study is to investigate appropriate methods to measure syntactic development of learners of Japanese as a second language. There exist some measures to calculate syntactic complexity based on the analysis of syntactic dependency structures. Jiang and Ouyang (2018) used MDD, Mean Dependency Distance, to measure syntactic complexity of essays written by Chinese learners of English, arguing that MDD captures learners' development. Komori et al. (2019) examined the validity of MDD with Chinese learners of Japanese, by comparing it to another measure, MHD, Mean Hierarchical Distance. However, their results were not consistent with the previous study: MHD distinguished the learning stages but MDD did not. Jing and Liu (2015) examined the cross-linguistic differences between MDD and MHD in terms of sentence length, but the results were inconclusive.

Second language learners' language skills are assumed to become closer to the native speakers' as they become more proficient. The closeness to native speakers can be regarded as the measure of second language development. If native speakers do not necessarily produce complex or longer sentences, complexity measures are not necessarily appropriate to measure the syntactic development. Recently, a new linguistic index called Omega (Ferrer-i-Cancho et al. 2022) has been proposed to measure the optimality of syntactic dependency distances. MHD is a structural index, i.e. it is independent from the linear ordering of the sentence, whereas MDD is an unnormalized linear order index that incorporates indirectly/implicitly, structural properties. Omega is a normalized hybrid index that incorporates structural and linear order aspects directly/explicitly in its definition.

The current study explores the applicability of Omega along with MDD and MHD to the analysis of syntactic development of second language learners. We refined the dataset of Komori et al.'s (2019), comparing Chinese learners of Japanese versus natives. Participants belong to three levels of competence: (1) second-year learners, (2) third-year learners and (3) Japanese native speakers. We performed pairwise comparisons between all three linguistic levels.

We find that all indices except MDD show a progressive increase with linguistic competence. In particular, when comparing one linguistic level against a higher level, there is always a significant increase in the value of the index, with the only exception of MDD on third-year learners versus natives, where no significant difference was found. Concerning dependency distance, Omega is statistically more powerful than MDD. We suspect this phenomenon could be due to the hybrid nature of Omega (i.e. it incorporates more information about a sentence than MDD) or to a capacity of Omega to normalize dependency distances in a way that preserves differences between speakers' levels.

We also contribute with a theoretical understanding of the distinct indices. We confirm that MDD and MHD are weakly correlated (although significantly) while Omega is strongly correlated with MHD and less correlated (in absolute value) with MDD. The strong correlation between Omega and MHD (that does not happen with MDD) is consistent with the explicit/direct incorporation of structural aspects in the definition of Omega.

REFERENCES

Ramon Ferrer-i-Cancho, Carlos Gómez-Rodríguez, Juan Luis Esteban and Lluís Alemany-Puig (2022). "Optimality of syntactic dependency distances". *Physical Review E* 105, 014308.

Jingyang Jiang and Jinghui Ouyang (2018) "Minimization and Probability Distribution of Dependency Distance in the Process of Second Language Acquisition". *Quantitative Analysis of Dependency Structures*. Ed. by Jingyang Jiang and Haitao Liu. Berlin/Boston: De Gruyter Mouton, pp. 167–190.

Yingqi Jing and Haitao Liu (2015) "Mean Hierarchical Distance: Augmenting Mean Dependency Distance". *Proceedings of the third international conference on dependency linguistics (Depling 2015)*, pp. 161–170.

Saeko Komori, Masatoshi Sugiura and Wenping Li. (2019) "Examining MDD and MHD as syntactic complexity measures with intermediate Japanese learner corpus data". *Proceedings of the Fifth International Conference on Dependency Linguistics (Depling, SyntaxFest 2019)*, pp. 130–135.

Multilevel phylogenetic model shows no evidence for dependency locality in Indo-European

Yingqi Jing, Joakim Nivre and Michael Dunn

Uppsala University

It has long been assumed that that grammars of languages have evolved in a way so that speakers can arrange linguistic units (e.g., word orders) to facilitate the processing, production and learning. Many functional principles have been proposed to explain the word order universals and diversity in the world's languages. On the one hand, the principle of dependency locality suggests that languages prefer to minimize the linear distances (dependency lengths) between the head and its dependents, such as a verb-medial placement, so as to facilitate identification of syntactic structure (Ferrer i Cancho 2004; Liu 2008; Temperley 2008; Futrell, Mahowald, and Gibson 2015). On the other hand, theories of predictability maximization and learning efficiency would favor a consistently initial or final placement of the verb or head (Vasishth and Lewis 2006; Ferrer i Cancho 2017; Culbertson, Smolensky, and Legendre 2012), since it is easier to predict and acquire the head-dependent relations though this placement increases the dependency lengths and causes anti-locality (Ferrer i Cancho and Gómez-Rodríguez 2021; Jing, Blasi, and Bickel 2022).

To better understand how languages balance these competing pressures and detect the evolutionary trends of syntactic change, we use Bayesian phylogenetic inference to estimate the rates of change for locality or anti-locality across different dependency types in Indo-European. Here we focus on the lengths of major lexical dependencies (e.g., verb→'nsubj'→noun, noun→'amod'→adjective, etc) in 45 dependency-annotated corpora from Universal Dependencies v2.10 (Zeman et al. 2022). Instead of fitting phylogenetic models for each individual dependency type, we have developed a multilevel phylogenetic Continuous-time Markov Chain model that can infer evolutionary rates at both population and group levels (Nalborczyk et al. 2019; Stan Development Team 2022). We further incorporate the uncertainty in tip states, tree topologies and branch lengths into our model (Jing, Widmer, and Bickel 2022), and compare the evolutionary change in real utterances with a baseline that randomly linearizes the head and its dependents (Ferrer i Cancho 2004; Futrell, Mahowald, and Gibson 2015).

Our results show no evidence for a general bias towards dependency locality at the population level. By contrast, we observe a slight bias against locality when compared with the random baseline (mean rate ratio for real data: 1.38 and 90% CI = [0.47, 3.17]; mean rate ratio for the random baseline: 1.75 and

90% CI = [0.48, 4.23]). We also find substantial differences across dependency types. Most of them, including subjects, obliques and subordinate clauses (adnominal, adverbial and complement clauses) reveal a consistent anti-locality effect that favors a faster transition towards longer dependencies than towards shorter dependencies, whereas only certain modifiers of nouns and adjectives (e.g., 'amod', 'nmod' and 'advmod') show weak evolutionary biases towards locality. There are no directional trends for objects and adverbial modifiers of verbs between real data and random baseline.

The similar overall rates towards locality and anti-locality challenge the universal constraint of dependency locality (Ferrer i Cancho 2004; Liu 2008; Temperley 2008; Futrell, Mahowald, and Gibson 2015). Surprisingly, we found that languages even show stronger evolutionary biases against locality than randomly putting dependents on either side of the head. This anti-locality effect is particularly prominent in heavy constituents like obliques and subordinate clauses. This casts doubt on the explanation of their orders via dependency locality. Instead, we suggest that they likely evolve to maximize the predictability of head-dependent relations (Ferrer i Cancho 2017) or due to the simplicity bias in learning (Culbertson, Smolensky, and Legendre 2012). The observed anti-locality in subjects can be related to the general subject or agent first principle (Greenberg 1963; Napoli and Sutton-Spence 2014), since the initial subjects are often separated from the verb by an intervening constituent. The potential locality effects are limited to specific modifiers of nouns and adjectives. They also weaken the explanation of dependency locality as a general principle for word order evolution.

The regularity of polysemy patterns in the mind: Computational and experimental data

Alizée Lombard *University of Fribourg* (alizee.lombard@unifr.ch)
Anastasia Ulicheva *Royal Holloway University of London* (anastasia.ulicheva@gmail.com)
Maria Korochkina *Royal Holloway University of London* (Maria.Korochkina@rhul.ac.uk)
Kathleen Rastle *Royal Holloway University of London* (Kathy.Rastle@rhul.ac.uk)

Keywords: polysemy; regularity; experimental semantics; WordNet

Statistical regularities characterise many language phenomena (e.g. spelling-sound or morphological regularity). Interestingly, regularities also apply to polysemy: when at least two non-synonymous words use the same type of semantic extension (Apresjan 1974), such as *pig* 'dirty person' and *wolf* 'lonely person'. However, these patterns have not received much attention in psycholinguistic research, probably because there has not been a straightforward way to operationalise the graded nature of this phenomenon. Instead, researchers have typically described polysemy in a categorical manner (regular vs. irregular, metaphor vs. metonymy) (Klepousniotou et al. 2012, Rabagliati & Snedeker 2013, Brocher et al. 2018, a.o.). In this study, we sought to determine whether speakers assimilate information about polysemic regularities and use this information productively.

We quantified the regularity of metaphor patterns in English (e.g., ANIMAL (source sense) → PERSON (target sense)). Two regularity metrics ($R_{1\text{-}2}$) were based on values extracted from WordNet (Fellbaum 1998, Miller 2005): (i) the number of words with the source sense (N_{S1})only and (ii) the number of words with both the source and the target senses (N_{S2}). Measures $R_{3\text{-}4}$ took into account word form frequency (f_w).

$$R_1 = N_{S2} \qquad R_2 = \frac{N_{S2}}{N_{S1}} \qquad R_3 = \sum_{w=1}^{N_{S2}} log(f_w) \qquad R_4 = \frac{\sum_{w=1}^{N_{S2}} log(f_w)}{\sum_{w=1}^{N_{S1}} log(f_w)}$$

To assess psychological validity of our measures, we conducted an experiment in which adult participants rated the plausibility of semantic neologisms (i.e., existing words used with a new sense). We predicted that the degree to which adults deemed semantic neologisms as plausible would be modulated by pattern regularity (Bastuji 1974, Renouf 2013). We created 5 neologisms per pattern (7 patterns in total). These were embedded into sentence frames such that they could be interpreted without ambiguity (1a). They were compared to 35 'illegal' neologisms that had the same target sense (e.g., 'location part') but were not part of any pattern (1b). The conditions (new vs. illegal) were matched on target word frequency and syntactic structure of the sentence frame.

(1) BODY PART→ OBJECT PART (*heart, head, leg*)

a. new I always slow down on the **knee** of the mountain road.
b. illegal I always meditate on the **milk** of the highest hill.

Thirty-two participants judged the plausibility of these sentences on a scale from 0 to 100. Our hypotheses were (a that the 'new' senses would be more likely to be judged as plausible than the 'illegal' senses; and (b) that plausibility of 'new' senses would increase in line with pattern regularity.

The data were analysed using generalised linear mixed effects models.

We found (a) that, on average, the 'new' patterns were more likely to be judged as plausible than the 'illegal' ones, and (b) that plausibility was positively correlated with each regularity measure. The best model fit was achieved with proportion-based measures (R_2 and R_4).

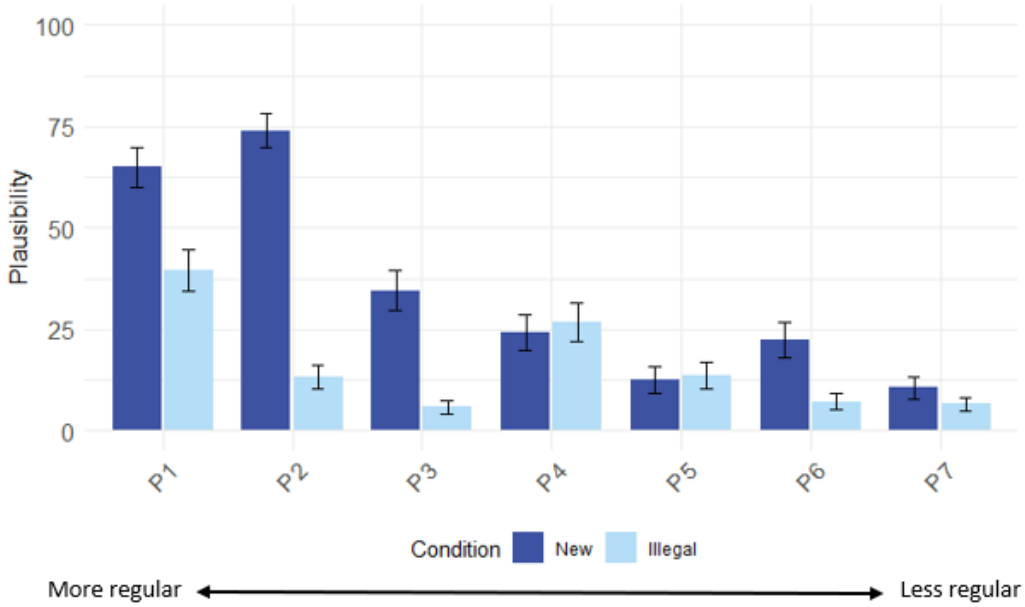


Figure : Average plausibility ratings (y-axis) per pattern (x-axis) and condition with 95% confidence intervals.

In summary, we pioneered an approach to quantification of graded regularity of polysemy patterns. We showed that this approach has psychological validity, and we hope that this work would pave the way for future studies of graded polysemy.

References

Apresjan, J. (1974). Regular Polysemy. *Linguistics*, 142, 5-32.

Bastuji, J. (1974). Aspects de la néologie sémantique. *Langages* 36, 6-19.

Brocher, A., Koenig, J. P., Mauner, G., and Foraker, S. (2018). About sharing and commitment: the retrieval of biased and balanced irregular polysemes. *Language, Cognition and Neuroscience*, 33(4), 443-466.

Fellbaum, Christiane. 1998. *Wordnet: An electronic lexical database*. Cambridge, MIT Press.

Klepousniotou, E., Pike, G. B., Steinhauer, K., and Gracco, V. (2012). Not all ambiguous words are created equal: An EEG investigation of homonymy and polysemy. *Brain and language*, 123(1), 11-21.

Miller, George A. 2005. Wordnet: A lexical database for english. *Communications of the ACM* 38(11). 39–41.

Rabagliati, H., and Snedeker, J. (2013). The truth about chickens and bats: Ambiguity avoidance distinguishes types of polysemy. *Psychological science*, 24(7), 1354-1360.

Renouf, A. (2013). A finer definition of neology in English: The life-cycle of a word. *Studies in Corpus Linguistics* 57, 177-208.

Van Heuven, W.J.B. & Mandera, P. & Keuleers, E. & Brysbaert, M. 2005. Subtlex-uk: A new and improved word frequency database for british english. *Quarterly Journal of Experimental Psychology* 67. 1176–1190.

Modelling semantic differentiation between near-synonyms with word2vec and t-SNE

Kaleigh Woolford (University of Toronto)

Keywords: near-synonymy, lexical choice, word2vec, t-SNE, collocations, word meaning

The current study, driven by the notion that semantic differences between near-synonyms reflect the conceptual structures that motivate language use (Glynn 2010), offers a novel method of disentangling the complex internal structure of near-synonyms by modelling their associations with contrasting semantic domains. While Desagulier (2014) proposes correspondence analysis to project frequency-based correlations between near-synonyms and their collocates onto a low-dimensional space, this method cannot determine whether collocates with similar distributions form a related semantic class. Subsequently, this approach fails to address whether contrasting collocational preferences reflect true patterns of semantic differentiation.

To move beyond this limitation, I combine neural network models (word2vec, van der Maaten and Hinton, 2008) and a dimensionality-reduction algorithm (t-SNE, t-Distributed Stochastic Neighbour Embedding, Mikolov et al. 2013) as a method of modelling not only the association between near-synonyms and their collocates, but also the relationship between those collocates. Through a case study of the near-synonym set *solely*, *exclusively*, *purely*, and *strictly* (Nevalainen 1991), I argue that this method reveals clear patterns of semantic differentiation, reflecting the underlying structures motivating their use.

The collocational preferences of *solely*, *exclusively*, *purely* and *strictly* were determined by submitting the co-occurrence frequencies for each adverb and its adjectival collocates (e.g., *solely responsible*) from the iWeb corpus (Davies and Kim 2019) to collostructional analysis (Stefanowitsch and Gries, 2003). Each adverb's top 15 adjectival collocates, along with each of those adjectives' 5 nearest semantic neighbours (extracted via word2vec) were then projected in space with t-SNE based on the similarity of their word2vec embeddings.

As Figure 1 shows, closely-related adjectives cluster together to form distinct semantic classes, such as 'secrecy' (*classified, confidential, anonymous* and *secret*) or 'responsibility' (*liable, negligent, culpable, responsible* and *accountable*). The near-synonym each adjective most distinctively collocates with is then visualized on top of these clusters, revealing almost every semantic domain is homogeneously associated with a single adverb; for instance, adjectives denoting 'secrecy' distinctively collocate with *strictly,* while those denoting 'responsibility' distinctively collocate with *solely*. As *solely*, *exclusively*, *purely* and *strictly* can thus be interpreted as operating across a set of coherent semantic domains, their distinctive collocational preferences can be taken to reflect the semantic differences that drive their use.

Overall, this finding demonstrates the utility of this method for teasing apart the complex internal structure of near-synonyms. The ability to evaluate whether near-synonyms are attracted to semantic domains, enabled here by a combination of word2vec and t-SNE, is key to revealing the underlying semantic structures that influence contrasting patterns of language use. More broadly, insights into these semantic structures are valuable not only for disentangling the phenomenon of near-synonymy, but also for other domains investigating factors influencing lexical choice such as language variation and change.

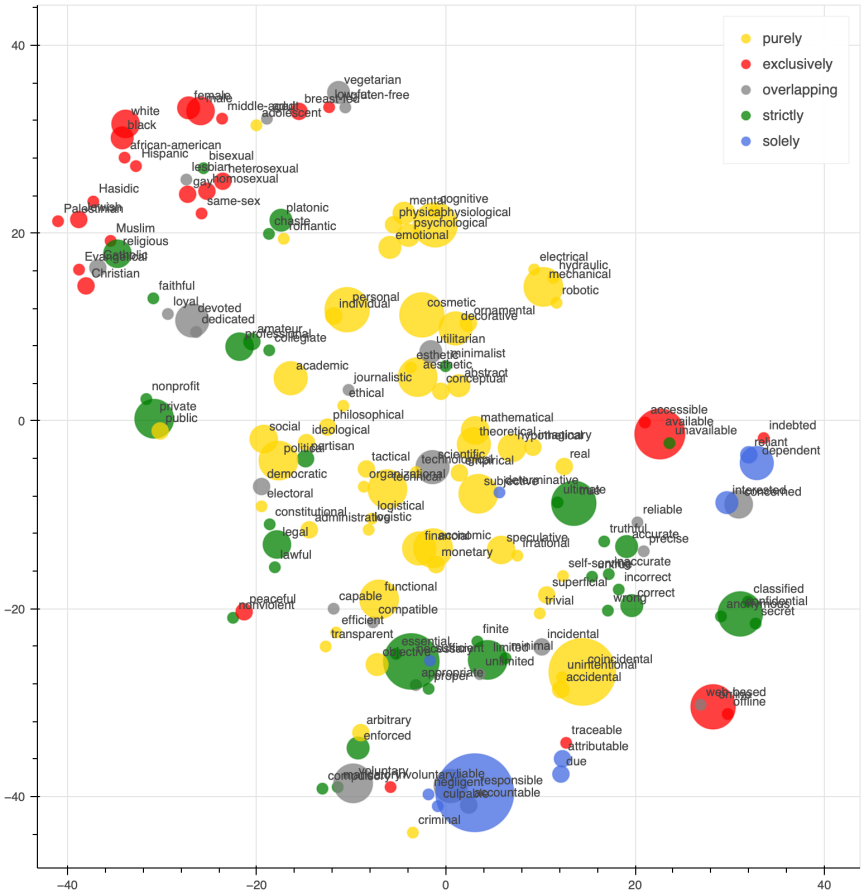


Figure 1. t-SNE visualization of s*olely, exclusively, purely* and *strictly*'s adjectival collocates on word2vec embeddings. Colour indicates which adverb an adjective is most distinctive of (grey indicating no distinctive preference). Size reflects frequency of an adjective across the 4 adverbs.

References

Davies, M and Kim, J. 2019. The advantages and challenges of "big data": Insights from the 14 billion word iWeb corpus. *Linguistic Research* 36(1): 1–34.

Desagulier, G. 2014. Visualizing distances in a set of near-synonyms. In D. Glynn and J.A Robinson, eds, *Corpus Methods for Semantics: Quantitative Studies in Polysemy and Synonymy*, volume 43, p. 145–178. John Benjamins.

Glynn, D. (2010). Synonymy, lexical fields, and grammatical constructions. A study in usage-based cognitive semantics. In H. Schmid and S. Handell (eds) *Cognitive Foundations of Linguistic Usage Patterns: Empirical Studies,* p. 89–118. DeGruyter Mouton.

van der Maaten, L. and Hinton, G. 2008. Visualizing data using t-SNE. *Journal of Machine Learning Research* 9(11): 2579–2605.

