#go awn: Sociophonetic variation in variant spellings on Twitter

Rachael Tatman
University of Washington
rctatman@uw.edu

While there is a long history of investigating sociophonetic variation in speech, it has been less studied in computer mediated communication contexts such as Twitter. The most obvious reason for this is that interactions in Twitter are text-based and therefore do not include acoustic information. Twitter users are, however, encoding sociophonetic information through their use of variant spellings, such as "awn" for "on". This study provides evidence that Twitter users in multiple dialect regions are using variant spellings to encode sociophonetic variation in a systematic way and that these variant spelling are sensitive to style shifting. The methodology used here may be used in future studies to determine the salience of sociophonetic variables.
Keywords: sociolinguistics; phonetics; social media; Twitter

1 Introduction

This study has two aims. The first is to see if it is possible to replicate findings based on speech data using Twitter data. Secondly, it will attempt to determine whether tweets that include sociophonetic variation are subject to style shifting. This will help to provide information on the nature of the connection between soicophonetic variation and variant spellings on Twitter.

2 Background

It may seem counter-intuitive to attempt to look at phonetic information in a predominantly written medium. Previous work, however, has suggested that the parallels between face-to-face and computer medicated communication (CMC) are robust. If this can be shown to extend to phonetic variation, then data from computer mediated communication can provide a new line of enquiry for sociophoneticians and a new way to verify findings.

Twitter data is an especially appealing area to investigate for a number of reasons. Though it has some drawbacks, the sheer volume of available data makes it an appealing area to investigate. Perhaps the most exciting quality of Twitter data is the high number of variant spellings used. While it seems clear that these variant

spellings are encoding *something*, whether or not they are actually representing phonetic variation in speech is an open question.

2.1 Computer mediated communication & variationist sociolinguistics

There have previously been relatively few variationist studies of computer mediated communication (CMC). Androutsopoulos (2006) suggests that there are two main reasons for this: demographic information on users of CMC is often unreliable or missing and there is a lack of phonetic/phonological information.

Even with these hurdles, however, a body of work on variation in computational contexts has emerged. The consensus in the literature so far strongly suggests that 1) systematic variation does exist in CMC and 2) it parallels variation found in speech. This has been shown across a number of domains, including text messaging, Internet Relay Chat (IRC), blogs, instant messaging (IM) and Twitter.

Thurlow and Brown (2003) found that non-standard spellings in text-messaging can be divided into a small number of distinct purposes including accent stylization. They argue that these nonstandard spellings are explicitly used to reflect speech and create a casual style that helps to form and maintain close social ties.

Linguistic variation has also been observed in Internet Relay Chat (IRC) contexts. Siebenhaar (2006) investigated the use of standard and dialect forms in Swiss-German chat rooms. He found that not only were Swiss dialect forms being used in the chat room contexts (somewhat surprising given their rarity in written forms elsewhere), but that their use was mediated by both the regional identity and age of speakers. More dialect forms were used in region-specific channels and by younger speakers.

Blogs are another area of CMC that show predictable variation. Using markers that have been identified in previous computational work as "male" or "female", Herring and Paolillio (2006) investigated the use of "gendered" features in weblogs or blogs. They found that, while there was variation in the use of these features, they were tied more closely to genre than gender. This shows the need for the careful application of sociolinguistic knowledge and techniques in investigating variation in CMC. Unprincipled data-mining or statistical feature extraction runs the risk of misidentifying the role of variants.

More recently, Tagliamonte and Denis (2008) found that variation in Instant Messaging (IM) not only mirrors that of the speech community but also exhibits the same ongoing linguistic changes. They note that for some features—such as the distribution of personal pronouns—instant messaging patterns much more strongly with speech than other written mediums.

