European Language Ecology and Bilingualism with English on Twitter

Steven Coats
University of Oulu, Finland
English Philology, Faculty of Humanities, 90014 University of Oulu, Finland
Email: steven.coats@oulu.fi

Abstract

Societal and demographic changes have contributed to increasing bi- and multilingualism in European countries in recent years, and communication on social media platforms such as Twitter reflects this linguistic diversity. While high rates of English use online have been attested for many European countries by survey research, relatively little work has quantified the extent to which English is used on social media in European contexts. In this study, English use and bilingualism with English in Europe are investigated on Twitter. A large corpus of Twitter messages with geographical metadata was created by accessing the Twitter APIs. After language detection and filtering, linguistic profiles for European countries were created and the behavior of bi- and multilingual users examined. The analysis supports some previous findings that suggest that a large-scale language shift towards English may be ongoing in Europe in some communicative domains. Geographical differences shed light on the dynamics of this process.

Keywords: Bilingualism, social media, Twitter, corpus linguistics, quantitative methods

1. Introduction and Background

Recent years have seen an increase in the relative prominence of computer-mediated communication (CMC) modalities such as texting, instant messaging, or posting on social media, and platforms such as Twitter have become multilingual sites with global representation (Mocanu et al. 2013; Leetaru et al. 2013). At the same time, population movements and changes in education and media consumption have contributed to an increasing bi- and multilingualization of local environments, particularly with English – trends that are particularly evident in online communicative domains in some European societies. Although national languages continue to receive reinforcement in education and media, bilingualism with English has become the norm for many within Europe, particularly for young people.

In this study, bi- and multilingualism with English are investigated by means of a quantitative analysis of Twitter messages with location metadata in order to establish a language ecology (Haugen 1972). The research poses the following questions: Which languages are favored by multilinguals on Twitter in Europe? How linguistically diverse are European societies on the platform, and what role does English play? And to what extent do national languages play a role in the discourse of European Twitter users? Addressing these questions may allow us to characterize European Twitter discourse in terms of a language ecology that can "tell us something about where [a] language stands and where it is going in comparison with other languages of the world" (Haugen 1972, p. 337).

In a first step, the linguistic behavior on Twitter of users who can be reliably located within European countries is examined according to country in order to provide an overview of the language ecoloy of Europe. In a second step, the aggregate network behavior of bi- and multilingual users is examined more closely: Which languages do multilinguals favor in which places? The structure of the network of multilinguals between languages can shed light on the relative status of English and national languages and, due to the prevailing demographics of Twitter users, perhaps provide an indication of middle- to long-term language shift for European societies.

2. Previous Work: Twitter Language and Multilingualism

A number of studies of CMC and Twitter language have investigated aspects of English, including phenomena such as the discourse functions of hashtags (Wikström 2014; Squires 2015), lexical innovation in American English (Eisenstein et al. 2014), African-American Vernacular English dialect on Twitter (Jørgensen, Hovy, and Søgaard 2015), grammatical variation in English-language Twitter from Finland and the Nordic countries (Coats 2016a; Coats 2016b), or the interaction between demographic parameters such as gender with lexical and grammatical features in American English (Bamann, Eisenstein, and Schnoebelen 2014).

Ronen et al. (2014) found that English plays an important central role in multilingual networks of Wikipedia editors, book translations, and Twitter users. Hale (2014) investigated global multilingual networks on Twitter, including the network associations of retweets and user mentions, and found that while most interaction networks are language-based and English is the most important single mediating language, other languages collectively represent a larger bridging force. Eleta and Golbeck (2014) examined the tweets of 92 multilingual Twitter users and showed that their language choice on the Twitter reflects the predominant language of their social networks. Kim et al. (2014) used Shannon Entropy to quantify linguistic diversity on Twitter in Switzerland, Quebec and Qatar. They created networks of mono-, bi- and multilinguals, and demonstrated that while English mediates between language communities, users of local languages have more influence. Topic selection may also influence language choice. Such findings have confirmed the status of English as the global *lingua franca*, but the dynamics of multilingualism in a large social media data

set from all of Europe has to our knowledge not yet been subject to research attention.

