Universität Zürich[UZH]

Geographisches Institut

Universitärer Forschungsschwerpunkt Sprache und Raum

Spatial analysis of lexical variation in German using Twitter

André Rodrigues[1] (Supervised by Ross Purves[1] & Hanna Ruch[2])
[1]Geographisches Institut, [2]UFSP Sprache und Raum

1. Introduction

Geography in social media

Location of a user can be derived from...

- Geotag
- Words with geographical affinity (Cheng et al. 2010)
- Topic (Eisenstein et al. 2010)
- Social ties between users (Rout et al. 2013)
- Message form, i.e. geolinguistic variation

Geolinguistic variation in Twitter

Relationship between user location and language use

- Identification of Spanish macrodialects using Twitter (Gonçalves & Sánchez 2014)

German as a pluricentric language

Pluricentric language: a language with an official status in more than one nation and variation in the linguistic norms between the centres/nations (Clyne 1992)

Regional variation in the German standard language between Germany, Austria and Switzerland (Eichhoff 1977-2000, Ammon et al. 2004)

- Lexis (e.g. D *Rad* – Oe *Radl* – CH *Velo*)
- Orthography (e.g. D, Oe *Fuß* – CH *Fuss*)
- Phonetics/phonology (e.g. lexical stress D, Oe SBB – CH SBB)
- Morphosyntax (e.g. diminutive D *-lein*, *-le*, *-chen* – Oe *-erl*, *-(e)l* – CH *-li*)
- Pragmatics (e.g. D *Hallo* – CH *Hoi*)

Aims

- Test suitability of German Twitter data for geolinguistic analysis
- Analyse geographic distribution of lemmas that are known to show regional variation in the German standard language

Research questions

- Is the use of helvetisms, austriazisms and teotonisms equally evident on Twitter as in newspapers and literature?
- Is there geographic variation in orthography?
- Can lexical accommodation (i.e. adaptation) between Twitter users be observed?

2. Methods

Tweets were collected from the Twitter API using a bounding box-based query with keywords defined from a list of frequent lemmas based on entries in *Variantenwörterbuch des Deutschen* (Ammon et al. 2004) and, for lemma frequencies, the *Deutsches Referenzkorpus* (Institut für Deutsche Sprache Programmbereich Korpuslinguistik 2013).

Tweets and associated information (e.g. anonymised user IDs, time, content, location, Tweet characteristics) were stored in a bespoke database (**Fig. 1**). 490596 tweets were collected (**Fig. 2**).

Bots (automatically generated texts, for example radio-station playlists, weather reports and advertising) were identifed and filtered using a set of heuristics based on location (Chu et al. 2012)

After filtering for Bots, and selecting Tweets located in three countries of interest, 352835 Tweets were retained for analysis.

Tweets corresponding to particular variants were mapped in space using density surfaces, and tested for variation using X^2 statistics at national level for Germany, Austria and Switzerland (**Fig. 3**).

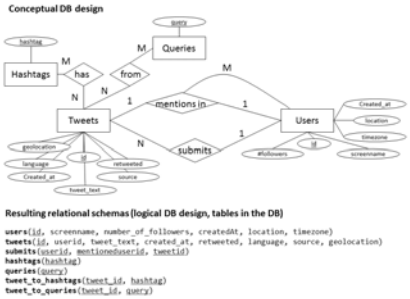

Figure 1: Conceptual design of database used to store Tweets and associated information

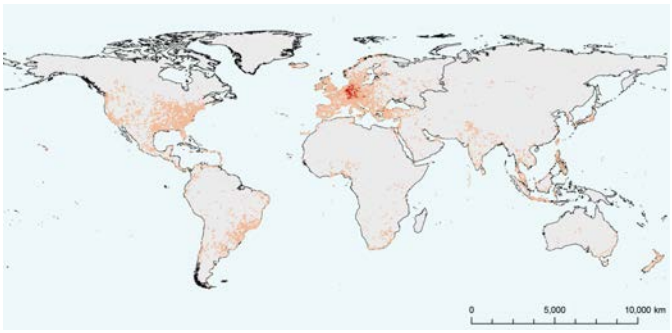

Figure 2: Origins of 490596 collected Tweets – 427084 (87%) were found in Germany, Austria or Switzerland (including bots) – 352835 were used in variant analysis

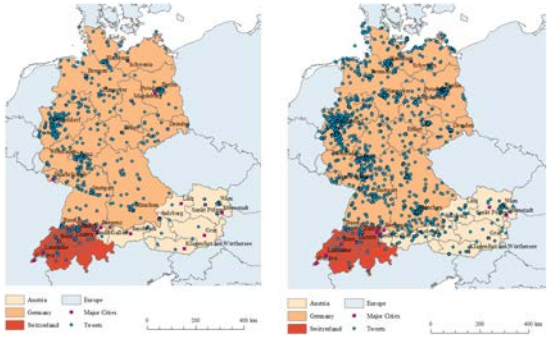

Figure 3: Geographical distribution of tweets containing *Ferien* (left) and *Urlaub* (right).

3. Results and Discussion

Geographical distribution of lemmas

- Expected differences in frequency observed for several lemmas (e.g. *Urlaub*, *Kiez*, *Fahrrad, Endspiel, Bundestag*)
- Othographical variation (e.g. *Fuss* in Switzerland, *Fuß* in Germany/Austria)
- Many lemmas and their regional variants did not generate enough tokens for statistical analysis despite large dataset

Method

Twitter data can be used to analyse how the pluricentric language German varies lexically in this social medium; however

- Very noisy data -> several steps of data cleaning required
- Few users generate majority of tweets
- Actual location of a user not necessarily in line with their origin
- Restrictions on data collection by Twitter (e.g. on user information -> analysis of accommodation challenging)

Future work

- Analysis of semantic fields to allow for exclusion of homonyms
- Part of speech tagging to allow for analysis of syntactic variation
- Use lemma list based on frequency of each of the regional variants to generate more tokens of interest
- Longer time period of data collection -> more tokens