Inducing a lexicon of sociolinguistic variables from code-mixed text

Philippa Shoemark[*]
p.j.shoemark@ed.ac.uk

James Kirby[†]
j.kirby@ed.ac.uk

Sharon Goldwater[*]
sgwater@inf.ed.ac.uk

[*]School of Informatics
University of Edinburgh

[†]Dept. of Linguistics and English Language
University of Edinburgh

Abstract

Sociolinguistics is often concerned with how variants of a linguistic item (e.g., *nothing* vs. *nothin'*) are used by different groups or in different situations. We introduce the task of inducing lexical variables from code-mixed text: that is, identifying equivalence pairs such as (*football*, *fitba*) along with their linguistic code (*football*→British, *fitba*→Scottish). We adapt a framework for identifying gender-biased word pairs to this new task, and present results on three different pairs of English dialects, using tweets as the code-mixed text. Our system achieves precision of over 70% for two of these three datasets, and produces useful results even without extensive parameter tuning. Our success in adapting this framework from gender to language variety suggests that it could be used to discover other types of analogous pairs as well.

1 Introduction

Large social media corpora are increasingly used to study variation in informal written language (Schnoebelen, 2012; Bamman et al., 2014; Nguyen et al., 2015; Huang et al., 2016). An outstanding methodological challenge in this area is the bottom-up discovery of sociolinguistic *variables*: linguistic items with identifiable variants that are correlated with social or contextual traits such as class, register, or dialect. For example, the choice of the term *rabbit* versus *bunny* might correlate with audience or style, while *fitba* is a characteristically Scottish variant of the more general British *football*.

To date, most large-scale social media studies have studied the usage of individual variant forms (Eisenstein, 2015; Pavalanathan and Eisenstein, 2015). Studying how a variable *alternates* between its variants controls better for 'Topic Bias' (Jørgensen et al., 2015), but identifying the relevant variables/variants may not be straightforward. For example, Shoemark et al. (2017b) used a data-driven method to identify distinctively Scottish terms, and then manually paired them with Standard English equivalents, a labour intensive process that requires good familiarity with both language varieties. Our aim is to facilitate the process of curating sociolinguistic variables by providing researchers with a ranked list of candidate variant pairs, which they only have to accept or reject.

This task, which we term *lexical variable induction*, can be viewed as a type of bilingual lexicon induction (Haghighi et al., 2008; Zhang et al., 2017). However, while most work in that area assumes that monolingual corpora are available and labeled according to which language they belong to, in our setting there is a single corpus containing code-mixed text, and we must identify both translation equivalents (*football, fitba*) as well as linguistic code (*football*→British, *fitba*→Scottish). To illustrate, here are some excerpts of tweets from the Scottish dataset analysed by Shoemark et al., with Standard English glosses in italics:[1]

1. need to come hame fae the football
   need to come home from the football
2. miss the fitba
   miss the football
3. awwww man a wanty go tae the fitbaw
   awwww man I want to go to the football

The lexical variable induction task is challenging: we cannot simply classify documents containing *fitba* as Scottish, since the *football* variant may also occur in otherwise distinctively Scottish texts, as in (1). Moreover, if we start by knowing only a few variables, we would like a way to learn what other likely variables might be. Had we not known

[1]Note that it is hard to definitively say whether tweets such as these are mixing English and Scots codes, or whether they are composed entirely in a single Scottish code, which happens to share a lot of vocabulary with Standard English.

Proceedings of the 2018 EMNLP Workshop W-NUT: The 4th Workshop on Noisy User-generated Text, pages 1–6
Brussels, Belgium, Nov 1, 2018.

the (*football, fitba*) variable, we might not detect that (2) was distinctively Scottish. Our proposed system can make identifying variants quicker and also suggest variant pairs a researcher might not have otherwise considered, such as (*football, fitbaw*) which could be learned from tweets like (3).

Our task can also be viewed as the converse of the one addressed by Donoso and Sanchez (2017), who proposed a method to identify geographical regions associated with different linguistic codes, using pre-defined lexical variables. Also complementary is the work of Kulkarni et al. (2016), who identified words which have the same form but different semantics across different linguistic codes; here, we seek to identify words which have the same semantics but different forms.

