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

Bettina Larl, Eva Zangerle
University of Innsbruck
E-mail: bettina.larl@uibk.ac.at, eva.zangerle@uibk.ac.at

Abstract

Languages, and thus Linguistics, have always been influenced by technological developments and new media forms and every development brought new methods and approaches of how language can or should be studied and explored. About 16% of the EU residents speak German as a native language and this makes it 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 boundary between the three standards.

Because of easy access and informal communication methods, more and more oral markers find their way 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 characteristics. Tweets can be geolocated, which means these language outputs can be assigned to the geographic location they were tweeted from.

To explore 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. This paper shows preliminary findings of hand sampling a random sample of 1,000,000 Tweets.

Keywords: Twitter, geolocation, German