Low Saxon corpus-based dialectometry

Janine Siewert
University of Helsinki
janine.siewert@helsinki.fi

In connection with our research project on diachronic and synchronic variation in Low Saxon, we investigate dialect similarity in 19th and 21st century Low Saxon based on data from Germany and the Netherlands. Traditionally, Low Saxon dialect classification has mostly been based on phonological and morphological traits, like the ones presented by Schröder (2004). In this study, however, we are focusing on the orthographic and the (morpho-)syntactic side and compare how these relate to each other and to the more traditional classifications.

The majority of our dataset is taken from the the LSDC dataset (Siewert et al., 2020), from relevant prose texts from Leopold and Leopold (1882)[1] and the Twentse Taalbank (van der Vliet, 2021). Our overall dataset covers eight dialect regions from the 19th, 20th and 21st century, but in this study, we use the 19th and 21st century data from the five major West Low Saxon dialect groups: Dutch North Saxon, German North Saxon, Dutch Westphalian, German Westphalian and Eastphalian. Overall, these consist of 34,460 sentences and 345,131 tokens from the 19th century, and 44,740 sentences and 740,849 tokens from the 21st century which we have converted to CoNLL-U format and automatically PoS tagged.

One interesting area to pay attention to with respect to dialect distance is the Dutch-German border. Like Goossens (2019) observed, the Low Saxon dialects along the border have started to diverge under the influence of the majority languages. According to him, this divergence is most pronounced at the lexical level, but convergence towards the majority language has also been attested in phonology, morphology and syntax. While studies on the divergence of dialects along the border often focus on the occurrence and frequency of particular traits based on interviews, cf. Smits (2011), this study addresses the overall (dis)similarity in prose texts.

Dialect similarity at the orthographical level based on character n-grams will be compared to dialect distance based on PoS tag sequences to investigate if these lead to different dialect groupings. Malmasi and Zampieri (2017) observed in their experiments for identifying Swiss German dialects that approaches based on character n-grams outperform word-based ones and, in their study on British dialects, Wolk and Szmrecsanyi (2016) have employed part-of-speech n-grams for corpus-based dialectometry concluding that this approach can achieve results comparable to manually selected features. We will combine these with clustering approaches on the one hand and principal component analysis (PCA) on the other hand.

As in the 19th century, school education and majority language media played a smaller role in

[1] Digitised by dbnl: `https://dbnl.nl/tekst/leop008sche00_01/`

everyday life compared with today, we hypothesize that the effect of language contact with Dutch and German is less visible in the morphology and syntax of 19th century Low Saxon, even though the border is probably already clearly discernable at the orthographic level. Therefore, we will investigate how these results compare to more modern data from the 21st century.

References

Ingrid Schröder. Niederdeutsch in der Gegenwart: Sprachgebiet – Grammatisches – Binnendifferenzierung. In Dieter Stellmacher, editor, *Niederdeutsche Sprache und Literatur der Gegenwart*, pages 35–97. Georg Olms Verlag, Hildesheim, Zürich and New York, 2004.

Janine Siewert, Yves Scherrer, Martijn Wieling, and Jörg Tiedemann. LSDC - a comprehensive dataset for low Saxon dialect classification. In *Proceedings of the 7th Workshop on NLP for Similar Languages, Varieties and Dialects*, page 25–35, Barcelona, Spain (Online), December 2020. International Committee on Computational Linguistics (ICCL). URL `https://www.aclweb.org/anthology/2020.vardial-1.3`.

Joh. A. Leopold and L. Leopold. *Van de Schelde tot de Weichsel*. J.B. Wolters, Groningen, 1882.

Goaitsen van der Vliet. Twentse taalbank. `http://www.twentsetaalbank.nl/`, 2021. Accessed: 2021-12-15.

Jan Goossens. „Dialektverfall" und „Mundartrenaissance" in Westniederdeutschland und im Osten der Niederlande. In Gerhard Stickel, editor, *Varietäten des Deutschen: Regional- und Umgangssprachen*, pages 399–404. De Gruyter, 2019. doi: doi:10.1515/9783110622560-023. URL `https://doi.org/10.1515/9783110622560-023`.

Tom Smits. Dialectverlies en dialectnivellering in nederlands-duitse grensdialecten. *Taal en Tongval*, 63(1):175–196, 2011.

Shervin Malmasi and Marcos Zampieri. German dialect identification in interview transcriptions. In *Proceedings of the Fourth Workshop on NLP for Similar Languages, Varieties and Dialects (VarDial)*, pages 164–169, 2017.

Christoph Wolk and Benedikt Szmrecsanyi. Top-down and bottom-up advances in corpus-based dialectometry. *The future of dialects: Selected papers from Methods in Dialectology XV*, 1:225, 2016.