We frame our task as a ranking problem, aiming to generate a list where the best-ranked pairs consist of words that belong to different linguistic codes, but are otherwise semantically and syntactically equivalent. Our approach is inspired by the work of Schmidt (2015) and Bolukbasi et al. (2016), who sought to identify pairs of words that exhibit gender bias in their distributional statistics, but are otherwise semantically equivalent. Their methods differ in the details but use a similar framework: they start with one or more seed pairs such as {(*he, she*), (*man, woman*)} and use these to extract a 'gender' component of the embedding space, which is then used to find and rank additional pairs.

Here, we replace the gendered seed pairs with pairs of sociolinguistic variants corresponding to the same variable, such as {(*from, fae*), (*football, fitba*)}. In experiments on three different datasets of mixed English dialects, we demonstrate useful results over a range of hyperparameter settings, with precision@100 of over 70% in some cases using as few as five seed pairs. These results indicate that the embedding space contains structured information not only about gendered usage, but also about other social aspects of language, and that this information can potentially be used as part of a sociolinguistic researcher's toolbox.

2 Methods

Our method consists of the following steps.[2]

Train word embeddings We used the Skip-gram algorithm with negative sampling (Mikolov et al., 2013) on a large corpus of code-mixed text to obtain a unit-length embedding $\boldsymbol{w}$ for each word in the input vocabulary V.[3]

Extract 'linguistic code' component Using seed pairs $S = \{(\boldsymbol{x_i}, \boldsymbol{y_i}), i = 1 \ldots n\}$, we compute a vector $\boldsymbol{c}$ representing the component of the embedding space that aligns with the linguistic code dimension. Both Schmidt and Bolukbasi et al. were able to identify gender-biased word pairs using only a single seed pair, defining the 'gender' component as $\boldsymbol{c} = \boldsymbol{w_{she}} - \boldsymbol{w_{he}}$. However, there is no clear prototypical pair for dialect relationships, so we average our pairs, defining $\boldsymbol{c} = \frac{1}{n}\sum_i \boldsymbol{x_i} - \frac{1}{n}\sum_i \boldsymbol{y_i}$.[4] We experiment with the number of required seed pairs in §5.

Threshold candidate pairs From the set of all word pairs in $V \times V$, we generate a set of candidate output pairs. We follow Bolukbasi et al. (2016) and consider only pairs whose embeddings meet a minimum cosine similarity threshold δ. We set δ automatically using our seed pairs: for each seed pair $(\boldsymbol{x_i}, \boldsymbol{y_i})$ we compute $\cos(\boldsymbol{x_i}, \boldsymbol{y_i})$ and set δ equal to the lower quartile of the resulting set of cosine similarities.

Rank candidate pairs Next we use $\boldsymbol{c}$ to rank the remaining candidate pairs such that the top-ranked pairs are the most indicative of distinct linguistic codes, but are otherwise semantically equivalent. We follow Bolukbasi et al. (2016),[5] setting $score(\boldsymbol{w_i}, \boldsymbol{w_j}) = \cos(\boldsymbol{c}, \boldsymbol{w_i} - \boldsymbol{w_j})$.

Filter top-ranked pairs High dimensional embedding spaces often contain 'hub' vectors, which are the nearest neighbours of a disproportionate number of other vectors (Radovanović et al., 2010). In preliminary experiments we found that many of our top-ranked candidate pairs included such 'hubs', whose high cosine similarity with the word vectors they were paired with did not reflect genuine semantic similarity. We therefore discard all pairs containing words that appear in more than m of the top-n ranked pairs.[6]

[2] Code is available at `github.com/pjshoemark/lexvarinduction`.

[3] In preliminary experiments we also tried CBOW and FastText, but obtained better output with Skip-gram.

[4] Bolukbasi et al. (2016) introduced another method to combine multiple seed pairs, using Principal Component Analysis. We compared it and a variant to our very simple difference of means method, and found little difference in their efficacy. Details can be found in the Supplement. All results reported in the main paper use the method defined above.

[5] See Supplement for comparison with an alternative scoring method devised by Schmidt (2015).

[6] The choice of $m \in \{5, 10, 20\}$ and $n \in \{5k, 10k, 20k\}$ made little difference, although we did choose the best pa-

3 Datasets

We test our methods on three pairs of language varieties: British English vs Scots/Scottish English; British English vs General American English; and General American English vs African American Vernacular English (AAVE). Within each data set, individual tweets may contain words from one or both codes of interest, and the *only* words with a known linguistic code (or which are known to have a corresponding word in the other code) are those in the seed pairs.

