Orthographic reflections of (ing): a Twitter-based corpus study

George Bailey - *University of Manchester*

tweet: @grbails ◆ email: george.bailey@postgrad.manchester.ac.uk

1. Background

Is the orthographic variation in (ing) conditioned by the same factors as the variation in speech?

Three aspects of its variation are considered:

- regional distribution: *-in* favoured in the north of the UK and in the southern US states (Labov 2001)
- grammatical conditioning: *-in* favoured in verbs, *-ing* in nouns (Houston 1985)
- effect of word token frequency: reportedly absent for the variation in speech (Abramowicz 2007)

2. Methodology

Wrote a Python script to access Twitter's streaming API and download Tweets, 140-character messages, in real-time

> So weird how like a potato can taste so different just by **cookin** it in an oven rather than **boiling** it – it's the same food, #themindboggles

2,000,000 tweets - 20,693,233 words

So_RB weird_JJ how_WRB like_IN a_DT potato_NN can_MD taste_VB so_RB different_JJ just_RB by_IN cookin_VBG it_PRP in_IN an_DT oven_NN …

52.955481 -1.152948

- Fully POS-tagged using twitie-tagger, with a reported 91% accuracy (Derczynski et al. 2013)
- Geotagged with latitude and longitude coordinates, later discretised into regions using polygonal boundary files and a point-in-polygon Python function (see **Figure 1**)

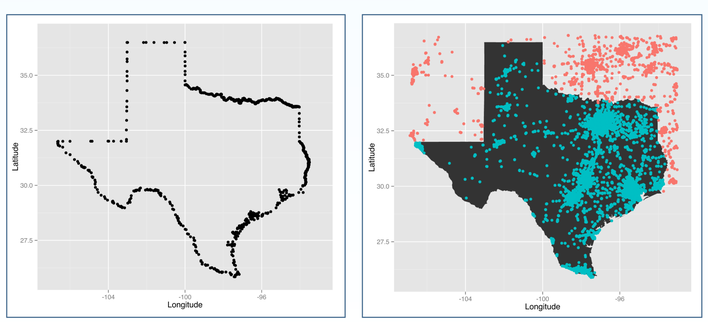

Figure 1: Polygonal state boundary of Texas (left), point-in-polygon functionality in Python (right)

3. Results and Discussion

3.1 Region

535,192 tokens of (ing), 4% of which (22,100) have been 'g-dropped'

- clearly much rarer in orthography than it is in speech

Figure 2 maps the polarity (+/-) of logistic regression log-odds by region

- positive log-odds reflect more *-in,* negative more *-ing*

Regional stratification clearly mirrors spoken (ing), as reported by Labov (2001), with more *-in* in the south of the US and the north of the UK

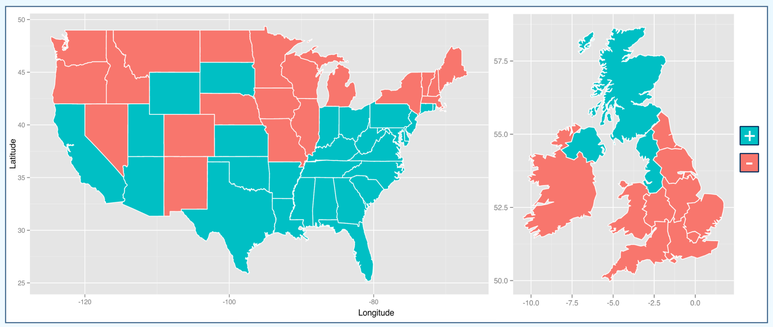

Figure 2: Regional stratification of *-in* in US states by polarity of logistic regression log-odds

To what extent does this reflect regional differences in socioeconomics?

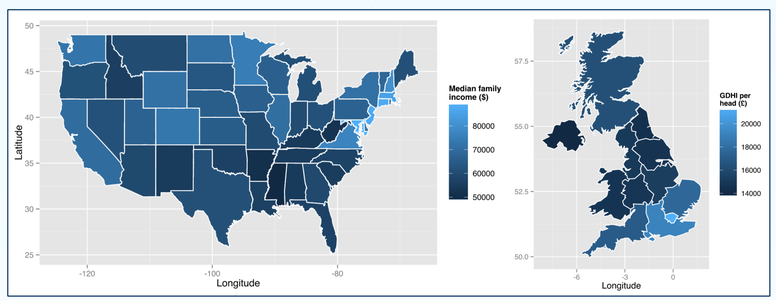

Figure 3: Measures of income by region (U.S. Census Bureau 2013; Office for National Statistics 2014)

3.2 Grammatical category

The 'nominal-verbal continuum' is absent here due to the surprisingly high rate of adjectival *-in* (see **Figure 4**)

However, this category is predominantly made up of expletives such as '**fucking**', '**fricking**', '**frigging**', '**freaking**' etc.