Twitter data has also been the focus of sociolinguistic investigations. Danescu-Niculescu-Mizil et al. (2011) used computational modelling to argue that Twitter users show accommodation, and that the degree of accommodation is influenced by the social network of the individuals involved. Bamman et al. (2014) also found that social network affects variation in the use of features linked with gender; the more an individual's network is made up of a single gender, the more

gender markers they are likely to use. Eisenstein (2015) found that the use of g-dropping and th-stopping was higher in conversations and areas with higher African American populations, mirroring findings from speech data.

What emerges from the literature is a strong trend: computer mediated communication strongly patterns with speech with regard to variation. Further, many of the sociolinguistic processes that have been observed in speech—language change, crossing, code-switching, style-shifting—have also been observed in CMC. This suggests that CMC is a fruitful area for sociolinguistic research. Twitter data, especially, offers an exciting area for future research.

There are many benefits to using Twitter data for sociolinguistic investigations. The most obvious is that there is a huge amount of Twitter data already available. There are over 500 million Tweets sent every day (About, 2015) and many of these are available for almost-instantaneous collection. Further, for the foreseeable future all Tweets will be archived at the Library of Congress (Osterberg, 2013). This is especially satisfying for researchers who are worried about reproducible research: since Twitter data is publicly available and archived all Twitter research is inherently reproducible. As an additional benefit to sociolinguistic studies, since Twitter is mainly used for peer-to-peer communication rather than research data collection, the effect of the Observer's Paradox (Labov, 1972) is minimized. Twitter presents one of the largest, richest and most accessible sources of linguistic data extant today.

However, there are also serious drawbacks to using Twitter data for sociolinguistic investigations. The first is that demographic data (age, geographic area, social class, etc.) is rarely available for users. One possible way to ameliorate this problem is by deducing demographic information from the content of tweets (Rao et al., 2010). There is also limited control over data production, making many sociolinguistic methods unusable. And the sheer amount of available data is as much of a drawback as it is a strength. Without a principled way to sample tweets it is impossible to extract meaningful insights from them. One final drawback, however, is perhaps the most difficult to overcome for sociophonetic research: there is no acoustic data available.

2.2 Variant spellings

That is not to say, however, that there is no phonetic information available. Twitter users, much like the text messengers investigated by Thurlow and Brown (2003), use a high proportion of variant spellings.

Variant spellings are non-standard orthographic representations of words. They are occasionally referred to as “dialect orthography” (Krapp, 1919) or “dialect respellings” (Preston 1985). There has been a resurgence of interest in variant spellings in the context of CMC. In contrast with the earlier use of non-standard orthography, which was almost always used to represent the speech of others, variant spellings in CMC are used at least sometimes to represent the speakers' own speech habits. Although as Dinkin (2014) points out, this is not always the case; some very common phonetic variables, such as (ing) are not

represented to the same degree in CMC.

Though it may not perfectly mirror speech, there is certainly more variation in spelling in CMC contexts is more variable than in other writing (Sebba, 2003). Variant spellings have also been observed to be doing social work in other informal written contexts. Androutsopoulos (2000) found that variant spellings in German punk fanzines encoded many of the same differences later observed in texting by Thurlow and Brown (2003), including phonetic and regional variation.

What has not been established, however, is whether the use of multiple variant spellings pattern with phonetic variation in speech. There seem to be two possibilities. The first is that variant spellings are in fact different lexical items and do not reflect sociophonetic variation. "Go awf", for example, is a set phrase, which seems to be used primarily by younger African American women as an expression of approval and solidarity. But many of the individuals who use this form prefer the spelling "off" in other contexts. Another example would be spellings like "hawt" which, if "aw" is used to represent /ɔ/, does not reflect the spoken production of individuals without the low back merger (Labov et al., 2005).

The second possibility is that variant spellings *are* representing phonetic variation, albeit perhaps imperfectly. If this is the case, we would expect phonetic variables that pattern together in speech to pattern together in variant spellings in Twitter. We would also expect them to show style-shifting. Finally, we would also expect them to occur in very low-frequency lexical items. This last quality can be readily observed—it is unlikely, for example, that "spelunkin" in the song title "Monster Spelunkin" (Tran & Velema, 2014) represents a separate lexical item from "spelunking"—and so will not be considered here. Very low frequency but intentional variant spellings, however, may be an interesting area for future work.

