Geolocating German on Twitter
Hitches and Glitches of Building and Exploring a Twitter Corpus

Bettina Larl and Eva Zangerle (University of Innsbruck, Austria)

About 16% of EU residents speak German as a native language, which makes it the most widespread language within the European Union. While German is the official language in Austria, Germany and Switzerland, the language differs widely in the three countries: 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 boundary between the three language standards. Additionally, to those national varieties, there are multiple varieties on the regional and dialectal spectrum.

Languages, and thus Linguistics, have always been influenced by technological developments and new forms of media. Each new development has brought new methods and approaches of how language can or should be studied and explored. Because of easy access and informal communication methods, increasing numbers of oral markers are being incorporated into written language. This is often showcased on social media platforms such as *Twitter.* Every tweet includes language output in the form of short messages that can contain different regional markers. Tweets can be geolocated, which means these language outputs can be assigned to the geographic location they were tweeted from.

This paper explores and describes the process of building a geotagged Twitter corpus of German tweets and the exploration of a preliminary sample. To research questions like "*Is there a connection between the language output and the geographic location tweets were sent from?*" and "*Could, for example, lexical varieties be allocated to a specific region by geolocation information provided in tweets?*" we are building a Twitter Corpus. The Corpus contains tweets collected via the Twitter streaming API, using a binding box around the rough approximation of the *Deutscher Sprachraum* and re-filtering the results for Tweets sent within Germany, Austria, Switzerland and South Tyrol/Italy. The data was gathered over a period of 24 months and more than 60.000.000 tweets were collected.

In this paper, we show and illustrate the way from data to corpus and how we address various challenges along the way.

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.

Bouvier, Gwen (2015): *What is a discourse approach to Twitter, Facebook, YouTube and other social media: connecting with other academic fields?* In: Journal of Multicultural Discourses, vol. 10, no. 2. 149-162.

Deppermann, Arnulf & Linke, Angelika (2010): *Sprache intermedial. Stimme und Schrift, Bild und Ton*. (Jahrbuch des Instituts für Deutsche Sprache 2009). Berlin: de Gruyter.

Gonçalves, Bruno & Sánchez, David (2014): *Crowdsourcing Dialect Characterization through Twitter.* PLoS ONE 9(11): e112074.

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

Saif, Hassan; He, Yulan; Fernández, Miriam; Alani, Harith (2014): *Semantic patterns for sentiment analysis of Twitter.* In: The Semantic Web – ISWC 2014, Springer International Publishing, 324–340.

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

Zappavigna, Michele (2015): *Searchable talk: the linguistic functions of hashtags.* In: Social Semiotics, vol. 25, no. 3. 274-291.