Mikolov, T., Sutskever, I., Chen, K., Corrado, G.S and Dean, J. 2013. Distributed representations of words and phrases and their compositionality. In *NeurIPS*, p. 3111–3119.

Nevalainen, T. 1991. *But, only, just: Focusing adverbial change in Modern English, 1500-1900*. Societé Néophilologiqué, Helsinki.

Stefanowitsch, A. and Gries, S. Th. 2003. Collostructions: Investigating the interaction of words and constructions. *International Journal of Corpus Linguistics* 8(2): 209–243.

Homomorphism, Voronoi tesselation, and lexical meaning

Hermann Moisl
Newcastle University, UK

Quantitative linguistics has applied mathematical concepts and methods to theoretical modelling of the formal aspects of language from phonetics through to syntax. To my knowledge this has not extended to modelling of linguistic meaning. The present discussion proposes a way of doing so..

Humans intuitively feel that they possess a head-internal meaningfulness, that is, an awareness of the self and its relationship to the perceived world which is independent of interpretation of one's behaviour by observers. This intuition is captured by the philosophical concept of intentionality (Jacob 2019; Morgan & Piccinini, 2018; Neander, 2017) which is used in present-day philosophy of mind to denote the 'aboutness' of mental states, '*the power of minds and mental states to be about, to represent, or to stand for, things, properties and states of affairs*' (Jacob, 2019).

In earlier work (Moisl 2021, 2022) I argued that the mathematical concept of homomorphism (Gowers et al 2008) is fundamental to physical implementation of human cognitive intentionality in general and linguistic meaning in particular. The present discussion proposes topology preservation using Voronoi tesselation as an implementation mechanism for homomorphism with respect to lexical meaning.

The discussion is in four main parts. The first part motivates the proposal on the basis of existing work in cognitive science which argues that intrinsic intentionality is a necessary condition for a mentalist theory of linguistic meaning. The second part outlines the mathematical concepts of homomorphism and Voronoi tesselation. The third shows how the self-organizing map (SOM), a type of artificial neural network, can be used to implement a homomorphism between high-dimensional linguistic input and its representation on a two-dimensional surface by learning a Voronoi tesselation of the input manifold. The fourth presents a computational example of such a SOM.

Keywords: intentionality, lexical meaning, homomorphism, Voronoi tesselation, self-organizing map

References

Aurenhammer, Franz; Klein, Rolf; Lee, Der-Tsai (2013). *Voronoi Diagrams and Delaunay Triangulations*. World Scientific

Gowers, T., Barrow-Green, J., Leader, I. (2008) *The Princeton companion to mathematics*, Princeton University Press

Jacob, P. (2019). Intentionality. *Stanford Encyclopedia of Philosophy*, (Spring 2019 Edition), ed. E. Zalta, URL = <https://plato.stanford.edu/archives/spr2019/entries/intentionality/>

Kohonen, T. (2001) *Self Organizing Maps*. 3rd ed. Berlin: Springer

Moisl, H. (2021) Implementation of intrinsic natural language lexical intentionality, *Academia Letters* 2021

Moisl, H. (2022) Dynamical systems implementation of intrinsic sentence meaning, *Minds and Machines* 32. DOI: 10.1007/s11023-022-09590-1.

Morgan, A., Piccinini, G. (2018). Towards a cognitive neuroscience of intentionality. *Minds and Machines,* 28, 119-139

Neander, K. (2017). *A mark of the mental: In defense of informational teleosemantics*. MIT Press

Two+-dimensional uncertainty estimates for frequency, dispersion, and association measures

Stefan Th. Gries
UC Santa Barbara & JLU Giessen

Many corpus-linguistic studies use quantitative methods and report statistical results. Apart from using 'general statistics', i.e. methods such as regression or principal component analysis that can be applied in virtually any quantitative research domain, there are several more specifically corpus-linguistic statistics such as different kinds of frequencies, measures of dispersion in corpora, or collocational and other association measures. However, most studies reporting frequencies, dispersions, and associations do not also provide what in many other quantitative fields is commonplace, namely measures of the variability that come with the results. For example and very subjectively, we are not sure we have ever seen a study that provided dispersion measures or Mutual Information collocate scores with a confidence interval (CI). What is more, if CIs are provided in corpus linguistics, they are often parametric, meaning they are based on *t*- or *z*-scores and, thus, involve incorrect assumptions such as the bag-of-words model for corpora, which usually makes them anticonservative.

In this paper, we aim to address this unfortunate gap and will do three things. First, we will demonstrate how parametric CIs can overestimate significance whereas CIs involving speaker-/file-based bootstrapping are more realistic (in terms of sampling) and thus fare better conceptually.

Second, we will exemplify the computation of bootstrapped CIs for observed corpus frequencies (logged or on the Zipfscale), dispersion values, and association measures based on words and patterns in the International Corpus of English and the Corpus of Historical American English; synchronic results for selected ditransitives and phrasal verbs and diachronic results for two words spanning multiple decades in American English indicate that relying on point estimates alone can lead to misleading interpretations that the current approach avoids.

Third, we will also argue that, ideally, researchers would not study any one of these dimensions (frequency, dispersion, association) in isolation, but in combination. For instance, we will argue that the association of words to patterns should in fact be measured such that frequency and association are kept separately so that we can distinguish all four combinations of high and low frequency and high and low association. We then exemplify how the combination of file-/speaker-based bootstrapping together with 90- or 95% data ellipses can represent the volatility of co-occurrence data in corpora in a visually straightforward way. This in turn allows for a revision of nearly all kinds of corpus analyses based on simple ranking of elements co-occurring and interpreting the top *n* items by identifying which items come with such high degrees of volatility that their different ranks should not be overinterpreted.

We therefore recommend that the field adopts a standard policy of having authors provide such non-parametric CIs for all corpus statistics.

Morphological Complexity in Lexical Networks

Petra Steiner
Technische Universität Darmstadt
Darmstadt, Germany
petra.steiner@tu-darmstadt.de

Abstract

In this paper, three hypotheses and their functional interplay in word formation are observed. These are a. the impact of the polylexy of lexemes on their word-formation activity, b. Hawkins Principle of Early Immediate Constituents and c. the impact of the hierarchy level of lexemes within lexical networks (Sambor, 2005) on their word-formation activity. While the first hypothesis has been corroborated for free and bound constituents of words (Krott, 2004; Steiner, 1995), the second hypothesis by Hawkins (1994) has been widely examined for syntactic levels (Hoffmann, 1999; Hawkins, 1999; Köhler, 2012, 138ff.) but less for morphology. The third hypothesis is derived from observations and hypotheses in psycholinguistics by Rosch's (1978) notions of superordinate level, basic level, and subordinate level in taxonomies which have predictable properties. For example, the basic level is used more frequently and has a higher degree of prototypicality which again leads to more common attributes with the lexemes of the next level.

As data, we use the German morphological trees database built by the tools of Steiner (2017). It combines the analyses of the German part of the CELEX database (Baayen et al., 1995) which were exploited by Steiner and Ruppenhofer (2018), and the annotated compounds from the GermaNet database (Henrich and Hinrichs, 2011). Figure 1 presents an example of a complex German compound, Figure 2 describes its depth structure. The left-branching structure is consistent with the principle of EIC as the head is at the end of the lexeme. We derive the lexical networks by searching through the GermaNet synsets which have the relation of hyperonymy. From definition chains such as *Währungsausgleichsfonds – Fonds – Geld, Geldanlage, Kapital, Finanzausstattung, Asset, Anlage, Besitz etc. – Vermögen, Finanzen – materieller Besitz – Besitz* we derive the hierarchy levels of the compounds.

The investigation indicates towards multiple factors for the construction of complex words. These are the requirement for specification, the need of reducing the memory effort, and furthermore the efficient semantic coverage of lexemes.

Keywords: morphological complexity, Principle of Early Immediate Constituents, polylexy, lexical networks, prototype semantics

References

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

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

John A. Hawkins. 1999. The relative order of prepositional phrases in English: Going beyond Manner–Place–Time. *Language Variation and Change* 11(3):231–266. https://doi.org/10.1017/S0954394599113012.

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

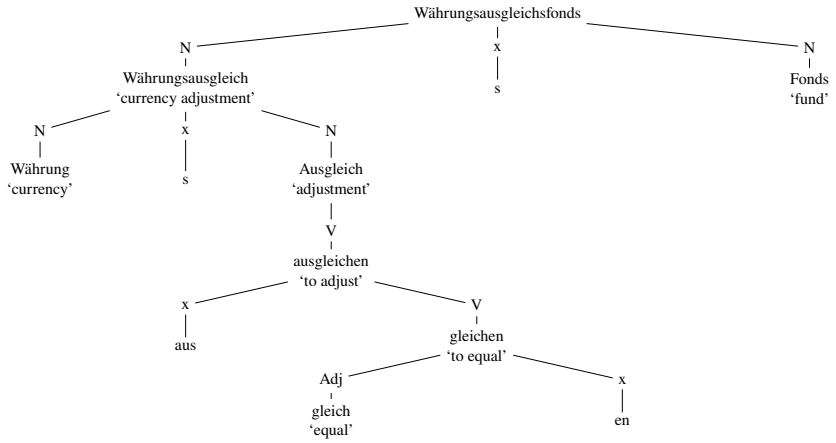


Figure 1: Morphological analysis of *Währungsausgleichsfonds* 'currency adjustment fund'

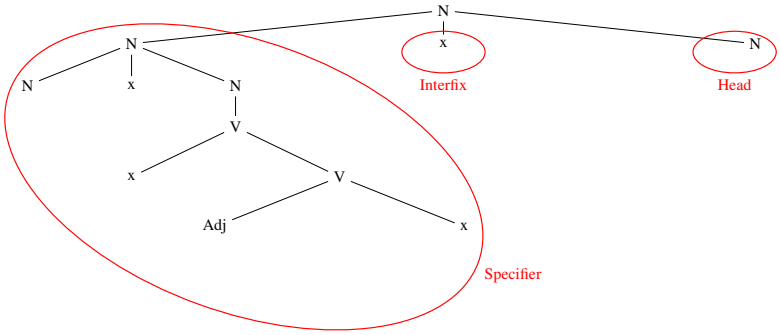


Figure 2: **Structure: (1 (2) (2) (2 (3 (4) (4 (5) (5))))) (1) (1)**

Christiane Hoffmann. 1999. Word Order and the Principle of "Early Immediate Constituents" (EIC). *Journal of Quantitative Linguistics* 6(2):108–116. https://doi.org/10.1076/jqul.6.2.108.4133.

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

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

Eleanor Rosch. 1978. Principles of categorization. In Eleanor Rosch and Barbara B. Lloyd, editors, *Cognition and Categorization*, Erlbaum, Hillsdale, NJ, pages 27–48.

Jadwiga Sambor. 2005. Lexical networks (Lexikalische Netze). In Reinhard Köhler, Gabriel Altmann, and Raĭmond Genrikhovich Piotrovskiĭ, editors, *Quantitative Linguistik - Quantitative linguistics*, de Gruyter, Berlin and New York, Handbücher zur Sprach- und Kommunikationswissenschaft - Handbooks of Linguistics and Communication Science, pages 447–458.

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

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

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

Object: Research proposal for an oral session to the *Quantitative Linguistics Conference* 2023 (https://wp.unil.ch/qualico2023/call-for-papers/)

Authors:

Maud Reveilhac, Lausanne University, Institute of social sciences, maud.reveilhac@unil.ch

Gerold Schneider, Zürich University, Institut für Computerlinguistik, gschneid@ifi.uzh.ch

Title: Measuring language complexity about European politics using different data sources and methods

Abstract (< 500 words - excluding keywords, authors affiliations and references)

From a quantitative linguistics perspective, we investigate changes in complexity, style and register (Biber et al., 1998) of political language in parliamentary debates. We use text analysis methods on a salient and divisive topic in the Swiss population and parliament: Europe, and European integration. More generally, the topic is often central for political campaigning (Hutter & Grande, 2014) and media coverage (Vliegenthart et al., 2008). It is also a complex topic on which citizens demand to receive clear and complete information, especially in light of the "democratic deficit" (Boomgaarden et al., 2010). However, as it mixes many sub-issues, the communication to the public of justification for decision-making can be very complex. This is further complicated by the fact that extreme voices often rely on this topic to advance their agenda (e.g., demonisation of the unelected Brussels elite). This leads to varied language use, from technical explanations requiring high linguistic complexity to populism with simple recipes and slogans. Political language has often been said to become simpler (Wyss et al., 2015; Tucker et al., 2020; Reveilhac & Schneider, 2023) – does this also hold for Swiss parliamentary debates? Our study addresses the following research questions: Has the language become simplified in Parliament (RQ1)? How does it compare to other sources of information (RQ2)? What are the most important indicators of linguistic complexity across the different sources (RQ3)?

This paper draws from Rauh's (2022) distinction of three components of language clarity and measures linguistic complexity using different textual properties and methods: 1) the concept of *language complexity* (e.g., sentence length, word length, number of syllables), 2) the concept of *language familiarity* (e.g., mean segment type-token ratio (Malvern & Richards, 2002), language frequency of words approximated by the average general word usage from Google Books corpus (Michel et al., 2010), reliance on pronouns to manage perceptions of in- and out-groups (Tyrkkö, 2016), and topical diversity), 3) the concept of *language accessibility* (e.g., degree of agency or verbal style from a verb-to-noun ratio, frequency of hedges and modals, and degree of passive language).

We answer RQ 1 and 2 by relying on aggregate descriptive statistics (in terms of mean and standard deviation) at the year level. The study compares how the prevalence of these different concepts differs in transcripts of Parliamentary debates (between 2011 and 2021) from other important sources of information for citizens' opinion formation, such as party press releases and the traditional print media (while distinguishing quality from tabloid newspapers). The expert language further considered as an additional source of information and the scientific discourse measured in abstracts of published papers referring to the Swiss context and those that stem from top science journals. To answer RQ3, variable importance analysis is conducted to assess which properties are most indicative of the type of source.

The comparison between parliamentary language and the scientific- and public-oriented languages provide us relevant indicators and benchmarks to assess the extent to which the parliamentarians' communication is accessible to the general public about a complex (and controversial) political issue.

Keywords: linguistic complexity, corpus linguistics, text analysis, comparative research, European politics

References:

Biber, D., S. Conrad, & R. Reppen. (1998). *Corpus Linguistics: Investigating Language Structure and Use*. Cambridge: Cambridge University Press.

Boomgaarden, H. G., Vliegenthart, R., de Vreese, C. H., & Schuck, A.R.T. (2010). News on the move: exogenous events and news coverage of the European Union. *Journal of European Public Policy*, 17(4), 506-526, doi: 10.1080/13501761003673294

Hutter, S., and E. Grande. (2014). Politicizing Europe in the National Electoral Arena: A Comparative Analysis of Five West European Countries, 1970–2010. *Journal of Common Market Studies,* 52(5), 1002–1018. doi: 10.1111/jcms.12133.

Malvern, D., & Richards, B. (2002). Investigating accommodation in language proficiency interviews using a new measure of lexical diversity. Language Testing, 19, 85-104.

Michel, J., Y K. Shen, A P. Aiden, A. Veres, M K. Gray, J P. Pickett, D. Hoiberg, et al. (2010). Quantitative Analysis of Culture Using Millions of Digitized Books. *Science*, 331(6014), 176–182. doi: 10.1126/science.1199644.

Rauh, C. (2022). Clear messages to the European public? The language of European Commission press releases 1985–2020, *Journal of European Integration*, doi: 10.1080/07036337.2022.2134860

Tucker, E. C., Capps, C. J., & Shamir, L. (2020). A data science approach to 138 years of congressional speeches. Heliyon, 6(8), e04417. doi: 10.1016/j.heliyon.2020.e04417

Tyrkkö, J. (2016). Looking for rhetorical thresholds: Pronoun frequencies in political speeches. Studies in Variation, Contacts and Change in English, 17.

Vliegenthart, R., Schuck, A. R., Boomgaarden, H. G., & De Vreese, C. H. (2008). News coverage and support for European integration, 1990–2006. International Journal of Public Opinion Research, 20(4), 415-439. doi: 10.1093/ijpor/edn044

Wyss, D., Beste, S., & Bächtiger, A. (2015). A decline in the quality of debate? The evolution of cognitive complexity in Swiss parliamentary debates on immigration (1968–2014). *Swiss Political Science Review*, 21 (4). 636-653.

Discourse Markers' Role in Syntactic Complexity of Sentence Structure: A Distance-driven Quantitative Case Study Based on TED Talks

Zheyuan Dai & Jianwei Yan
Department of Linguistics, Zhejiang University

Abstract: The role of Discourse Markers (DMs) is a traditional and ever-lasting linguistic topic. Discussions of DMs generally circle pragmatics and discourse coherence (e.g., Blakemore 1992, 2006; Fraser 1996, 1999; Redeker 1991). Nevertheless, despite various efforts focusing on the role of DMs, studies seldom investigated this issue from syntactic perspectives and downplayed the syntactic role of DMs with qualitative analysis (e.g., Heine 2013; Sakita 2013; Zwicky 1985). In short, DMs are dispensable in terms of syntactic structure (Maschler and Schiffrin 2015: 192). However, we have some reservations about this stance. According to the dependency-based view (Melčuk 2014; Tesnière 2015), a syntactic organization is like a drama, involving a performance with its actors and circumstances (Tesnière 2015:97). In light of this, a DM would play the role of being part of the syntactic components in the drama. Therefore, in the present study, we attempt to quantitatively reexamine this issue, investigating the syntactic interpretations of the roles of DMs.

Under the theoretical framework of dependency grammar (Tesnière 2015), the present study quantitatively explores the syntactic role of Discourse Markers (DMs) by taking *and*, *but,* and *so* in the utterances of TED talks as the research objects. The metrics adopted in this study include dependency distance (DD) and mean dependency distance (MDD). DD refers to the linear distance between two linguistic units with a syntactic relationship (Heringer et al. 1980; Hudson 1995; Jiang and Liu 2015), and MDD is the DD average of certain dependency types, sentences, texts, etc. (Liu 2008; Wang and Liu 2017; Yan and Liu 2022). Since the proposal of the psychological reality of syntactic dependency structures (Hudson 2003), empirical evidence demonstrates these indicators' correlation with working memory capacity (Liu 2008; Niu and Liu 2022). Thus, they are measures of syntactic complexity as well as the memory burden of language processing, reflecting the dynamic cognitive load of language (Hudson 1995; Liu 2008). With the guidance of such theories, on the one hand, we may quantitatively probe into the syntactic relationship between DMs and the rest of the sentence; on the other hand, we may also revisit the underlying processing mechanism of DMs through text-based analysis.

Based on the distance-driven measures of DD and MDD, we found that the target sentences (sentences with a DM at the initial position) are syntactically more complex than the common sentences without any DMs. Moreover, the followed-up sentences after the initial-posited DMs are syntactically more complex than common sentences due to the specific structures, namely, complement clauses and adverbial clauses modifiers. The findings demonstrate the syntactic significance of DMs to natural sentences, hinting at the relationship between DMs' syntactic role and human cognitive mechanism. Specifically, the sentence-initial DMs are prone to co-

occur with complex syntax, denoting the upcoming difficulties with heavier processing burdens. In other words, DMs may serve as an early warning mechanism for higher processing difficulty.

Keywords: Discourse Markers; Quantitative Analysis; Dependency Distance; Syntactic Complexity; TED Talks

Reference

Blakemore, Diane. 1992. *Understanding utterances*. Oxford: Blackwell.

Blakemore, Diane. 2006. Discourse markers. In Laurence R. Horn & Gregory L. Ward (eds.), *The Handbook of Pragmatics*, 221-240. Oxford & Malden: Blackwell.

Fraser, Bruce. 1996. Pragmatic markers. *Pragmatics* 6: 167-190.

Fraser, Bruce. 1999. What are discourse markers?. *Journal of Pragmatics* 31(7): 931-952.

Heine, Bernd. 2013. On discourse markers: Grammaticalization, pragmaticalization, or something else? *Linguistics* 51(6): 1205-1247.

Heringer, Hans-Jürgen, Bruno Strecker & Rainer Wimmer. 1980. *Syntax*: *Fragen-Lösungen-Alternativen*. München: Wilhelm Fink Verlag.

Hudson, Richard. 1995. *Measuring Syntactic Difficulty*. Manuscript. London: University College London.

Hudson, Richard. 2003. *The psychological reality of syntactic dependency relations*. MTT2003, Paris.

Jiang, Jingyang & Haitao Liu. 2015. The effects of sentence length on dependency distance, dependency direction and the implications–based on a parallel English–Chinese dependency treebank. *Language Sciences* 50: 93-104.