BrEng/Scottish For our first test case, we combined the two datasets collected by Shoemark et al. (2017a), consisting of complete tweet histories from Aug-Oct 2014 by users who had posted at least one tweet in the preceding year geotagged to a location in Scotland, or that contained a hashtag relating to the 2014 Scottish Independence referendum. The corpus contains 9.4M tweets.

For seeds, we used the 64 pairs curated by Shoemark et al. (2017b). Half are discourse markers or open-class words (*dogs, dugs*), (*gives, gees*) and half are closed-class words (*have, hae*), (*one, yin*). The full list is included in the Supplement.

BrEng/GenAm For our next test case we recreated the entire process of collecting data and seed variables from scratch. We extracted 8.3M tweets geotagged to locations in the USA from a three-year archive of the public 1% sample of Twitter (1 Jul 2013–30 Jun 2016). All tweets were classified as English by langid.py (Lui and Baldwin, 2012), none are retweets, none contain URLs or embedded media, and none are by users with more than 1000 friends or followers. We combined this data with a similarly constructed corpus of 1.7M tweets geotagged to the UK and posted between 1 Sep 2013 and 30 Sep 2014.

To create the seed pairs, we followed Shoemark et al. (2017b) and used the Sparse Additive Generative Model of Text (SAGE) (Eisenstein et al., 2011) to identify the terms that were most distinctive to UK or US tweets. However, most of these terms turned out to represent specific dialects *within* each country, rather than the standard BrEng or GenAm dialects (we discuss this issue further below). We therefore manually searched through the UK terms to identify those that are standard BrEng and differ from GenAm by spelling only, and paired each one with its GenAm spelling variant, e.g. (*color, colour*), (*apologize, apologise*), (*pajamas, pyjamas*). This process involved looking through thousands of words to identify only 27 pairs (listed in the Supplement), which is a strong motivator for our proposed method to more efficiently increase the number of pairs.

GenAm/AAVE While creating the previous dataset, we noticed that many of the terms identified by SAGE as distinctively American were actually from AAVE. To create our GenAm/AAVE seed pairs, we manually cross-referenced the most distinctively 'American' terms with the AAVE phonological processes described by Rickford (1999). We then selected terms that reflected these processes, paired with their GenAm equivalents, e.g. (*about, bou*), (*brother, brudda*). The full list of 19 open-class and 20 closed-class seed pairs is included in the Supplement.

4 Evaluation Procedure

We evaluate our systems using Precision@K, the percentage of the top K ranked word pairs judged to be valid sociolinguistic variables. We discard any seed pairs from the output before computing precision. Since we have no gold standard translation dictionaries for our domains of interest, each of the top-K pairs was manually classified as either valid or invalid by the first author.

For a pair to be judged as valid, (a) each member must be strongly associated with one or the other language variety, (b) they must be referentially, functionally, and syntactically equivalent, and (c) the two words must be ordered correctly according to their language varieties, e.g. if the seeds were (BrEng, GenAm) pairs, then the BrEng words should also come first in the top-K output pairs.

Evaluation judgements were based on the author's knowledge of the language varieties in question; for unfamiliar terms, tweets containing the terms were sampled and manually inspected, and cross-referenced with urbandictionary.com and/or existing sociolinguistic literature.

Our strict criteria exclude pairs like *(dogs, dug)* which differ in their inflection, or *(quid, dollar)* whose referents are distinct but are equivalent across cultures. In many cases it was difficult to judge whether or not a pair should be accepted, such as when not all senses of the words were interchangable, e.g. BrEng/GenAm *(folk, folks)*

rameters for each language pair: $m = 20$, $n = 20k$ for BrEng/Scottish; $m = 5$, $n = 5k$ for GenAm/AAVE; and $m = 10$, $n = 5k$ for BrEng/GenAm.