- the disproportionate use of *-in* likely reflects the extremely informal register of these words

Excluding '**fucking**' somewhat strengthens the evidence for grammatical conditioning

- verbs favour *-in* at a statistically-significant level (log-odds=0.41, $p < 0.01$)

The effect is also more prominent in the US, which could again reflect phonological (ing)

- reports of a grammatical category effect are inconsistent in the UK (see Bailey 2015, cf. Tagliamonte 2004)

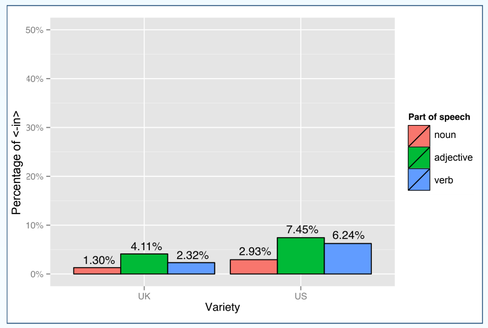

Figure 4: Rates of *-in* by grammatical category and variety

3.3 Lexical frequency

Figure 5 reveals a negative trend between word frequency and *-in* rate

- infrequent words show more *-in*
- counter to what we may expect?

This might reflect stylistic properties of this particular medium

These low frequency, high *-in* rate words tend to be 'online slang' terms like '**nuttin**' (which also shows th-stopping), '**pimpin**', '**frickin**', and '**ballin**'

These are of course rare in terms of standard, externally-sourced frequency measures (SUBTLEX), but their use signals an informal style that favours g-dropping

- covert prestige and identity construction?

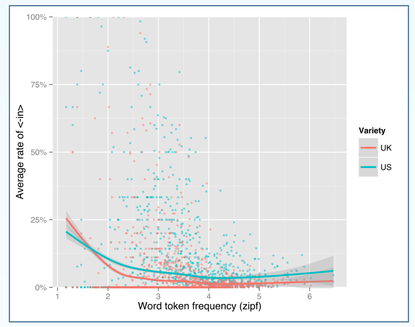

Figure 5: Rates of *-in* by zipf-frequency and variety

4. Conclusion

The results of this study have both **methodological** and **theoretical** implications:

It has shown the validity of employing innovative techniques in data collection using online social media

- Twitter is a rich source of natural language data, and its geographic metadata facilitates studies of regional stratification at an unprecedented level of detail

It suggests parallels between (ing)'s variation at the grapheme and phoneme level, where the non-standard spelling of <-ing> clusters is possibly phonologically-motivated

- the probability of a speaker typing <-in> could be influenced by their baseline rate of [ɪn] in speech, as well as other factors that condition the phonological variation
- that said, the effect of word token frequency is a reminder that the style of discourse used on online social media is not directly comparable to speech, and warrants investigation in and of itself

References

Abramowicz, Ł. 2007. Sociolinguistics meets exemplar theory: frequency and recency effects in (ing). *University of Pennsylvania Working Papers in Linguistics* 13(2), 27-37.

Bailey, G. 2015. *Social and internal constraints on (ing) in northern Englishes.* Unpublished MA dissertation, University of Manchester.

Derczynski, L., A. Ritter, S. Clarke, & K. Bontcheva. 2013. Twitter part-of-speech tagging for all: overcoming sparse and noisy Data. In Angelova, G., K. Bontcheva, & R. Mitkov (eds.), *Proceedings of Recent Advances in Natural Language Processing,* 198-206. Shoumen, Bulgaria: INCOMA Ltd.

Houston, A. 1985. *Continuity and change in English morphology: the variable (ING).* Ph.D dissertation, University of Pennsylvania.

Labov, W. 2001. *Principles of linguistic change, vol. 2: social factors.* Oxford: Blackwell.

Office for National Statistics. 2014. *Annual estimates of NUTS1 regional Gross Disposable Household Income (GDHI).* London: Office for National Statistics.

Tagliamonte, S. 2004. Somethi[ŋ]'s goi[n] on! Variable (ing) at ground zero. In Gunnarsson, B., L. Bergström, G. Eklund, S. Fridell, L. Hansen, A. Karstadt, B. Nordberg, E. Sundgrenand & M. Thelander (eds.), *Language Variation in Europe: Papers from the Second International Conference on Language Variation in Europe (ICLaVE 2),* 390-403. Uppsala: University Press.

U.S. Census Bureau. 2013. *Selected economic characteristics 2009-2013: American community survey 5-year estimates.*