Liu, Haitao. 2008. Dependency distance as a metric of language comprehension difficulty. *Journal of Cognitive Science* 9(2): 159-191.

Maschler, Yael & Deborah Schiffrin. 2015. Discourse Markers: Language, Meaning, and Context. In Deborah Tannen, Heidi E. Hamilton & Deborah Schiffrin (eds.), *The Handbook of Discourse Analysis* (ed.2), 189-221. West Sussex: John Wiley & Sons.

Mel'čuk, Igor. 2014. Dependency in language. In Kim Gerdes, Eva Hajičová & Leo Wanner (eds.), *Dependency Linguistics: Recent Advances in Linguistic Theory Using Dependency Structures*, 189-221. Amsterdam: John Benjamins.

Niu, Ruochen & Haitao Liu. 2022. Effects of syntactic distance and word order on language processing: An investigation based on a psycholinguistic treebank of English. *Journal of Psycholinguistic Research* 51(5): 1043-1062.

Redeker, Gisela. 1991. Linguistic markers of discourse structure. *Linguistics* 29: 1139-1172.

Sakita, Tomoko I. 2013. Discourse markers as stance markers: Well in stance alignment in conversational interaction. *Pragmatics & Cognition* 21(1): 81-116.

Tesnière, Lucien. 2015. *Elements of Structural Syntax.* Translated by Timothy Osborne & Sylvain Kahane. Amsterdam: John Benjamins.

Wang, Yaqin & Haitao Liu. 2017. The effects of genre on dependency distance and dependency direction. *Language Sciences* 59:135-147.

Yan, Jianwei & Haitao Liu. 2022. Semantic roles or syntactic functions: The effects of annotation scheme on the results of dependency measures. *Studia Linguistica* 76(2):406-428.

Zwicky, Arnold M. 1985. Clitics and particles. *Language* 61: 283-305.

Uncovering the Relationships Among Slavic Languages: A Lexical Diversity Analysis

Chenliang Zhou[1], Junyi Xu[2*]

[1]Department of Linguistics, Zhejiang University, Hangzhou, China
[2]School of Literature, Zhejiang University, Hangzhou, China
*** Correspondence:** 12004022@zju.edu.cn
Keywords: Quantitative Typology, Slavic Classification, Lexical Diversity, Morphological Continuum, Geographical Distribution

Currently, there are 7,151 recognized languages in use worldwide (Eberhard et al., 2022). The field of linguistics has long aimed to uncover the underlying patterns and regularities across these manifold languages. Through decades of research, linguists have come to understand the significant diversity that exists at multiple levels across languages. This diversity highlights the importance of language classification. In order to fully grasp the diversity of languages, it is essential to consider their phylogenetic relationships - a key characteristic of their properties and categories. As a result, genealogical classification is widely considered to be a reliable method to language classification.

One of the well-established example is the traditional classification scheme for Slavic languages, which separates the Slavic group into East, West, and South Slavic branches (Robins, 1973; Corbett & Comrie, 1993; Hymes, 1993; Campbell & Poser, 2008). This scheme is widely accepted in historical and comparative linguistics.

Recent advancements in computational techniques have also allowed for more efficient and effective analysis of large language datasets, further expanding our understanding of the complexity and diversity of human language. In this study, we propose a novel linguistic classification method based on quantitative typology, utilizing a large-scale multilingual parallel corpus. This approach aims to eliminate the influence of covariates such as text genre and semantic content in cross-language comparisons, resulting in more valid language classification results.

Specifically, we investigate the following questions:

1. Can lexical diversity metrics reflect the morphological types, specifically the degree of “analytism-synthetism” of morphology?

2. What is the difference between our quantitative classification scheme based on lexical diversity and the genealogical classification of Slavic languages?

3. How does the classification of Slavic languages based on lexical diversity relate to the geographical distribution of languages?

To answer these questions, we model the type-token relationships of each Slavic parallel text and calculate characteristic values of lexical diversity to approximate the morphological complexity of the language. We then use this information to perform automatic clustering of languages based on integrated lexical diversity metrics. Our results indicate that lexical diversity metrics can accurately reflect a language’s position on the “analytism-synthetism” continuum and that automatic clustering based on these metrics in Slavic languages can effectively reflect their genealogical classification (see Fig.1(a)). Additionally, we observe a monotonic increasing trend in the geographical distribution of Slavic languages from southwest to northeast (see Fig.1(b1-b3)), consistent with patterns found in previous studies (Bentz, 2018; Nichols & Bentz, 2018).

The methodological approach taken in this study is data-driven and theoretically independent, making it amenable to computer processing. Our findings contribute to a better understanding of corpus-based typology and could offer new insights into the understanding of language as a human-driven complex adaptive system.

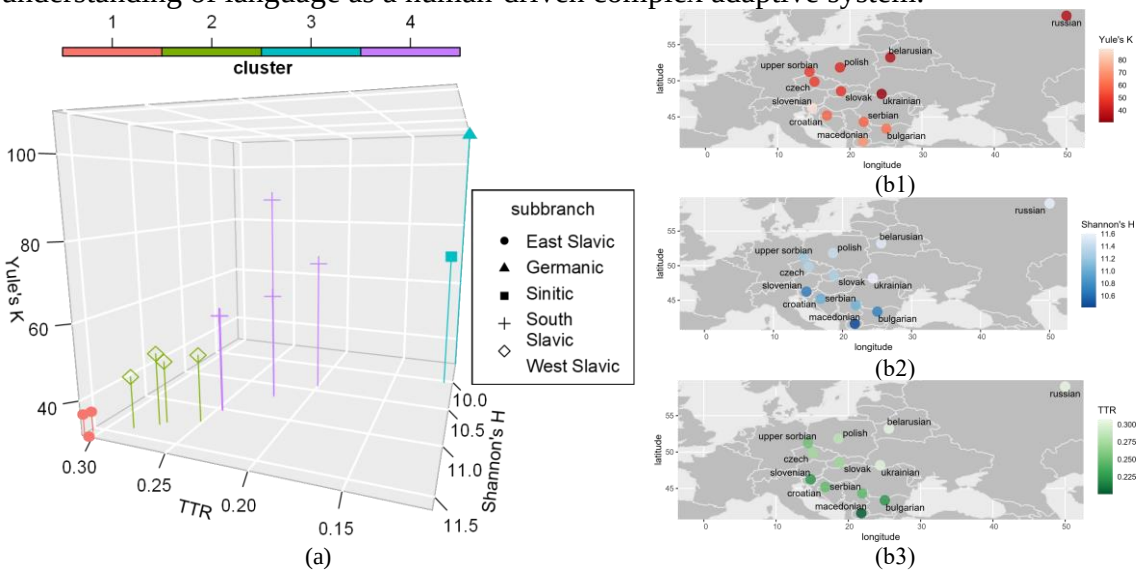


Fig. 1 (a) The multidimensional clustering based on lexical diversity metrics; the geographical distribution of lexical diversity metrics: (b1) Yule's K, (b2) Shannon's H, (b3) TTR.

References

Bentz, C. (2018). Adaptive Languages: An Information-Theoretic Account of Linguistic Diversity. In *Adaptive Languages* (Vol. 316). De Gruyter Mouton. https://doi.org/10.1515/9783110560107

Campbell, L., & Poser, W. J. (2008). *Language classification: History and method*. Cambridge University Press.

Corbett, P. G., & Comrie, P. B. (1993). *The Slavonic Languages*. Routledge. https://doi.org/10.4324/9781136861376

Eberhard, D. M., Simons, G. F., & Fennig, C. D. (2022). *Ethnologue: Languages of the World* (25th ed.). SIL International. https://www.ethnologue.com/

Hymes, D. H. (1993). Genetic Classification: Retrospect and Prospect. *Anthropological Linguistics*, *35*(1/4), 21–37. https://www.jstor.org/stable/30028240

Nichols, J., & Bentz, C. (2018). Morphological complexity of languages reflects the settlement history of the Americas. In K. Harvati, G. Jäger, & H. Reyes-Centeno (Eds.), *New Perspectives on the Peopling of the Americas*. Kerns Verlag. https://helda.helsinki.fi/handle/10138/312062

Robins, R. H. (1973). The History of Language Classification. In H. M. Hoenigswald (Ed.), *Diachronic, areal, and typological Linguistics* (pp. 3–42). De Gruyter. https://doi.org/10.1515/9783111418797-003

Ján Mačutek[1], Emmerich Kelih[2], Michaela Koščová[3]
[1]Mathematical Institute, Slovak Academy of Sciences / Constantine the Philosopher University in Nitra
[2]Institute of Slavonic Studies, University of Vienna
[3]Mathematical Institute, Slovak Academy of Sciences

A quantitative approach to noun declension in Slavic languages

Key words: declension, frequency, noun, Slavic, morphology

The presentation focuses on the noun declension in Slavic languages (Czech, Russian, Slovak, and Slovene). These languages have a relatively rich inflectional morphology, where the case is mostly expressed by variety of suffixes added to the basic word form (see Haspelmath and Sims 2010 for the relation between basic word forms and inflected word forms in general). For the sake of comparability of results obtained we use texts from a Slavic parallel corpus. We thus generalize findings from Mačutek and Čech (2013) and Mačutek et al. (2023).

We propose a method for the quantification of morphophonetic changes of nouns with respect to their basic forms. We show that the magnitude of change correlates negatively with frequency of word forms, i.e. words with lesser changes occur more often. In addition to morphophonetic changes, the categories of grammatical gender (masculine, feminine, and neuter) and animacy (being a subcategory of gender, cf. Klenin 2009) play a very important role, whereby animacy seems to have a decisive impact on frequency behaviour.

The results give us the possibility to incorporate morphologic features into a synergetic language model, where word frequency, word length, grammatical case and gender, and animacy seem to influence each other.

References

Haspelmath, M., Sims, A.D. (2010). *Understanding Morphology*. London: Routledge.

Klenin, E. (2009). Animacy, personhood. In: Kempgen, S., Kosta, P., Berger, T., Gutschmidt, K. (eds.), *Die slavischen Sprachen. Ein internationales Handbuch zu ihrer Geschichte, ihrer Struktur und ihrer Erforschung. Band 1* (pp. 152-161). Berlin, New York: de Gruyter.

Mačutek, J., Čech, R. (2013). Frequency and declensional morphology of Czech nouns. In: Obradović, I., Kelih, E., Köhler, R. (eds.), *Methods and Applications of Quantitative Linguistics* (pp. 59-68). Beograd: Akademska Misao.

Mačutek, J., Koščová, M., Kelih, E., Čech, R. (2023). Frequency and morphological behaviour of nouns in Czech and Russian. *Bohemistyka* 23(1), 110-118.

Acknowledgment

Supported by research projects APVV SK-AT-20-0003 (J. Mačutek, E. Kelih, M. Koščová) and VEGA 2/0096/21 (J. Mačutek, M. Koščová).

Distribution of syntactic functions in different styles and genres

Miroslav Kubát[1], Radek Čech[1], Xinying Chen[2]

[1] University of Ostrava

[2] Xi'an Jiaotong University

Keywords: stylometry, syntax, genre, Czech, corpus

Abstract

Quantitative linguistics is concerned with analyzing the distribution of various linguistic units in order to understand general patterns of language use for decades. A notable example is Zipfian rank-frequency distribution of words and other units across different languages. However, language usage differs based on the communication situation and the type of text (style, genre, etc.). Therefore, it is worthwhile to study the differences of distribution of linguistic units between different types of texts.

We aim to study the distribution of syntactic functions in different styles and genres in Czech language. First, we find a suitable model for our data. Second, we study how the parameters of the model and the overall fit expressed by coefficient of determination R^2 varies across different styles and genres. Distribution of syntactic functions in general and their behavior in different text types have not been much studied yet especially because of lacking sufficient data. However, contemporary corpus linguistics provide language material with satisfactory syntactic annotation nowadays.

The dataset of this research comes from Czech National Corpus, namely the corpus SYN2020 (Křen et al. 2020). It is a large balanced corpus of contemporary written Czech consisting of three main text types (fiction, non-fiction, journalism) and various genres (e.g. novel, short story, poetry, drama, administrative texts, humanities, social sciences, newspapers, leisure magazines) with the total size of 100 million words. The corpus is lemmatized, morphologically and syntactically annotated (Jelínek et al. 2021). Syntactic annotation was performed using a parser from the NeuroNLP toolkit trained on the data of the Prague Dependency Treebank (Bejček et al. 2012) and the fiction corpus FicTree (Jelínek 2017).

The results show that frequencies of syntactic functions fit well the exponential probability distribution $y = a\, e^{-bx}$. More specifically, coefficient of determination $R^2 \geq 0.95$ in all genres (with exception of poetry where $R^2 = 0.937$). In terms of obtained values of the parameters (a, b), we can see quite clear pattern distinguishing genres. Genres such as poetry, drama, novels, short stories belonging to fiction tend to have lower values of parameters (a, b) while non-fiction genres like administrative texts or professional literature have higher values. Journalism texts like newspapers and leisure magazines are somewhere between fiction and non-fiction literature. We can therefore conclude that it is important to pay attention to the language material in terms of various text types when studying probability distributions of syntactic functions (and perhaps other linguistic units in general). The comparison of parameters seems to be a suitable approach for stylometric research.

References

Bejček, E., Panevová, J., Popelka, J., Straňák, P., Ševčíková, M., Štěpánek, J., Žabokrtský, Z. (2012). Prague Dependency Treebank 2.5 – a revisited version of PDT 2.0. In: *Proceedings of the 24th International Conference on Computational Linguistics (Coling 2012)*. Mumbai, pp. 231–246.

Jelínek, T. (2017): FicTree: a Manually Annotated Treebank of Czech Fiction. In: Hlaváčová, J. (ed.), *Proceedings of the 17th Conference on Information Technologies - Applications and Theory (ITAT 2017)*, pp. 181–185. http://ceur-ws.org/Vol-1885/181.pdf

Jelínek, T., Křivan, J., Petkevič, V., Skoumalová, H., Šindlerová, J. (2021): SYN2020: A new corpus of Czech with an innovated annotation. In: K. Ekštein, F. Pártl, M. Konopík (eds.), *Text, Speech, and Dialogue. TSD 2021. Lecture Notes in Computer Science*, vol. 12848. Cham: Springer, pp. 48–59.

Křen, M., Cvrček, V., Henyš, J., Hnátková, M., Jelínek, T., Kocek, J., Kováříková, D., Křivan, J., Milička, J., Petkevič, V., Procházka, P., Skoumalová, H., Šindlerová, J., Škrabal, M. (2020). *SYN2020: representative corpus of contemporary written Czech*. Institute of the Czech National Corpus, Faculty of Arts, Charles University in Prague. Available at http://www.korpus.cz.

Acknowledgment

Miroslav Kubát and Radek Čech were supported by the Czech Science Foundation (GAČR), project No. 22-20632S.

Menzerath-Altmann's law versus Menzerath's law as a criterion of complexity in communication

Iván González Torre[a,b], Łukasz Debowski[c], Antoni Hernández-Fernández[d,e]

[a]*Language and Speech Laboratory. Universidad del País Vasco/Euskal Herriko Unibertsitatea, c/ Justo Vélez de Elorriaga, 1, Vitoria-Gasteiz, 01006, Basque Country, Spain*
[b]*Departamento de Matemática Aplicada, Universidad Politécnica de Madrid, Avda. Puerta de Hierro, 2-4, Madrid, 28040, Spain*
[c]*Institute of Computer Science, Polish Academy of Sciences, ul. Jana Kazimierza 5, Warszawa, 01-248, Poland*
[d]*Societat Catalana de Tecnologia, Secció de Ciències i Tecnologia, Institut d'Estudis Catalans, C/Carme 47, Barcelona, 08001, Catalonia, Spain*
[e]*Institut de Ciències de l'Educació, Universitat Politècnica de Catalunya, Av. Doctor Marañón 44-50, Edifici P, Planta 3, Campus Sud, Barcelona, 08028, Catalonia, Spain*

Abstract

Menzerath's law is a quantitative linguistic law which states that the longer is a linguistic construct, the shorter are its constituents, on average. Gabriel Altmann developed a more precise mathematical formula called subsequently *Menzerath-Altmann's law* (MAL) [1]:

$$y = \alpha m^{\beta} \exp(-\gamma m) \tag{1}$$

where y is the constituent size, m is the size of the linguistic construct, and α, β, γ are empirical parameters. MAL has been explained as a manifestation of complex behavior and appeared in many biological systems [2]. In this work we study MAL for constructs being word tokens and constituents being syllables, measuring its length in graphemes [3].

First, we derive the exact form of MAL for texts generated by the memoryless source with three emitted symbols, which can be interpreted as a null model: this null model complies with Menzerath's law, but this is not the case with MAL, which predicts an inverted regime for sufficiently range constructs (i.e., the longer is a word, the longer are its syllables) not found in the memoryless source. To support empirically this claim, we analyze 21 languages from the

Preprint submitted to Qualico 2023 *January 5, 2023*

Standardized Project Gutenberg and we show the presence of the inverted regime not exhibited neither by the null model nor by Menzerath's law. This fact is also relevant to the study of animal communication, where these and other linguistic laws are being studied, often proposing complexity in these systems (see review in [2]).

Secondly, whereas the memoryless source is able to reproduce Menzerath's law, it does so at the expense of predicting that all consonant clusters within a word are equally long on average: we show that this prediction is not satisfied by human languages: the mean size of a syllable strongly depends on its position in the word. The independence of elements that compose a linguistic unit is a strong assumption that have been previously addressed not true in general [4]. We report the distribution of syllable sizes with respect to their position in the word, which might be related with the emerging MAL [3].

In conclusion, our results indicate that Menzerath's law in terms of correlations could be a spurious observation, while complex patterns in communication systems should be rather attributed to specific forms of MAL.

Keywords: Menzerath's law, Menzerath-Altmann's law, Memoryless source, Standardized Gutenberg Corpus, Syllable size, Linguistic laws

---

Funding: This work has been funded by the project PRO2023-S03 HERNANDEZ from Secció de Ciències i Tecnologia, Institut d'Estudis Catalans (https://www.iec.cat/).

References

[1] Gabriel Altmann. Prolegomena to menzerath's law. *Glottometrika*, 2(2):1–10, 1980.

[2] Stuart Semple, Ramon Ferrer i Cancho, and Morgan L. Gustison. Linguistic laws in biology. *Trends in Ecology Evolution*, 37(1):53–66, 2022.

[3] Iván G. Torre, Łukasz Debowski, and Antoni Hernández-Fernández. Can menzerath's law be a criterion of complexity in communication? *PLOS ONE*, 16(8):1–21, 08 2021.

[4] Iván G. Torre, Bartolo Luque, Lucas Lacasa, Christopher T. Kello, and Antoni Hernández-Fernández. On the physical origin of linguistic laws and lognormality in speech. *Royal Society Open Science*, 6(8):191023, 2019.

Modelling Menzerath's Law with Gaussian Copula

Jiří Milička
Institute of the Czech National Corpus, Faculty of Arts, Charles University, Prague

Menzerath's Law [3], also known as the Menzerath-Altmann Law [1], is a well-established relationship between the length of a construct and the average length of its constituents. A recent study [2] has shown that even simple stochastic processes can exhibit Menzerathian behavior. However, the model presented in the study does not fit the real-world data. This presents an opportunity to identify simple stochastic processes that can accurately model Menzerath's Law in the real world.

To study this phenomenon, we will shift our perspective slightly. Instead of focusing on the relationship between the length of a construct and the average length of its constituents, we will examine the dependence of the length of constructs (measured in terms of their constituents) on the length of the same constructs measured in terms of their subconstituents. For example, we will investigate how the number of syllables in words is related to the number of phonemes in the same words. This may appear to be a different approach, but the Menzerath-Altmann Law can still be calculated from this joint distribution. In fact, Menzerath himself measured this joint distribution in his original publication on the topic [3, p. 96].

As shown in Fig. 1, the original data collected by Paul Menzerath contains empty spaces that cannot be filled due to the definition of the relationship. For instance, there are no words with three syllables and two phonemes. However, by looking at the number of boundaries between segments rather than the number of segments directly, we can simplify the joint distribution. This is a simple transformation that can be reversed at the end. It turns out that a simple Gaussian Copula is a good model for this joint probability (Figure 1, right), thus providing a stochastic process that can effectively model Menzerath's Law (Fig. 2 top, one hundred random samples from Gaussian copula).

This stochastic process not only works well for Menzerath's original dataset, but also for many other datasets measured at different language levels and in various languages (e.g. Czech phoneme-morpheme-word level, see Fig. 2 bottom). The only input required for the Gaussian copula model is the marginal distributions (e.g. distribution of number of syllables in words and distribution of the number of phonemes in words) and the coefficient of correlation between the two variables.

