Workshop: Computational models of diachronic language change

Organizers:
Stefania Degaetano-Ortlieb*, Lauren Fonteyn+, Pauline Krielke*, Elke Teich*
*Saarland University, +Leiden University

While the study of diachronic language change has long been firmly grounded in corpus data analysis, it seems fair to state that the field has been subject of a ‘computational turn’ over the last decade or so, computational models being increasingly adopted across several research communities, including corpus and computational linguistics, computational social science, digital humanities, and historical linguistics.

The core technique for the investigation of diachronic change are distributional models (DMs). DMs rely on the fact that related meanings occur in similar contexts and allow us to study lexical-semantic change in a data-driven way (e.g. as argued by Sagi et al. 2011), and on a larger scale (e.g. as shown on the Google NGram corpus by Gulordava & Baroni 2011). Besides count-based models (e.g. Hilpert & Saavedra 2017), contextualized word embeddings are increasingly employed for diachronic modeling, as such models are able to encode rich, context-sensitive information on word usage (see Lenci 2018 or Fonteyn et al., 2022 for discussion).

In previous work, DMs have been used to determine laws of semantic change (e.g. Hamilton et al. 2016b, Dubossarsky et al. 2017) as well as develop statistical measures that help detect different types of change (e.g. specification vs. broadening; cultural change vs. linguistic change; Hamilton et al. 2016a, Del Tredici et al. 2019). DMs have also been used to map change in specific (groups of) concepts (e.g. ‘racism’, ‘knowledge’; see Sommerauer & Fokkens 2019 for a discussion). Further studies have suggested ways of improving the models that generate (diachronic) word embeddings to attain these goals (e.g. Rudolph & Blei 2018).

Existing studies and projects focus on capturing and quantifying aspects of semantic change. Yet, over the past decade, DMs have also been shown to be useful to investigate other types of change in language use, including grammatical change. Within the computational and corpus linguistic communities, for example, Bizzoni et al. (2019, 2020) have shown an interdependency between lexical and grammatical changes and Teich et al. (2021) use embeddings to detect (lexico-) grammatical conventionalization (which may lead to grammaticalization). Within diachronic linguistics, the use of distributional models is focused on examining the underlying functions of grammatical structures across time (e.g. Perek 2016, Hilpert and Perek 2015, Gries and Hilpert 2008, Fonteyn 2020, Budts 2020). Specifically targeting historical linguistic questions, Rodda et al. (2019) and Sprugnoli et al. (2020) have shown that computational models are promising for analyzing ancient languages, and McGillivray et al. (2022) highlight the advantages of word embeddings (vs. count-based methods) while also pointing to the challenges and the limitations of these models.

A common concern across these different communities is to better understand the general principles or “laws” of language change and the underlying mechanisms (analogy, priming, processing efficiency, contextual predictability as measured by surprisal, etc.). In the proposed workshop, we will bring together researchers from relevant communities to talk about the unique promises that computational models hold when applied to diachronic data as well as the specific challenges they involve. In doing so, we will identify common ground and explore the most pressing problems and possible solutions. The program of the workshop will include talks by both invited speakers and open call for paper presentations.

Specific questions will concern:

Model utility: How can we capture change in language use beyond lexical-semantic change, e.g. change in grammatical constructions, collocations, phraseology?

Model quality: How can we evaluate computational models of historical language stages in absence of native-speaker 'gold standards'? To what extent does the quality of historical and diachronic corpora affect the performance of models?
Model analytics: How do we transition from testing the reliability of models to employing them to address previously unanswered research questions on language change? How can we detect and "measure" change? What are suitable analytic procedures to interpret the output of models?

References:

Bizzoni, Y., Degaetano-Ortlieb, S., Menzel, K., Krielke, P., and Teich, E. (2019). "Grammar and meaning: analysing the topology of diachronic word embeddings". In Proceedings of the 1st International Workshop on Computational Approaches to Historical Language Change, ACL, Florence, Italy, pp. 175–185.

Bizzoni, Y., Degaetano-Ortlieb, S., Fankhauser, P., and Teich, E. (2020). "Linguistic variation and change in 250 years of English scientific writing: a data-driven approach". Frontiers in Artificial Intelligence, 3.

Budts, S. (2020). "A connectionist approach to analogy. On the modal meaning of periphrastic do in Early Modern English". Corpus Linguistics and Linguistic Theory, 18(2), pp. 337–364.

Del Tredici, M., Fernández, R., and Boleda, G. (2019). "Short-term meaning shift: A distributional exploration." In Proceedings of the Conference of the North American Chapter of the Association for Computational Linguistics (NAACL): Human Language Technologies, Minneapolis, Minnesota, USA, pp. 2069–2075.

Dubossarsky, H., Weinshall, D., and Grossman, E. (2017). "Outta control: laws of semantic change and inherent biases in word representation models". In Proceedings of the Conference on Empirical Methods in Natural Language Processing (EMNLP), Copenhagen, Denmark, pp. 1136–1145.

Fonteyn, L. (2020). "What about grammar? Using BERT embeddings to explore functional-semantic shifts of semi-lexical and grammatical constructions." Computational Humanities Research CEUR-WS, pp. 257–268.

Fonteyn, L., Manjavacas, E., and Budts, S. (2022). "Exploring Morphosyntactic Variation & Change with Distributional Semantic Models". Journal of Historical Syntax, 7(12), pp. 1–41.

Gries, S. T., and Hilpert, M. (2008). "The identification of stages in diachronic data: variability-based Neighbor Clustering". Corpora, 3(1), pp. 59–81.

Gulordava, K., and Baroni, M. (2011). "A distributional similarity approach to the detection of semantic change in the Google Books Ngram corpus". In Proceedings of Geometrical Models for Natural Language Semantics (GEMS), EMNLP, Edinburgh, United Kingdom, pp. 67–71.

Hamilton, W. L., Leskovec, J., and Jurafsky, D. (2016a). "Cultural shift or linguistic drift? comparing two computational models of semantic change". In Proceedings of the Conference on Empirical Methods in Natural Language Processing, EMNLP, Austin, Texas, USA, pp. 2116–2121.

Hamilton, W. L., Leskovec J., and Jurafsky, D. (2016b). "Diachronic word embeddings reveal statistical laws of semantic change". In Proceedings of Morphosyntactic Variation & Change with DSMs, 54th Annual Meeting of the Association for Computational Linguistics, ACL, Berlin, Germany, pp. 1489–1501.

Hilpert, M., and Saavedra, D.C. (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), pp. 393–424.

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

Lenci, A. (2018). "Distributional Models of Word Meaning". Annual Review of Linguistics, 4, pp. 151–171.

McGillivray, B., Jenset, G.B., Salama, K. (2022). "Investigating patterns of change, stability, and interaction among scientific disciplines using embeddings". Nature Humanities and Social Science Communication 9, 285.

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

Rodda, M.A., Probert, P., and McGillivray, B. (2019). “Vector space models of Ancient Greek word meaning, and a case study on Homer”. TAL Traitement Automatique des Langues, 60(3), pp. 63–87.

Rudolph, M., and Blei, D. (2018). “Dynamic embeddings for language evolution”. In Proceedings of the 2018 World Wide Web Conference (WWW ’18), Lyon, France, pp. 1003–1011.

Sagi, E., Kaufmann, S., and Clark, B. (2011). “Tracing semantic change with Latent Semantic Analysis”. Current Methods in Historical Semantics, 73, pp. 161–183.

Sommerauer, P., and Fokkens, A. (2019). “Conceptual Change and Distributional Semantic Models: An Exploratory Study on Pitfalls and Possibilities”. In Proceedings of the 1st International Workshop on Computational Approaches to Historical Language Change, Florence, Italy, pp. 223–233.

Sprugnoli, R., Moretti, G., and Passarotti, M. (2020). “Building and Comparing Lemma Embeddings for Latin. Classical Latin versus Thomas Aquinas”. IJCoL. Italian Journal of Computational Linguistics, 6(6-1), pp. 29–45.

Teich, E., Fankhauser P., Degaetano-Ortlieb, S., and Bizzoni, Y. (2021). “Less is More/More Diverse: On the Communicative Utility of Linguistic Conventionalization”. Frontiers in Communication, 5.