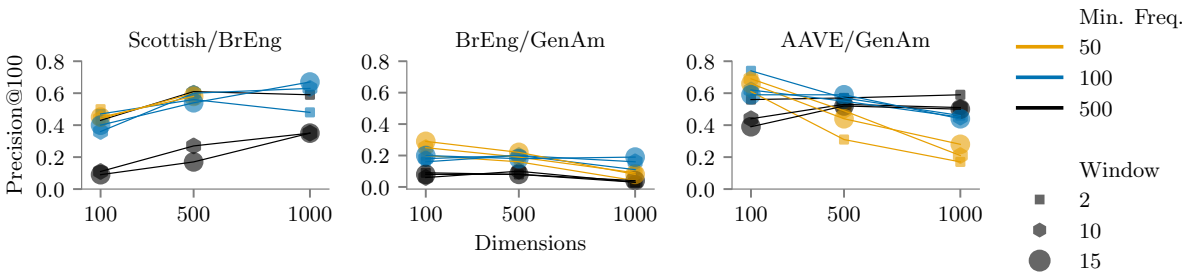


Figure 1: Precision@100 for various Skip-gram hyperparameter settings.

works for the 'people' sense of *folk*, but not the adjectival sense: *(folk music, *folks music)*. The BrEng/GenAm dataset also yielded many pairs of words that exhibit different frequencies of usage in the two countries, but where both words are part of both dialects, such as *(massive, huge)*, *(vile, disgusting)*, and *(horrendous, awful)*. We generally marked these as incorrect, although the line between these pairs and clear-cut lexical alternations is fuzzy. For some applications, it may be desirable to retrieve pairs like these, in which case the precision scores we report here are very conservative.

5 Results and Discussion

We started by exploring how the output precision is affected by the hyperparameters of the word embedding model: the number of embedding dimensions, size of the context window, and minimum frequency below which words are discarded. Results (Figure 1) show that the context window size does not make much difference and that the best scores for each language use a minimum frequency threshold of 50-100. The main variability seems to be in the optimal number of dimensions, which is much higher for the BrEng/Scottish data set than for GenAm/AAVE. Although the precision varies considerably, it is over 40% for most settings, which means a researcher would need to manually check only a bit over twice as many pairs as needed for a study, rather than sifting through a much larger list of individual words and trying to come up with the correct pairs by hand. Results for BrEng/GenAm are worse than for the other two datasets, for reasons which become clear when we look at the output.

Table 1 shows the top 10 generated pairs for each pair of language varieties, using the best hyperparameters for each of the datasets. The top 100 are given in the Supplement. According to our strict evaluation criteria, many of the output pairs for the BrEng/GenAm dataset were scored as incorrect. However, most of them are actually sensible, and examples of the kinds of grey areas and cultural analogies (e.g., (*amsterdam, vegas*), (*bbc, cnn*)) that we discussed in §4. These types of pairs likely predominate because BrEng and GenAm are both standardized dialects with very little difference at the lexical level, so cultural analogies and frequency effects are the most salient differences.

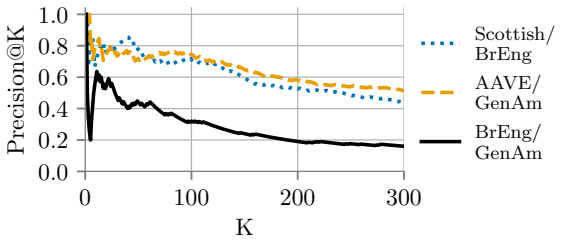


Figure 2: Precision@K from K=1 to 300 for each language variety pair.

BrEng / Scottish	BrEng / GenAm	GenAm / AAVE
now / noo	mums / moms	the / tha
what / whit	*dunno / idk*	with / wit
wasnt / wis	*yeh / yea*	getting / gettin
cant / canny	*shouting / yelling*	just / jus
would / wid	*quid / dollars*	*and / nd*
doesnt / disny	learnt / learned	making / makin
cant / cannae	favour / favor	*when / wen*
going / gonny	sofa / couch	looking / lookin
want / wanty	advert / commercial	something / somethin
anyone / embdy	adverts / commercials	going / goin

Table 1: Top 10 ranked variables for each language pair (invalid variables in italics).

To show how many pairs can be identified effectively, Figure 2 plots Precision@K as a function of K $\in \{1..300\}$. For BrEng/Scottish and GenAm/AAVE, more than 70% of the top-100 ranked word pairs are valid. Precision drops off fairly slowly, and is still at roughly 50% for these two datasets even when returning 300 pairs.

To assess the contribution of the 'linguistic code' component, we compared the performance of our system with a naïve baseline which does not use the 'linguistic code' vector $\boldsymbol{c}$ at all. Since translation equivalents such as *fitba* and *football* are likely

	Baseline	Our Method
BrEng / Scottish	0.00	0.71
BrEng / GenAm	0.04	0.32
GenAm / AAVE	0.08	0.74

Table 2: Precision@100 for our method and the baseline for each language pair.

