DYLEN: Diachronic Dynamics of Lexical Networks

Andreas Baumann
Department of English and American Studies, University of Vienna, Austria
andreas.baumann@univie.ac.at

Julia Neidhardt
Faculty of Informatics, TU Wien, Austria
julia.neidhardt@ec.tuwien.ac.at

Tanja Wissik
Austrian Centre for Digital Humanities, Austrian Academy of Sciences, Austria
tanja.wissik@oeaw.ac.at

Abstract

In this contribution we present a use case of the application of big language data and digital methods such as natural language processing, machine learning, and network analysis in the fields of digital humanities and linguistics by characterizing and modeling the diachronic dynamics of lexical networks. The proposed analysis will be based on two corpora containing 20 years of data with billions of tokens.

2012 ACM Subject Classification Human-centered computing → Social network analysis; Computing methodologies → Natural language processing; Computing methodologies → Machine learning

Keywords and phrases language change, language resources, natural language processing, network analysis, big data

Funding The project *Diachronic Dynamics of Lexical Networks (DYLEN)* is funded by the ÖAW go!digital Next Generation grant (GDNG 2018-020).

1 Background and Research Aims

Evidently, languages are constantly subject to change. For example, on the word level, new items enter the vocabulary (i.e. the lexical system) of a language, others cease to be used by speakers, and some established words may change their meaning.

Characterizing and modeling these dynamics has a broad field of applications including linguistics, natural language processing, digital humanities, artificial intelligence, computer sciences and cognitive sciences. In the project *Diachronic Dynamics of Lexical Networks* we therefore want to investigate, 1) how and why lexical systems of natural languages change, thereby considering social factors such as influential individuals as well as cognitive factors [3, 6, 11]; and 2) how language change in the lexical domain can be measured. Here, approaches such as corpus analysis and statistical analysis of word-frequency trajectories are typically employed in the field of diachronic linguistics (i.e. the analysis of language over time). Figure 1, for example, shows frequency trajectories of two lexical innovations. Recently, however, network-based approaches [1] have become increasingly important in this context [16, 9, 10, 4].

The advantage of network-based approaches for the analysis of lexical dynamics is that they allow to study the semantic properties of words in addition to word frequency, since the meaning of a word is closely related with its context, i.e. other words it co-occurs with frequently. So, we can track lexical innovations (i.e. new words) introduced by influential individuals (politicians) and systematically analyze contextual, i.e., semantic, changes of these words.

More specifically, our project focuses on the following research questions:

LDK 2019 - Posters Track.
Editors: Thierry Declerck and John P. McCrae

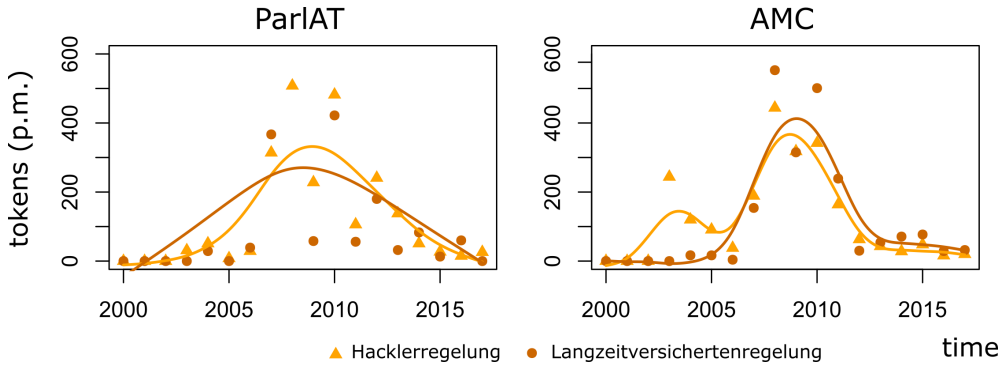

Figure 1 Frequency trajectories of two competing Austrian German terms, "Hacklerregelung" and "Langzeitversichertenregelung" (long-term insurance regulation). Both terms show a frequency increase during the observation period in ParlAT and AMC. Do they also undergo contextual change?