3 Case studies

In order to investigate the distinction above, data was collected for two dialect areas: the American South and Scotland. The following data-gathering procedure was used for both.

First, a well-studied sociophonetic variable with a clear alternate spelling was chosen. Then, high frequency words with that variable were selected. Next, tweets containing the variant spelling forms of the high frequency words were sampled using the public Twitter API (Application Program Interface) and the Twitter R package (Gentry & Gentry, 2014). Finally, the tweets selected this way were sorted by hand to remove tweets were the target words occurred in URL's, user names, foreign words or typos. Any other variant spellings in remaining tweets were then marked by hand

3.1 Southern American and African American English

3.1.1 Methodology

The target variable in for the American South was distinction between /ɑ/ and /ɔ/, with /ɔ/ spelled "aw". The inclusion of "w" in the spelling is probably to highlight the rounding distinction. The lack of a low back merger is a sociolinguistic marker of Southern American English and African American English (Labov et al., 2005). For merged speakers, it is not possible to guess which is /ɔ/, which results in over-application of the spelling in words like "hawt" or "dawg".

The 100 most frequent English words with this distinction were selected using the Corpus of Contemporary American English (Davies, 2008) for frequency data and the CMU Pronunciation Dictionary (Weide, 1998) for pronunciation data. Future work should make use of more nuanced sources for /ɔ/ forms, as the CMU dictionary returned several words— "or", "for" and "your"—which were not good candidates for distinguishing speakers without a merger as the target vowel is prerhotic.

3.1.2 Results

However, the remaining targets items—on, all, also, want, and because—were all found spelled with "aw"—awn, awl, awlso, wawnt, becawse. In addition, of the 74 filtered tweets with one of these target variant spellings, 50% contained more than one variant spelling. And these variant spellings were also encoding sociophonetic variation, including th-stopping, g-dropping, r-lessess, cluster reduction and /ai/-monophthongization. These are summarized in Illustration 1. These features are consistent with those found in Southern English and, with the exception of /ai/-monophthongization, African American English (Labov & Boburg, 2005).

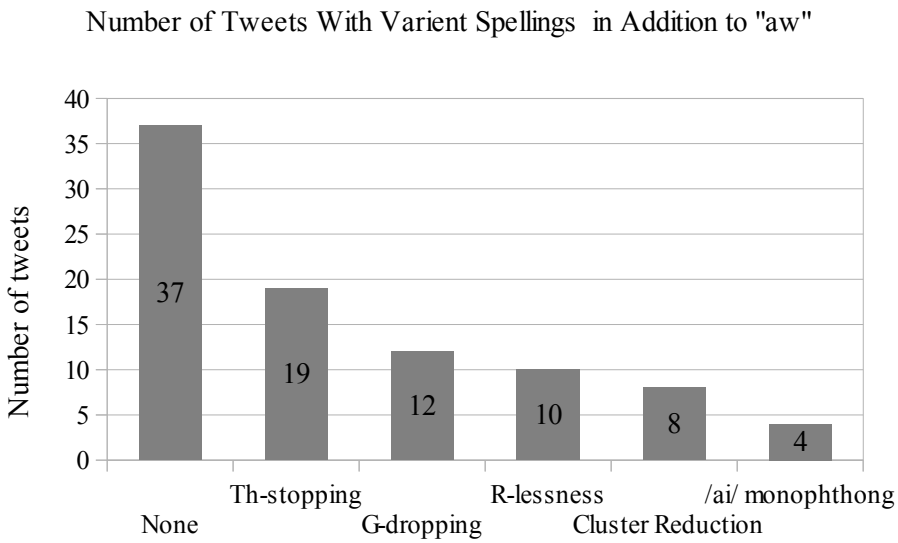

Illustration 1: Other phonetic variables encoded with variant spellings in tweets containing the "aw" variant spelling in /ɔ/ words.