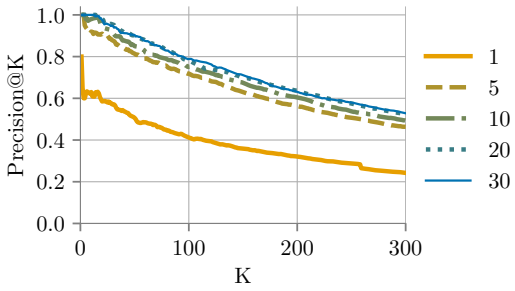
to be very close to one another in the embedding space, it is worth checking whether they can be identified on that basis alone. The baseline ranks all unordered pairs of words in the vocabulary just by their cosine similarity, $cos(\boldsymbol{w_i}, \boldsymbol{w_j})$. Since this baseline gives us no indication of which word belongs to which language variety, we evaluated it only on its ability to correctly identify translation equivalents (i.e. using criteria (a) and (b), see §4), and gave it a free pass on assigning the variants to the correct linguistic codes (criterion (c)). Results are in Table 2. Despite its more lenient evaluation criteria, the baseline performs very poorly. Perhaps if we were starting with a pre-defined set of words from one language variety which were known to have equivalents in the other, then simply returning their nearest neighbours might be sufficient. However, in this more difficult setting where we lack prior knowledge about which words belong to our codes of interest, an additional signal clearly is needed.

Finally, we examined how performance depends on the particular seed pairs we used and the number of seed pairs. Using the BrEng/Scottish and GenAm/AAVE datasets, we subsampled between 1 and 30 seed pairs from our original sets. Over 10 random samples of each size, we found similar average performance using just 5 seed pairs as when using the full original sets (see Figure 3). Performance increased slightly when using only open-class seed pairs: P@100 rose to 0.77 for Scottish/BrEng and 0.75 for GenAm/AAVE (cf. 0.71 and 0.74 using all the original seed pairs). These results indicate our method is robust to the number and quality of seed pairs.

6 Conclusion

Overall, our results demonstrate that sociolinguistic information is systematically encoded in the word embedding space of code-mixed text, and that this implicit structure can be exploited to identify sociolinguistic variables along with their linguistic code. Starting from just a few seed variables, a simple heuristic method is sufficient to identify a large number of additional candidate pairs with precision of 70% or more. Results are somewhat sensitive to different hyperparameter settings but even non-optimal settings produce results that are likely to save time for sociolinguistic researchers. Although we have so far tested our system only on varieties of English[7], we expect it to perform well with other pairs of language varieties which have a lot of vocabulary overlap or are frequently code-mixed *within* sentences or short documents, including code-mixed languages as well as dialects. This framework may also be useful for identifying variation across genres or registers.

(a) Scottish/BrEng



(b) AAVE/GenAm

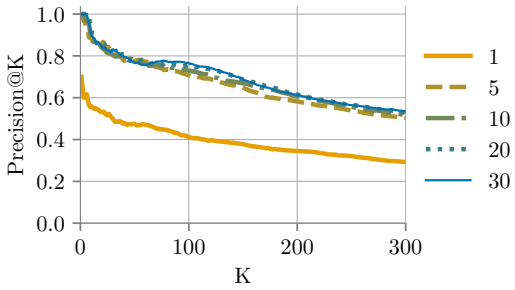


Figure 3: Mean Precision@K curves for different sized samples from the original seed pair lists. Each curve is averaged across 10 random samples of n seed pairs, for $n \in \{1, 5, 10, 20, 30\}$.

Acknowledgments

This work was supported in part by the EPSRC Centre for Doctoral Training in Data Science, funded by the UK Engineering and Physical Sciences Research Council (grant EP/L016427/1) and the University of Edinburgh.

[7]The Scots language, while not a variety of Modern English, developed from a dialect of Old English and in practise is often inextricably mixed with Scottish English.

References

David Bamman, Jacob Eisenstein, and Tyler Schnoebelen. 2014. Gender identity and lexical variation in social media. *Journal of Sociolinguistics*, 18(2):135–160.

Tolga Bolukbasi, Kai-Wei Chang, James Zou, Venkatesh Saligrama, and Adam Kalai. 2016. Man is to computer programmer as woman is to homemaker? Debiasing word embeddings. In *Proceedings of the 30th International Conference on Neural Information Processing Systems*, NIPS'16, pages 4356–4364. Curran Associates Inc.