A Diachronic Analysis of Using Sentiment Words in Scandinavian Literary Texts from 1870-1900

Ali Al-Laith, Kirstine Nielsen Degn, Bolette Sandford Pedersen, Daniel Hershcovich, Jens Bjerring-Hansen
University of Copenhagen, Denmark

Diachronic corpora, or collections of texts spanning a significant time period, are useful computational linguistics tools for studying language change and evolution. They can be used to investigate changes in vocabulary [1, 2], grammar [3], and usage patterns over time [4]. Additionally, they can be used to analyze the development of different language varieties, and dialects [5, 6]. They can also be used to understand how language is used in different contexts and how language use changes in response to social, cultural, and historical factors [7, 8, 9, 10]. Other potential applications of diachronic corpora in computational linguistics include the creation of language processing tools and systems that consider the historical context in which a text was produced [11].

To track the cultural development in society through literature analysis, one can study the themes and ideas present in the literature over time and look for trends, and changes [12]. This includes examining shifts in how these themes and ideas are presented and changes in the style and form of literature and subjects addressed. It is also essential to consider the social, political, and economic context in which the literature was produced, as these factors can influence the culture and development of society [13]. There are several ways to track the use of emotional language over time in literature [14, 15]. One method is to conduct a content analysis of the text, in which the frequency of emotional words and phrases is counted [16]. Another approach is to use thematic analysis, which involves examining the themes related to emotions in the text and how they are presented [17, 18]. A third option is to employ sentiment analysis, which uses computational tools to analyze the emotional content of the text through natural language processing algorithms or the use of dictionaries or lexicons of emotional words and phrases [19, 20].

Given the large collection of diachronic literary texts that is currently available, we expect to see variations in the usage of sentiment-bearing words in different time periods and in relation to the shifting discussions and themes over time. In this research, we examine the evolution of sentiment words' use in the MEMO corpus, a collection of almost 900 Danish and Norwegian novels from the latter part of the 19th century [21].

A dynamic BERTopic model is a powerful tool for analyzing the evolution of topics in a col- lection of documents over time. It uses transformers and class-based TF-IDF to identify clusters of words and phrases representing the main topics discussed in the corpus. It also incorporates important words in the topic descriptions for improved interpretability. By tracking the use of sentiment words, the dynamic BERTopic model allows us to gain a deeper understanding of the changes and developments in the discussions over time. To further analyze these patterns, we em- ploy the Danish Sentiment Lexicon (DDS)1 [22, 23] to identify any changes in the use of sentiment words over time.

This research aims to track the evolution of sentiment towards a specific topic over time and the evolution of which words are used to express sentiment. The goal is to understand how public sentiment or attitudes towards the topic have changed, identify trends and patterns in the way the topic is discussed, and provide historical context that helps explain how the topic has been represented.

Keywords— Sentiment Analysis, Sentiment Lexicon, Topic Modeling, Scandinavian Literature, Diachronic Corpora, Danish Text, Norwegian Text

1https://github.com/dsldk/danish-sentiment-lexicon

References

P. Cassotti, P. Basile, M. de Gemmis, and G. Semeraro, "Analysis of lexical semantic changes in corpora with the diachronic engine.," in CLiC-it, 2020.

T. McEnery and A. Wilson, “Diachronic Corpora in the Study of Vocabulary Change: A Review,” Corpus Linguistics and Linguistic Theory, vol. 11, no. 2, pp. 203–226, 2015.

C. Mair, “Tracking ongoing grammatical change and recent diversification in present-day standard english: the complementary role of small and large corpora,” in The changing face of corpus linguistics, pp. 355– 376, Brill, 2006.

I. Renau and R. Nazar, “Automatic extraction of lexical patterns from corpora,” in En EURALEX International Congress: Lexicography and Linguistic Diversity, pp. 823–830, 2016.

A. Karjus, R. A. Blythe, S. Kirby, and K. Smith, “Challenges in detecting evolutionary forces in language change using diachronic corpora,” arXiv preprint arXiv:1811.01275, 2018.

A. Jatowt and K. Duh, “A framework for analyzing semantic change of words across time,” in IEEE/ACM Joint Conference on Digital Libraries, pp. 229–238, IEEE, 2014.

M. Hilpert, “The great temptation: What diachronic corpora do and do not reveal about social change,” Corpora and the Changing Society: Studies in the Evolution of English. Amsterdam and Philadelphia: John Benjamins, pp. 3–27, 2020.

G. M. Alessi and A. Partington, “Modern diachronic corpus-assisted language studies: methodologies fro tracking language change over recent time.,” 2020.

M. Liakata and P. Rayson, “Using Diachronic Corpora to Study Language Change and Evolution: A Review,” Corpus Linguistics and Linguistic Theory, vol. 8, no. 2, pp. 227–250, 2012.

S. Kemmer and E. Zaretsky, “Diachronic Corpus-Based Approaches to the Study of Semantic Change: A Review,” Corpus Linguistics and Linguistic Theory, vol. 11, no. 1, pp. 29–62, 2015.

M. Piotrowski, “Natural language processing for historical texts,” Synthesis lectures on human language technologies, vol. 5, no. 2, pp. 1–157, 2012.

M. L. Jockers and D. Mimno, “Significant themes in 19th-century literature,” Poetics, vol. 41, no. 6, pp. 750–769, 2013.

G. Blix, “The Social Role of Literature in Society,” Acta Universitatis Upsaliensis, vol. 11, no. 3, pp. 53–63, 1986.

J. Petzold and M. Dickinson, “Tracking Emotional Language in Literature over Time: A Corpus-Based Approach,” Corpus Linguistics and Linguistic Theory, vol. 8, no. 2, pp. 267–291, 2012.

T. L. C. Jockers and J. D. Porter, “A Quantitative Approach to Tracking Emotional Language in Literary Texts over Time,” Literary and Linguistic Computing, vol. 29, no. 1, pp. 115–132, 2014.

S. Fiedler and K. Kunz, “The Use of Content Analysis in the Study of Emotional Language in Literary Texts,” Methods of Empirical Linguistics, vol. 21, no. 1, pp. 17–38, 2014.

S. Fiedler and K. Kunz, “Thematic Analysis of Emotional Language in Literary Texts: A Diachronic Corpus-Based Study,” Corpus Linguistics and Linguistic Theory, vol. 5, no. 2, pp. 195–224, 2009.

S. Fiedler and K. Kunz, “Thematic Analysis of Emotional Language in Literary Texts: A Review of Methodological Approaches,” Methods of Empirical Linguistics, vol. 18, no. 1, pp. 1–23, 2012.

M. Mu¨ller and A. Panchenko, “Sentiment Analysis of Emotional Language in Literary Texts: A Compar- ative Study of Machine Learning and Dictionary-Based Approaches,” Corpus Linguistics and Linguistic Theory, vol. 7, no. 2, pp. 227–250, 2011.

S. Fiedler and K. Kunz, “Sentiment Analysis of Emotional Language in Literary Texts: A Review of Methodological Approaches,” Methods of Empirical Linguistics, vol. 21, no. 1, pp. 17–38, 2014.

J. Bjerring-Hansen, R. D. Kristensen-McLachlan, P. Diderichsen, and D. H. Hansen, “Mending fractured texts. a heuristic procedure for correcting OCR data,” 2022.

S. Nimb, S. Olsen, B. S. Pedersen, and T. Troelsg˚ard, “A thesaurus-based sentiment lexicon for danish: The danish sentiment lexicon,” in Proceedings of the Thirteenth Language Resources and Evaluation Conference, pp. 2826–2832, 2022.

B. S. Pedersen, S. Nimb, and S. Olsen, “Dansk betydningsinventar i et datalingvistisk perspektiv.,” Danske Studier, 2021.