An example of a tweet with a number of variables can be seen in (1). Variant spellings of interest are italicized.

(1) *Hype hayed* fo*ah d*at bec*aw*se it was 8 bucks fo*ah* 2 yea*hs* and w *d*a jets i like *readin* about *d*a prospects o*g*ay (JPG 2015)
"I paid for that because it was eight bucks for two years, and with the Jets [American football team] I like reading about the prospects, okay?"

Note that it includes r-lessness in "foah" and "yeahs", th-stopping in "dat" and "da" and g-dropping in "readin". There are also some variant spellings that are not encoding sociophonetic variation, such as "w" for "with" and lower case "I". Finally, there are some spellings where the intended interpretation is not entirely clear, such as "ogay" for "okay" or "hype hayed" for "I paid". Although the second "h" could be a representing aspiration, the first one is puzzling.

3.1.3 Discussion

It does appear that Twitter users are using multiple variant spellings together to encode phonetic variables that are consistent with those found in African American and Southern English. However, there are some issues with the data already discussed. For one, there is limited geographic data available. Only one tweet was geocoded, and though it was from Louisiana, that hardly shows that the bulk of these tweets were from the South or areas with a high proportion of African American residents. The use of variant spellings in this way could also be limited

to one dialect area. Convergent findings from another dialect area are necessary to validate these findings and methodology.

3.2 Scottish English

3.2.1 Methodology

The sampling variable for the Scottish English dataset was the [du] vowel, which is produced [de] (Stuart-Smith 2004) and commonly spelt "dae". (Including in educational materials produced by Education Scotland, the national body for education assessment in Scotland [Education Scotland 2015]). All the words containing [u] were selected from among the 50 most frequent English words (Davies 2011). Tweets containing these words— "who", "do", "you" and "to"—with their variant spellings— "whae", "dae", "yae" and "tae"—were then sampled using the same code as before. After sorting, 45 tweets remained.

3.2.2 Results

The use of variant spellings was even more prevalent in this sample. 84% percent of the tweets contained more than one variant spelling and there was an average of three variant spellings per tweet. A summary of number of variant spellings can be found in Illustration 2. In addition, they encoded sociophonetic variables associated with Scottish Standard English, including, [u] → [ʉ], [ai] → [æ], [l] vocalization, lack of [ʊ], [ɔ], and [ei] → [i] (Stuart-Smith 2004, Renni 2001). A number of these variables can be seen in example 2, with the variant spellings italicized.

(2) *dae ye* ever look back *oan* how much time *ye* wasted *oan* someone nd wonder why *nae*one punched u in the *heed* (bj 2015)
"Do you ever look back on how much time you wasted on someone and wonder why no one punched you in the head?"

Note that [ɔ] in "on" here is spelled "oa" where, presumably, the "o" also represents the presence of rounding. Other variables include [u] fronting in "ye", and [ei] surfacing as [i] in "heed".

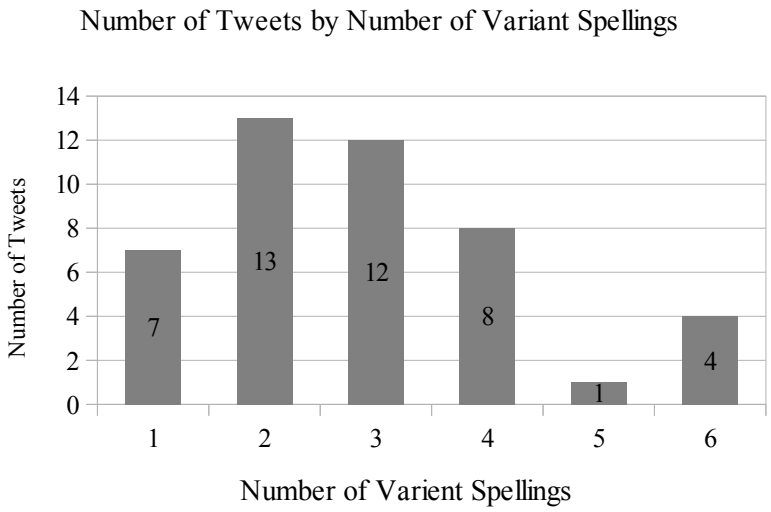

Illustration 2: Frequency of variant spelling counts in tweets. Note that while the average was three, some tweets contained as many as six.