Key words: Menzerath's Law, Menzerath-Altmann Law, Gaussian copula

References

[1] Gabriel Altmann. Prolegomena to Menzerath's law. *Glottometrika*, 2:1–10, 1980.

[2] Iván G. Torre, Łukasz Dębowski, and Antoni Hernández-Fernández. Can Menzerath's law be a criterion of complexity in communication? *Plos one*, 16(8):e0256133, 2021.

[3] Paul Menzerath. *Die Architektonik des deutschen Wortschatzes*, volume 3. F. Dümmler, 1954.

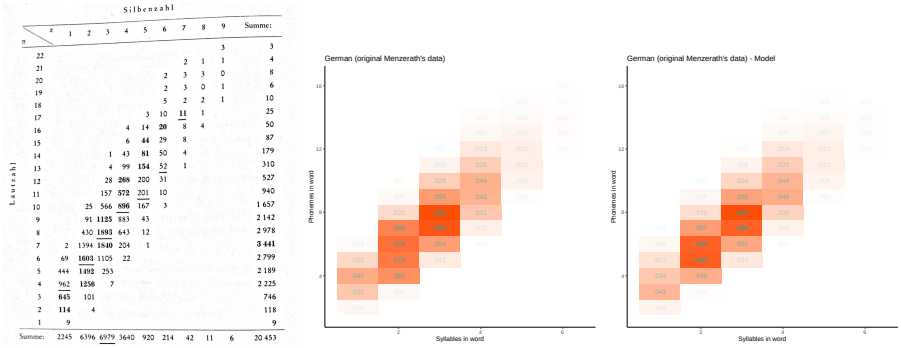


Figure 1: Original Menzerath's joint distribution of his dataset(from Menzerath [3, 96])). The chart on right represents the Gaussian copula model of the data.

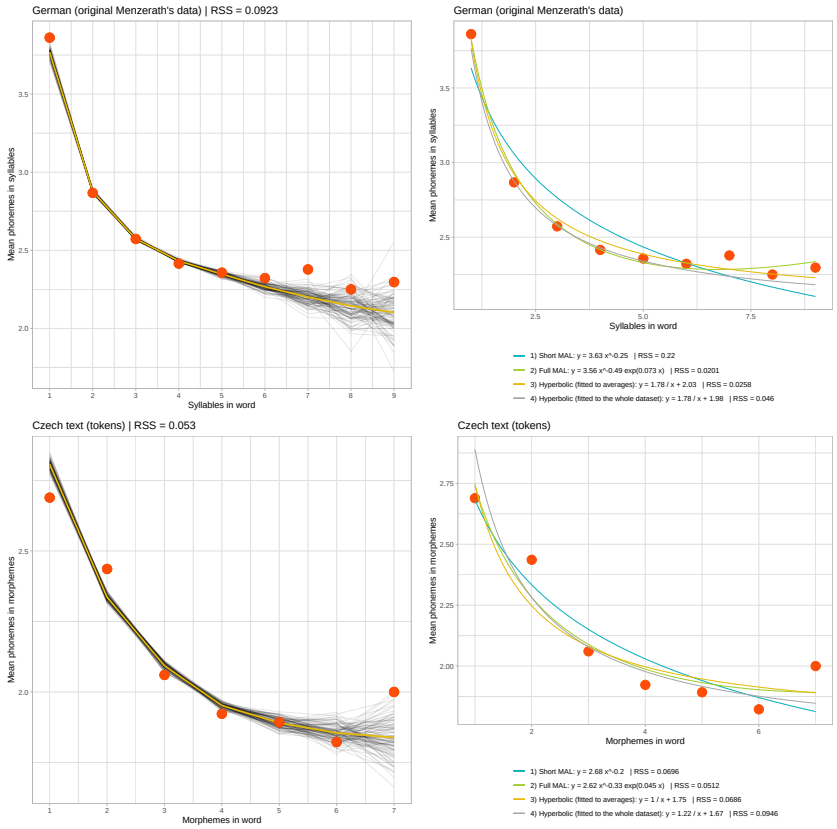


Figure 2: Original Menzerath's data made into the traditional Menzerath's Law visualization. They are modelled using Gaussian copula (left) and traditional models (right). Bottom charts represent Menzerath's law captured for phoneme-morpheme-word level in a Czech text (Krysař by Viktor Dyk).

Principled Analytic Corrections of Zipf's Law

Łukasz Dębowski* Iván González Torre†‡

The aim of the paper is to derive and verify principled corrections to Zipf's law for texts of an arbitrary size. Our derivation rests on three assumptions: The first assumption is the urn model which states that word frequency distributions for shorter texts look as if the word tokens were sampled at random from a given longer text. The second assumption is that we have an exact analytic formula for the vocabulary growth function. The third assumption is that the formulae obtained for smaller text sizes (interpolation) can be continued analytically for larger texts (extrapolation). These three assumptions were progressively developed by Khmaladze [3], Baayen [1], Milička [4], and Davis [2]. Our contribution is to derive analytic formulae for the rank-frequency function.

To fix the notation, suppose that we count words in texts. For each word w and a text $\mathbf{t} = (t_1, t_2, ..., t_n)$ we define the frequency of the word as $f(w) := \sum_{i=1}^{n} \mathbf{1}\{t_i = w\}$, where $\mathbf{1}\{\text{true}\} := 1$ and $\mathbf{1}\{\text{false}\} := 0$. The number of types is $v := \sum_{w:f(w)>0} 1$ and the number of tokens is $n := \sum_{w:f(w)>0} f(w)$. The frequency spectrum is $(v_1, v_2, ...)$, where v_k is the number of types with frequency k, namely, $v_k := \sum_{w:f(w)=k} 1$. We may express $v = \sum_{k=1}^{\infty} v_k$ and $n = \sum_{k=1}^{\infty} k v_k$. Moreover, the inverse rank-frequency function is $r_f = v - \sum_{k=1}^{f-1} v_k$.

Consider a text of length $n' < n$ sampled from the text of length n. According to the urn model [1, 4, 2], the expected number of types v' and the expected frequency spectrum $(v'_1, v'_2, ...)$ for the text of length n' are

$$v' = v - \sum_{k=0}^{\infty} v_k \frac{\binom{n-n'}{k}}{\binom{n}{k}} \approx g\left(\frac{n'}{n}\right), \qquad g(x) := v - \sum_{k=1}^{\infty} v_k (1-x)^k, \tag{1}$$

$$v'_l = \sum_{k=l}^{\infty} v_k \frac{\binom{n'}{l}\binom{n-n'}{k-l}}{\binom{n}{k}} \approx g_l\left(\frac{n'}{n}\right), \qquad g_l(x) := \sum_{k=l}^{\infty} v_k \binom{k}{l} x^l (1-x)^{k-l}, \tag{2}$$

where the approximation is valid for $k \ll n', n$ since $\binom{n}{k} := \frac{n!}{k![n-k]!} \approx \frac{n^k}{k!}$. Moreover, as observed by Davis [2], functions $g_l(x)$ can be evaluated by taking derivatives of the vocabulary growth function $g(x)$. In particular, for a given function $g(x)$, we may evaluate the expected inverse rank-frequency function as

$$r'_f = v' - \sum_{k=1}^{f-1} v'_k \approx g(x) + \sum_{k=1}^{f-1} \frac{(-x)^k}{k!} \frac{d^k g(x)}{dx^k}, \quad x := \frac{n'}{n}. \tag{3}$$

*Institute of Computer Science, Polish Academy of Sciences, ul. Jana Kazimierza 5, 01-248 Warszawa, Poland.

†Language and Speech Laboratory, Universidad del País Vasco/Euskal Herriko Unibertsitatea, c/ Justo Vélez de Elorriaga, 1, Vitoria-Gasteiz, 01006, Spain.

‡Departamento de Matemática Aplicada, Universidad Politécnica de Madrid, Avda. Puerta de Hierro, 2-4, Madrid, 28040, Spain.

Example 1. *Let the vocabulary growth function be given by Herdan-Heaps' law* $g(x) = vx^{\beta}$ *for some* $\beta \in (0,1)$. *Then* $v'_k \approx v'(-1)^{k+1}\binom{\beta}{k}$, *where* $\binom{r}{l} := \frac{r(r-1)\ldots(r-l+1)}{l!}$ *for a real number* r. *It can be proved easily that*

$$\frac{r'_f}{v'} \approx \frac{r_f}{v} \approx 1 - \sum_{k=1}^{f-1} \frac{v_k}{v} = \binom{f-\beta-1}{f-1}. \tag{4}$$

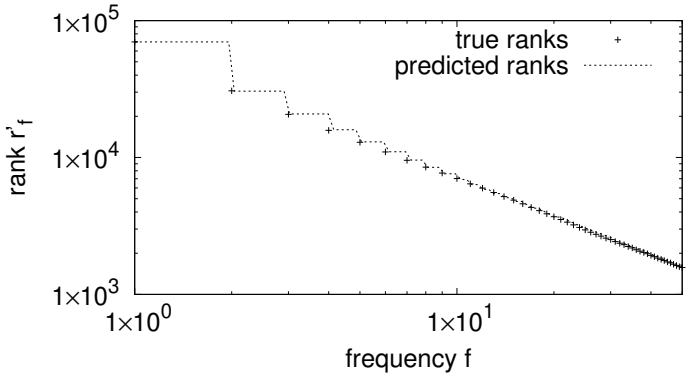
In this case, the normalized ranks r'_f/v' *do not depend on text size* n'. *For* $f \to \infty$, *formula (4) tends to Zipf-Mandelbrot's law* $r_f \propto 1/f^{\beta}$.

Example 2. *Let the vocabulary growth function be given by Davis' ansatz* $g(x) = v\frac{x\log x}{x-1}$ *[2]. Then for the text size* $n' = n$, *we have Lotka's law* $v_k \approx \frac{v}{k(k+1)}$ *and Zipf's law* $r_f \approx v/f$. *What Davis [2] did not show, we have*

$$r'_f \approx v \cdot \frac{\log x - \sum_{j=1}^{f-1}(1-1/x)^j/j}{(1-1/x)^f} \left[= v\sum_{j=0}^{\infty}\frac{(1-1/x)^j}{j+f} \text{ if } x \geq \frac{1}{2}\right]. \tag{5}$$

Formula (5) corrects the ideal Zipf law $r_f \approx v/f$ *for an arbitrary text size.*

Example 3. *Consider the linear combination* $g(x) = A\frac{x\log x}{x-1} + Bx^{\beta}$. *Then the inverse rank-frequency function is a linear combination of formulae (5) and (4). Below we present how the model fits to Shakespeare's First Folio/35 Plays. The fitted parameters are:* $A = 33187.8, B = 19506.3, \beta = 0.719654, x = 1.63819$.



We plan a more extensive empirical verification. The take-away is that it suffices to assume a certain analytic vocabulary growth function $g(x)$ to derive the inverse rank-frequency function r'_f for an arbitrary text size.

References

[1] R. H. Baayen. *Word frequency distributions*. Dordrecht: Kluwer Academic Publishers, 2001.

[2] V. Davis. Types, tokens, and hapaxes: A new heap's law. *Glottotheory*, 9 (2):113–129, 2018.

[3] E. Khmaladze. The statistical analysis of large number of rare events. Technical Report MS-R8804. Centrum voor Wiskunde en Informatica, Amsterdam, 1988.

[4] J. Milička. Type-token & hapax-token relation: A combinatorial model. *Glottotheory*, 2(1):99–110, 2009.

Getting creative: A Neural Network approach to predicting child utterances in 12 typologically diverse languages

Olivier Rüst[1], Marco Baroni[2], Sabine Stoll[1]
[1]University of Zürich, Department of Comparative Language Science, [2]Universitat Pompeu Fabra

Keywords: language acquisition, typologically diverse languages, neural network

How do children go from partially productive speech formulas to fully productive language? Research until now has focused on a very narrow definition of such formulas, lacks cross-linguistic support and has never investigated how adults use such formulas in comparison (Hartmann et al., 2021; McCauley & Christiansen, 2019). We introduce a novel, generalized method to investigate this phenomenon, based on the idea of predictability of new utterances based on previously used and heard ones.

We ask to what degree child utterances can be predicted by previously used and heard utterances. So far, research has relied on predefined operations and patterns to reconstruct utterances. Here we introduce a new approach using neural language models, which recognize probabilistic co-occurrences of words or morphemes.

The underlying idea is that highly formulaic language is easier to predict than a creative one due to more predictable co-occurrences of words. We investigate the emergence of productive language in 12 typologically extremely diverse languages (61 children aged 0;7-6;0) from the ACQDIV database (Moran et al., 2019), which allows for language-wide generalizations.

In Study 1, we train neural language models (LSTMs) on longitudinal subsets of the same size of our corpora, to evaluate predictability of subsequent utterances. We predict *child* utterances by previous *child* and *adult* utterances separately. Model performance is evaluated with perplexity. Perplexity measures the (lack of) predictability, it is the standard measure to quantify language models' behaviour. **We hypothesize** that perplexity increases for both predictions as the child grows, since more schematic utterances are easier to predict than fully creative ones. Moreover, for child utterances we expect her previous utterances to be a better predictor than adult utterances, as adult utterances are more creative than child utterances. **We find** a statistically credible general increase in model perplexity in all 12 languages (β=35.55, *CI*: 34.19, 36.91) with Bayesian multilevel models. This indicates that children use less formulaic and more creative utterances as they grow older. We also show that child utterances are a better predictor than adult utterances.

In Study 2, we train language models as in study 1, but evaluate whether predictability of child utterances approaches adult levels over time, i.e., if adult and child utterances become similar in terms of their adherence to speech formulas. We predict *child* utterances by previous *child* utterances and *adult* utterances by previous *adult* utterances. **We hypothesize** that predictability of adult utterances stays relatively stable, predictability of child utterances decreases and thus slowly converge over time. **We find** a stable predictability of adult utterances, a progressively worse prediction of child utterances, and the convergence of predictability of adult and child utterances across all 12 languages (see figure 1; β=-22.37, *CI:* -27.10, -17.58). We interpret this as evidence that children rely continuously less on speech formulas and converge to adult levels over time.

This paper introduces a generalized method to test the trajectory from item-specific use to full productivity. We show that there is a similar learning trajectory in children from a linguistically extremely diverse background.

Word count: 500

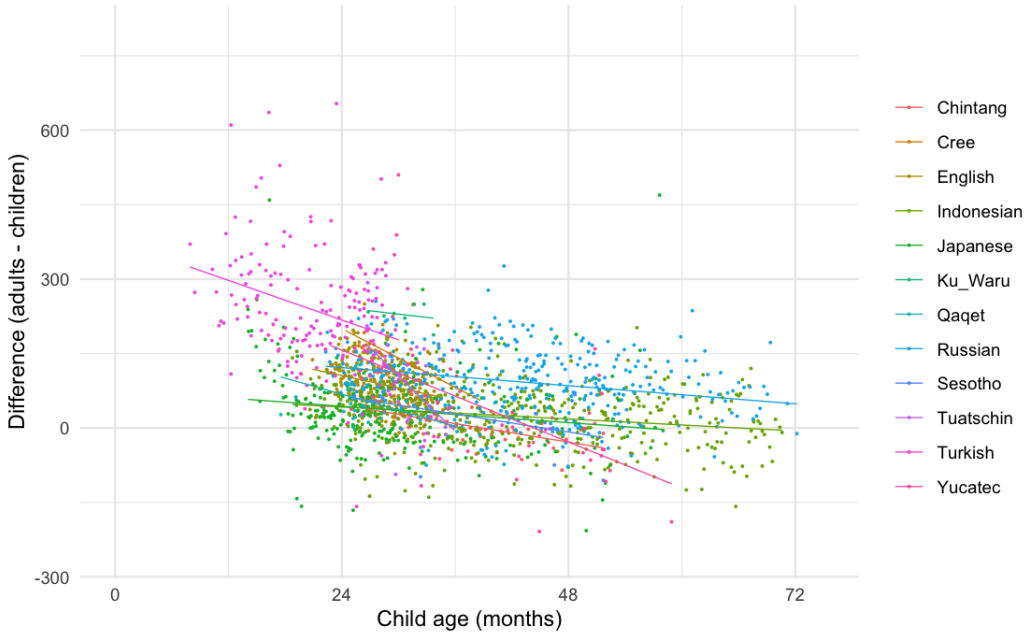


Figure 1: Results of study 2: Difference in utterance predictability (adult - children) over time.

References

Hartmann, Stefan, Nikolas Koch & Antje Endesfelder Quick. 2021. The traceback method in child language acquisition research: identifying patterns in early speech. *Language and Cognition* 13(2). 227–253.

McCauley, Stewart M & Morten H Christiansen. 2019. Language learning as language use: A cross-linguistic model of child language development. *Psychological review* 126(1). 1.

Moran, Steven, Robert Schikowski, Danica Pajović, Cazim Hysi & Sabine Stoll. 2019. The ACQDIV Corpus: a comparative longitudinal language acquisition corpus. Version 1.0.

Entropic analyses of the Voynich Manuscript using a diverse cross-linguistic corpus and neural networks

Julia Łukasiewicz-Pater [1], Ximena Gutierrez-Vasquez [2], Christian Bentz [3]

Keywords: Voynich manuscript, transliterations, entropy, language model

Introduction The Voynich manuscript is an alleged medieval codex shrouded in mystery. It remains undeciphered, despite decades of efforts. There is no consensus in terms of its origin, which language or code it was written in, and whether it is not simply a hoax.

One method to compare Voynichese with natural languages is the application of entropic measures which reflect the amount of information carried by a given linguistic symbol (e.g. a character, or an n-gram). Several analyses in this direction were carried out, and it seems that a consensus is emerging: compared to many natural language texts, Voynichese is more predictable at the character level, i.e. its entropy (at the unigram and bigram level) is surprisingly low. Such results were achieved e.g. by Zandbergen (2010)[4], Bennet (1976) – who compared Voynichese to i.a. four European languages –, and most recently by Bowern and Lindemann (2021) who contrasted it with a wider variety of scripts.

Our research question is whether we are able to replicate these results while applying a greater variety of textual resources, transliterations of Voynich, and entropy estimators.

Analyses

- Data
  - We use all three (near) complete and openly available transliterations (TT: Takeshi Takahashi; v101: Voynich 101 by Glen Claston; ZL: Zandbergen-Landini). Two of them (TT and ZL) are based on the so-called Extensible Voynich Alphabet (EVA), while the v101 alphabet is a genuinely different attempt at capturing unique characters[5].
  - We examine so-called Voynichese A and B (alleged underlying languages of the manuscript) separately.
  - We use a cross-linguistic corpus for 89 typologically diverse languages and overall more than 20k texts of different registers and styles (Moran *et al.* 2022).
- Methods
  - We apply two methods of estimating entropies: (1) Maximum Likelihood (ML), (2) calculating entropy rates based on estimations of the probability of n-gram sequences with a feedforward neural network language model.

All data and code for entropic analyses, as well as further analyses, e.g. of the distribution of word frequencies, can be found in our GitHub repository (`https://github.com/christianbentz/VoynichQuantLing`).

Results Figure 1 displays the results of our entropic analyses. What is most striking is the apparent influence of the transliteration used. Namely, the v101 transliterations clearly fall into the range of natural language texts. For instance, in terms of unigram entropy (ML), they fall right next to Fijian, Quechua, and Paiwan, while the TT and ZL transliterations (both based on EVA) are further shifted to outlier values, and fall outside the range of natural languages in terms of bigram entropies (ML). This indicates that the entropic results strongly depend on the conceptualization of what a single character constitutes in Voynichese. As there is no definitive answer as to which transliteration variant is the most appropriate, not acknowledging different approaches to this matter may introduce bias and thus blur the true picture of the manuscript. Unless there is good reason to exclude v101 transliteration, our result should be kept in mind when trying to find a coding scheme/generative process which fits the manuscript.