Computational linguistic modelling of the temporal dynamics of scientific communication: a quantitative corpus study on the journal Nature

Gard Jenset[1], Isabell Landwehr[2], Barbara McGillivray[3] and Stefania Degaetano-Ortlieb[2]
[1]Springer Nature Group, [2] Saarland University, [3]King's College London and The Alan Turing Institute

We trace the linguistic evolution of English written scientific communication within the journal Nature, one of the world's leading multidisciplinary science journals, published since 1869. Our study applies computational models for diachronic linguistic analysis to investigate the statistical distribution of lexical and lexical-semantic features in a collection consisting of over 230,000 titles and abstracts from articles published in the journal Nature between 1869 and 2022, accessed via the Dimensions database (Hook et al. 2018).

We dynamically model changes in scientific language use over time. This overcomes the limitations of working with raw frequencies which tend to highlight only high-frequency features, disregarding low-frequency items (e.g. Biber and Gray 2016; Moskowich and Crespo 2012; Rissanen et al. 1997; Teich et al. 2016). We compare changes in probability distributions of individual lexical, grammatical, and semantic features with relative entropy as a measure of divergence for entire sets of features (e.g. all lemmas, parts of speech etc.), allowing for a comprehensive coverage of frequency bands. The dynamicity of the model is achieved by sliding over the timeline and continuously comparing adjacent time spans. The more a distribution of a feature changes over time, the higher the divergence will be, indicating changes in use. The sum of all features' divergence at a particular point in time gives an overall estimate of how much current language use is distinct from past practices, i.e. if a large number of features shows an increase in divergence over a time span, this will indicate a period of change. In terms of interpretability of the model, we are not only able to detect periods of change in a data-driven fashion, but can attribute these changes to sets of linguistic features that contribute to them. In addition, drawing on title and abstract embeddings for Nature articles using Google's Universal Sentence Encoder, we measure the trends in similarity between articles over time.

Previous work on the publications of The Royal Society of London (Degaetano-Ortlieb and Teich 2019, Degaetano-Ortlieb 2021) has proven the adaptability of applying dynamic divergence models to investigate change in scientific language use, showing specialisation trends at the lexical level and at the same time grammatical conventionalization trends. Sun et al. (2021) show similar results employing word embeddings methods. Research using embedding technologies applied to the labels of scientific disciplines rather than to the linguistic content has also found evidence for disciplines undergoing a process of global convergence combined with local specialisation (McGillivray et al. 2022). Previous work on Nature (Monastersky and Van Noorden 2019a) has shown specialisation of particular keywords in individual titles and abstracts. Our overarching question is whether these trends can be found for the journal Nature at scale, indicating general mechanisms of change in language use which contribute to the formation of the English scientific register. In addition, we are interested in changes that might be an indication of journal-specific linguistic features, especially considering the leading position of Nature in the scientific research landscape, as well as the journal's shift in focus over time (Monastersky and Van Noorden 2019a). We investigate the following sub-questions: (a) Can we observe similar/diverging diachronic trends between Nature and The Royal Society corpus, i.e. can we detect lexical and lexical-semantic diversification and grammatical conventionalization in Nature? (b) While we would assume similar diverging trends at the lexical level (new discoveries and technical advancement call for new linguistic expressions), do we encounter journal-specific trends at the grammatical and semantic level, and if so, are these disparate trends or do some trends start off in one journal and are picked up later in the other? Here we assume, besides grammatical trends indicating terminology formation processes, also changes in grammatical features that indicate text structuring functions (e.g. introductory linguistic material such as prepositional phrases or discourse markers) and those that meet expressive needs given extra-linguistic pressures, such as passive voice usage during periods of increased experimental work).

References

Biber, Douglas & Bethany Gray. 2016. Grammatical complexity in academic English: Linguistic change in writing. Studies in English Language. Cambridge, UK: Cambridge University Press.

Degaetano-Ortlieb, S. (2021). Measuring informativity: The rise of compounds as informationally dense structures in 20th century Scientific English. In Elena Soave and Douglas Biber (eds.), Corpus Approaches to Register Variation, chapter 11, John Benjamins Publishing Company, pp. 291-312.

Degaetano-Ortlieb, Stefania and Teich, Elke. "Toward an optimal code for communication: The case of scientific English" Corpus Linguistics and Linguistic Theory, vol. 18, no. 1, 2022, pp. 175-207.

Hook, D.W., Porter, S.J. and Herzog, C., 2018. Dimensions: building context for search and evaluation. Frontiers in Research Metrics and Analytics, 3, p.23.

McGillivray, B., Jenset, G.B., Salama, K. and Schut, D. 2022. Investigating patterns of change, stability, and interaction among scientific disciplines using embeddings. Humanities and Social Sciences Communications 9, 285. https://doi.org/10.1057/s41599-022-01267-5

Monastersky, Richard & Van Noorden, Richard. 2019a. 150 years of Nature: a data graphic charts our evolution. Nature: 575, 22-23. https://doi.org/10.1038/d41586-019-03305-w

Monastersky, Richard & Van Noorden, Richard. 2019b. 150 years of Nature: a data graphic charts our evolution. Supplementary information: Methodology. Nature: 575. https://www.nature.com/magazine-assets/d41586-019-03305-w/17345736 (last accessed date: 23 December 2022).

Moskowich, Isabel & Begona Crespo (eds.). 2012. Astronomy Playne and simple: The writing of science between 1700 and 1900. Amsterdam/Philadelphia: John Benjamins.

Rissanen, Matti, Merja Kytö & Kirsi Heikkonen (eds.). 1997. English in transition: Corpus-based studies in linguistic variation and genre analysis. Berlin: Mouton de Gruyter.

Sun, K., Liu, H. and Xiong, W., 2021. The evolutionary pattern of language in scientific writings: A case study of Philosophical Transactions of Royal Society (1665–1869). Scientometrics, 126(2), pp.1695-1724.

Teich, Elke, Stefania Degaetano-Ortlieb, Peter Fankhauser, Hannah Kermes & Ekaterina Lapshinova-Koltunski. 2016. The linguistic construal of disciplinarity: A data mining approach using register features. Journal of the Association for Information Science and Technology (JASIST) 67(7). 1668–1678.

Quantifying Changes in English Noun Compound Productivity and Meaning

Maximilian Maurer & Chris Jenkins & Filip Miletic & Sabine Schulte im Walde
Universität Stuttgart

Combinations of words are considered to be multi-word expressions (MWEs) if they are semantically idiosyncratic to some degree, i.e., the meaning of the combination is not entirely (or even not at all) predictable from the meanings of the constituents [Sag et al., 2002, Baldwin and Kim, 2010]. MWEs subsume multiple morpho-syntactic types, including noun compounds such as flea market, which have been explored extensively and across research disciplines from synchronic perspectives [Reddy et al., 2011, Bell and Schäfer, 2013, Schulte im Walde et al., 2013, Salehi et al., 2014, 2015, Schulte im Walde et al., 2016, Cordeiro et al., 2019, Alipoor and Schulte im Walde, 2020, i.a.], but state-of-the-art studies are lacking large-scale distributional approaches towards diachronic models of noun compound meaning.

The current study goes beyond the restricted synchronic concept of compound semantics and provides a novel diachronic perspective on meaning changes and compositionality (i.e., meaning transparency) of English noun compounds. We specifically investigate the diachronic evolution of the productivity of compound constituents relative to their degree of compositionality, relying on an established gold standard dataset with human compositionality ratings by Reddy et al. [2011] and a cleaned version of the English diachronic corpus CCOHA [Alatrash et al., 2020]. Given that type and token frequencies and probabilities, type-token ratios, entropy, etc. represent key concepts in determining quantitative properties of corpora as well as regarding individual word types and co-occurrences, we compute a range of statistical measures to quantify changes in productivity. These include Baayen's Large Number of Rare Events (LNRE) measures [Baayen, 2001], which have become a standard in statistical estimation of productivity, as well as measures that represent textual constants and therefore smooth the effect of different text lengths. For example, Tweedie and Baayen [1998] showed that with the exception of two measures, K suggested by Yule [1944] and Z suggested by Orlov [1983], all constants systematically change as a function of the text length.