3.2.3 Discussion

Taken together these two case studies provide evidence that speakers are using spelling variation to reflect sociophonetic variation. This suggests that variant spellings may provide a rich area of enquiry for sociophoneticians. It also shows that the methodology outlined above can be usefully applied to sample tweets that highlight sociophonetic variation—especially variation that is highly salient.

4 Style Shifting

While these case studies are suggestive that speakers are using variant spellings in the same way they do phonetic variables in speech, it is still not conclusive evidence. In order to further demonstrate this, tweets from a single user of Scottish English was sampled and examined to determine whether style-shifting was affecting the use of variant spellings.

Twitter user BradleyKirkwood (https://twitter.com/BradleyKirkwood) is a frequent user of variant spellings and also uses many markers of Scottish identity. His location is noted as "Scotland|Ayershire" his cover photo shows fans of the Rangers (a Scottish football team) cheering in the stands during a match (as of April 23, 2015).

His 100 most recent tweets were downloaded on April 23, 2015. They were then marked for variant spellings in the same way as the earlier samples. Most of the tweets in the sample, 64%, used at least one variant spelling.

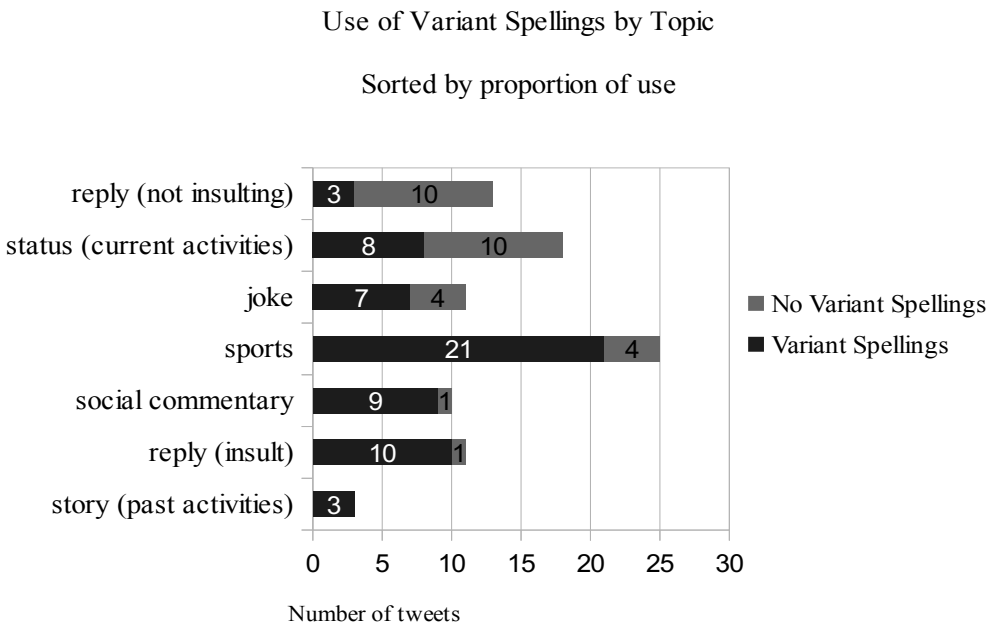

Illustration 3: Twitter user @BradleyKirkwood's rate of use of variant spellings by topic area (as determined by a content analysis)

A content analysis was performed on these tweets and they were then sorted into the topic categories shown in Illustration 3. As can be seen, the rate of use of variant spellings varies dramatically between domains. Even with the relatively small sample size, these differences were large enough to be significant, χ^2 (6, N = 91) = 25.53, p <.001. It is perhaps not surprising that tweets discussing sports were not only very common but also contained a high proportion use of variant spellings. One example can be seen in (3), below.