Other studies have used surveys to investigate online exposure to and use of languages, their relative status in various media or communicative contexts, and attitudes towards them in Europe (e.g. the *Eurobarometer* surveys conducted by the European Commission or Leppänen et al. 2011 for Finland). Increasing knowledge of English has cemented the language's "hypercentral" position within the language ecology of Europe (Swaan 2001; Soler-Carbonell 2016), and there may be evidence that English has now displaced some local languages in certain functional domains in some European societies (Görlach 2002, p. 16; for a discussion see the contributions in Linn 2016).

Few large-scale studies of aggregate online language use in Europe, however, have been based on documented usage, and empirical research into aggregate use on Twitter has typically offered only an overview of language frequencies. Additionally, while language-use profiles at country level for Twitter data exist (e.g., Mocanu et al. 2013; Leetaru et al. 2013; Magdy et al. 2014; Graham, Hale, and Gaffney 2014), relatively few studies focus specifically on bi- or multilingualism.

3. Methods

Corpus-based and NLP methods were employed in the study. They comprised the collection of data online, filtering of data, quantification of multilingualism, and the construction and visualization of language networks.

3.1. Data Collection

Over 140 million tweets with `place` attributes from European countries or territories were collected from the Twitter Streaming API from November 2016 until June 2017 using the *Tweepy* library in Python (Roesslein 2015). From this "seed" dataset of tweets by 2.9 million users, the tweets of those with at least 20 tweets and at least 50% of tweets from a single country (654,676 users) were retained for analysis.[1] In total, the data used for analysis comprised over 69.8 million tweets from 55 European countries or territories.

3.2. Data Filtering and Language Detection

Not all tweet user messages are composed by humans: A substantial proportion of tweets is generated automatically by apps or bots that interact with the Twitter API (Haustein et al. 2016). Because many apps post content that is not user-composed but rather consists of automatically-generated text, filtering tweets by the `source` value can reduce the amount of noise in the data set. A manual analysis of a selection of tweets showed that widely-used Twitter apps such as "Twitter for iPhone" or the Twitter Web Client (i.e. www.twitter.com) were less likely to broadcast automatically-generated text than were some infrequently-used apps. For this reason, the data was filtered to retain only those tweets broadcast by the following apps: Twitter Web Client, Twitter for iOS, Twitter for iPhone, Twitter for Android, Twitter for Windows Phone, Twitter for Instagram, Tweetbot for iOS, and Tweetbot for iPhone. Tweets with these sources collectively comprised over 87% of all those by European users.

A consideration of bi- and multilingualism on the Twitter platform critically depends on accurate characterization of the language of individual tweets, but automatic language detection of tweets can pose difficulties. Character sequences present in URL addresses, usernames, hashtags, emojis, and non-standard orthography can create problems for automatic language detection algorithms, as they rarely correspond to items in the lexicons of natural languages. Even after removing such sequences, very short texts are not handled well by language detection algorithms (Figure 1). To increase detected language accuracy, the data was therefore filtered to include tweets that exhibited three-way agreement between the native Twitter language detection algorithm and the algorithms `langid` (Lui and Baldwin 2014) and `compact language detector 2` (Sites 2014) after removal of URLs, usernames, hashtags, and emojis. For some less-widely-used languages not identified by all three algorithms, such as Faroese, Nynorsk, Albanian, or Somali (among others), two-way language identification or identification by a single algorithm with a high probabilistic accuracy value was used to assign languages to tweets.[2]

	text	langlangid
12260	normalee, ili dzeperice :)	it
12272	luicrede divincere mann saràcosì! referendum trascu...	it
12397	gol dla 2:1!	it
12459	koje crno malo :))	it
12736	no hej	it
154	ssshes so besyitful	de
251	popieram:-)	de
344	gesehen? fantastische serie.	de
518	i am bored	de
773	a legend	de

Figure 1: Language misidentification on short texts by `langid`

3.3. Quantification of Bilingualism Strength

A user in the dataset was determined to be bilingual for languages i, j if he or she had authored at least 10% of the total number of tweets in each of the two languages. The connection strength between languages i, j was quantified on the basis of all users with the phi coefficient, calculated from a contingency table (Table 1 and Equation 1).