In terms of empirical findings, we hypothesise that the current-language degree of compositionality differs for compounds with high- vs. low-productive constituents [Jurafsky et al., 2001, Hilpert, 2015, i.a.]. That is, we expect to find distinct analogical temporal development patterns for compositional compounds (such as maple tree, prison guard, climate change) in comparison to more idiosyncratic compounds (such as flea market, night owl, melting pot), with regard to modifier as well as head productivity. Our results constitute an important step towards a better understanding of compound semantics over time, as well as a reference point for future work deploying other modeling approaches on the same topic.

References

Reem Alatrash, Dominik Schlechtweg, Jonas Kuhn, and Sabine Schulte im Walde. CCOHA: Clean Corpus of Historical American English. In Proceedings of the 12th Language Resources and Evaluation Conference, pages 6958–6966, Marseille, France, 2020.

Pegah Alipoor and Sabine Schulte im Walde. Variants of Vector Space Reductions for Predicting the Compositionality of English Noun Compounds. In Proceedings of the 12th Language Resources and Evaluation Conference, pages 4379–4387, Marseille, France, 2020.

R. Harald Baayen. Word Frequency Distributions. Kluwer Academic Publishers, 2001.

Timothy Baldwin and Su Nam Kim. Multiword expressions. In Nitin Indurkhya and Fred J. Damerau, editors,
Handbook of Natural Language Processing, pages 267–292. CRC Press, Boca Raton, USA, 2010.

Melanie J. Bell and Martin Schäfer. Semantic Transparency: Challenges for Distributional Semantics. In Proceedings of the IWCS Workshop on Formal Distributional Semantics, pages 1–10, Potsdam, Germany, 2013.

Silvio Cordeiro, Aline Villavicencio, Marco Idiart, and Carlos Ramisch. Unsupervised Compositionality Prediction of Nominal Compounds. Computational Linguistics, 45(1):1–57, 2019.

Martin Hilpert. From hand-carved to computer-based: Noun-Participle Compounding and the Upward Strengthening Hypothesis. Cognitive Linguistics, 26(1):1–36, 2015.

Daniel Jurafsky, Alan Bell, Michelle Gregory, and William D. Raymond. Probabilistic Relations between Words: Evidence from Reduction in Lexical Production. In Joan Bybee and Paul Hopper, editors, Frequency and the Emergence of Linguistic Structure, Typological Studies in Language, pages 229–254. John Benjamins, Amsterdam / Philadelphia, 2001.

Y. K. Orlov. Ein Modell der Häufigkeitsstruktur des Vokabulars. In Studies on Zipf's Law, pages 154–233. Brockmeyer, Bochum, 1983.

Siva Reddy, Diana McCarthy, and Suresh Manandhar. An Empirical Study on Compositionality in Compound Nouns. In Proceedings of the 5th International Joint Conference on Natural Language Processing, pages 210– 218, Chiang Mai, Thailand, 2011.

Ivan A. Sag, Timothy Baldwin, Francis Bond, Ann Copestake, and Dan Flickinger. Multiword Expressions: A Pain in the Neck for NLP. In Proceedings of the Conference on Intelligent Text Processing and Computational Linguistics, Mexico City, Mexico, 2002.

Bahar Salehi, Paul Cook, and Timothy Baldwin. Using Distributional Similarity of Multi-way Translations to Predict Multiword Expression Compositionality. In Proceedings of the 14th Conference of the European Chapter of the Association for Computational Linguistics, pages 472–481, Gothenburg, Sweden, 2014.

Bahar Salehi, Paul Cook, and Timothy Baldwin. A Word Embedding Approach to Predicting the Compositionality of Multiword Expressions. In Proceedings of the Conference of the North American Chapter of the Association for Computational Linguistics/Human Language Technologies, pages 977–983, Denver, Colorado, USA, 2015.

Sabine Schulte im Walde, Stefan Müller, and Stephen Roller. Exploring Vector Space Models to Predict the Compositionality of German Noun-Noun Compounds. In Proceedings of the 2nd Joint Conference on Lexical and Computational Semantics, pages 255–265, Atlanta, GA, USA, 2013.

Sabine Schulte im Walde, Anna Hätty, and Stefan Bott. The Role of Modifier and Head Properties in Predicting the Compositionality of English and German Noun-Noun Compounds: A Vector-Space Perspective. In Proceedings of the 5th Joint Conference on Lexical and Computational Semantics, pages 148–158, Berlin, Germany, 2016.

Fiona J. Tweedie and R. Harald Baayen. How Variable May a Constant be? Measures of Lexical Richness in Perspective. Computers and the Humanities, 32:323–352, 1998.

G. Udny Yule. The Statistical Study of Literary Vocabulary. Cambridge University Press, 1944.

Modeling sound change to reconstruct protowords

Neige Rochant[1] and Marc Allassonnière-Tang[2]
[1]Sorbonne Nouvelle University, National Museum of Natural History[2]

We present the preliminary results of a linguistically informed probabilistic model of articulatory-motivated sound change. The model uses Markov chains whose probabilities of transition between two sounds are based on sound change universals. The model also considers the frequency of phonemes worldwide and the linguistic area or family, which allows the model to account for changes that are rare cross-linguistically, but expected in a specific area or family. The model implements rates of context-blind and context-sensitive change. The former are the absolute probability for a segment *x* to be replaced by a segment *y* or deleted regardless of the phonotactic context. They are implemented via a flexible architecture of decision trees. The latter are the conditional probability for a segment *x* to be replaced by a segment *y* (by assimilation or dissimilation), deleted, inserted or metathesized, depending on the phonotactic context.

Both types of change rates are based on linguistic theories. As an example of theory shaping context-blind change, learner/listener-oriented theories of sound change (Ohala 1981; Blevins 2007; Hale, Kissock, and Reiss 2013) suggest that some acoustic signals are more likely to be misparsed by the hearer, which in turn would lead to mispronouncing at production and hence result in a higher probability of sound change at the language level. Therefore, we deduced a general model, supported by Blevins (2004), with higher transition rates between articulations that are more likely to yield both similar acoustic signals and similar visual cues, viz. that are closer in the vocal tract (e.g. retroflexes and alveolars) and that differ by fewer laryngeal features (e.g. [p] and [b] as opposed to [pʰ] and [b]). In addition, transitions from stops or voiceless segments to fricatives, approximants or voiced segments are assigned a higher rate than the reverse, following the general intuition proven by Bybee and Easterday (2019) that lenition occurs more often than fortition. Furthermore, context-sensitive change rates will be key to approximating a realistic model. We implement sound change tendencies analyzed, e.g., by Blevins (2004), such as velar palatalization before high-front vowels and compensatory lengthening.

This paper contributes to the field of historical linguistics by introducing a transparent model for predicting sound change and inferring sound changes in the past. The model has theoretical applications for testing hypotheses about parameters affecting sound change, e.g. the weight of articulatory vs. analogical motivation. The model is expected to approximate protowords and infer time depths of language families by being calibrated based on known historical splits between languages, as is also performed with phylogenetic methods. Performance at these tasks, which require working with word lists of related languages, depends on the right calibration of sound change theories, e.g. the theory of regular sound change (Osthoff and Brugmann 1878) and the theory of lexical diffusion (Schuchardt 1885; Bybee 2002; 2007). For example, the model could apply reverse change to all instances of a phonological structure in the wordlist at once, while applying a higher lenition and deletion rate in frequent words to control for a Zipfian frequency distribution (Zipf 1935; Strauss, Grzybek, and Altmann 2007; Pierrehumbert 2001).

To assess the performance of the model, it is tested on languages families with abundant information on protowords (e.g. Indo-European).

References

Blevins, J. (2004). Evolutionary phonology: The emergence of sound patterns. Cambridge: Cambridge University Press.

(2007). “Interpreting Misperception: Beauty is in the ear of the beholder”. In: Experimental Approaches to Phonology. Ed. by S., P. Speeter Beddor, and M. Ohala. Oxford: Oxford University Press, pp. 144–154.

Bybee, J. (2002). “Word frequency and context of use in the lexical diffusion of phonetically conditioned sound change”. In: Language variation and change 14.3, pp. 261–290.

(2007). “Word Frequency in Lexical Diffusion and the Source of Morphophonological Change”. In: Frequency of Use and the Organization of Language. Oxford: Oxford University Press.