(3) @RangersFC @LiviFCOfficial thats what happens when *ye* play lee mcculloch, tell mccall *a* said he's *tae* write that *doon* (BradleyKirkwood 2015)
"That's what happens when you play Lee Mcclloch (Socttish football player), tell McCall (Rangers football team manager) I said he's to write that down."

BradleyKirkwood's variant spellings seemed to be connected with both regional identity as well as covert prestige and possibly affiliation with the working class. Given earlier findings (i.e. Trudgill 1972) it would seem likely that male Twitter users would use more variant spellings associated with regional variation than Twitter users of other genders.

As it stands, however, this result suggests that this Twitter user is using variant spellings to show domain-based style shifting (Fishman 1967). The fact that his use of variant spellings fluctuates also suggests that they are sociolinguistic markers or stereotypes, rather than indicators (Labov 1972). The use of variant spellings during style shifting is a rich area for future research.

5 Conclusion

This investigation provides evidence that variant spelling on Twitter are encoding phonetic variation. These variant spellings were found to encode clusters of known phonetic features associated with two separate dialect regions. In addition, at least for one Twitter user, the use of variant spellings varies by domain. The fact that the use of variant spellings is susceptible to style shifting provides convergent evidence for the claim that variant spellings pattern closely with spoken language phonetic variation.

This study has been a very cursory examination of these phenomena, however, and many questions remain unanswered: Are Twitter users accurately reflecting their own speech through the use of variant spellings? Or are they more likely to use variant spellings to show reported speech or crossing? In terms of style-shifting, are some variables are more likely to occur in some contexts? Are variant spellings as common in languages other than English? Are there similar effects found in languages without phonetic writing systems? Encoding sociophonetic information in peer-to-peer computer mediated communication is still very much a new area for exploration.

Though there is much that remains unknown, in the small samples considered here the use of variant spellings is principled. While there may certainly be fixed lexical forms of certain variants (consider “awn” and “oan” in the American and Scottish studies respectively) the fact that they pattern together and reflect spoken language variation suggest that variant spellings are more than just new lexical forms. Twitter users are using variant spellings to encode their knowledge of sociophonetic variation.

6 Acknowledgements

This material is based upon work supported by the National Science Foundation Graduate Research Fellowship under Grant No. DGE-1256082. I would also like to thank the audience at NWLC, the UW Phonetics Lab and Aaron Dinkin for their assistance. Any remaining mistakes and omissions are my own.

References

About. (2015). Retrieved May 4, 2015, from: https://about.twitter.com/company

Androutsopoulos, J. (2006). Introduction: Sociolinguistics and computer-mediated communication. *Journal of Sociolinguistics, 10*(4), 419–438.

Androutsopoulos, J. K. (2000). Non-standard spellings in media texts: The case of