[1] For this data, correlation between the center of the `place` bounding box and the precise GPS coordinates from the `coordinates` object, if both were present, was found to be quite high ($= 0.992$). For this reason, the `place` field was considered an accurate indication of true user location when posting a tweet.

[2] The presence of unique vocabulary markers (Ljubešić, Fišer, and Erjavec 2014) can be used to collect tweets in less-used languages, but the method is not applicable to the detection of already-collected tweets.

Table 1: Contingency Table for Number of Bilinguals

	$language_i$	$\sim language_i$	
$language_j$	O_{11}	O_{12}	$= R_1$
$\sim language_j$	O_{21}	O_{22}	$= R_2$
	$= C_1$	$= C_2$	$= N$

$$\phi_{ij} = \frac{(O_{11}O_{22} - O_{12}O_{21})}{\sqrt{R_1 R_2 C_1 C_2}} \qquad (1)$$

ϕ is equivalent to Pearson's product-moment correlation coefficient for occurrence frequence of two binary variables, and ranges in value from -1 to 1. Positive values indicate the language pairs are more strongly connected than would be expected based on the prevalence of the languages in the multilingual dataset.

A t-statistic was calculated to test the significance of the correlation between languages according to the formula

$$t_{ij} = \frac{\phi_{ij}\sqrt{D-2}}{\sqrt{1-\phi_{ij}^2}} \qquad (2)$$

where $D = max(R_1, C_1)$.

A multilingualism network for Europe was created in which node size corresponds to the number of multilingual users for a language and edge width corresponds to the strength of the connection (number of bilinguals) for a language pair. The network for those nodes and edges with at least 10 bilinguals and t-test p-values < 0.05 was retained (Figure 2). Network relationships were visualized using the R packages *igraph* and *visNetwork* (Csardi and Nepusz 2006; Thieurmel 2016). Additionally, nodes and vertices were annotated with information about total number of users, total number of bilinguals, and average message length.

4. Results

In terms of overall language representation in the European Twitter bi- and multilingualism network, English is the most prevalent language, with approximately 30% of all tweets in English. For Europe as a whole, a network of 42 languages and 68 edges describes the statistically significant bilingual links (Figure 2). English clearly plays the most important role: it is connected to almost all of the languages in the network. Other languages with large numbers of users, such as French, Spanish, Turkish, and Russian, have multiple connctions to other languages.

It should be remarked that the bi- and multilingual network only accounts for productive language use (i.e. authorship of discourse in a particular language), not passive understanding of languages. In an additional step, follower and friend statistics will be used to estimate these values as well.

Multilingual networks were also created for individual European countries. In them, English serves as as a bridge between linguistic communities, but the principal national