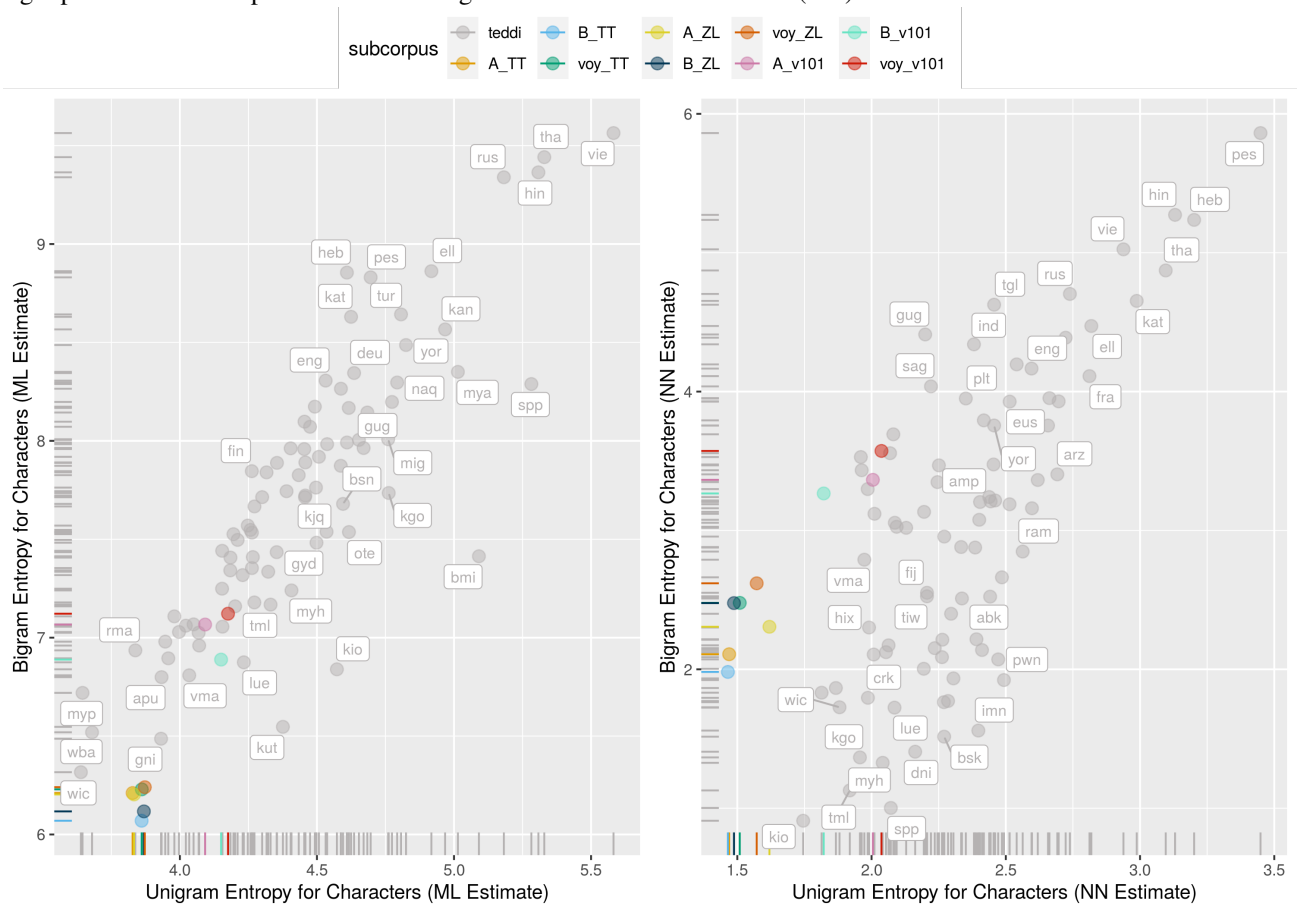
References • **Bennet, W. R.** (1976). *Scientific and Engineering Problem-Solving with the Computer*. NJ: Prentice Hall. • **Bowern, C. L. and Lindemann, L.** (2021). The linguistics of the Voynich Manuscript. *Annual Review of Linguistics*, **7**(1), 285–308. • **Moran, S., Bentz, C., Gutierrez-Vasques, X., Pelloni, O., and Samardzic, T.** (2022). TeDDi sample: Text data diversity sample for language comparison and multilingual NLP. In *Proceedings of the Thirteenth Language Resources and Evaluation Conference*, pp. 1150–1158. European Language Resources Association. • **Zandbergen, R.** (2010). The voynich manuscript. `http://www.voynich.nu/index.html`. Accessed: 2022-12-19.

[1]University of Warsaw
[2]University of Zurich
[3]University of Tübingen

[4]http://www.voynich.nu/a2_char.html

[5]See `http://www.voynich.nu/transcr.html`

Figure 1: The left panel gives unigram and bigram entropies calculated using Maximum Likelihood (ML), whereas the right panel shows entropies calculated using a feedforward neural network (NN).



Text Analysis Using Convolutional Neural Networks with Multi-Head Attention

Guaresi M., Haris S., et Vanni L.

Côte d'Azur University / CNRS (BCL - UMR 7320)

Abstract

In recent years, deep learning approaches have shown remarkable performances in many tasks related to natural language processing (automatic translation, generation, prediction...). These methods get their strength from their ability to automatically extract information from training data (Wallace et al., 2020). Trained models are based on an abstraction of the data allowed by different types of architecture such as Convolutional Neural Networks (CNN) or Transformers. Convolutional networks are well known for their performance in detecting local saliency (collocation, n-grams, repeated segments) in texts (vanni et al. 2018). Distant relations between words are better detected by the self-attention mechanisms used by Transformers (Vaswani et al. 2017). Data abstraction also depends on the training task. Automatic generation and translation of texts are mostly based on language modeling, whereas classification identifies contrasts between texts.

In this contribution, we propose a text analysis approach based on the complementarity of CNN and Multi-Head Attention layers to explore new linguistic markers. Using a text classification task in a corpus-driven approach (Tognini-Bonelli, 2002), we combine the automatic feature extraction provided by CNN (local saliency) and self-attention scores (long distance dependencies) to detect complex linguistic patterns as markers responsible for the classification decision. Through this hybrid architecture, we hypothesize that deep neural networks can identify significant features combining the syntagmatic and paradigmatic axes of texts (Lapesa et al., 2014). To capture complex patterns, we add multi-channel encoding of the data to map words into three levels: graphical form, lemma, and part of speech.

To test our approach, we apply this model to different types of political and media corpora. We focus on French presidential speeches and on gendered representations in movie scripts. We will demonstrate that the deep patterns derived from the trained models can be used to describe unknown features of current presidential discourse (such as the care discourse) or to identify complex stereotypes of female characters' lines in films during mixed and non-mixed interactions.

Keywords : Text analysis, Deep learning, Convolutional Neural Networks, Transformers, Multi-Head Attention, Corpus-driven approach, Political discourse, Gender representations

References

L. Vanni, M. Ducoffe, D. Mayaffre, F Precioso, D. Longrée, V. Elango, N. Santos, J. Gonzalez, L. Galdo, and C. Aguilar. 2018. Text deconvolution saliency (tds): a deep tool box for linguistic analysis. In 56th Annual Meeting of the Association for Computational Linguistics (ACL), pages 548–557, Melbourne.

Vaswani, Ashish and Shazeer, Noam and Parmar, Niki and Uszkoreit, Jakob and Jones, Llion and Gomez, Aidan N and Kaiser, \L ukasz and Polosukhin, Illia. 2017. Attention is All you Need. In Advances in Neural Information Processing Systems 30 (NIPS 2017), California.

Gabriella Lapesa, Stefan Evert, and Sabine Schulte im Walde. 2014. Contrasting syntagmatic and paradig- matic relations: Insights from distributional seman- 706
tic models. In Proceedings of the Third Joint Conference on Lexical and Computational Semantics (*SEM 2014), pages 160–170, Dublin, Ireland. Association for Computational Linguistics and Dublin City University.

Elena Tognini-Bonelli. 2002. Corpus linguistics at work. Computational Linguistics, 28:583-583

Capturing Distinctiveness: Transparent Procedures to Escape a Pervasive Black-Box Propensity

Matilde Trevisani

Dipartimento di Scienze Economiche, Aziendali, Matematiche e Statistiche "Bruno de Finetti" – Università degli studi di Trieste, email: matilde.trevisani@deams.units.it

Arjuna Tuzzi

Dipartimento di Filosofia, Sociologia, Pedagogia e Psicologia applicata – Università degli studi di Padova, email: arjuna.tuzzi@unipd.it

Abstract

In many quantitative linguistics applications scholars are interested in identifying a set of linguistic features which proves distinctive for a text or a class of texts with reference to a corpus or a model. Ordinary features are lexical-based elements (words, multi-words, n-grams, lemmas), part of speech categories, or further phonetic and morphosyntactic phenomena. Moreover, in many applications a set of distinctive features is selected *a priori* in order to achieve a qualitative reading of the texts or to be exploited in text clustering, topic modelling or content mapping tasks.

Classification based on supervised Machine Learning (ML) algorithms is commonly used to classify texts (test set) on the basis of training data (training set). Thanks to large amounts of available, mixed, undifferentiated, multilevel, multilayer, and multipurpose features, ML generally provides an effective way to discriminate among existing classes and, then, to ascribe each new text to one of them. Although the accuracy of classification is often highly satisfactory, the distinctive features of each class remain only seldom explainable and transparent.

The need to move from *black-box* procedures to explainable methods is at the basis of the distinction between ML and Statistical Learning (SL) approaches. Both SL and ML exploit data to make predictions but SL aims at a more in-depth understanding of data structures and relations among variables. From this perspective, SL methods capable of identifying the distinctive features of each class should interact with the solutions offered by ML algorithms in order to achieve a description in terms of linguistic similarities and differences.

The umbrella term *keyness* is often used in text analysis to refer to different measures that reveal to what extent a word can be considered distinctive of a text or a text class. Many measures have been developed to meet the requirements of different perspectives, e.g. term frequency-inverse document frequency (TFIDF), log-likelihood and odd ratios, p-values based on the hypergeometric model (to mention just a few) and methods for keyword extraction, distance-based measures as well as solutions provided by Bayesian approaches and generative (topic) models.

Starting from an established corpus of institutional speeches (corpus of End-of-Year Addresses of the Italian Presidents of the Republic 1949-2022) arranged by President-classes, this study explored the concept of *keyness* to highlight the strengths and weaknesses of different approaches, their consistency (overlapping) and how they can be applied in practice, particularly when working with large corpora. As most procedures are grounded on the observation of the occurrences reported in a term-document matrix (TDM), where terms represent features and documents represent texts or classes, most measures should tackle data normalization and dispersion problems (e.g. a linguistic feature should not be considered distinctive of a text as a whole when it occurs only within a specific portion, or of an entire class when it occurs only in one

or a limited number of its texts). This work also shows to what extent procedures that exploit equal-sized text chunks samples and tailor-made normalizations of raw frequencies (with related diagnostic measures) play a fundamental role in improving results.

References

Gabrielatos, C. (2018). Keyness Analysis: nature, metrics and Techniques. In Taylor C. and Marchi A. *Corpus Approaches To Discourse: A Critical review.* pp. 225-258. Oxford: Routledge.

Gries, S. T. (2021). *Statistics for linguistics with R: A practical introduction*. Walter de Gruyter.

Hastie, T., Tibshirani, R., & Friedman, J. (2009). *The Elements of Statistical Learning: Data Mining, Inference, and Prediction* (Second Edition). New York: Springer-Verlag.

James, G., Witten, D., Hastie, T., & Tibshirani, R. (2021). *An Introduction to Statistical Learning with Applications in R* (Second edition). New York: Springer.

Lebart, L., Salem, A., & Berry, L. (1998). *Exploring textual data*. Dordrecht: Kluwer Academic Publication.

Linguistic image of selected decimal classification categories in large bibliographies. Comparative analysis of representative languages of Central Europe and Scandinavia

Lars Johnsen[1], Adam Pawłowski[2] and Tomasz Walkowiak[3]

[1]National Library of Norway
[2]University of Wrocław
[3]Wrocław University of Technology

Large bibliographies are not only an information retrieval tool, but also a linguistic resource that can be used the same way as a text corpus. They have the advantage of very accurate and systematically prepared metadata, including the time of creation of a given micro-text (title), author (with the possibility of gender recognition), place and - very importantly - subject category. Such categories are coded with keywords, but also with the decimal universal classification symbols. So, in the perspective of linguistics, decimal categories correspond to semantic ones. The subject of the paper will be the analysis of corpora, consisting of titles of publications (books), included in selected categories of the universal decimal classification. Large bibliographic databases, functionally corresponding to national bibliographies, will be used as material. Such corpora are created by national libraries and sometimes made available through an API interface. The Polish and Czech databases were considered representative for Central Europe, while the Norwegian, Finnish, Danish and Swedish databases represent languages of Scandinavia. The database that in some sense spans Central Europe and the resources of Western European countries is the German bibliography. Among other things, it includes publications from the territories of Central Europe that lie beyond the borders of today's Germany. The volumes of databases that can be generated from large bibliographies are entirely sufficient to allow general conclusions to be inferred from them. The German database, for example, consists of about 5 million records, which, multiplied by the average length of a title, forms a serious text corpus. The essence of the study will be to compare the representation of certain subject fields (e.g. medicine, theology, mathematics, metaphysics, politics, etc.) in different languages. Titles in a large corpus conceptualize these subject areas in some sense, and the large volume of data allows the use of quantitative methods for vocabulary extraction. However, it is not obvious how this conceptualization

takes place in different languages. A comparative study of bibliographic corpora will precisely determine whether the sets of representative lexemes of each category are the same or different in different languages. As a working hypothesis, we accept the statement that the conceptualization of subject areas, the core of which is scientific writing, will be similar in different languages, despite cultural or religious differences. Extraction of distinctive lexemes will be carried out using class-based TF-IDF and multilingual word embeddings.

Word count: 498 (abstract proper only, excluding references, title, affiliation information)

Bayesian and frequentist approaches to explaining (and predicting) morphosyntactic variation in East Asia using social media data

Wilkinson Daniel Wong Gonzales
Department of English
The Chinese University of Hong Kong
Hong Kong SAR, China

Variation is ubiquitous in (natural) language (Weinreich et al. 1968; Labov 1972), and being able to explain and predict it has long been the holy grail of linguists and industry analysts alike. For linguists, the implications of such a 'solution' lean towards the theoretical. Understanding how variation is systematic sheds light on the properties of language, challenging notions of 'real' language as being inherently mono-stylistic and being divorced from social processes. For NLP-based industry professionals, an understanding of how variation works can result in less recognition and processing errors and improve product quality.

This paper develops statistical models of variation in East Asia, focusing on the morphosyntax of Hong Kong English (HKE) and Philippine English (PhilE) using Twitter data. It attempts to estimate the relationship between language-external factors and three salient and documented morphosyntactic features that exhibit variation in both varieties (Setter et al. 2010). Using the 123-million-word TCOEHK (HK corpus) and the 135-million-word TCOPE (PhilE corpus), I investigate whether time (i.e., year) and geography (i.e., district, region) constrain the variation in (1) the use of double comparatives (e.g., ***more*** *happi**er*** vs. *happi**er,*** ***more*** *happy*), (2) the use of infinitive in situations that require the participle (e.g., *have* ***eat*** vs. *haven* ***eaten***), and (3) the use of *will* and *shall* as modals that indicate future time reference (e.g., *he* ***will*** *know* vs. *he* ***shall*** *know*). For the last feature, I also test whether imputed age and sex – computationally derived using Deep Learning methods (Wang et al. 2019) – can explain and predict variation.

The results of my frequentist regression analyses reveal that time conditions the pattern involving double comparatives in HKE (β= -0.05, SE = 0.01, p<0.001) and the use of modals in PhilE (β= -0.06, SE = 0.02, p<0.001). Geography constrains modal-related patterns in PhilE (β= -0.48, SE = 0.14, p<0.001). Variation in modals can additionally be explained by age (β= -3.29, SE = 1.60, p<0.05). Sex by itself cannot predict modal variation (β= 23.64, SE = 27.6, p=0.392), but appear to do so only in specific Philippine regions (β= 23.64, SE = 27.6, p=0.392). Bayesian regression counterpart models using the Markov chain Monte Carlo (MCMC) method were also created for the purposes of comparison. The posterior distributions reveal some divergences from the frequentist models. For example, in the modal model, only geography (median = -0.37, 89% HDI = [-0.60, -0.09], pd = 99.93%, ps=0.92, BF=13.37) and age (median = 1.36, 89% HDI = [0.47, 1.86], pd = 100%, ps=1, BF=9.98e+09) can explain and reliably predict variation. There is strong evidence that time does not condition variation in PhilE modals (median = 0.001, 89% HDI = [-0.01,0.02], pd= 54%, ps=0, BF=0.02).

My Bayesian and frequentists results converge on the fact that time and geography are robust predictors of morphosyntactic variation in PhilE and HKE, showing that language-external factors influence and can be used to predict certain forms of morphosyntactic variation in East Asia, reflecting the sociolinguistic complexities in the region. The results have implications to linguistics and beyond.

Keywords: morphosyntactic variation, Bayesian and frequentist regression, sociolinguistic variation and change; East Asian Englishes, social media text mining, Markov chain Monte Carlo method, Deep Learning methods

References

LABOV, WILLIAM. 1972. The social motivation of a sound change. *Sociolinguistic patterns*, 251–265. New York: Academic.

SETTER, JANE.; CATHY S. P. WONG.; and BRIAN HOK-SHING CHAN. 2010. *Hong Kong English*. Dialects of English. Edinburgh: Edinburgh University Press.

WANG, ZIJIAN.; SCOTT A. HALE.; DAVID ADELANI.; PRZEMYSLAW A. GRABOWICZ.; TIMO HARTMANN.; FABIAN FLÖCK.; and DAVID JURGENS. 2019. Demographic Inference and Representative Population Estimates from Multilingual Social Media Data. *The World Wide Web Conference*.2056–2067. doi:10.1145/3308558.3313684.

WEINREICH, URIEL.; WILLIAM LABOV.; and MARVIN I. HERZOG. 1968. Empirical foundations for a theory of language change. *Directions for historical linguistics*, ed. by Winfred P. Lehmann and Yakov Malkiel, 100. Austin: University of Texas Press.

Submitted to QUALICO 2023

Assessing gender impact on paralinguistic accommodation in French WhatsApp conversations

Prakhar Gupta, Department of Language and information sciences, University of Lausanne
Elisa Pellegrino, Department of Computational linguistics, University of Zurich
Leyla Benkais, Department of Language and information sciences, University of Lausanne
Aris Xanthos, Department of Language and information sciences, University of Lausanne

Face-to-face (F2F) communication uses a wide range of means for expressing socio-emotional content, prevalently in the paralinguistic and extralinguistic domains, including prosody, facial expressions, gestures, and proxemics (e.g. Reilly & Seibert, 2003). In text-based computer-mediated communication (CMC), these cues are obviously missing, which has led to the emergence of new communicative devices which function as substitutes for F2F paralinguistic cues. While the earliest of these devices relied on orthographic and typographic conventions (Carey, 1980), the use of various graphical devices (or "graphicons", cf. Herring & Dainas, 2017) has consistently increased since the early 1980's. In particular, emojis (e.g. "☺") have become very popular in interpersonal instant messages (e.g. Dürscheid & Siever, 2017) and evidence suggests that they have taken over several communicative functions previously assigned to other paralinguistic cues (Pavalanathan & Eisenstein, 2016). The use of various cue types has also been found to differ significantly as a function of various sociodemographic variables, notably gender (e.g. Oleszkiewicz et al., 2017; Prada et al., 2018).

In conversational settings, paralinguistic cues are observed to undergo mutual or unidirectional accommodative adjustments between dialogue partners, resulting in increased or decreased similarity in paralinguistic (e.g. prosody, cf. Pardo et al., 2022) and/or extralinguistic information (e.g. facial expression and body movements, cf. Lakin, 2013; Dijksterhuis & Bargh, 2001), with remarkable inter and intra-speaker variability depending on the functions served by these adjustments and on speaker-specific characteristics. Several studies have examined the effect of speakers' gender on accommodation, with varying results depending on factors with which gender interacted in the examined conversational setting (e.g. conversational role, dialogue partners, cf. Pardo et al., 2018). It is not uncommon, however, to observe mutual accommodation in mixed gender pair (Levitan et al. 2012; Bilous & Krauss, 1988) that exceed convergence in same gender pair (e.g. Levitan et al. 2012).

Interestingly, similar patterns of mixed gender acoustic convergence have been observed in CMC. In a large corpus of private social media messages by Flemish teenagers, Hilte et al. (2022) observed that girls and boys adopt a more similar style in mixed gender talks, in terms of frequency and quality of paralinguistic cues. Whether this pattern of accommodation can also be attested at a more mature age is unclear. In adults, accommodation between--and within--gender groups may manifest itself in more diverse ways as opposed to younger age, as the interactional dynamics which determine the degree of assertiveness or affiliativeness of males and females can be more varied (Palomares et al. 2016).

With these premises, in this contribution, we examine patterns of paralinguistic accommodation in same and mixed gender pairs across different age groups in a large corpus of WhatsApp chats between French-speaking users living in Switzerland. The features analyzed include the frequency of emojis, emoticons, letter repetitions, and punctuation sequences. An approach based on the probabilistic framework proposed by Danescu-Niculescu-Mizil et al. (2011) is used to assess the extent to which mixed gender convergence is attested in adults' use of paralinguistic cues.