Bybee, J. and S. Easterday (2019). “Consonant strengthening: A crosslinguistic survey and articulatory proposal”.
In: Linguistic Typology 23.2, pp. 263–302. doi: doi:10.1515/lingty-2019-0015.

Hale, M., M. Kissock, and C. Reiss (2013). “An I-Language approach to phonologization and lexification”. In: Handbook of Historical Phonology. Ed. by P. Honeybone and J. Salmons. Oxford: Oxford University Press.

Ohala, J. J. (1981). “The listener as a source of sound change”. In: Papers from the parasession on language and behavior. Ed. by C. S. Masek, R. A. Hendrick, and M. F. Miller. Chicago: Chicago Linguistic Society, pp. 178––203.

Osthoff, H. and K. Brugmann (1878). Morphologische Untersuchungen auf dem Gebiete der indogermanischen Sprachen. Vol. 3. S. Hirzel.

Pierrehumbert, J. B. (2001). “Exemplar dynamics: Word frequency”. In: Frequency and the emergence of linguistic structure 45.137, pp. 10–1075.

Schuchardt, H. (1885). Ueber die Lautgesetze: gegen die Junggrammatiker. R. Oppenheim.

Strauss, U., P. Grzybek, and G. Altmann (2007). “Word length and word frequency”. In: Contributions to the science of text and language: word length studies and related issues. Ed. by P. Grzybek. Vol. 31.

Zipf, G. K. (1935). The psycho-biology of language: an introduction to dynamic philology. Boston: Houghton Mifflin.

A computerized investigation of Albanian diachronic phonology

Clayton Marr
The Ohio State University

Computerized forward reconstruction, or CFR (Sims-Williams, 2018), offers an automatic and systematic means of testing hypotheses about the chronology of sound change in a language. While computing the effects of historical sound changes over millennia for thousands of etyma is laborious and extremely time-consuming, this task is accomplished within seconds by a CFR system such as DiaSim, which was created for not only evaluating hypothesized relative chronologies of sound changes, or "diachronic cascades", but also "debugging them" by reporting statistics on how errors pattern (Marr and Mortensen, 2020). As a test case, past work applied this system to the phonological evolution of Latin into French, and a CFR-enabled "debugging" procedure improved accuracy from a 3.2% baseline for a cascade based on the 1934 received view to 84.9%. In the process, various proposals in the post-1934 literature on French were supported by the fact that they were independently produced as part of a systematic debugging process using DiaSim that was undertaken without reference to them (Marr and Mortensen, 2022), while the endeavor also may have revealed a new regular sound change in Old French, which was ultimately robustly supported by additional data (Marr, 2023b). However, as French boasts both a large corpus since medieval times and extensive past research, the experiment with French was more of a "laboratory run" to test the validity of the approach of debugging a language's historical phonology via CFR, a prelude to bringing it into the field as an investigative technique.

This paper will bring in CFR to tackle Albanian diachronic phonology, starting with the Latin stratum of the its lexicon. Given the lack or loss of attestation of Albanian before the 15th century and its status as the only surviving member of its branch of Indo-European (Rusakov, 2018), reconstruction of Albanian diachronic phonology, and thus of Proto-Albanian, has always leaned heavily on the outcomes of strata of loanwords in Albanian from better-attested sources (Orel, 2000). Of these, the Latin layer (Çabej, 1962; Bonnet, 1998) is by far the most significant. Latin loanwords are more numerous than inheritance from Proto-Indo-European, Proto-Albanian is dated in relation to the time of contact with Latin, and Albanian diachronic phonology is in a large part an exercise in generalization from analyses of the outcomes of ancient Latin loans (Orel, 2000; Demiraj, 2006; Rusakov, 2017; De Vaan, 2018), though with significant contributions from Albanian historical dialectology (Curtis, 2018) and the other "layers". Nevertheless, issues do remain that concern the Latin layer of Albanian, such as rival etymologies between imperial-era Latin loans and later Romance loans (Bonnet, 1998), and these have potential implications for the reconstruction of Proto-Albanian, and the greater mysteries of the language's history within the Balkans (Friedman and Joseph, 2022). Thus, an evaluation and debugging of the received view on Albanian diachronic phonology as applied to its largest single pillar, the Latin stratum, offers both a new approach to an old but still vexing problem, and a step for CFR as an empirical method, between the curated "lab" case of French, and the "field" of understudied languages and language families.

This endeavor will apply DiaSim to CLEA, a dataset compiled in 2020–2022 and to be released with this paper, of 1007 Albanian etyma of ancient Latin origin as asserted by at least one of a set of reputed references (Bonnet, 1998; Orel, 1998, 2000; De Vaan, 2018; Topalli, 2017; Çabej, 1986), and will work from a base cascade representing the views of Orel (2000) and De Vaan (2018). The same debugging process as Marr and Mortensen (2022) will be applied, with accuracy reported for modern Albanian outcomes, and discussion of any systematic patterning of errors and possible solutions proposed.

Keywords: computerized forward reconstruction, diachronic phonology, Albanian, Latin

References

Anamali, S. and Prifti, K. (2002). Historia e popullit shqiptar, volume 1. Botimet Toena. AShSh and IGJL (2006). Fjalor i gjuhës shqipe. Online at: https://fjalorthi.com/Fjalorthi.com. Bonnet, G. (1998). Les mots latins de l'albanais. L'Harmattan.

Brüch, J. (1922). Lateinische etymologien. Indogermanische Forschungen, 40(1):196–247.

Bufli, G. and Rocchi, L. (2021). A historical-etymological dictionary of turkisms in albanian (1555-1954).

Çabej, E. (1953). Grupet nd, ng në gjuhën shqipe. Buletin për Shkencat Shoqërore, 4.

Çabej, E. (1962). Zur Charakteristik der lateinischen Lehnworter im Albanischen. Revue de Linguistique (Bucureşti), 7:161–199.

Camarda, D. (1864). Saggio di grammatologia comparata sulla lingua albanese. Livorno: Succesore di Egisto Vignozzi.

Çabej, E. (1986). Studime gjuhësore, volume I–VIII. Prishtinë: Rilindja.

Curtis, M. C. (2012). Slavic-Albanian language contact, convergence, and coexistence. The Ohio State University. Ph.D. dissertation.

Curtis, M. C. (2018). 98. The dialectology of Albanian. In Handbook of Comparative and Historical Indo-European Linguistics, pages 1800–1811. De Gruyter Mouton.

Dashi, B. (2013). Italianismi nella lingua albanese. Edizioni Nuova Cultura.

De Vaan, M. (2004). PIE *e in Albanian. Die Sprache. Zeitschrift für Sprachwissenschaft, 44(1):70–85.

De Vaan, M. (2018). 95. The phonology of Albanian. In Handbook of Comparative and Historical Indo-European Linguistics, pages 1732–1749. De Gruyter Mouton.

Demiraj, B. (2018). 100. The evolution of Albanian. In Handbook of Comparative and Historical Indo-European Linguistics, pages 1812–1815. De Gruyter Mouton.

Demiraj, B. and Dayan, P. (1997). Albanische Etymologien: Untersuchungen zum albanischen Erbwortschatz, volume 134. Rodopi.

Demiraj, S. (2004). Gjuhësi Ballkanike. Akademia e Shkencave e Republikës së Shqipërisë, Instituti i Gjuhësisë dhe i Letërsisë.

Demiraj, S. (2006). Albanian. In The Indo-European Languages. Routledge.

Ducellier, A. (1981). La façade maritime de l'Albanie au moyen age, Durazzo et Valona du XIe au XVe siècle. Greece, Thessaloniki: Institute for Balkan Studies.

Fine, J. V. and Fine, J. V. A. (1994). The late medieval Balkans: A critical survey from the late twelfth century to the Ottoman conquest. University of Michigan Press.

Friedman, V. A. and Joseph, B. D. (2022). The Balkan Languages. Cambridge University Press. Gjinari, J. (1989). Dialektet e gjuhës shqipe. Tiranë: Akademia e Shkencave e RPS të Shqipërisë. Haarmann, H. (1972). Der lateinische Lehnwortschatz im Albanischen. Hamburg: Helmut
Buske.

