Leiwand Oida[1]: Geolocating Regional Linguistic Variation of German on Twitter

Bettina Larl[1] & Eva Zangerle[2]
[1] Linguistics, University of Innsbruck, [2] Databases and Information Systems, University of Innsbruck
Bettina.Larl@uibk.ac.at

Keywords: web corpus, geolocation, language variation

Twitter has been used for collecting language data and linguistic research in a variety of languages. (Goncalves & Sànchez 2014; Eisenstein, O'Connor, Smith & Xing 2014; Yuan, Guo, Kasakoff, Grive 2016). The proposed poster demonstrates the process of building a large Twitter corpus containing geolocated Tweets from the Deutscher Sprachraum (German language area) and investigates how German language varieties are used on Twitter.

German is the widest spread language within the European Union. German is a pluricentric language with three standard varieties: German Standard German, Swiss Standard German and Austrian Standard German. The official borders between Germany, Austria and Switzerland also form the official boundaries between the three standards. In addition to those national varieties, there are multiple varieties on the regional and dialectal spectrum. (Ammon 2015; Clyne 1992)

Easy access and its open API has made Twitter a popular source of data for research in various scientific fields and Twitter data shows great potential for linguistic research in multiple areas of expertise. Of particular interest for this poster are the tracking and exploring of regional linguistic variation of German on Twitter: Is there, for example, a connection between the language output and the geographic location tweets were sent from? To address such questions, a Twitter corpus of geotagged German Tweets within the Deutscher Sprachraum has been built. (Larl & Zangerle 2017)

This poster explores and describes the process of building the geotagged Twitter corpus of German tweets as well as giving a first glimpse into version.1 of the corpus.

The corpus version.1 currently contains tweets collected over a period of 30 (+1) months (January 2015 to June 2017).[2]

The Tweets were collected using the public Twitter Streaming API. 85,810,255 geolocated Tweets could be retained within a geographic rectangle (5.85, 46.016667 and 17.1, 55.016667) that covers the Deutscher Sprachraum. These tweets were re-filtered by removing those geolocalised outside of Austria, Germany, Switzerland or Italy (South Tyrol). Twitter's own language detector found 71 different languages within this data set. Subsequently, the corpus was filtered to only retain Tweets that were identified as German. The data was further refined by removing Tweets with missing latitude and/or longitude coordinates or other such deficiencies. In total 18,645,263 German Tweets, sent from within Austria, Germany, Switzerland and the German speaking part of Italy South Tyrol, could be processed and added to the corpus. The data has been tokenised with the SoMaJo-Tokeniser (Proisl, Uhrig 2016) and tagged with the SoMeWeTa-Tagger (Proisl 2018). The Metadata consists of coordinates, town name, country, date, time, ID. Within the corpus you can find Tweets from 452.501 individual users.

The corpus includes texts, hyperlinks and emoticons, as those can be seen as linguistic features. (Beißwenger 2015)

This poster describes the process from data to corpus and explains the various challenges that were encountered

[1] [Viennese; eastern Austria] Awesome, Dude!

[2] The collection process is still ongoing but will end at the end of June 2018. This will result in introducing another 12 months of Tweets – Tweets sent from July 2017 to June 2018 – to the corpus. This corpus version.2 will then contain Tweets with a character limit of 140 and such with a character limit of 280.

along the way. Furthermore, a first version of the corpus on CQP-web (restricted access only!) will be available for preview on sight.

References

Ammon, Ulrich; Bickel, Hans; Ebner, Jakob; et. al. [ed.] (2004). *Variantenwörterbuch des Deutschen: Die Standardsprache in Österreich, der Schweiz, Deutschland, Liechtenstein, Luxemburg, Ostbelgien und Südtirol.* Walter de Gruyter: Berlin.

Ammon, Ulrich (2015). *Die Stellung der deutschen Sprache in der Welt.* Walter de Gruyter: Berlin/München/Boston.