Keywords
Accomodation, paralinguistic cues, emoji, emoticon, WhatsApp, chats, CMC, probabilistic framework

References

Bilous, F. R., & Krauss, R. M. (1988). Dominance and accommodation in the conversational behaviors of same- and mixed-gender dyads. *Language and Communication,* 8: 183–194

Carey, J. (1980). Paralanguage in Computer Mediated Communication. *18th Annual Meeting of the Association for Computational Linguistics*, pp. 67–69.

Danescu-Niculescu-Mizil, C., Gamon, M., & Dumais, S.T. (2011). Mark my words! Linguistic style accommodation in social media. *ArXiv, abs/1105.0673*.

Dijksterhuis, A., & Bargh, J. A. (2001). The perception–behavior expressway: Automatic effects of social perception on social behavior. In M. P. Zanna (Ed.), *Advances in experimental social psychology,* Vol. 33 (pp. 1–40). Academic Press.

Dürscheid, C., & Siever, C. M. (2017). Jenseits des Alphabets – Kommunikation mit Emojis. *Zeitschrift für germanistische Linguistik,* 45(2): 256–285.

Herring, S.C., & Dainas, A.R. (2017). "Nice Picture Comment!" Graphicons in Facebook Comment Threads. *Hawaii International Conference on System Sciences*.

Hilte, L., Vandekerckhove, R. & Daelemans, W. (2022). Linguistic Accommodation in Teenagers' Social Media Writing: Convergence Patterns in Mixed-gender Conversations, *Journal of Quantitative Linguistics*, 29(2): 241–268.

Levitan, R., Gravano, A, Willson, L.,Benus, S., Hirschberg, J. and Nenkova, A. (2012). Acoustic-Prosodic Entrainment and Social Behavior. *Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pp. 11–19, Montreal, Canada, June 3-8, 2012.

Oleszkiewicz, A., Karwowski, M., Pisanski, K., Sorokowski, P., Sobrado, B., & Sorokowska, A. (2017). Who uses emoticons? Data from 86702 Facebook users. *Personality and Individual Differences*, *119*: 289–295.

Palomares, N., Giles, H., Soliz, J., & Gallois, C. (2016). Intergroup Accommodation, Social Categories, and Identities. In H. Giles (Ed.), *Communication Accommodation Theory: Negotiating Personal Relationships and Social Identities across Contexts* (pp. 123-151). Cambridge: Cambridge University Press.

Pardo, J. S., Pellegrino, E., Dellwo, V., Möbius, B. (2022). Special issue: Vocal accommodation in speech communication. *Journal of Phonetics,* 95, 101196, ISSN 0095-4470.

Pardo, J. S., Urmanche, A., Wilman, S., Wiener, J., Mason, N., Francis, K., & Ward, M. (2018). A comparison of phonetic convergence in conversational interaction and speech shadowing. *Journal of Phonetics,* 69: 1–10.

Pavalanathan, U., & Eisenstein, J. (2016). More emojis, less :) The competition for paralinguistic function in microblog writing. *First Monday,* 21(11).

Prada, M., Rodrigues, D. L., Garrido, M. V., Lopes, D., Cavalheiro, B., & Gaspar, R. (2018). Motives, frequency and attitudes toward emoji and emoticon use. *Telematics and Informatics*, 35(7): 1925–1934.

Reilly, J., & Seibert, L. (2003). Language and emotion. In R. J. Davidson, K. R. Scherer, & H. H. Goldsmith (Eds.), *Handbook of Affective Sciences* (pp. 535–558). Oxford University Press.

Disambiguating adverbs within a quantitative approach. Identification and annotation of polysemy

Corinne Rossari[1], Cyrielle Montrichard[1], Claudia Ricci[1]

[1]University of Neuchâtel

Keywords: polysemy, textometric approach, adverbs, annotation, corpus linguistics

Lexical polysemy is a real stumbling block for any digital approach to discourse. Our presentation aims to address this issue in relation to a category particularly affected by polysemy, adverbs, using quantitative methods and more specifically textometric tools (TXM, cf. Heiden *et al.* 2010) on French corpora from different genres (press, encyclopedia, political discourse). It is well known that the categories of adverbs, whether based on semantic or syntactic criteria, are particularly porous, and that the differences in the use of the same adverb can be so great that the forms seem to be more homonymous than polysemous. It is difficult to see a common point between the use of *seulement* ('only'), where the adverb seems to function as a discourse connector similar to *mais* ('but') (1), and the use where it functions as a focalization adverb similar to *uniquement* ('exclusively') (2).

1. Je viens de recevoir [un] nouveau mobile[…]. **Seulement** je ne peux ni appeler ni recevoir d'appel ? (Internet)

   I just received a new mobile phone […]. Only, I can neither call nor receive calls?

2. […] 15 % **seulement** des Brésiliens sont syndiqués. (Le Monde, 2010)

   Only 15% of Brazilians are unionized.

Such a pragmatic function is liable to concern all types of adverbs: a manner, a modal, or a temporal adverb can for instance all be used similarly to connectives:

3. J'ai mon cousin qui vient de Roumanie, **clairement**$_{\text{manner}}$ [**forcément**$_{\text{modal}}$/**maintenant**$_{\text{temporal}}$] il a pas de papiers (Internet)

   I have a cousin coming from Romania, clearly [necessarily/now] he has no papers

To address polysemy in adverbs, we will consider two complementary statistical methods, selecting, within Molinier & Levrier's (2000) list of conjunctive and disjunctive adverbs, those that, in addition to their conjunctive or enunciative function, can endorse at least another function.

(i) For each adverb, we will distinguish two positions that usually bring out a discrepancy in its semantic value: the initial position and the position in which the adverb is in the vicinity of the verb. We will manually annotate the function of the adverb in each position and examine the statistical correlations with its syntactic position, in order to assess for each adverb the possibility to proceed to an automatic annotation.

(ii) We will examine the specific linguistic environments characterizing the adverbs in each of these two positions. The more different these environments are, the more we can identify a divergence of meaning.

This second method will confirm or not the suitability of proceeding with an automatic annotation, based on the correlation between the position of the adverb and its semantics, making visible the fact that some adverbs show more discrepancy according to their syntactic position than others.

Our methodology, which combines (i) a statistically adequate correlation between the position and the semantic value and (ii) a sufficiently different specific environment of the adverb in each position, will contribute to designing a systematic annotation of the polysemous adverbs.

References :

Heiden S, Magué J.-P., and Pincemin B. (2010). TXM : Une plateforme logicielle open-source pour la textométrie – conception et développement. *JADT 2010 : 10th International Conference on the Statistical Analysis of Textual Data*. Rome, Italie. URL: http://halshs.archives-ouvertes.fr/docs/00/54/97/79/PDF/Heiden_al_jadt2010.pdf

Molinier, Ch. & Levrier, F. (2000). *Grammaire des adverbes. Description des formes en -ment.* Genève/Paris : Droz.

Syntactic strategies for null and pronominal subjects: a quantitative study

Giuseppe Samo (University of Geneva)

Keywords: Quantitative Syntax, Typology, Generative Grammar

In this presentation, we test different linguistic membership assignment classes of languages with respect to the strategies for null and prononimal subjects. To reach this goal, we follow the spirit of Quantitative Computational Syntax (Merlo 2016 and related works), which explores large-scale databases and simple computational methods to test linguistic proposals.

The linguistic facts and a syntactic account (generative) Languages vary in realizing prononimal subjects in the relevant contexts. English (among many others) is a typical case of a language in which the subject is always overt (but see Haegeman (2013), for specific registers and syntax) e.g. **(she) is writing the article*). On the other hand, in Italian (cf. Rizzi 1982), pronominal subjects can be omitted (*(lei) sta scrivendo l'articolo* 'she is writing the article'), but they can be overt when the subject is topical/contrastive/focal (cf. Rizzi (2018); Calabrese (1986); e.g. *Mario$_i$ ha incontrato Luigi$_j$ e ($\emptyset_i$ / lui$_j$) ha parlato dell'articolo* 'Mario$_i$ met Luigi$_j$ and ($\emptyset_i$ / he$_j$)) talked about the article'. A third type (but see also Holmberg and Roberts (2013) for finer descriptions) is represented by languages like Chinese (cf. Huang 1984; Frascarelli and Casentini 2019), in which the subject can be omitted if it is topical/given (e.g. *(wo) lai le* '(I)'m coming'). Proposals in generative approaches reduce the differences in terms of syntactic activation of just two functional projections. English does not allow any drop, Italian if the subject is in a I position (pro-drop languages), Chinese if the subject is in a C position (topic-drop).

The classification of WALS The World Atlas of Language Structures (WALS, Haspelmath et al. 2005) provide a fine-classification for a set of 711 languages (see the details in Dryer 2013). Languages may have (i) *subject affixes on verb* and therefore might not require the overt expression of the prononimal subject, (ii) *obligatory pronouns in subject positions*, (iii) *optional pronouns in subject positions*, (iv) *subject clitics on variable host*, (v) *subject pronouns in different positions*, or being (vi) *mixed*.

Materials & Methods We shall discuss a series of studies. In all these studies, we perform our count extracting naturally occurring examples (and counts) from the treebanks annotated under the guidelines of the Universal Dependencies (Zeman et al. 2022). We retrieved our data *via* count.grew.fr. We followed the set of languages in (Samo, 2021, table 2, 116-117), which lists set of languages in which at least one UD treebank is available as well the description in the relevant chapter (101) in the WALS. We extract our data from 44 languages, exploring the biggest Universal Dependencies (Zeman et al. (2022)) treebank for every language, avoiding parallel treebanks. Dedicated queries targetting inflected verb without subjects, verbs with prononimal and XP subjects have been implemented.

Results of study 1 and future studies Results of the first study, testing WALS labels are summarized in Figure 1. For example, we observe that *optional pronouns in subject position* might be investigated in detail to map crosslinguistic differences.

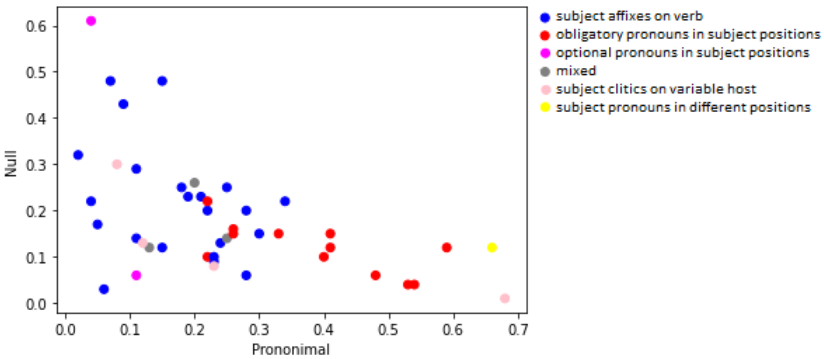


Figure 1: Distributions of pronominal and nulla subject across treebanks and types of languages according to the WALS.

Bibliography

Calabrese, A. (1986). Pronomina: some properties of the italian pronominal system. *Mit working papers in linguistics 8*(1), 46–1986.

Dryer, M. S. (2013). Expression of pronominal subjects. In M. S. Dryer and M. Haspelmath (Eds.), *The World Atlas of Language Structures Online*. Leipzig: Max Planck Institute for Evolutionary Anthropology.

Frascarelli, M. and M. Casentini (2019). The interpretation of null subjects in a radical pro-drop language: Topic chains and discourse-semantic requirements in chinese. *Studies in Chinese Linguistics 40*(1), 1–45.

Haegeman, L. (2013). The syntax of registers: Diary subject omission and the privilege of the root. *Lingua 130*, 88–110.

Haspelmath, M., M. S. Dryer, D. Gil, and B. Comrie (2005). *The world atlas of language structures*. OUP Oxford.

Holmberg, A. and I. Roberts (2013). The syntax–morphology relation. *Lingua 130*, 111–131.

Huang, C.-T. J. (1984). On the distribution and reference of empty pronouns. *Linguistic inquiry*, 531–574.

Merlo, P. (2016). Quantitative computational syntax: some initial results. *IJCoL. Italian Journal of Computational Linguistics 2*(2-1).

Rizzi, L. (1982). Negation, wh-movement, and the null subject parameter. *An Annotated Syntax Reader 169*.

Rizzi, L. (2018). Subjects, topics and the interpretation of pro. In *From sounds to structures: Beyond the Veil of Maya*. De Gruyter.

Samo, G. (2021). N-merge systems in adult and child grammars: a quantitative study on external arguments. *Quaderni di Linguistica e Studi Orientali 7*, 103–130.

Zeman, D., J. Nivre, M. Abrams, and et alia. (2022). Universal dependencies 2.11. LINDAT/CLARIAH-CZ digital library at the Institute of Formal and Applied Linguistics (ÚFAL), Faculty of Mathematics and Physics, Charles University.

Quantitative analysis of interviews in cooperation contexts: a stylometric profiling of relevant psychological processes

Alessandro Meneghini[1], Valentina Rizzoli[2] and George Markopoulos[3]

[1]Department of Philosophy, Sociology, Education and Applied Psychology, University of Padova
[2]Department of Communication and Social Research, Sapienza University of Rome
[3]Department of Linguistics, National & Kapodistrian University of Athens

This contribution is rooted within the area of sustainable development, especially for what concerns social development and reduction of inequalities, and within the capacity building approach, over the premise that development can be attained only when a common goal is pursued by relevant stakeholders that have the capabilities to work on it. Within this field, a key communicative component is the capability to involve multiple stakeholders and to define common local goals that can be pursued. Therefore, the focus of this contribution is to identify linguistic statistical markers functioning as indicators of psychosocial processes that come into play in the definition of such common goals. This research relies on Construal Level Theory – a psychosocial theory that conceptualizes the level of abstraction of an object in terms of perceived distance from the self - to categorize several conceptions of social development as high-level or low-level construal, following the assumption that the closer an "object" is in term of social or psychological distance (low level), the more concretely it is mentally construed; the further away it is (high level), the more abstractly it is construed. Therefore, having the possibility to guide the identification of personal ideas of development goals as low-level vs high-level construal using quantitative measures and statistical markers could be a key tool within the cooperation contexts, where the possibility of having several visions needs to be considered - possibly combining them. The research aim is to identify specific stylometric features associated with each construal level within text produced by people working in cooperation contexts, using statistical measures like the Part-of-Speech frequencies, word length, and sentence length. The research corpus has been composed of 20 semi-structured interviews with employees of organizations working in social development and local businesses, using questions aimed at generating high-level vs low-construal level material on topics

such as goals in development, ideal and actual strategies on how they pursue these goals as citizens vs as member of these organization, their perception of criticalities. The interviews have been transcribed and the resulting text has been divided into two categories, high vs low construal levels, according to the sections of the interviews. All text has then been processed with Python to identify the following measures: Part-of-Speech frequencies, word length and sentence length. These measures underwent a MANOVA to analyze the effect of each stylistic specificity between the two construal levels. The results will focus on the presentation of the previously indicated measures and the outcome of the MANOVA, expecting a relevance of POS tags adjectives within the high construal text, and nouns and verb tags within the low construal text. This could act as a first preliminary step in the construction of a text-based categorization system of these processes from a linguistic perspective, as well as an enrichment of the relevant psychological theory.

CapekDraCor database and some aspects of quantitative linguistic analysis of the Čapek brothers' plays

Petr Pořízka, Palacký University Olomouc, Czech Republic

Abstract

This contribution aims at presenting the CapekDraCor corpus database, a part of the international DraCor project (Drama Corpora, https://dracor.org/) containing all ten plays by Karel and Josef Čapek. We can find most of these texts in the Czech National Corpus (CNC), but they are imported into the CNC without detailed segmentation that would adequately consider the specifics of the plays and their structure. Subsequent quantitative analyses would, unfortunately, be biased for this reason, as neither the individual text layers nor the data and metadata are not distinguished from each other.

The DraCor project is a corpus database that includes texts of plays from several European languages (currently 15 sub-databases). DraCor thus represents a unique platform (data and tools) for the analysis of literary texts, which have their specificities in terms of structuring (multi-layered texts) and linguistic characteristics (standing at the borderline between written and spoken language – metatextual comments versus character dialogues).

From the processing point of view, the CapekDraCor database represents a new source of data for the linguistic analysis of the Čapek brothers' dramas: the texts are segmented into the relevant sub-layers: (1) text (character dialogues); (2) metatext and indexing of its type (situational or authorial comments, stage directions); the proper names preceding the characters' lines are also processed separately (as metatext). This data format allows a more precise, targeted analysis of the text and its subcomponents.

The talk also presents the first results of a quantitative analysis of selected linguistic phenomena of Čapek's plays through selected methods of quantitative text analysis. These methods include a quantitative content analysis focusing on thematic and semantically orientated keywords in texts, which aims at a more detailed or comprehensive description of the main topics of Čapek's dramas. Specifically, this involves classifying the extracted words into different lexical-semantic classes, expressing their lexical dispersion, and modelling the interrelationships of these prominent text units statistically by the degree of association (i.e., indicating the importance of the linkage). We further compare the dramas using stylometric multidimensional scaling of the data. This method allows us to analyse the relationships between literary characters from different linguistic perspectives within a specific work, but also to compare all texts with each other and to express the degree or nature of linguistic similarity or difference: i.e., to see in which linguistic phenomena the texts correspond or differ.

Keywords: computational literary analysis, DraCor database, drama, prominent text units, stylometry

References

DraCor database. Available from www: https://dracor.org/

Fischer, Frank, et al. (2019). Programmable Corpora: Introducing DraCor, an Infrastructure for the Research on European Drama. In *Proceedings of DH2019: "Complexities"*, Utrecht University, doi:10.5281/zenodo.4284002.

Pořízka, P.: *The Function of Proper Nouns in Quantitative Analysis of Dramas* (A case study of Karel Čapek's plays), ICOS2021 – 27th International Congress of Onomastic Sciences, Kraków, Poland (accepted manuscript, will be published in the proceedings in the first half of 2023)

Are All Languages Equally Complex?:
Information Theory-based Method to Measure the Overall Complexity of a Language

Takuto Nakayama, Keio University, Japan

Keywords: *the equi-complexity, the overall linguistic complexity, information theory, Shannon entropy*

It has long been accepted by linguists that no language is simpler or more complex than another. This belief is called 'the equi-complexity of language'. However, until the end of the 20th century, linguists had neither attempted to examine whether this statement is true nor reached any consensus on how to measure complexity, especially an overall linguistic complexity including multiple simultaneous facets. This research aims to propose a method based on information theory to compare the overall linguistic complexity of two languages and to demonstrate this method using English and Japanese texts.

This research defines linguistic complexity as an entropy of Shannon's information theory (Shannon, 1948), which is an unpredictability of what linguistic sequence (e.g., characters or words) will appear next, and examines how entropies behave as the number of components of a sequence increases (from 1 to 100 in the pilot study below). As another barometer, this research uses a value called redundancy, which is the ratio of how much lower one entropy is than the maximum entropy. This redundancy value is required to standardize the entropies, which are partially dependent on the number of total linguistic sequences in a text. The two values are determined by the equations in the PDF file (see PDF file: 1. Equations). Applying these equations results in a vector that consists of the exponents of the power regression curves of the entropies and the redundancies of a subdomain of a language. The terms 'character', 'word', and 'part of speech' are the language subdomains used in the pilot study introduced below. To visualize how similar each text is, this research employs a hierarchical clustering analysis using Euclid distances between the vectors. The advantage of this method is that it can cover any linguistic subdomains

a researcher may want to use. Being able to consider not only one value but a vector with several values leads to a better understanding of the overall complexity of a text.

The English and Japanese texts for this pilot study were collected from online databases of digitized literature within the public domain. The study analyzed five texts in each language (10 in total) in which the Japanese texts are translated versions of the English ones. It is determined that the texts cannot be appropriately distinguished in their languages by the entropies; given redundancies, the texts are shown to be almost categorized in their languages (see PDF file: “2. Pilot Study”). This result suggests that the multi-domain complexities of the text are not significantly different from one another unless the number of types of sequences in the texts is considered. Further research will require collecting more data to make the analysis more precise and making comparisons with other types of obviously simpler texts, such as children’s books and texts for non-native readers.

Refernence

Shannon, C., E. (1948). A mathematical theory of communication. *Bell System Technical Journal*, *27*(3), 379–423.

