VARIATION IN LANGUAGE USE ACROSS SOCIAL VARIABLES
A Data-driven Approach

STEFANIA DEGAETANO-ORTLIEB
SAARLAND UNIVERSITY

We present a data-driven approach to study language use over time according to social variables (henceforth SV), considering also interactions between different variables. Besides sociolinguistic studies on language variation according to SVs (e.g., Weinreich et al. 1968, Bernstein 1971, Eckert 1989, Milroy and Milroy 1985), recently computational approaches have gained prominence (see e.g., Eisenstein 2015, Danescu-Niculescu-Mizil et al. 2013, and Nguyen et al. 2017 for an overview), not least due to an increase in data availability based on social media and an increasing awareness of the importance of linguistic variation according to SVs in the NLP community.

In our project, we aim at investigating possible linguistic profiles of social variables, their interaction and their development over time. For this, we use the Old Bailey Corpus (OBC; Huber et al. 2012), a diachronic corpus of the *Proceedings of the Old Bailey*, London's criminal court, ranging from 1720-1913. The OBC is annotated with linguistic information (e.g., tokens, parts of speech) and includes annotation of SVs (e.g., age, gender, and social class of the speaker based on the HISCO standard). The utterances in the corpus amount at approx. 18 million tokens.

To detect linguistic patterns of variation at different linguistic levels (e.g., lexical, syntactic) according to SVs, we use Kullback-Leibler Divergence (KLD; Kullback and Leibler 1951, Fankhauser et al. 2014). KLD measures the distance between two probability distributions based on linguistic units (e.g., morphemes, words, parts of speech). It calculates the average amount of additional bits needed to encode linguistic units of a distribution A (e.g., male) by using an optimal code of a distribution B (e.g., female). The more additional bits are needed, the more distinct or distant A and B are. Moreover, by KLD ranking we can inspect typical features of a given distribution (e.g., features typical of female vs. male). This allows us to combine macro- and micro-analytic perspectives.

We present analyses on variation based on interactions between SVs. Consider, for example, differences between male and female based on social class. The biggest difference is found between male of higher vs. female of lower class. This difference is seen overall (KLD 0.23) but also over the time period investigated (1720-1913). Within gender, difference between female of higher and lower class is stronger (KLD 0.112) than between male of higher and lower class (KLD 0.08). Again, this tendency is reflected over time. Moreover, we also inspect features contributing to these differences. For example, female of lower class are distinguished by self-reference (e.g., *I*) and lexis covering roles (e.g., *master*, *servant*, *mistress*) as well as locations (e.g., *kitchen*, *home*, *stairs*), while females of higher class are distinguished by a more prominent use of prepositions (e.g., *of*, *for*, *in*), conjunctions (e.g., *and*, *but*) and determiners (e.g., *a*, *this*) reflecting a more elaborate style.

In the talk, we will present our methodology as well as macro- and micro-analytic analyses on linguistic variation based on social variables.

References

Bernstein B. 1971, *Class, Code and Control: Volume 1 Theoretical Studies towards a Sociology of Language*, Routledge Taylor & Francis Group, London and New York.

Danescu-Niculescu-Mizil C., Sudhof M., Jurafsky D., Leskovec J. and Potts C. 2013, A Computational Approach to Politeness with Application to Social Factors, in *Proceedings of the 51st Annual Meeting of the Association for Computational Linguistics*, volume 1, Sofia, Bulgaria, pp. 250-259.

Eckert P. 1989, *Social Categories and Identity in the High School*, Teachers College Press, New York.
Eisenstein J. 2015, *Written Dialect Variation in Online Social Media*, in Boberg C., Nerbonne J. and Watt D. (eds.), *Handbook of Dialectology*, Wiley.
Fankhauser P., Knappen J. and Teich E. 2014, *Exploring and Visualizing Variation in Language Resources*, in *Proceedings of the Ninth International Conference on Language Resources and Evaluation* (LREC'14), Reykjavik, Iceland, pp. 4125-4128.
Huber M., Nissel M., Maiwald P. and Widlitzki B. 2012, *The Old Bailey Corpus. Spoken English in the 18th and 19th centuries*. www.uni-giessen.de/oldbaileycorpus
Kullback S. and Leibler R.A. 1951, *On Information and Sufficiency*, in "The Annals of Mathematical Statistics" 22 [1], pp. 79-86.
Milroy J. and Milroy L. 1985, *Linguistic Change, Social Network and Speaker Innovation*, in "Journal of Linguistics" 21 [2], pp. 339-384.
Nguyen D., Dogruöz A.S., Rosé C.P. and de Jong F. 2016, *Computational Sociolinguistics: A Survey*, in "Computational Linguistics" 42 [3], pp. 537-593.
Weinreich U., Labov W. and Herzog M.I. 1968, *Empirical foundations for a theory of language change*, in Lehmann W.P. and Malkiel Y. (eds.), *Directions for Historical Linguistics: A Symposium*, pp. 95-188, University of Texas Press, Austin.