Gonzalo Donoso and David Sanchez. 2017. Dialectometric analysis of language variation in Twitter. In *Proceedings of the Fourth Workshop on NLP for Similar Languages, Varieties and Dialects*, VarDial'17, pages 16–25. Association for Computational Linguistics.

Jacob Eisenstein. 2015. Systematic patterning in phonologically-motivated orthographic variation. *Journal of Sociolinguistics*, 19(2):161–188.

Jacob Eisenstein, Amr Ahmed, and Eric P. Xing. 2011. Sparse additive generative models of text. In *Proceedings of the 28th International Conference on International Conference on Machine Learning*, ICML'11, pages 1041–1048. Omnipress.

Aria Haghighi, Percy Liang, Taylor Berg-Kirkpatrick, and Dan Klein. 2008. Learning bilingual lexicons from monolingual corpora. In *Proceedings of the 46th Annual Meeting of the Association for Computational Linguistics on Human Language Technologies*, ACL'08: HLT, pages 771–779. Association for Computational Linguistics.

Yuan Huang, Diansheng Guo, Alice Kasakoff, and Jack Grieve. 2016. Understanding U.S. regional linguistic variation with Twitter data analysis. *Computers, Environment and Urban Systems*, 59:244 – 255.

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*, W-NUT'15, pages 9–18. Association for Computational Linguistics.

Vivek Kulkarni, Bryan Perozzi, and Steven Skiena. 2016. Freshman or fresher? quantifying the geographic variation of language in online social media. In *Proceedings of the Tenth International Conference on Web and Social Media*, ICWSM'16, pages 615–618. AAAI Press.

Marco Lui and Timothy Baldwin. 2012. langid.py: An off-the-shelf language identification tool. In *Proceedings of the ACL 2012 system demonstrations*, ACL'12, pages 25–30. Association for Computational Linguistics.

Tomas Mikolov, Ilya Sutskever, Kai Chen, Greg Corrado, and Jeffrey Dean. 2013. Distributed representations of words and phrases and their compositionality. In *Proceedings of the 26th International Conference on Neural Information Processing Systems - Volume 2*, NIPS'13, pages 3111–3119. Curran Associates Inc.

Dong Nguyen, Dolf Trieschnigg, and Leonie Cornips. 2015. Audience and the use of minority languages on Twitter. In *Proceedings of the Ninth International Conference on Web and Social Media*, ICWSM'15, pages 666–669. AAAI Press.

Umashanthi Pavalanathan and Jacob Eisenstein. 2015. Audience-modulated variation in online social media. *American Speech*, 90(2):187–213.

Miloš Radovanović, Alexandros Nanopoulos, and Mirjana Ivanović. 2010. Hubs in space: Popular nearest neighbors in high-dimensional data. *Journal of Machine Learning Research*, 11(Sep):2487–2531.

John R Rickford. 1999. *African American vernacular English: Features, evolution, educational implications*, chapter 1. Blackwell Malden, MA.

Ben Schmidt. 2015. Rejecting the gender binary: a vector-space operation. http://bookworm.benschmidt.org/posts/2015-10-30-rejecting-the-gender-binary.html.

Tyler Schnoebelen. 2012. Do you smile with your nose? stylistic variation in Twitter emoticons. *University of Pennsylvania Working Papers in Linguistics*, 18(2):14.

Philippa Shoemark, James Kirby, and Sharon Goldwater. 2017a. Topic and audience effects on distinctively scottish vocabulary usage in Twitter data. In *Proceedings of the Workshop on Stylistic Variation*, pages 59–68. Association for Computational Linguistics.

Philippa Shoemark, Debnil Sur, Luke Shrimpton, Iain Murray, and Sharon Goldwater. 2017b. Aye or naw, whit dae ye hink? Scottish independence and linguistic identity on social media. In *Proceedings of the 15th Conference of the European Chapter of the Association for Computational Linguistics: Volume 1, Long Papers*, EACL'17, pages 1239–1248. Association for Computational Linguistics.

Meng Zhang, Haoruo Peng, Yang Liu, Huan-Bo Luan, and Maosong Sun. 2017. Bilingual lexicon induction from non-parallel data with minimal supervision. In *Proceedings of the Thirty-First AAAI Conference on Artificial Intelligence*, AAAI'17, pages 3379–3385. AAAI Press.