Analyzing Japanese texts with evaluation of randomness in binary expression

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

1: Institute of Particle and Nuclear Studies, High Energy Accelerator Research Organization (KEK)
2: National Institute for Japanese language and Linguistics

Statistical analyses are powerful methods for extracting quantitative features in the languages and genres of the text such a newspaper, journal, literature, etc. In such studies, a character is treated as the minimum unit normally. Alternatively, each character can be expressed with binary codes like UTF-8 and SJIS. Since the text consists of the characters, the distributions of the text with those of the binary expression is also expected to contain information on features in the languages and genres of the text.

It is essential to analyze the binary sequences in the development of a random number generator, which is a crucial technology for the encrypted communication, to evaluate the randomness of its output quantitively. One such method is the Borel normality [1, 2]. An infinite binary sequence is the Borel normal if every binary string appears in the sequence with 2^{-n} for a string of length n. If the sequence is of finite length, the condition of the Borel normality can be expressed with the inequality, with respect to which the randomness of the binary sequence can be evaluated. Given that the binary codes of the characters contain fixed binary sequences, the binary expression of the text cannot be perfect random. However, the level of randomness and the shape of the distribution of the Borel normality may contain information about the language and genre of the text, and accordingly, the text may be classified using the randomness of its binary sequence.

We study the application of the analysis methods used in the development of a random number generator to analyses of the texts. As the first trial, the Borel normality was investigated for Japanese texts to study how its distribution behaves depending on the different genre of the text. For this study, we used the version 1.1 of the Balanced Corpus of Contemporary Written Japanese (BCCWJ), which was developed in the National Institute for Japanese Language and Linguistics in Japan. After converting the Japanese texts into the binary expressions such as UTF-8 and SJIS, the Borel normality is calculated. The periodic patterns are observed in the distributions with UTF-8 code that indicate features of its prefix in a character. In addition, the distributions of the texts in which each character is shuffled

randomly show clear differences from the original texts. Our aim is to define quantitative indices to evaluate the features of the texts from these measurements.

This study is the first attempt to adopt the analysis methods of the random numbers for the studies on the texts, and the same approach is applicable to any language. Our next challenge would be to investigate the variation of the Borel normality between different languages as well as to apply other methods to evaluate the randomness.

[1] Abbott, A. A., Calude, C. S., Dinneen, M. J., & Huang, N. (2019). Experimentally probing the algorithmic randomness and incomputability of quantum randomness. Physica Scripta, 94(4), 045103.
[2] Aolis, A., Angulo Martinez, A. M., Ramirez Alarcon, R., Cruz Ramirez, H., U'Ren, A. B., & Hirsch, J. G. (2015). How random are random numbers generated using photons? Physica Scripta, 90(7), 074034.

Keywords: Borel normality, Japanese, binary expression, BCCWJ

Title: Text Covering Efficiency and Word Tier Analysis for the proposal of vocabulary learning order and the analysis of text genres

Author: Tatsuhiko Matsushita

Affiliation: National Institute for Japanese Language and Linguistics

Keywords: Text Covering Efficiency, Word Tier Analysis, vocabulary learning order, genre analysis, Japanese medical texts

Abstract:
L2 vocabulary learning is a significant burden for learners, and the exploration of efficient vocabulary learning order is one of the applications where quantitative linguistics can potentially contribute to second language education. In this study, to demonstrate the usefulness of an index titled Text Covering Efficiency (TCE) to perform Word Tier Analysis (WTA) proposed by Matsushita (2012)[1], TCE of various lexical groups in a corpus of Japanese medical texts was calculated, to exemplify in what order medical vocabulary can be learned most efficiently. It will also be shown that lexical differences among various text genres can be clarified by WTA using TCE, and that differences in relative importance of different groups of words according to the purpose of learning.

Domain-specific words such as academic words (Coxhead, 2000) are often extracted for efficient vocabulary learning in a genre. Text coverage has been used for evaluating these groups of words (Coxhead, 2000; Hyland & Tse, 2007); however, it is not appropriate for comparing the efficiency between grouped words when the numbers of words are different between the groups. To solve this problem, Matsushita (2012) developed TCE which is the mean text coverage per unit number of words of each group of words. Matsushita also tested the validity of TCE by applying it to various genres, such as conversation, literary texts, newspapers, introductory academic texts, specialized texts in humanities, social sciences, and natural sciences. The result showed that TCE clearly indicates the efficiency in gaining text coverage, and thus it is useful for deciding a more efficient learning/teaching order of words depending on the different purposes of learning. In addition, TCE is a robust index by which different lexical features in different

[1] The term “word tier” in Matsushita (2012) means different concept from the one used in Burch and Egbert (2022).

genres can be clarified as well. For example, such an analysis allows you to say things like, "Learning the intermediate Common Academic Words is 6.2 times more efficient in covering social science texts than learning other words at the same level, and 8.3 times more efficient than learning the advanced common academic words". TCE is relatively easy to calculate, and the results can be easily applied with relatively little distortion due to different corpus sizes. If word lists such as basic/academic/technical vocabulary for various genres are analyzed by a vocabulary frequency profiler such as AntWordProfiler (Anthony, 2022), for example, the word tiers can be shown by calculating the TCE easily.

In this study, as an example of a proposal of vocabulary learning order, it will be presented based on a corpus of Japanese medical books, in what order students studying medicine in Japanese should learn different types of vocabulary most efficiently.

Although the analysis presented here is based on Japanese as an example, methods such as TCE and WTA are applicable to any language. Furthermore, by incorporating these methods into a word frequency profiler such as J-LEX (Suganaga and Matsushita, 2013), such analysis can be facilitated.

(495 words)

References

Anthony, L. (2022). AntWordProfiler (Version 2.0.1) [Computer Software]. Tokyo, Japan: Waseda University. Available from https://www.laurenceanthony.net/software

Coxhead, A. (2000). A new academic word list. *TESOL Quarterly, 34*(2), 213–238.

Hyland, K., & Tse, P. (2007). Is there an "Academic Vocabulary"? *TESOL Quarterly, 41*(2), 235–253.

Matsushita, T. (2012). In what order should learners learn Japanese vocabulary? A corpus-based approach (PhD thesis). Victoria University of Wellington. Downloaded from https://researcharchive.vuw.ac.nz/xmlui/handle/10063/4476

Suganaga, Y. and Matsushita, T. (2013). J-LEX: An Online Lexical Analyzer of Japanese Texts. Available from http://www17408ui.sakura.ne.jp/index.html

Trump's Simple Language: His Idiolect or Global Trend? Exploring Lexical Sophistication in U.S. Presidential Discourses

Woonhyung Chung
(Yonsei University, South Korea)

Abstract

Background: In the field of political discourse analysis, there has been an ongoing debate as to whether Donald Trump's use of simple language was a sign of his low intellect, or a global trend of dumbing down language (Conway, & Zubrod, 2022; Reyes, 2020). While no conclusive answer has been provided regarding this issue, previous studies have mostly focused on simple lexical measures, such as lexical density (i.e., proportion of content words) and lexical diversity (i.e., number of word types) (Savoy, 2017; Wang, & Liu, 2018). As an attempt to readdress this issue in a more sophisticated manner, this study adopted finer-grained lexical sophistication indices for comparing Donald Trump with his three predecessors (William Clinton, George Bush, and Barack Obama) and his successor (Joe Biden) in terms of their word usage. Lexical sophistication is defined as a text's lexical difficulty beyond surface-level features of lexicon (Crossley, & McNamara, 2012). Indices of lexical sophistication gauge how many difficult words are used in a text by means of word frequency and range, a word's psycholinguistic properties, and its relation to other words. These indices have been identified as reliable indicators of the depth and breadth of one's lexical knowledge (Eguchi, & Kyle, 2020). Thus, analyses of lexical sophistication are expected to provide an ideal opportunity for testing whether lexical simplification is specific to Donald Trump or reflects a potential trend of American political discourse.

Method: This study constructed a 290,000-word corpus that consisted of the five current and former Presidents' interviews. For a comprehensive assessment of the lexical properties of the data, the study employed both traditional lexical diversity index (e.g., mattr50) and lexical sophistication indices, including academic words and psycholinguistic property of words (e.g., the proportion of concrete words). The lexical diversity index was measured using the Tool for the Automatic Analysis of Lexical Diversity 1.3.1 (Kyle, Crossley, & Jarvis, 2021), and the lexical sophistication indices were measured using the Tool for the Automatic Analysis of Lexical Sophistication 2.2 (Kyle, Crossley, & Berger, 2018). For each of these indices, this study conducted a one-way ANOVA to inspect statistical differences among the texts.

Results and discussion: The mean index scores for each President are presented in Figure. The mattr50 scores showed the greatest lexical diversity for Obama and the lowest diversity for Trump.

The other three Presidents did not show any statistical differences. When investigating the lexical sophistication indices, both Trump and Biden used fewer academic words than the other Presidents. Furthermore, these two Presidents used more concrete words than the others, showing a lower degree of lexical sophistication. The use of unsophisticated words of the two Presidents suggests some evidence of a recent trend toward dumbed-down language in politics. This study provides a promising framework to analyze lexical sophistication in political figures' language use, calling upon further research to investigate whether the trend of dumbing down language is also found in other politicians.

***Keywords*:** lexical complexity, lexical diversity, lexical sophistication, political discourses, U.S. Presidents

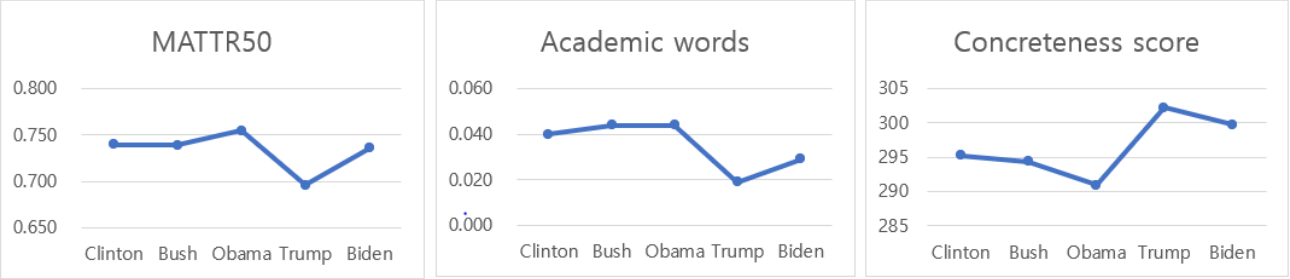


Figure. MATTR50, normed frequency of academic words, and concreteness score

References

Conway, L. G., & Zubrod, A. (2022). Are U.S. Presidents becoming less rhetorically complex? Evaluating the integrative complexity of Joe Biden and Donald Trump in historical context. *Journal of Language and Social Psychology, 41*(5), 613-625.

Crossley, S. A., & McNamara, D. S. (2012). Predicting second language writing proficiency: The roles of cohesion and linguistic sophistication. *Journal of Research in Reading 35*(2). 115–135.

Eguchi, M., & Kyle, K. (2020). Continuing to explore the multidimensional nature of lexical sophistication: The case of oral proficiency interviews. *The Modern Language Journal, 104*(2), 381-400.

Kyle, K., Crossley, S. A., & Berger, C. (2018). The tool for the automatic analysis of lexical sophistication (TAALES): version 2.0. *Behavior Research Methods, 50*(3), 1030-1046.

Kyle, K., Crossley, S. A., & Jarvis, S. (2021). Assessing the validity of lexical diversity indices using direct judgements. *Language Assessment Quarterly, 18*(2), 154-170.

Reyes, A. (2020). I, Trump: The cult of personality, anti-intellectualism and the Post-Truth era. *Journal of Language and Politics, 19*(6), 869-892.

Savoy, J. (2017). Trump's and Clinton's style and rhetoric during the 2016 presidential election, *Journal of Quantitative Linguistics, 25*(2), 168-189.

Wang, Y. Q., & Liu, H. T. (2018). Is Trump always rambling like a fourth–grade student? An analysis of stylistic features of Donald Trump's political discourse during the 2016 election. *Discourse and Society, 29*(3), 299–323.

Community-specific Context Typicality as a determinant of lexical variation

Barend Beekhuizen & Kaleigh Woolford
University of Toronto

Social factors like age and gender affect lexical choice between near-synonyms (Labov 1972; Tagliamonte 2008). However, the linguistic context of the near-synonyms, known to affect lexical choice too (Sinclair 2004), may constitute a confound in studies considering social factors. Here, we present a corpus-derived measure that reconceptualizes linguistic contexts as factors of interest to the sociolinguist: socially stratified differences in the use of linguistic contexts suggest that the choice of *what* is said matters when studying lexical variation, i.e., *how* it is said. We demonstrate this for English intensifiers.

We gathered a corpus of spontaneously produced language from a Reddit community where users frequently self-identify with age and gender, and extracted all intensifiers (N=47,749). We next developed measures of how (1) typical linguistic contexts are for certain social groups and

(2) whether the typicality differs across groups. First, we determined, for each token in a test sample of 39,749 tokens, if the token would be typical for each of eight age/gender bins (four age bins {16-19, 20-23, 24-27, 28-31} by two attested genders {male, female}). We did so by computing a contextualized vector (using BERT; Devlin et al. 2018) for each token, and, per age/gender bin *b*, measuring the token vector's average similarity to the *K=30* most similar tokens in a 1000-token held-out sample from *b* (cf. Yates 2005 for analogous measures in phonology). This value tells us how typical that token would be for *b*: if the most-similar tokens in *b* are highly similar, the token reflects a context that is typical for users in *b*. Second, a Principal Component Analysis of the eight age/gender bin typicality values lets us detect patternings in the variation between the bins. The PCA's first three components form interpretable scales from tokens typical for *X* to ones typical for *Y*:

- [PC1] *X* = every bin, *Y* = no-one (general typicality)
- [PC2] *X* = older users, *Y* = younger users (age typicality)
- [PC3] *X* = men, *Y* = women (gender typicality)

If the social (age, gender) typicality of the context affects lexical choice, we expect them to predict lexical choice over and above user age and gender: e.g., younger users may produce comparably more intensifiers associated with older users in contexts associated with older users than in 'younger' contexts. We applied logistic regressions with each of the seven most frequent intensifiers (*very, really, pretty, so, extremely, super*, *pretty*) as dummy-coded dependent variables, with five independent variables (general typicality, age typicality, gender typicality, user age, user gender).

Results correspond to Tagliamonte's (2008) observations: women use more *so*, men more *pretty*; older speakers more *very*, younger speakers more *really*. Importantly, in all cases, gender typicality and age typicality are significant predictors in the same directions over and above age and gender: e.g., female-typical contexts display more *so*, regardless of user gender (Figure 1), and 'older' contexts display more *very*, over and above user age (Figure 2), confirming that the social typicality of contexts is a factor both measurable and worth exploring further.

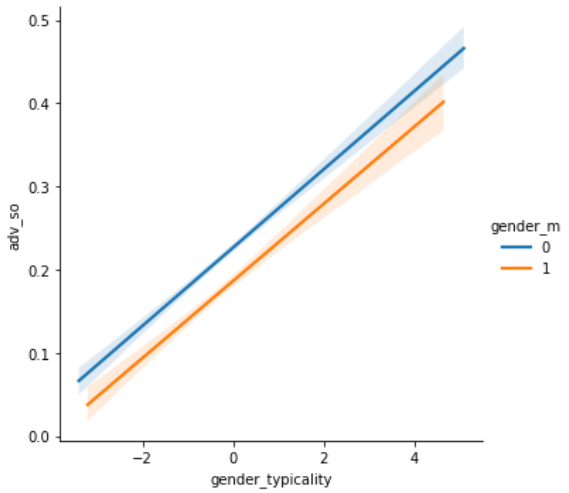


Figure 1: Proportions lexical choice for so *over gender typicality, broken down by gender.*

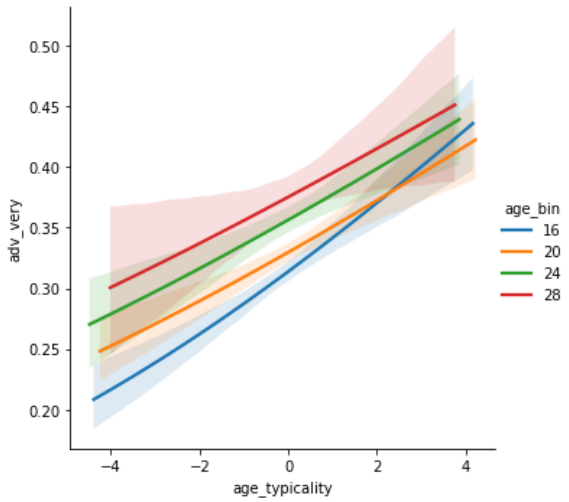


Figure 2: Proportions lexical choice for very *over age typicality, broken down by age.*

References

Devlin, J., Chang, M. W., Lee, K., & Toutanova, K. (2018). Bert: Pre-training of deep bidirectional transformers for language understanding. arXiv preprint arXiv:1810.04805.

Labov, W. 1972. *Sociolinguistic Patterns*. University of Pennsylvania Press, Philadelphia.

Sinclair, J. (2004). *Trust the text: Language, corpus and discourse*. Routledge.

Tagliamonte, S. A. (2008). So different and pretty cool! Recycling intensifiers in Toronto, Canada. *English Language & Linguistics*, 12(2), 361-394.

Yates, M. (2005). Phonological neighbors speed visual word processing: evidence from multiple tasks. *Journal of Experimental Psychology: Learning, Memory, and Cognition*, 31(6), 1385.

Title: Part-of-speech proportion as an index of formality and informality: The case of Japanese

Keywords: formality, informality, index, part-of-speech, Japanese

Author: Tatsuhiko Matsushita

Affiliation: National Institute for Japanese Language and Linguistics

Abstract:
This study uses a Japanese corpus to show that the ratio(s) of the particular part(s) of speech (POS) in the total number of words is a useful index for the degree of formality/informality of a text and that it can be used to analyze text genres.

This study used the book and Internet forum portions of the Balanced Corpus of Contemporary Written Japanese (BCCWJ) 2009 monitor version (NINJAL, 2009) consisting of approximately 33 million words[1].

Using the Nippon Decimal Classification system used for book classification as a guide, the entire corpus was divided into 10 fields: literary works (LW), languages, linguistics and philosophy (LP), history and ethnology (HE), arts and other humanities (AH), politics and law (PL), economics and commerce (EC), sociology, education and other social issues (SE), science and technology (ST), biology and medicine (BM), and Internet Q&A forums (IF). Texts in these 10 domains were then analyzed with the morphological analyzer MeCab Ver. 0.98 (Kudo, 2009) and the morphological dictionary UniDic Ver.1.3.11 (Den et al., 2009) with POS tagger to determine the proportion of each part of speech in the total number of words (i.e., text coverage).

The results showed that the proportions of the four parts of speech (suffixes, nouns, verbal nouns, and conjunctions - tentatively referred to as "A-set POS") in the 10 domains were all highly correlated with each other, ranging from r=.67 to .86. When the 10 domains were ordered from left to right, from lowest to the highest proportion of suffixes, the proportions of these four parts of speech were all linearly, increasingly aligned with each other in a right-

[1] Of the completed version of the BCCWJ (NINJAL, 2011), the book corpus is approximately 50 million words, and the 2009 monitored version accounts for about two-thirds, which is almost identical to the completed version in its nature.

to-left direction. Similarly, the seven parts of speech (particles, adverbs, adjectives, auxiliary verbs, verbs, pronouns, and interjections - tentatively called "B-set POS") were linearly, decreasingly aligned from left to right, and showed high negative correlations with the ratios of the A-set POS, ranging from r=-.55 to -.92. No such pattern was observed for the other three parts of speech (adjectival nouns, pre-noun adjectivals, and prefixes), and there was little correlation with the ratios of the other parts of speech.

Of importance here is that all parts of speech are clearly divided into three groups: those showing positive correlation, those showing negative correlation, and those showing no correlation at all, and that the ratios of particular parts of speech line up linearly. As far as the arrangement of the 10 genres sorted by the proportions of parts of speech in the A and B sets is concerned, the A set will be an index of the degree of formality, while the B set will be an index of the degree of informality.

The dissimilarities with some previous studies and the comparison with some studies on languages other than Japanese will further be discussed in the presentation.

(499 words)

*This study is extracted from and added to Chapter 4 of the presenter's doctoral thesis (Matsushita, 2012).

Reference

Den, Y., Yamada, A., Ogura, H., Koiso, H., & Ogiso, T. (2009). UniDic (digitized dictionary for morphological analysis) 1.3.11. Downloaded from http://www.tokuteicorpus.jp/dist/

Kudo, T. (2009). *MeCab (morphological analyzer) 0.98*. Downloaded from http://mecab.sourceforge.net/

Matsushita, T. (2012). In what order should learners learn Japanese vocabulary? A corpus-based approach (PhD thesis). Victoria University of Wellington. Downloaded from https://researcharchive.vuw.ac.nz/xmlui/handle/10063/4476

NINJAL (The National Institute for Japanese Language). (2009). Balanced Corpus of Contemporary Written Japanese (2009 monitor version). (Available by application at that time).

