CS 8803 Replication Project Proposal

Yuval Pinter, Ian Stewart

November 17, 2017

1 Introduction

We propose to replicate Shoemark *et al.*'s [SSS$^+$17] study of the Scottish independence referendum that took place in September 2014 [Mul14]. The authors found that Twitter users who openly supported Scottish independence were more likely to use Scottish words than those who openly opposed independence, and that UK Twitter users in general are more likely to use Scottish in non-referendum tweets than in referendum tweets. The study is important because it explores the political implications of language variation, which has been acknowledged as a marker of cultural and personal identity but less acknowledged as a marker of political identity.

For our replication project, instead of focusing on the relationship between Scots and UK politics, we will look at the relationship between the Catalan language and Spanish politics. In October 2017, the semi-autonomous region of Catalonia held a referendum on independence from Spain, and 92% of respondents voted for independence. [Fot17] We are interested in testing the relationship between regional language use and political attitudes. Similar to Shoemark *et al.*, we expect to find a positive correlation between the probability of using region-specific language (in this case, Catalan language) and the probability of supporting Catalonian independence.

2 Replication strategy

Data We have already begun mining data related to the Catalan independence debate using a set of hashtags developed from iterative querying of Twitter[1]. As a proof of concept, it appears that a significant proportion of these tweets are written in Catalan (see "Methods"), which suggests that there is enough language variation to support the analysis. Our final data will be extracted from a 1% sample of Twitter data similar to Shoemark *et al.*'s data, following their procedure of first mining from a geographic region and then extracting relevant hashtags that cooccur with a "seed" set.

Methods We will use the same methods as Shoemark *et al.*, except for the method of identifying language variation. Instead of identifying individual words that mark language variation, we propose to identify the language of each tweet and use Catalan versus Spanish as the language variables. We will use the *langid* package[2] [LB12] to identify the language of each tweet. Initial tests show that when restricting the output to high-confidence predictions, *langid* is capable of differentiating Spanish,

[1] Full hashtag list: #DemocraciaMarcaEspaña, #CataluñaLibre, #RepúblicaCatalana, #Independencia-Cataluña, #CatalunyaNoEstasSola (pro); #NoALosPaísesCatalanes, #CataluñaEsEspaña, #BarcelonaEsEspaña, #EspañaUnida, #EspanaNoSeRompe (anti).

[2] https://github.com/saffsd/langid.py

Catalan and irrelevant tweets with high accuracy. Manual inspection of a small sample of tweets also suggests that code-switching between Spanish and Catalan is not a concern, and that the number of users who tweet in both Spanish and Catalan is not small.

Metrics Our main goal is to replicate both Experiments 1 and 2 from the original paper, which have straightforward hypothesis tests comparing the probabilities of using variant and standard language. These results are binary (significantly different or not) which makes them easy to replicate. We may not be able to replicate the exact amount of data collected, but as long as we have a sufficient coverage of the different data combinations to test then the statistical power of the hypothesis tests should be comparable to the original study.

Risks The biggest risk is the potential for insufficient data, which could happen during the geographic data mining step due to the sparsity of geotagged data on Twitter. One workaround would be identifying geotagged data based on user location rather than tweet location, which would be a slight deviation from Shoemark *et al.*'s methodology but would provide a wider coverage of data. Another workaround would be to access an alternative data source, such as historical Twitter data scraped from the site through the *GetOldTweets* module[3].

Limits We may not be able to replicate the second experiment, which requires comparing each Twitter user's referendum tweets with control tweets. Data sparsity again may prevent us from collecting a sufficient control sample without resorting to extreme measures such as scraping each Twitter user's timeline to search for control tweets. In the worst case of being unable to replicate this experiment, we still believe that replicating the first experiment on its own is valuable because of how it extends Shoemark *et al.*'s findings to a bilingual situation.

3 Proposed timeline

We propose the following timeline and division of labor among team members to complete our replication project.

1. Nov. 17 - Nov. 24: Re-collect and filter data.
   - Yuval - Language ID for tweets; implement user filtering and data cleaning, produce unified dataset for both experiments.
   - Ian - Mine tweets by region (Spain and Catalonia): use per-city mining in GetOldTweets and in Twitter archive.
2. Nov. 24 - Dec. 1: Run analysis.
   - Yuval - Experiment 1; if insufficient data then consider broader set of referendum-related hashtags.
   - Ian - Experiment 2; if insufficient control data for users then mine additional control tweets.
3. Dec. 2 - Dec. 6: Write-up, potentially try stretch goal.

[3] https://github.com/Jefferson-Henrique/GetOldTweets-python

- Yuval - Write; stretch goal 2
- Ian - Write; stretch goal 1

Stretch goals The experiments in Shoemark *et al.* are straightforward to implement and verify, which leaves some room for stretch goals. One interesting extension (1) would be to look at the role of audience design in determining Catalan versus Spanish usage, by comparing each user's broadcast tweets with their @-reply tweets. Similar to prior work [NTC15], we may find that bilingual users are more likely to use Catalan when they have a smaller intended audience.

Another stretch goal (2) would be to try and bridge the gap between the linguistic variables in the Scottish experiment, where usage context is somewhat controlled for by focusing on parallel terms (Table 3), and in ours, where only language ID is extracted. Consequently, we can look for meaningful individual words in their Spanish and Catalan forms and compare statistics based on their appearance to those obtained by the much coarser tweet language variable we're using in our reproduction.

References

[Fot17] Alasdair Fotheringham. Catalan independence referendum: Region votes overwhelmingly for secession from Spain. *Independent*, 2017.

[LB12] Marco Lui and Timothy Baldwin. langid. py: An off-the-shelf language identification tool. In *ACL*, pages 25–30, 2012.

[Mul14] Tom Mullen. The Scottish independence referendum 2014. *Journal of Law and Society*, 41(4):627–640, 2014.

[NTC15] Dong Nguyen, Dolf Trieschnigg, and Leonie Cornips. Audience and the Use of Minority Languages on Twitter. In *ICWSM*, pages 666–669, 2015.

[SSS+17] Phillippa Shoemark, Debnil Sur, Luke Shrimpton, Iain Murray, and Sharon Goldwater. Aye or naw, whit dae ye hink? Scottish independence and linguistic identity on social media. In *EMNLP 2017*, pages 1239–1248, 2017.