1. **How and why do lexical systems change?**
   - What is the role of influential innovators (e.g. politicians) in lexical change?
   - What determines the successful spread of lexical innovations?
   - Can we disentangle social factors from cognitive factors in lexical change?
2. **How can lexical change be measured?**
   - Does network science give more detailed answers about language change than traditional frequency based methods?
   - Which computational method is most suitable to analyze the evolution of lexical networks through time?
   - How can we enrich the digital-humanities toolbox with the output of the project?

2 Used Data Sets

As data sets we use two diachronically layered big text corpora available for Austrian German: the Austrian Media Corpus (AMC), containing more than 20 years of journalistic prose [15] and the ParlAT corpus, covering the Austrian parliamentary records of the last 20 years [21]. The journalistic prose included in the Austrian Media corpus comprises Austrian press agency releases, most Austrian periodicals such as all daily national newspapers as well as a large number of the major weekly and monthly magazines, in total 53 different newspapers and magazines.

Moreover, the Austrian Media Corpus contains also transcripts of Austrian television news programs, news stories and interviews [15]. In total, the AMC contains 10.500 million tokens with 40 million wordforms and 33 million lemmas. The ParlAT corpus contains the stenographic records, in German called "Stenographische Protokolle" from the XX to the XXV legislative period (1996 – 2017). So they are not transcripts of recordings but shorthand records. The corpus size is 75 million tokens with over 0.6 million word forms and 0.4 million lemmas [21]. Both corpora are tokenized, part-of-speech tagged and lemmatized.

Crucially, the two corpora cover lexical innovations both directly in the linguistic output of politicians as well as indirectly in media texts. Thus, the two corpora provide an ideal testing ground for the hypotheses outlined above.

3 Approach and Expected Outcome

To address the questions mentioned in section 1, we analyze the above described data sets, namely the Austrian Media Corpus (AMC), and the ParlAT corpus. In addition, we will provide an easy-to-use online tool to enable researchers to do diachronic analyses of lexical networks by themselves. Our approach requires the following steps, which are schematically depicted in Figure 2:

1. **NLP pre-processing and data model development:** For both corpora (i.e. AMC and ParlAT) a number of data pre-processing steps have already been conducted, i.e. tokenization, part-of-speech tagging, segmentation, lemmatization, named-entity (NE) recognition. Parts of the existing NE recognition will be enhanced using machine learning and semantic knowledge bases, e.g. Wikidata [19]. Furthermore, we will introduce a comprehensive data model combining both corpora and all metadata. In addition we will compile a list of relevant Austrian politicians as we want to analyze their impact on language change.
2. **Network construction and description:** A systematic procedure will be defined to 1) construct different co-occurrence networks (i.e. networks, where nodes represent identified entities, e.g. politicians, as well as nouns, verbs or adjectives and edges represent the co-occurrence of these nodes in a sentence, paragraph or document) for different time intervals (i.e. all documents within a week, a month, a year, etc.); and 2) extract basic properties (e.g. number of nodes/edges, clustering, centrality) to describe the networks. Together with frequency of occurrence, these properties can be interpreted cognitively and semantically [9, 10].
3. **Network analyses and comparisons:** In-depth analyses of the resulting networks will be conducted using network analysis and visualization. As the number of networks is assumed to be quite large, an approach will be developed to systematically compare these networks over time and across the two corpora. Therefore, different methods from network analysis, machine learning and statistical modeling will be tested. This will allow to identify relevant parameters (e.g. network properties) to capture diachronic developments.
4. **Modeling diachronic developments:** Statistical models including time-series analysis with generalized additive models and time-series clustering techniques for analyzing the co-evolution of parameters (see 3.) in multiple networks will be employed.
5. **Interactive web application:** A web-based interactive tool will be developed that retrieves the constructed networks and allows to explore, analyze and visualize them.