NINJAL (The National Institute for Japanese Language and Linguistics). (2011). Balanced Corpus of Contemporary Written Japanese. (Available by application).

Quantifying meaning differences between English clippings and their source words

Martin Hilpert, David Correia Saavedra, & Jennifer Rains
Université de Neuchâtel

This paper uses corpus data and methods of distributional semantics in order to study English clippings such as *dorm* (< *dormitory*), *memo* (< *memorandum*), or *quake* (< *earthquake*). We investigate whether systematic meaning differences between clippings and their source words can be detected. Alber and Lappe (2012: 314) observe that semantic questions have received relatively little attention in the study of clippings, and they remark that systematic studies of meaning in truncatory processes are virtually absent. The present paper tries to address that gap.

Our analysis is based on a sample of 50 English clippings and their source word counterparts. Pairs such as *cardio-cardiovascular*, *chemo-chemotherapy*, and *intro-introduction* are analyzed in terms of their collocational behavior. Each of the clippings is represented by a concordance of 100 examples in context that were gathered from the Corpus of Contemporary American English (Davies 2008). We compare clippings and their source words both at the aggregate level, and in terms of comparisons between individual clippings and their source words.

The aggregate comparisons reveal general distributional asymmetries that suggest a difference relating to involved vs. informational text production (Biber 1988). Clippings have a relatively greater tendency to appear in texts with contextual elements such as first or second person pronouns, demonstratives, or contractions. Clippings thus appear to be preferred in contexts in which there is substantial common ground between speaker and hearer, which aligns with the notion that clippings signal familiarity with the ideas that are conveyed (Wierzbicka 1984, Katamba 2005).

For the individual comparisons between clippings and their source words, we draw on the distributional semantic method of token-based semantic vector spaces (Hilpert and Correia Saavedra 2020). The method allows us to pinpoint aspects of meaning that are specifically associated with a clipping, rather than its source word, and vice versa, while also revealing how their respective meanings overlap. Our findings show that clippings such as *chemo* and

cardio are semantically distinct from their source words, but we also document cases such as *fridge*, in which the collocational profile of the clipping is indistinguishable from that of the source word.

We interpret these findings against the theoretical background of research on communicative efficiency (Levshina and Lorenz 2022). We discuss whether speakers' choices between clippings and their source words are motivated by meaning, rather than by efficiency. For cases in which clippings overlap semantically with their source words, we documented facets of meaning that are preferentially or even exclusively expressed by one of the two alternatives. We take these results as an indication that the use of English clippings is primarily motivated by semantic factors.

Alber, Birgit, and Sabine Arndt-Lappe. 2012. Templatic and subtractive truncation. In J. Trommer (ed.), *The Phonology and Morphology of Exponence – the State of the Art,* 289-325. Oxford: Oxford University Press.

Biber, Douglas. 1988. *Variation across Speech and Writing*. Cambridge: Cambridge University Press.

Davies, Mark. 2008. The Corpus of Contemporary American English (COCA). Available online at https://www.english-corpora.org/coca/.

Hilpert, Martin and Correia Saavedra, David. 2020. Using token-based semantic vector spaces for corpus-linguistic analyses: From practical applications to tests of theoretical claims. *Corpus Linguistics and Linguistic Theory* 16/2, 393-424. https://doi.org/10.1515/cllt-2017-0009

Katamba, Francis. 2005. *English Words*. 2nd edition. New York: Routledge.

Levshina, Natalia and David Lorenz. 2022. Communicative efficiency and the Principle of No Synonymy: Predictability effects and the variation of *want to* and *wanna*. *Language and Cognition* 14/2, 249-274. https://doi:10.1017/langcog.2022.7

Wierzbicka, Anna. 1984. Diminutives and depreciatives: Semantic representation for derivational categories. *Quaderni di semantica* 5(1), 123-130.

Incidence- and abundance-based measures to assess rivalry in word formation

Justine Salvadori, Rossella Varvara, Richard Huyghe
University of Fribourg (Switzerland)

Keywords: affix rivalry, nominalization, deverbal suffix, French, similarity measure

Affix rivalry occurs between affixes that have equivalent semantic functions and can therefore compete in the formation of derivatives (see e.g. Lindsay & Aronoff 2013; Arndt-Lappe 2014; Fradin 2019). However, equivalence may be established only between some of the functions of polyfunctional derivational processes. According to Lieber (2016), for example, English suffixes *-ation* and *-al* can both derive event (*conversation*, *portrayal*) and result (*coloration*, *acquittal*) nouns, but only the former can be used to derive instrument (*decoration*) and agent (*administration*) nouns. The possibility for rival affixes to not be strictly equivalent entails a gradient notion of morphological competition. A continuum of rivalry can be postulated, ranging from no rivalry in case of semantic disjunction to full rivalry in case of semantic identity. This gradient nature of affix rivalry calls for an appropriate, i.e. quantified, assessment. Ideally, a coefficient of competition should be provided so that different situations of rivalry can be compared both within languages and cross-linguistically.

This study introduces two measures drawn from studies in ecology that can be used to assess degrees of rivalry between polyfunctional affixes: the Sørensen index (Sørensen 1948), which considers the proportion of shared functions between rival affixes; and the Percentage similarity coefficient (as a complement to the Percentage difference index proposed by Odum 1950), which is based on the realization frequency of functions. Two complementary measures, Balanced richness (for the Sørensen index) and Balanced abundance (for the Percentage similarity coefficient), are also provided to further analyze differences in the structure of shared functionality. For instance, they can be used to identify nestedness, i.e. when the functions of an affix A are a subset of the functions of an affix B, and overlap, i.e. when two affixes A and B have functions in common but also specific functions that are not covered by the other suffix (Plag 1999; Guzmán Naranjo & Bonami 2023; a.o.).

We explore the potential of the four measures using simulated data, before applying them to real linguistic material, viz. a sample of 600 nouns formed with 6 nominalizing suffixes in French. To identify functions, all derivatives are analyzed according to a double semantic typology combining referential and relational information. Overall, the results show that the four measures can highlight different facets of similarity relationships and that they complement each other accordingly. As incidence-based measures, the Sørensen and Balanced richness indices allow in-depth investigation of functionality structures. Their interpretation can reveal principles of association between semantic functions that influence the proportion of functions shared between two affixes. As abundance-based measures, the Percentage similarity and Balanced abundance indices can weight functional rivalry by realization frequency and shed a different light on the sharing of functions. The comparison between the two types of measures informs on the architecture of rivalries and on the congruence or discrepancy between the number of shared functions and the number of derivatives that instantiate these functions.

References

Arndt-Lappe, Sabine. 2014. Analogy in suffix rivalry: The case of English *-ity* and *-ness*. *English Language & Linguistics* 18(3). 497–548.

Fradin, Bernard. 2019. Competition in derivation: What can we learn from French doublets in *-age* and *-ment*? In *Competition in Inflection and Derivation*, Franz Rainer, Francesco Gardani, Wolfgang U. Dressler & Hans Christian Luschützky (eds), 67-93. Cham: Springer.

Guzmán Naranjo, Matías & Bonami, Olivier. 2023. A distributional assessment of rivalry in word formation. *Word Structure* 16(1).

Lieber, Rochelle. 2016. *English Nouns: The Ecology of Nominalization*. Cambridge: Cambridge University Press.

Lindsay, Mark & Aronoff, Mark. 2013. Natural selection in self-organizing morphological systems. In *Morphology in Toulouse. Selected Proceedings of Décembrettes 7*, Nabil Hathout, Fabio Montermini & Jesse Tseng (eds), 133-153. München: Lincom Europa.

Odum, Eugene P. 1950. Bird populations of the Highlands (North Carolina) Plateau in relation to plant succession and avian invasion. *Ecology* 31: 587–605.

Plag, Ingo. 1999. *Morphological Productivity: Structural Constraints in English Derivation.* Berlin: Mouton de Gruyter.

Sørensen, Thorvald A. 1948. A method of establishing groups of equal amplitude in plant sociology based on similarity of species content and its application to analyses of the vegetation on Danish commons. *Kongelige Danske Videnskabernes Selskabs Biologiske Skrifter* 5: 1–34.

Probabilistic Regularity in Translation: A Quantitative Description of Dependency Treebank of Academic Abstracts

Yan Liang
Xi'an Jiaotong University

Abstract: Translational language, the language of translated texts, is distinct from both the source and the target language. Research on the linguistic characteristics of translational language is important for a deeper understanding of probabilistic translation regularities. Although there are many studies on translational language (Laviosa, 2002; Klaudy, 2004), they present contradictory results duo to both methodological limitations and theoretical confusion (Puurtinen, 2003; Jantunen, 2001). On one hand, the adoption of global indices, such as lexical density, type-token ratio, etc., may obscure the internally dynamic features of translational language. On the other hand, the study of translation universals may conceal the probabilistic features of translational language.

This study investigates the syntactic and typological properties of translational language by adopting the two main indices of dependency grammar, mean dependency distance (MDD) and dependency direction, with an aim to reveal the probabilistic features of translational language: standardization. Dependency grammar assumes that language is a conceptual network and a sentence is a network of nodes, and the nodes are all connected by syntactic dependencies that consist of a head and a dependent (Tesnière, 1959; Ninio, 2006; Hudson, 2010; Liu, 2009). Dependency distance, the linear distance between the dependent and its head, can be used to measure syntactic complexity and language processing cost (Temperley, 2007: 301). Dependency direction can serve as a valuable indicator for the study of typological property of translational language (Dryer, 1997; Liu, 2010). A comparable dependency treebank, consisting of translated and non-translated English abstracts, was built and quantitatively described.

The results show that (1) there presents significant difference between the MDD of translational English and that of non-translational English, which suggests the effect of standardization; (2) the MDD of translational English is within the threshold of four; (3) translational English has a different dependency direction distribution from non-

translated language. These findings show the effect of standardization and suggest that a quantitative method is valid for a systemic description of translational language.

Keywords: probabilistic regularity; quantitative method; dependency treebank; academic abstracts

References

Dryer, M., 1997. On the 6-way word order typology. Studies in Language 21, 69-103.

Hudson, R., 2010. An Introduction to Word Grammar. Cambridge University Press, Cambridge.

Jantunen, J., 2001. Synonymity and lexical simplification in translation: a corpus-based approach. Across Languages and Cultures 1: 97-112.

Klaudy, K., 2004. Explicitation. In Baker M., Saldanha G. (Eds.), Routledge Encyclopedia of Translation Studies. Routledge, London and New York, pp. 104-108.

Laviosa, S., 2002. Corpus-based Translation Studies. Rodopi, Amsterdam and New York.

Liu, H., 2009. Dependency Grammar: From Theory to Practice. Science Press, Beijing.

Liu, H., 2010. Dependency direction as a means of word-order typology: a method based on dependency treebanks. Lingua 120, 1567-1578.

Ninio, A., 2006. Language and the Learning Curve: A New Theory of Syntactic Development. Oxford University Press, Oxford.

Puurtinen, T., 2003. Genre-specific features of translationese? Linguistic differences between translated and non-translated Finnish children's literature. Lit. Linguist. Comput. 18 (4), 389-406.

Tesnière, L., 1959. Eléments de la syntaxe structurale. Klincksieck, Paris.

Temperley, D., 2007. Minimization of dependency length in written English. Cognition 105 (2), 300-333.

Author index

Practical information

Timetables may be subject to last-minute changes, to be announced at the conference

Venue
The conference activities take place on the ground floor of the Anthropole Building of the Dorigny site of the University of Lausanne (UNIL), in rooms A (ANT-1031) and B (ANT-1129) - see the map.
Metro Stop : **UNIL-Chamberonne**, metro m1. For access to the Dorigny site *from Lausanne railway station*, railway passengers take metro m2 (one stop *Lausanne-Gare* to *Lausanne-Flon)*, then metro m1 from *Lausanne-Flon* towards *Renens*.

Desk
The registration desk stands in front of room A. It opens the first day (28.6) at 08:30, (one hour before the Opening session), as well as between the Conference sessions.

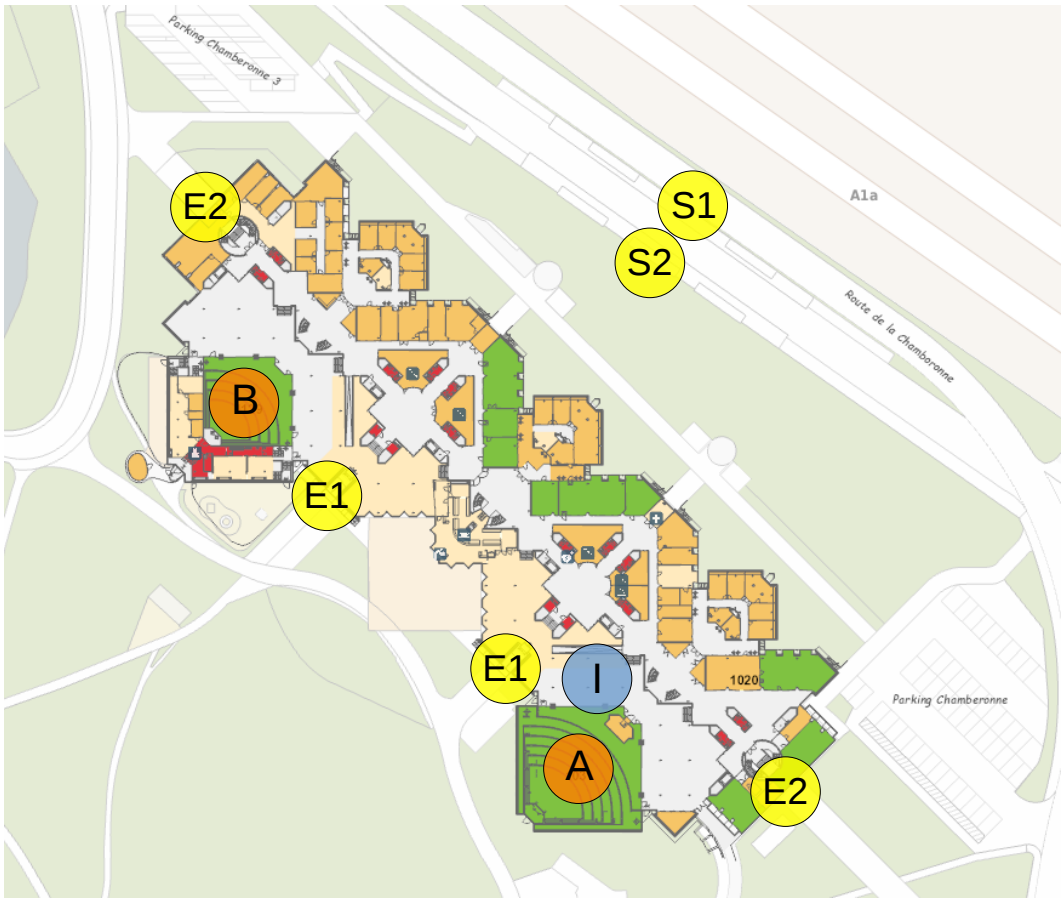
Wifi
The university *eduroam* wifi network is available by following the guidelines of the home institution of the participants.
Alternatively, a dedicated conference wifi is available: select the wifi “public-unil”, click on “Event-Login”, and use the username *qualico2023* and the password *qualico2023*.

Lunches
Lunches take place at *Unithèque* (see the map), where some tables are reserved for the conference. Two vouchers are provided each day, up to a maximal amount of 14.- CHF for the meal + drink (external guest rate), plus max. 3.- CHF for a coffee or else after the lunch; extra costs are borne by the participants.

Welcome drink
A welcome drink is organized the first day (28.6) from 17:30 to 18:30 in front of room A.

Social Event
The social event (29.6; inscriptions are closed) is in two parts:
From 6:30 pm to 7:45pm : the ***Lavaux-Express*** road train tour (https://www.lavauxexpress.ch/en/). It is a round trip starting from *Lutry*'s shore.
Various forms of public transport will be used to get there from the conference venue (make sure you to have your ticket with you). The group journey takes around 40 minutes, including a 10-minute walk.
At about 8.45 pm (back to Lausanne by return public transportation): **Dinner at *Le Milan*** (https://www.lemilan.ch/). *Le Milan* is located near Lausanne railway station and serves good italian food. Drinks are at the charge of the participants.

Contact address
qualico2023@unil.ch



S1 Subway M1 stop
UNIL-Chamberonne
Direction Renens

S2 Subway M1 stop
UNIL-Chamberonne
Direction Lausanne

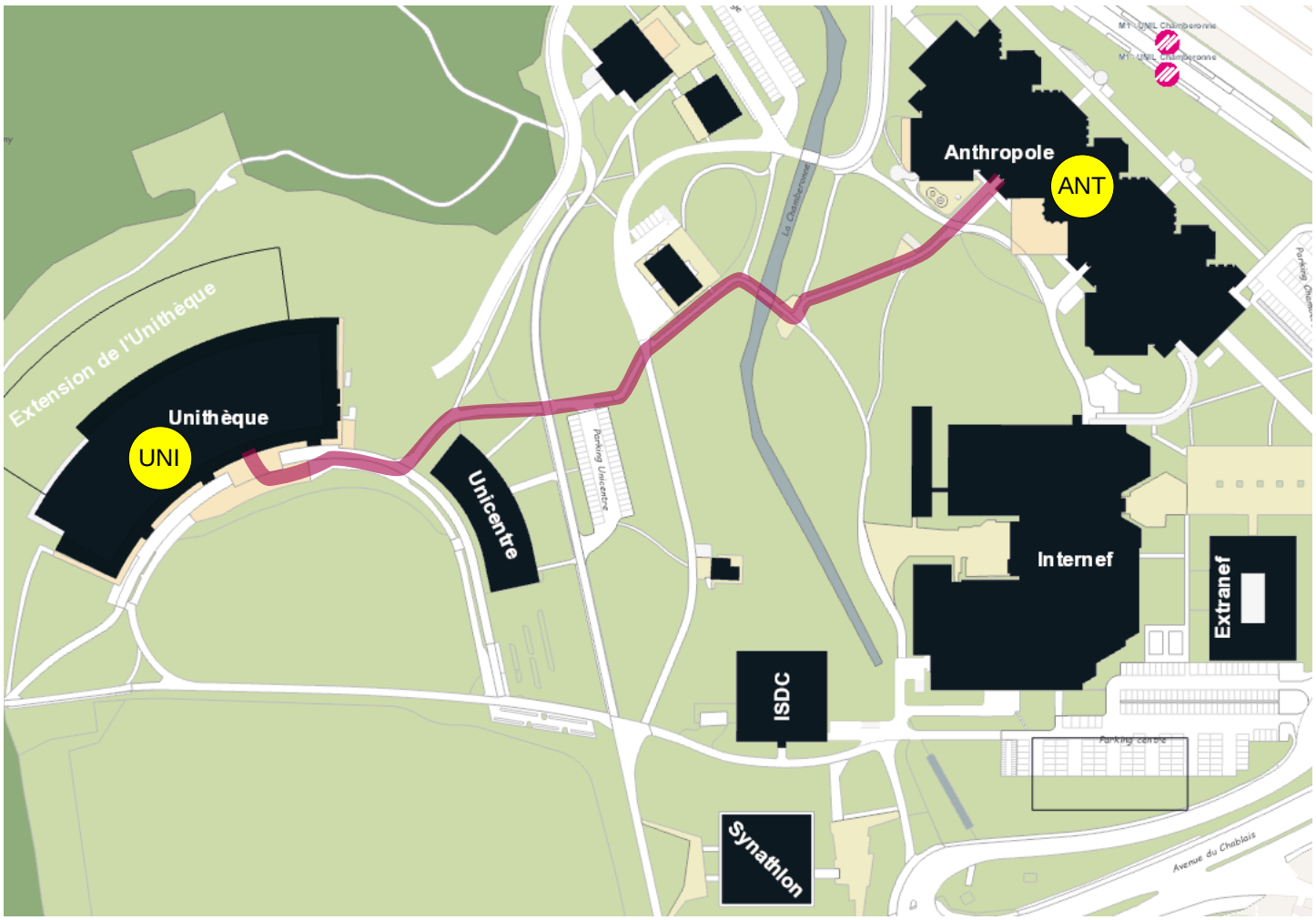
E1 Entrance/Exit
1st floor

E2 Entrance/Exit
2nd floor

I Welcome desk
Posters
1st floor

A Room A
ANT-1031
1st floor

B Room B
ANT-1129
1st floor



UNI Unithèque – Lunch place

ANT Anthropole – Conference Venue