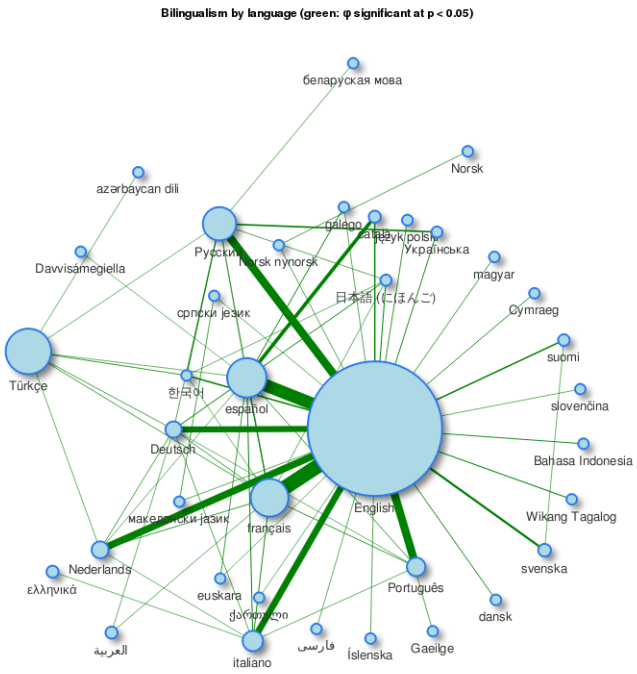

Figure 2: Bilingualism Network

language(s) figure prominently. Nonetheless, high levels of bilingualism with English, when considered in light of previously reported demographic statistics (European Commission 2012), may indicate a shift towards English and away from traditional languages in Europe, particularly for some minority languages with official status.

References

Bamann, David, Jacob Eisenstein, and Tyler Schnoebelen (2014). "Gender Identity and Lexical Variation in Social Media". In: *Journal of Sociolinguistics* 18.2, pp. 135–160.

Coats, Steven (2016a). "Grammatical feature frequencies of English on Twitter in Finland". In: *English in Computer-mediated Communication: Variation, Representation, and Change*. Ed. by Lauren Squires. Berlin: De Gruyter, pp. 179–210.

– (2016b). "Grammatical frequencies and gender in Nordic Twitter Englishes". In: *Proceedings of the 4th conference on CMC and social media corpora for the humanities*. Ljubljana, pp. 12–16. URL: http://nl.ijs.si/janes/wp-content/uploads/2016/09/CMC-conference-proceedings-2016.pdf.

Csardi, Gabor and Tamas Nepusz (2006). "The igraph software package for complex network research". In: *InterJournal* Complex Systems, p. 1695. URL: http://igraph.org.

Eisenstein, Jacob et al. (2014). "Diffusion of Lexical Change in Social Media". In: *PLoS ONE* 9.1.

Eleta, Irene and Jennifer Golbeck (2014). "Multilingual use of Twitter: Social networks at the language frontier". In: *Computers in Human Behavior* 41, pp. 424–432.