Haspelmath, M. (2009). Lexical borrowing: Concepts and issues. Loanwords in the world's languages: A comparative handbook, pages 35–54.

Helbig, R. (1903). Die italienischen Elemente im Albanesischen. JA Barth.

Huld, M. E. (1979). An etymological glossary of selected Albanian items. University of California, Los Angeles.

Huld, M. E. (1986). Accentual stratification of Ancient Greek loanwords in Albanian. Zeitschrift für vergleichende Sprachforschung, 99(2. H):245–253.

Hyllested, A. and Joseph, B. D. (2022). 13. Albanian. In Olander, T., editor, The Indo-European Language Family, pages 227–248. Cambridge University Press.

Janda, R. D. and Joseph, B. D. (2003). Reconsidering the canons of sound-change. In Historical Linguistics 2001: Selected Papers from the 15th International Conference on Historical Linguistics, Melbourne, 13-17 August 2001, volume 237, page 205. John Benjamins Publishing.

Jokl, N. (1984). Sprachliche Beiträge zur Paläo-Ethnologie der Balkanhalbinsel:(zur Frage der ältesten griechisch-albanischen Beziehungen), volume 29. Austrian Academy of Sciences
Press.

Jorgaqi, K. (2001). Ndikimi i italishtes në letërsinë e vjetër shqipe: shek. XVI-XVII, Botimet Toena, Tiranë.

Joseph, B. D. (2020). Language contact in the Balkans. The handbook of language contact, pages 537–549.

Kore, M. K. (2013). Similarities between Albanian and Romanian in the entire language sub- systems. Mediterranean Journal of Social Sciences, 4(2):175–175.

Lafe, G. (2000). Der italienische Einfluß auf das Albanische. Zweiter Teil. Wörterbuch der Italianismen im Albanischen, in Ponto-Baltica, 10:31–120.

Landi, A. (1989). Gli elementi latini nella lingua albanese. Napoli: Edizioni Scientifiche Italiane.

Marr, C. (2023a). The angevin–albanian element in the albanian lexicon.

Marr, C. (2023b). A regular velar onset voicing rule for 12th century French. Submitted for publication.

Marr, C. and Mortensen, D. R. (2020). Computerized forward reconstruction for analysis in diachronic phonology, and Latin to French reflex prediction. In Proceedings of LT4HALA 2020-1st Workshop on Language Technologies for Historical and Ancient Languages, pages 28–36.

Marr, C. and Mortensen, D. R. (2022). Large-scale computerized forward reconstruction yields new perspectives in French diachronic phonology. Diachronica.

Matzinger, J. (2018). 98. The lexicon of Albanian. In Handbook of Comparative and Historical Indo-European Linguistics, pages 1788–1800. De Gruyter Mouton.

Mazaudon, M. and Lowe, J. B. (1993). Regularity and exceptions in sound change. In Annual Conference of the Linguistic Society of Belgium.

Meyer, G. (1891). Etymologisches Wörterbuch der albanesischen Sprache. Strasbourg: Trübner. Meyer-Lübke, W. (1935). Romanisches etymologisches Wörterbuch. Heidelberg: Carl Winter. Miklosich, F. (1870). Albanische Forschungen: I. Die slavischen Elemente im Albanischen.
Wien: Karl Gerold's Sohn; Denkschr. Akad. Wien XIX.

Miklosich, F. (1871). Albanische Forschungen: II. Die romanischen Elemente im Albanischen.
Wien: Karl Gerold's Sohn; Denkschr. Akad. Wien XIX.

Newmark, L. (1998). Albanian-English Dictionary. Oxford University Press, USA. Online at: http://www.seelrc.org:8080/albdict/Online Albanian Dictionary.

Niedermann, M. (1905). Contributions à la critique et à l'explication des gloses latines, vol- ume 1. Attinger.

Ölberg, H. M. (1972). Griechisch-albanische Sprachbeziehungen. Institut für vergleichende Sprachwissenschaft der Universität.

Orel, V. (1998). Albanian etymological dictionary. Brill.

Orel, V. Ė. (2000). A concise historical grammar of the Albanian language: reconstruction of Proto-Albanian. Brill.

Pokorny, J. (1959). Indogermanisches etymologisches Wörterbuch, volume 1. Bern, München: Francke Verlag.

Prifti, E. (2012). Aromanian elements of an Albanian argot.

Rusakov, A. (2017). Albanian. The Indo-European languages, pages 552–608.

Rusakov, A. (2018). 94. The documentation of Albanian. In Handbook of Comparative and Historical Indo-European Linguistics, pages 1716–1731. De Gruyter Mouton.

Schuhardt, H. E. M. (1866-8). Der Vokalismus des Vulgärlateins I-III. Leipzig: B. G. Teubner.

Seliščev, A. M. and Olesch, R. (1931). Slavjanskoe naselenie v Albanii. Böhlau.

Sims-Williams, P. (2018). Mechanising historical phonology. Transactions of the Philological Society, 116(3):555–573.

Svane, G. (1992). Slavische Lehnwörter im Albanischen, volume 67. Aarhus Universitetsforlag. Thumb, A. (1909). Altgriechische elemente des albanischen. Indogermanische Forschungen,
26:1–20.

Topalli, K. (2017). Fjalor etimologjik i gjuhës shqipe. Qendra e Studimeve Albanologjike, Instituti i Gjuhësisë dhe i Letërsisë.

Uhlisch, G. (1964). Neugriechische Lehnwörter im Albanischen. Typescript: Berlin.

Vătăşescu, C. I. (1997). Vocabularul de origine latină din limba albaneză în comparaţie cu româna. PhD thesis, Editura Universităţii din Bucureşti.

Vicario, F. (1993). L'influsso, lessicale veneto in albanese. Balkan-Archiv (NF), 17(18):1992. Ylli, X. (2000). Das slavische Lehngut im Albanischen: Teil 2: Ortsnamen. Peter Lang International Academic Publishers.

The LSCD Benchmark - A testbed for diachronic word meaning tasks

Dominik Schlechtweg
Universität Stuttgart

Lexical Semantic Change Detection (LSCD) is a field of NLP that studies methods automating the analysis of changes in word meanings over time. In recent years, this field has seen much development in terms of models, datasets and tasks (Schlechtweg et al., 2020). This has made it hard to keep a good overview of the field. Additionally, with the multitude of possible options for preprocessing, data cleaning, dataset versions, model parameter choice or tuning, clustering algorithms, and change measures a shared testbed with common evaluation setup is needed in order to precisely reproduce experimental results. Hence, we present a benchmark repository implementing evaluation procedures for models on most available LSCD datasets. We hope that the resulting benchmark by standardizing the evaluation of LSCD models and providing models with near-SOTA performance can serve as a starting point for researchers to develop and improve models. The benchmark allows for a wide application and testing of models by focusing on multilingual models and their evaluation on several languages.

Models solving the LSCD task often employ sub-models solving other related lexical semantic tasks like Word Sense Induction (WSI, Navigli, 2009) or Word-in-Context (WiC, Pilehvar & Camacho- Collados, 2020). Performance on these tasks can be evaluated separately contributing to optimization of individual model components and to facilitation of error analysis. However, existing data sets for the latter two tasks are usually synchronic, which makes it hard to compare different sub-models and select optimal ones for the LSCD task that requires good performance on diachronic data. Hence, we exploit existing, richly annotated LSCD datasets as evaluation data for WSI and WiC in a diachronic setting. Using the same data sets for evaluation of WSI, WiC and LSCD has the additional advantage that performance on the meta task LSCD can be directly related to performance on the subtasks WSI and WiC, as it can be assumed that performance on the subtasks directly determines performance on the meta task. We aim to stimulate transfer between the fields of WSI, WiC and LSCD by providing a repository allowing for evaluation on all these tasks with shared model components.

References

Dominik Schlechtweg, Barbara McGillivray, Simon Hengchen, Haim Dubossarsky, Nina Tahmasebi. 2020. SemEval-2020 Task 1: Unsupervised Lexical Semantic Change Detection. Proceedings of the 14th International Workshop on Semantic Evaluation.
Roberto Navigli. 2009. Word sense disambiguation: a survey. ACM Computing Surveys.
Mohammad Taher Pilehvar and Jose Camacho-Collados. 2019. WiC: The Word-in-Context Dataset for Evaluating Context-Sensitive Meaning Representations. Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies.