German fanzines. *Journal of Sociolinguistics*, *4*(4), 514–533.
Bamman, D., Eisenstein, J. & Schnoebelen, T. (2014). Gender identity and lexical variation in social media. *Journal of Sociolinguistics*, *18*(2), 135–160.
bj [beccajamiesonx]. (2015, Apr 19). dae ye ever look back oan how much time ye wasted oan someone nd wonder why naeone punched u in the heed [Tweet]. Retrieved from: https://twitter.com/beccajamiesonx/status/589927462903095296
Bradley Kirkwood [BradleyKirkwood]. (2015, April 15). @RangersFC @LiviFCOfficial thats what happens when ye play lee mcculloch, tell mccall a said he's tae write that doon. [Tweet]. Retrieved via Twitter API.
Danescu-Niculescu-Mizil, C., Gamon, M. & Dumais, S. (2011). Mark my words!: Linguistic style accommodation in social media. In *Proceedings of the 20th international conference on World wide web* (745–754). ACM.
Davies, M. (2008). The corpus of contemporary American English: 425 million words, 1990-present.
Davies, M. (2011). *Word frequency data from the Corpus of Contemporary American English* (COCA).
Dinkin, A. J. (2014) *A phonological variable in a textual medium:(ing) in online chat,* presented at Change and Variation in Canada 8, Kingston, Ontario.
Education Scotland. (2015). Scots spelling. *Knowledge of Language*. Retrieved June 1, 2015, from: http://www.educationscotland.gov.uk/knowledgeoflanguage/scots/writinginscots/scotsspelling/index.asp
Eisenstein, J. (2015). Systematic patterning in phonologically-motivated orthographic variation. *Journal of Sociolinguistics*, *19*(2), 161–188.
Fishman, J. A. (1967). Bilingualism with and without diglossia; diglossia with and without bilingualism. *Journal of Social Issues*, *23*(2), 29–38.
Gentry, J. & Gentry, M. J. (2014). *Package 'twitteR'*.
Herring, S. C. & Paolillo, J. C. (2006). Gender and genre variation in weblogs. *Journal of Sociolinguistics*, *10*(4), 439–459.
JPG [gencoj]. (2015, Feb 06). @OrdioMongo @___OJ___ hype hayed foah dat becawse it was 8 bucks foah 2 yeahs and w da jets i like readin about da kid prospects ogay [Tweet]. Retrieved from: https://twitter.com/gencoj/status/563726906491957248
Krapp, G. P. (1919). *The pronunciation of standard English in America*. Oxford University Press, American Branch.
Labov, W. (1972). *Sociolinguistic patterns*. Philadelphia: University of Pennsylvania Press
Labov, W., Ash, S. & Boberg, C. (2005). *The atlas of North American English: Phonetics, phonology and sound change.* Walter de Gruyter.
Osterberg, G. (2013, January). Update on the Twitter archive at the Library of Congress. In *Library of Congress* (Vol. 4).
Preston, D. R. (1985). The Li'l Abner syndrome: Written representations of speech. *American speech*, 328–336.
Rao, D., Yarowsky, D., Shreevats, A. & Gupta, M. (2010, October). Classifying latent user attributes in Twitter. In *Proceedings of the 2nd international*

workshop on Search and mining user-generated contents (37–44). ACM.
Rennie, S. (2001). The electronic Scottish national dictionary (eSND): Work in Progress. *Literary and Linguistic Computing*, *16*(2), 153–160.
Schilling-Estes, N. (1998). Investigating "self-conscious" speech: The performance register in Ocracoke English. *Language in Society*, *27*(1), 53–83.
Sebba, M. (2003). Spelling rebellion. *Pragmatics & beyond new series* (151–172), John Benjamins Publishing Company.
Siebenhaar, B. (2006). Code choice and code-switching in Swiss-German internet relay chat rooms. *Journal of Sociolinguistics*, *10*(4), 481-506.
Stuart-Smith, J. (2004). Scottish English: Phonology. *A Handbook of Varieties of English*, *1*, 47–67.
Tagliamonte, S. A. & Denis, D. (2008). Linguistic ruin? LOL! Instant messaging and teen language. *American speech*, *83*(1), 3–34.
Thurlow, C. & Brown, A. (2003). Generation Txt? The sociolinguistics of young people's text-messaging. *Discourse Analysis Online*, *1*(1), 30.
Tran, A. & Velema, S. (2014) Monster spelunkin'. *Monster buddies soundtrack.*
Trudgill, P. (1972). Sex, covert prestige and linguistic change in the urban British English of Norwich. *Language in Society*, *1*(2), 179–195.
Weide, R. L. (1998). *The CMU pronouncing dictionar*y. Retrieved from: http://www.speech.cs.cmu. edu/cgibin/cmudict