European Commission (2006). “Europeans and their languages: Special Eurobarometer 243”. In: URL: http://ec.europa.eu/public_opinion/archives/ebs/ebs_243_en.pdf.
– (2011). “User language preference online: Flash Eurobarometer 313”. In: URL: http://ec.europa.eu/public_opinion/flash/fl_313_en.pdf.
– (2012). “Europeans and their languages: Special Eurobarometer 386”. In: URL: http://ec.europa.eu/public_opinion/archives/ebs/ebs_386_sum_en.pdf.
Graham, Mark, Scott A. Hale, and Devin Gaffney (2014). “Where in the World Are You? Geolocation and Language Identification in Twitter”. In: *The Professional Geographer* 66.4, pp. 568–578. URL: http://dx.doi.org/10.1080/00330124.2014.907699.
Görlach, Manfred (2002). *Still More Englishes*. Amsterdam: John Benjamins.
Grosjean, François (2008). “Studying bilinguals: Methodological and conceptual issues”. In: *Handbook of bilingualism*. Ed. by Tej K. Bhatia and William C. Ritchie. Malden, MA: Wiley-Blackwell, pp. 32–63.
Hale, Scott (2014). “Global connectivity and multilinguals in the Twitter network”. In: *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems*. ACM. New York, pp. 833–842.
Haugen, Einar (1972). “The ecology of language”. In: *The ecology of language*. Ed. by Einar Haugen and Anwar Dil. Palo Alto: Stanford University Press, pp. 325–339.
Haustein, Stefanie et al. (2016). “Tweets as impact indicators: Examining the implications of automated ”bot” accounts on Twitter”. In: *Journal of the Association for Information Science and Technology* 67.1, pp. 232–238. ISSN: 2330-1643. DOI: 10.1002/asi.23456. URL: http://dx.doi.org/10.1002/asi.23456.
Jørgensen, Anna Katrine, Dirk Hovy, and Anders Søgaard (2015). “Challenges of studying and processing dialects in social media”. In: *Proceedings of the ACL 2015 Workshop on Noisy User-generated Text*. Association for Computational Linguistics. Stroudsburg, PA, pp. 9–18. URL: http://aclweb.org/anthology/W15-4302.
Kim, Suin et al. (2014). “Sociolinguistic Analysis of Twitter in Multilingual Societies”. In: *Proceedings of the 25th ACM Conference on Hypertext and Social Media*. HT ’14. Santiago, Chile: ACM, pp. 243–248. URL: http://doi.acm.org/10.1145/2631775.2631824.
Leetaru, Kalev H. et al. (2013). “Mapping the global Twitter heartbeat: The geography of Twitter”. In: *First Monday* 18.5/6.
Leppänen, Sirpa et al. (2011). *National Survey on the English Language in Finland: Uses, meanings and attitudes (= Studies in Variation, Contacts and Change in English, Volume 5)*. Helsinki: Varieng.
Linn, Andrew, ed. (2016). *Investigating English in Europe: Contexts and agendas*. Berlin and Boston: De Gruyter Mouton.
Ljubešić, Nikola, Darja Fišer, and Tomaž Erjavec (2014). “TweetCaT: a tool for building Twitter corpora of smaller languages”. In: *LREC*.
Lui, Marco and Timothy Baldwin (2012). “Langid.py: An off-the-shelf language identification tool”. In: *50th Proceedings of the Association for Computational Linguistics*. Association for Computational Linguistics. Stroudsburg, PA, pp. 25–30.
– (2014). “Accurate language identification of Twitter messages”. In: *Proceedings of the 5th Workshop on Language Analysis for Social Media (LASM) EACL 2014*. Association for Computational Linguistics. Stroudsburg, PA, pp. 17–25.
Magdy, Amr et al. (2014). “Exploiting geo-tagged tweets to understand localized language diversity”. In: *Proceedings of Workshop on Managing and Mining Enriched Geo-Spatial Data*. GeoRich’14. Snowbird, UT, USA: ACM, 2:1–6. URL: http://doi.acm.org/10.1145/2619112.2619114.
Mocanu, Delia et al. (2013). “The Twitter of babel: Mapping world languages through microblogging platforms”. In: *PLoS ONE* 8.4.
Roesslein, Joshua (2015). *Tweepy*. Python programming language module. URL: https://github.com/tweepy/tweepy.
Ronen, Shahar et al. (2014). “Links that speak: The global language network and its association with global fame”. In: *PNAS* 111.52, E5616–E5622.
Sites, Dick (2014). *Compact language detector 2*. R package version 1.0.2. URL: https://github.com/CLD2Owners/cld2.
Soler-Carbonell, Josep (2016). “English in the language ecology of Europe”. In: *Investigating English in Europe: Contexts and agendas*. Ed. by Andrew Linn. Berlin and Boston: De Gruyter Mouton, pp. 53–58.
Squires, Lauren (2015). “Twitter: Design, discourse, and implications of public text”. In: *The Routledge Handbook of Language and Digital Communication*. Ed. by Alexandra Georgakopoulou and Tereza Spilioti. London and New York: Routledge, pp. 239–256.
Swaan, Abram De (2001). *Words of the world: The global language system*. Cambridge: Polity.
Thieurmel, Benoit (2016). *visNetwork: Network Visualization using ’vis.js’ Library*. R package version 1.0.2. URL: https://CRAN.R-project.org/package=visNetwork.
Wikström, Peter (2014). “#srynotfunny: Communicative functions of hashtags on Twitter”. In: *SKY Journal of Linguistics* 27, pp. 127–152.