Model evaluation for diachronic semantics: A view from Portuguese and Spanish

Amaral, Patrícia*; Hu, Hai**; Tian, Zuoyu*; Kübler, Sandra*
*Indiana University; **Shanghai Jiao Tong University

For research on semantic change that spans over several centuries, assessing the accuracy of embeddings comes with two challenges: (i) native speakers who can provide judgments about meaning are not available, and (ii) historical corpora are often much smaller than contemporary datasets, which raises issues of model accuracy (Hellrich, 2019; Hu et al., 2021). This paper presents the lessons learned from developing intrinsic evaluations to test the quality of distributional models used to investigate semantic change in Medieval Spanish and Portuguese. For Spanish we experimented on a 7 million word corpus (Chronicles corpus, with texts from 13th-16th c.) (Hu et al., 2021) and for Portuguese on a ca. 2,5 million token corpus, CIPM, with texts from 12th-16th c. (Tian et al., 2021).

The lessons learned include the following: 1) We cannot use tests developed for modern languages/corpora off the shelf, since the tests' vocabulary (e.g., capitals of the world, country names and currencies) does not overlap with that of the historical corpus.

We cannot use tests developed for other historical corpora without adaptations since those corpora tend to be restricted to specific domains, which also leads to a lack of overlap in vocabulary.

We need to account for spelling and morphological variation, which are important features of many Medieval corpora. For the historical Spanish corpus, e.g., we had to delete the test "adjective to adverbs" from contemporary Spanish (Cardellino, 2016), which maps an adjective to its corresponding adverb inmente, since the variability of forms of adverbs in Medieval Spanish would have resulted in more than one possible target form, including multi-word expressions (Company and Flores Da´vila, 2014). Instead, we added tests for several types of inflection (verbal morphology, gender and number in adjectives). The morphology tests were generated by using vocabulary based on the frequency counts from the Chronicles corpus. A summary of our analogy test is given in Table 1.

If the corpora are very small, using analogy tests alone may not provide enough information. Our work on the Portuguese corpus shows that using different tests that include a range of relations is important. The tests we created include: word similarity, outlier detection, and coherence assessment (see Table 2 for a summary). The latter is based on Zhao et al. (2018), who proposed a new evaluation method for assessing the quality of domain-specific word embedding models. They assume that the neighbors of a given word embedding should have the same characteristics of that word (e.g. neighbors of drug names should be drug names). In the Portuguese corpus, names of people and places are frequent, thus we can assess coherence by reporting the percentage of neighbors generated for a proper noun that were also proper nouns.

To summarize: Given the importance of register in research on semantic and syntactic change, as well as orthographic and morphological variation in historical corpora, specific tests are re- quired for a proper assessment of distributional models in studies of semantic change. Overall, assessment of word embeddings for historical research must meet the following criteria: appropriateness (corpus vocabulary is taken into account), sustainability (i.e. not requiring extensive expert input), comprehensiveness (tasks target different types of relations, i.e. syntactic, semantic, morphological), and complementarity (avoiding the biases of individual methods).

Source	Category	Example	#Questions
MTS	Morphology nouns: kinship terms	padre madre : hijo hija	506
	Morphology verbs: third person singular	comer come : ir va	650
	Morphology verbs: infinitive to participle	saber sabido : tomar tomado	1190
	Morphology verbs: gerund to participle	sabiendo sabido : tomando tomado	1190
ours	Morphology adj.: singular to plural	negra negras : rica ricas	992
	Morphology adj.: singular to plural	negro negros : rico ricos	992
	Morphology adj.: masc to fem	negro negra : negros negras	992
	Morphology adj.: masc to fem	negros negras : ricos ricas	992
	Morphology nouns : singular to plural	casa casas: capilla capillas	1332
	Morphology/Semantics: antonyms	feliz infeliz : posible imposible	42
	Semantics: antonyms	cerca lejos : bien mal	342
Total			9220

Table 1: Structure of our analogy test; MTS denotes the analogy test from Mikolov et al. (2013), translated into Spanish.

Test	Categories	#Questions
Analogy Test	nouns: gender; nouns: singular to plural; verbs: 1st person singular to 3rd person singular; verbs: 3rd person singular to 3rd person plural; verbs: infinitive to 3rd person singular; verbs: infinitive to gerund	2994
Word Similarity	synonymous; related (not synonymous); not related	97
Outlier Detection	body parts; Christianity; color; food; geography; parts of buildings; titles/professions; war	512
Coherence Assessment	proper nouns (names of people and places)	25

Table 2: Summary of the benchmark for assessing word embeddings generated for Medieval Portuguese

References

Cardellino, C. (2016). Spanish Billion Words Corpus and embeddings. Online at https://crscardellino.github.io/SBWCE/; retrieved August 2019.

Company, C. and Flores Dávila, R. (2014). Adverbios en mente. In Company, C., editor, Sintaxis histórica de la lengua española. Tercera parte: Adverbios, preposiciones y conjunciones. Relaciones interoracionales, pages 1195–1340. Fondo de Cultura Económica y Universidad Nacional Autónoma de México.

Hellrich, J. (2019). Word Embeddings: Reliability and Semantic Change. PhD thesis, Jena University Language and Information Engineering Lab.

Hu, H., Amaral, P., and Kübler, S. (2021). Word embeddings and semantic shifts in historical Spanish: Methodological considerations. Digital Scholarship in the Humanities, 37(2):441–461.

Mikolov, T., Chen, K., Corrado, G., and Dean, J. (2013). Efficient estimation of word representations in vector space.

In Proceedings of International Conference on Learning Representations (ICLR), Scottsdale, AZ.

Tian, Z., Jarrett, D., Escalona Torres, J., and Amaral, P. (2021). BAHP: Benchmark of assessing word embeddings in historical Portuguese. In Proceedings of the 5th Joint SIGHUM Workshop on Computational Linguistics for Cultural Heritage, Social Sciences, Humanities and Literature, pages 113–119, Punta Cana, Dominican Republic (online). Association for Computational Linguistics.

Zhao, M., Masino, A. J., and Yang, C. C. (2018). A framework for developing and evaluating word embeddings of drug-named entity. In Proceedings of the BioNLP 2018 workshop, pages 156–160, Melbourne, Australia. Association for Computational Linguistics.

Using simulated data to evaluate models of Indo-European vocabulary evolution

Philipp Rönchen[1], Oscar Billing[1], and Tilo Wiklund[2]
1Department of Linguistics and Philology, Uppsala University, Sweden
2Chief Data Scientist, UAB Sensmetry, Vilnius, Lithuania

In the last two decades the project of using data from the lexicon of modern languages to make inferences about historical language stages, though long envisioned (Hymes 1960, Embleton 1986), has been gaining steam. Gray and Atkinson (2003), Bouckaert et al. (2012) and Chang et al. (2015) use increasingly sophisticated methods to estimate the age of Indo-European, however the results of the earlier studies run counter to the established majority opinion in historical linguistics (Pronk, 2022) and Chang et al.'s methodology gives a different result. This raises the question how different computational models can be validated (see Nakhleh et al. 2005, Ritchie and Ho 2019, Jäger 2019a and 2019b)

Ideally one would like to evaluate computational methods using held-out data sets and test cases in which the correct inferences are known. However, compared to other disciplines like biology, the amount of lexical data available in data bases is very limited and the precise history of most language families in the world is unknown, leaving only a few quite shallow families as potential test cases. Moreover, it is not clear whether the success of a computational model on a language family from one part of the world should generalise to other families, since different evolutionary mechanisms might have operated. To work around the lack of data available for validation, Greenhill et al. (2009), Murawaki (2015) and Bradley (2016) simulate data sets which they use to evaluate computational methods.