Ammon, Ulrich (1995). *Die deutsche Sprache in Deutschland, Österreich und der Schweiz. Das Problem der nationalen Varietäten.* Walter de Gruyter: Berlin/New York.

Barbaresi, Adrien (2016). Collection and Indexing of Tweets with a Geographical Focus. In: *Proceedings of the 4th Workshop on Challenges in the Management of Large Corpora (CMLC). Tenth International Conference on Language Resources and Evaluation (LREC 2016)*, pp.24-27.

Beißwenger, Michael (2015). Sprache und Medien: Digitale Kommunikation. In: *Studikurs Sprach- und Textverständnis. E-Learning-Angebot der öffentlichrechtlichen Universitäten und Fachhochschulen und des Ministeriums für Innovation, Wissenschaft und Forschung (MIWF) des Landes Nordrhein-Westfalen.*

Beißwenger, Michael, Horsmann, Tobias, Zesch, Torsten (2017). Part-of-speech Tagging for Corpora of Computer-mediated Communication: A Case Study on Finding Rare Phenomena. In: Fišer, Darja, Beißwenger, Michael (Eds.): *Investigating Computer-Mediated Communication: Corpus-Based Approaches to Language in the Digital World.* Ljubljana: Ljubljana University Press (Translation Studies and Applied Linguistics), pp. 192-219.

Bouvier, Gwen (2015). What is a discourse approach to Twitter, Facebook, YouTube and other social media: Connecting with other academic fields? *Journal of Multicultural Discourses*, 10(2), pp. 149-162.

Clyne, Michael (1992). German as a Pluricentric language. In: Clyne, Michael [ed.] (1992): *Pluricentric Languages. Differing Norms in Different Nations*. Mouton de Gruyter: Berlin, New York, pp. 117-148.

Eisenstein, J., O'Connor, B., Smith, N. A., & Xing, E. P. (2014). Diffusion of lexical change in social media. *PloS one*, 9, e113114.

Goncalves, Bruno & Sànchez, David (2014). Crowdsourcing Dialect Characterization through Twitter. *PLoS one,* 9(11), e112074.

Larl, Bettina & Zangerle, Eva (2017). Geolocating German on Twitter. Hitches and glitches of building and exploring a Twitter corpus. *9th International Corpus Linguistics Conference, 24 to Friday 28 July 2017, University of Birmingham.*

Proisl, Thomas (2018). SoMeWeTa: A Part-of-Speech Tagger for German Social Media and Web Texts. In: *Proceedings of the Eleventh International Conference on Language Resources and Evaluation (LREC 2018)*. Miyazaki: European Language Resources Association (ELRA), pp. 665–670.

Proisl, Thomas, Peter Uhrig (2016). SoMaJo: State-of-the-art tokenization for German web and social media texts. In: *Proceedings of the 10th Web as Corpus Workshop (WAC-X) and the EmpiriST Shared Task.* Berlin: Association for Computational Linguistics (ACL), pp. 57–62.

Scheffler, Tatjana (2014). A German Twitter Snapshot. In: *Proceedings of LREC, Reykjavik, Iceland.*

Storrer, Angelika (2013). Sprachstil und Sprachvariation in sozialen Netzwerken. In: Frank-Job, Barbara, Mehler, Alexander, Sutter, Tilmann [ed.] (2013): *Die Dynamik sozialer und sprachlicher Netzwerke. Konzepte, Methoden und empirische Untersuchen an Beispielen des WWW*. Springer Fachmedien: Wiesbaden, pp. 331-366.

Yuan, Hauang, Guo, Diansheng, Kasakoff, Alice, Grive Jack (2016). Understanding U.S. regional linguistic variation with Twitter data analysis. *Computers, Environment and Urban Systems*, 59, pp. 244-255.

Zappavigna, Michele (2015). Searchable talk: the linguistic functions of hashtags. *Social Semiotics*, 25(3), pp. 274-291.