Analysing Spatial Distribution of Linguistic Variables in Geoencoded Tweets from Croatia, Bosnia, Montenegro and Serbia

Nikola Ljubešić,[*] Tanja Samardžić,[†] Maja Miličević[‡]

[*] Department of Knowledge Technologies, Jožef Stefan Institute
Jamova cesta 39, SI-1000 Ljubljana
nikola.ljubesic@ijs.si

[†]CorpusLab
University of Zurich
Plattenstrasse 54, CH-8032 Zurich
tanja.samardzic@uzh.ch

[‡] Faculty of Philology
University of Belgrade
Studentski trg 3, SR-11000 Belgrade
m.milicevic@fil.bg.ac.rs

1. Introduction

Twitter has recently become a popular medium for spatial analysis of language (Doyle, 2014; Jørgensen, et al, 2015), given that (1) it has an open, high-quality API, and (2) a small, but significant percentage of the messages it contains is geoencoded.

In this demo we will present a tool which is currently under development (https://github.com/scopes-reldi/geotweet) that enables researchers interested in spatial variation of language to define a geographic perimeter of interest, collect data from the Twitter streaming API published in that perimeter, filter the obtained data by language and location, define and extract variables of interest, and analyse the extracted variables by one spatial statistic and two spatial visualisations.

We will demonstrate the tool on the area and a selection of languages spoken in former Yugoslavia. By defining the perimeter, the languages and a series of linguistic variables of interest we will illustrate the tool's data collection, processing and analysis capabilities. The linguistic variables we focus on are those that are known to vary in our area of interest, more specifically across the highly similar languages of Croatian, Bosnian, Montenegrin and Serbian.

The only previous work on Twitter data for the linguistic areas mentioned above that we are aware of is Ljubešić and Kranjčić (2015), where the focus is on discriminating between the languages. Our intention in this paper is to address a new topic, namely, the spatial distribution of specific linguistic phenomena often recognised as characteristics of the languages / varieties in this area. The two main goals of our research activities in this domain are (1) further development of the methodology for analysing linguistic variation via geoencoded social media, and (2) a comparison of actual data with the distributions expected on the basis of the literature, or widely accepted beliefs.

The remainder of this abstract gives an overview of the tool functionality.

2. Data Collection

In order to start the data collection, the user needs to enter his/her Twitter API credentials (obtained from the Twitter Developer site) and define the geographic perimeter of interest. Once started, the data collection component communicates with the Public Twitter Streaming API and stores the messages satisfying the perimeter criterion into a relational database.

3. Data Processing

There are two main functionalities of the data processing module: data filtering and variable extraction.

3.1. Data Filtering

Currently there are three user filtering techniques implemented in the tool: filtering by the minimum number of posts collected from a user, filtering by the country in which most of a user's tweets were posted, and filtering by the language of the majority of a user's tweets.

3.2. Variable Extraction

There are three main mechanisms for defining variables to be extracted. The first one enables extracting metadata from the Status objects, e.g. number of retweets, posting time, whether the tweet is a reply to another tweet etc. The second mechanism enables defining a variable based on a lexicon / token list, while the third one allows the user to define the variable via regular expressions. During the demo session each of the mechanisms will be showcased on a series of example variables relevant for the languages in question.

The final result of the data processing module is a simple tab-delimited file that can be used either in the data analysis module, which is described next, or in some other tool chosen by the user.

4. Data Analysis

The analysis module consists of three functionalities: point visualisation, spatial trend detection, and the identification of dominant regions per variable level.

4.1. Point Visualisation

The point visualisation functionality allows the user to gain an initial impression of the spatial distribution of all levels of a linguistic feature. Each tweet containing a value for the inspected variable is represented on a map as a point, with the value of the variable level encoded by colour. The text of the tweet can be obtained by clicking on the point.

4.2. Spatial Trend Detection

The spatial trend detection functionality comprises a measure that quantifies the spatial dependency in the data, often referred to as spatial autocorrelation. We compare the spatial distances as computed between all tweets of one linguistic feature (expected distances) with the distances as calculated for each feature level separately (observed distances). Aggregating these two sets of distances into what we call a relative distance measure allows us to distinguish feature levels that are spatially clustered (observed distance < expected distance) from levels that are scattered in space (observed distance > expected distance).

4.3. Dominance Maps

Dominance maps visualise the dominant levels of a variable throughout a map. They are particularly useful when many measurements are available and points start to overlap, making by-point visualisation hard to decipher.

All three analysis functionalities will be showcased during the demonstration on variables extracted with the data processing module.

5. References

Gabriel Doyle. 2014. Mapping dialectal variation by querying social media. In Proceedings of the 14th Conference of the European Chapter of the Association for Computational Linguistics, pages 98–106, Gothenburg, Sweden, April. Association for Computational Linguis- tics.

Anna Jørgensen, Dirk Hovy, and Anders Søgaard. 2015. Challenges of studying and processing dialects in social media. In Proceedings of the Workshop on Noisy User- generated Text, pages 9–18, Beijing, China, July. Association for Computational Linguistics.

Nikola Ljubešić and Denis Kranjčić. 2015. Discriminating between Closely Related Languages on Twitter. Informatica, 39(1):1–8.