We create a large number of simulated data sets to evaluate the inferences of Chang et al. (2015) and Bouckaert et al. (2012) on Indo-European. Our data sets are specifically tailored to the methodologies of Chang et al. and Bouckaert et al. and try to mimic different plausible (though hypothetical) pre-histories of Indo-European, including loan events, a tree topology not too far from the consensus view in historical linguistics, and varying lexical change rates. We employ the computational fact that it is much easier to create realistic models for simulating data then it is to make inferences from existing data (see Kelly and Nicholls 2017 for the difficulties involved in constructing an inference method that allows for loans).

Both Chang et al.'s and Bouckaert et al.'s methodologies fail to correctly infer the age of Indo-European that was used to create our simulated data sets. We believe this warrants more investigation in the validity of different computational models.

References

Bouckaert, R., Lemey, P., Dunn, M., Greenhill, S. J., Alekseyenko, A. V., Drummond, A. J., Gray, R. D., Suchard, M. A., and Atkinson, Q. D. (2012). Mapping the origins and expansion of the Indo-European language family. Science, 337(6097):957–960.

Bradley, S. (2016). Synthetic language generation and model validation in BEAST2. arXiv preprint arXiv:1607.07931.

Chang, W., Cathcart, C., Hall, D., and Garrett, A. (2015). Ancestry-constrained phylogenetic analysis supports the Indo-European steppe hypothesis. Language, 91(1):194–244.

Embleton, S. M. (1986). Statistics in historical linguistics, volume 30. Brockmeyer.

Gray, R. D. and Atkinson, Q. D. (2003). Language-tree divergence times support the Anatolian theory of Indo-European origin. Nature, 426(6965):435–439.

Greenhill, S. J., Currie, T. E., and Gray, R. D. (2009). Does horizontal transmission invalidate cultural phylogenies? Proceedings of the Royal Society B: Biological Sciences, 276(1665):2299–2306.

Hymes, D. H. (1960). Lexicostatistics so far. Current anthropology, 1(1):3–44.

Jäger, G. (2019a). Computational historical linguistics. Theoretical Linguistics, 45(3-4):151–182.

Jäger, G. (2019b). Model evaluation in computational historical linguistics. Theoretical Linguistics, 45(3-4):299–307.

Kelly, L. J. and Nicholls, G. K. (2017). Lateral transfer in stochastic dollo models. The Annals of Applied Statistics, 11(2):1146–1168.

Murawaki, Y. (2015). Spatial structure of evolutionary models of dialects in contact. Plos one, 10(7):e0134335.

Nakhleh, L., Warnow, T., Ringe, D., and Evans, S. N. (2005). A comparison of phylogenetic reconstruction methods on an IE dataset. Transactions of the Philological Society, 3(2):171–192.

Pronk, T. (2022). Indo-European secondary products terminology and the dating of Proto-Indo-Anatolian. Journal of Indo-European Studies, 49(1&2):141–170.

Ritchie, A. M. and Ho, S. Y. (2019). Influence of the tree prior and sampling scale on Bayesian phylogenetic estimates of the origin times of language families. Journal of Language Evolution, 4(2):108–123.

Evaluating historical word embeddings: strategies, challenges and pitfalls

Oksana Dereza, Theodorus Fransen and John P. McCrae
University of Galway, Insight Centre for Data Analytics

When it comes to the quantitative evaluation of word embeddings, there are two main strategies: extrinsic, i.e. using pre-trained embeddings as input vectors in a downstream ML task, such as language modelling, and intrinsic, i.e. through analogy and similarity tasks that require special datasets (Bakarov, 2018).

Extrinsic evaluation

Language modelling seems to be the easiest way to evaluate historical word embeddings, since it is language independent, scalable and does not require dataset creation. Hypothetically, using pre-trained embeddings must lower the perplexity of a language model, even if these embeddings were trained on a different period of the same language. However, language modelling, as well as the majority of modern NLP tasks, is not very relevant to historical linguistics, so we might want to find a better downstream task or turn to intrinsic evaluation.

Intrinsic evaluation

There are two major tasks used for intrinsic evaluation of word embeddings: similarity and analogy. The **similarity task** consists in comparing similarity scores of two words yielded by an embedding model to those calculated based on experts' judgment. We did not explore this option, because it requires too much manual work by definition. The **analogy task** is simply asking an embedding model "What is to **a′** as **b** is to **b′** ?", and expecting **a** as an answer. Analogy datasets can be created automatically or semi-automatically if there exists a comprehensive historical dictionary of a language in question in machine readable format or a WordNet.

Traditionally, analogy datasets are based on pairwise semantic proportion and therefore every question has a single correct answer. Given the high level of variation in historical languages, such a strict definition of a correct answer seems unjustified. Therefore, in our Early Irish analogy dataset we follow the authors of BATS (Gladkova et al., 2016) providing several correct answers for each analogy question and evaluating the performance with set-based metrics, such as an average of vector offset over multiple pairs (3CosAvg).

Our dataset consists of 4 parts: morphological variation and spelling variation subsets were automatically extracted from eDIL (eDIL, 2019), while synonym and antonym subsets are translations of correspondent BATS parts proofread by 4 expert evaluators. However, the scores that Early Irish embedding models achieved on the analogy dataset were low enough to be statistically insignificant. Such a failure may be a result of the following problems:

The highest inter-annotator agreement score (Cohen's kappa) between experts was 0.339, which reflects the level of disagreement in the field of historical Irish linguistics. It concerns such fundamental questions as "What is a word? Where does it begin and end? What is a normalised spelling of a word at a particular stage of the language history?", which was discussed in (Doyle et al., 2018) and (Doyle et al., 2019) regarding tokenisation. It is arguable that it might be true for historical linguistics in general.

There is a lack of standardisation in different resources for the same historical language. For example, ~65% of morphological and spelling variation subsets, retrieved from eDIL, were not present in the whole Early Irish corpus retrieved from CELT (CELT, 1997), on which the biggest model was trained. As for synonym and antonym subsets, ~30% are missing in the corpus. Although our embedding models used subword information and were able to handle unknown words, such a discrepancy between the corpus, on which they were trained, and the historical dictionary, which became the source for the evaluation dataset, seriously affected the performance. This discrepancy originates from different linguistic views and editorial policies used by different text editors, publishers and resource developers throughout time.

References

Bakarov, A. (2018). A Survey of Word Embeddings Evaluation Methods. ArXiv:1801.09536 [Cs]. http://arxiv.org/abs/1801.09536

Bojanowski, P., Grave, E., Joulin, A., & Mikolov, T. (2016). Enriching Word Vectors with Subword Information. ArXiv:1607.04606 [Cs]. http://arxiv.org/abs/1607.04606

CELT: Corpus of Electronic Texts. (1997). University College Cork. http://www.ucc.ie/celt

Doyle, A., McCrae, J. P., & Downey, C. (2018). Preservation of Original Orthography in the Construction of an Old Irish Corpus. Proceedings of the LREC 2018 Workshop "CCURL2018 – Sustaining Knowledge Diversity in the Digital Age", 67–70.

Doyle, A., McCrae, J. P., & Downey, C. (2019). A Character-Level LSTM Network Model for Tokenizing the Old Irish text of the Würzburg Glosses on the Pauline Epistles. Proceedings of the Celtic Language Technology Workshop, 70–79. https://www.aclweb.org/anthology/W19-6910

EDIL 2019: An Electronic Dictionary of the Irish Language, based on the Contributions to a Dictionary of the Irish Language (Dublin: Royal Irish Academy, 1913-1976). (2019). www.dil.ie

Gladkova, A., Drozd, A., & Matsuoka, S. (2016). Analogy-based detection of morphological and semantic relations with word embeddings: What works and what doesn't. Proceedings of the NAACL Student Research Workshop, 8–15. https://doi.org/10.18653/v1/N16-2002

Kutuzov, A., Øvrelid, L., Szymanski, T., & Velldal, E. (2018). Diachronic word embeddings and semantic shifts: A survey. ArXiv:1806.03537 [Cs]. http://arxiv.org/abs/1806.03537

Schlechtweg, D., McGillivray, B., Hengchen, S., Dubossarsky, H., & Tahmasebi, N. (2020). SemEval-2020 Task 1: Unsupervised Lexical Semantic Change Detection. ArXiv:2007.11464 [Cs]. http://arxiv.org/abs/2007.11464