The technical implementation, which will build on an existing prototype [7], will mainly be based on Python and appropriate libraries [5, 8, 12, 14], on Neo4j [20] to store the network and on software for big data analysis , e.g. Apache Spark [23], Hadoop Yarn [17, 18], HDFS [17]. Gephi [2] will be used to visualize the graphs, and R for the statistical analyses [13, 22].

We expect our project, which has to face specific challenges such as NE recognition for Austrian German and the analysis of two large-scale diachronic corpora, to contribute to the understanding of the role that influential speakers and other linguistic factors play in lexical change by analyzing big amounts of language data. Since we cover both the linguistic output of influential speakers (ParlAT) as well as their linguistic reflex (AMC), we can test if lexical innovations introduced by these individuals behave differently than other lexical innovations. This allows us to disentangle social effects from cognitive effects in the process of lexical spread. For example, by analyzing the evolution of the clustering coefficients of

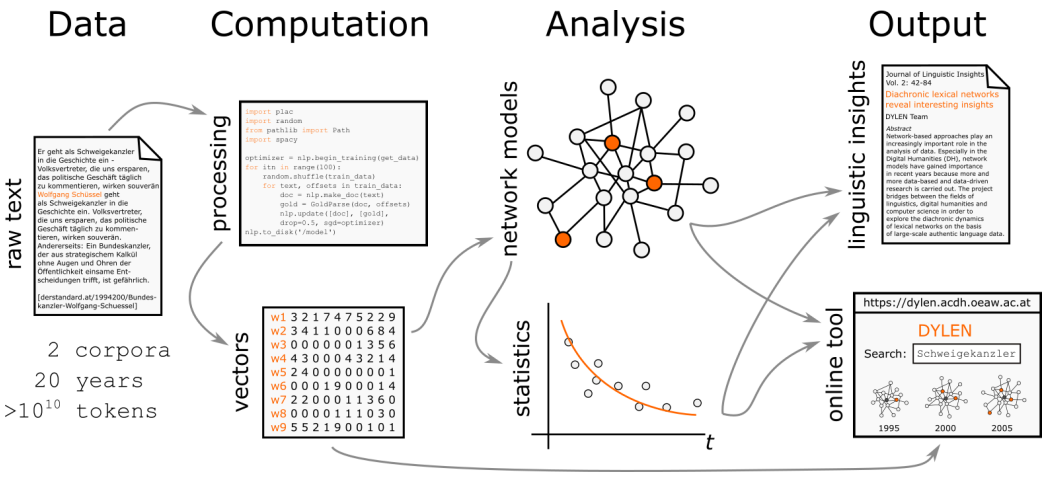

Figure 2 Work flow in the DYLEN project. Lexical networks are generated from diachronically layered corpus data. Network properties of lexical items, such as semantic neighborhood density, are then investigated across time to derive insights into semantic change.

networks around lexical innovations, we can test if increase in frequency is accompanied by semantic widening effects; a correlation which is expected given results from research on language change [3, 6, 11].

We also seek to foster network theory as a suitable tool to analyze and make sense of diachronic language data in the linguistic research community.

References

1 Albert-László Barabási. *Network science*. Cambridge university press, 2016.

2 Mathieu Bastian, Sebastien Heymann, and Mathieu Jacomy. Gephi: an open source software for exploring and manipulating networks. In *Third international AAAI conference on weblogs and social media*, 2009.

3 Joan Bybee. *Language, usage and cognition*. Cambridge University Press, 2010.

4 Heng Chen, Xinying Chen, and Haitao Liu. How does language change as a lexical network? an investigation based on written chinese word co-occurrence networks. *PloS one*, 13(2):e0192545, 2018.

5 Gabor Csardi, Tamas Nepusz, et al. The igraph software package for complex network research. *InterJournal, Complex Systems*, 1695(5):1–9, 2006.

6 Nick C Ellis, Matthew Brook O'Donnell, and Ute Römer. The processing of verb-argument constructions is sensitive to form, function, frequency, contingency and prototypicality, 2014.

7 Gabriel Grill, Julia Neidhardt, and Hannes Werthner. Network analysis on the austrian media corpus. In *VSS 2017 - Vienna young Scientists Symposium*, pages 128–129, 2017.

8 Aric Hagberg, Pieter Swart, and Daniel S Chult. Exploring network structure, dynamics, and function using networkx. Technical report, Los Alamos National Lab.(LANL), USA, 2008.

9 William L Hamilton, Jure Leskovec, and Dan Jurafsky. Cultural shift or linguistic drift? comparing two computational measures of semantic change. In *Proceedings of the Conference on Empirical Methods in Natural Language Processing. Conference on Empirical Methods in Natural Language Processing*, volume 2016, pages 2116–2121. NIH Public Access, 2016.

10 William L Hamilton, Jure Leskovec, and Dan Jurafsky. Diachronic word embeddings reveal statistical laws of semantic change. *arXiv preprint arXiv:1605.09096*, 2016.

11 Martin Hilpert and Florent Perek. Meaning change in a petri dish: constructions, semantic vector spaces, and motion charts. *Linguistics Vanguard*, 1(1):339–350, 2015.

12 Eric Jones, Travis Oliphant, and Pearu Peterson. SciPy: Open source scientific tools for Python. 2014.

13 Pablo Montero, José A Vilar, et al. TSclust: An R package for time series clustering. *Journal of Statistical Software*, 62(1):1–43, 2014.

14 Travis E Oliphant. *A guide to NumPy*, volume 1. Trelgol Publishing USA, 2006.

15 Jutta Ransmayr, Karlheinz Mörth, and Matej Ďurčo. *AMC (Austrian Media Corpus) - Korpus-basierte Forschungen zum Österreichischen Deutsch*, pages 27–38. Verlag der Österreichischen Akademie der Wissenschaften, 2017.

16 Eyal Sagi, Stefan Kaufmann, and Brady Clark. Tracing semantic change with latent semantic analysis. *Current methods in historical semantics*, 73:161–183, 2011.

17 Konstantin Shvachko, Hairong Kuang, Sanjay Radia, and Robert Chansler. The hadoop distributed file system. In *2010 IEEE 26th symposium on mass storage systems and technologies (MSST)*, pages 1–10. IEEE, 2010.

18 Vinod Kumar Vavilapalli, Arun C Murthy, Chris Douglas, Sharad Agarwal, Mahadev Konar, Robert Evans, Thomas Graves, Jason Lowe, Hitesh Shah, Siddharth Seth, et al. Apache Hadoop Yarn: Yet another resource negotiator. In *Proceedings of the 4th annual Symposium on Cloud Computing*, page 5. ACM, 2013.

19 Denny Vrandečić and Markus Krötzsch. Wikidata: a free collaborative knowledge base. 2014.

20 J Webber. A programmatic introduction to Neo4j in: Proceedings of the 3rd annual conference on systems, programming, and applications: Software for humanity, 217–218. *ACM*, 2012.

21 Tanja Wissik and Hannes Pirker. ParlAT beta corpus of austrian parliamentary records. In Darja Fišer, Maria Eskevich, and Franciska de Jong, editors, *Proceedings of the LREC2018 Workshop ParlaCLARIN*. European Language Resources Association, 2018.

22 Simon N Wood. *Generalized additive models: an introduction with R*. Chapman and Hall/CRC, 2017.

23 Matei Zaharia, Mosharaf Chowdhury, Michael J Franklin, Scott Shenker, and Ion Stoica. Spark: Cluster computing with working sets. *HotCloud*, 10(10-10):95, 2010.