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

This paper employs innovative corpus methodology to investigate orthographic reflections of the phonological variable (ing); specifically, it seeks to determine whether or not the non-standard <-in> spelling shows the same grammatical and regional patterning as the corresponding spoken variant /ɪn/. The study draws data from microblogging site Twitter to construct a corpus consisting of 2,000,000 'tweets' from North America and the United Kingdom, fully tagged for part-of-speech and geotagged with latitude/longitude coordinates.

The final dataset includes over 500,000 tokens of (ing). Frequencies of *-in* (represented orthographically as <-in> and <-in'>) are normalised by region (UK) and state (US), and plotted visually using polygonal boundary files in R to produce maps colour-coded by *-in* frequency. The results seem to suggest regional patterning that is largely reflective of (ing)'s phonological variation; that is, in the UK, *-in* is more prominent in the North West of England and particularly in Scotland, and in the US it is favoured in the southern states (see Labov 2001). The statistical significance of this result is confirmed in Rbrul multivariate analysis (Johnson 2009), with the polarity of each state's log-odds showing a particularly clear regional dichotomy (see Figure 1).

Evidence for the nominal-verbal continuum is also present, with nouns and adjectives showing comparably low rates of *-in,* and slightly higher frequencies found for verbs. The influence of lexical frequency is also investigated, taking frequency counts along the Zipf-scale (van Heuven et al. 2014) from the SUBTLEX-UK corpus. An effect of token frequency, though previously rejected for the phonological variable (Abramowicz 2007), appears here in a counter-intuitive direction; there is an indication that it is the *less* frequent words, usually computer-mediated 'online slang' terms such as *nuttin, pimpin* and *frickin*, that are more subject to g-dropping.

This study reveal fairly strong parallels between (ing)'s variation at the grapheme and phoneme level, prompting a more nuanced understanding of the relationship between non-standard orthography and the phonological processes that motivate it. It also makes a methodological contribution in assessing the viability of Twitter in corpus-creation; this social media platform proves to be a rich source of natural language data for corpus-based variationist studies, and its geographic metadata can facilitate investigations of regional stratification at an unprecedented level of detail and efficiency.

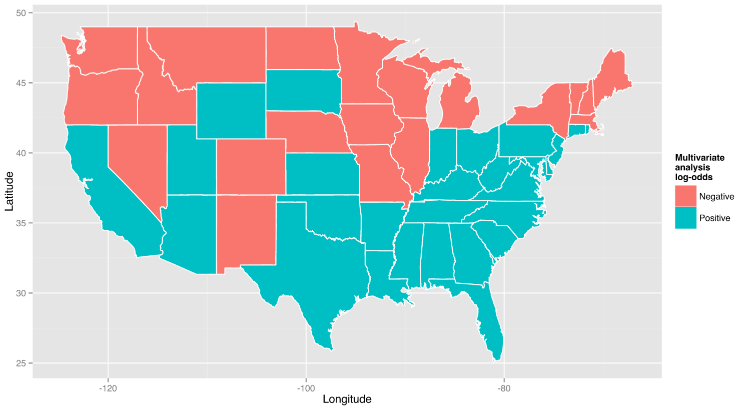

Figure 1: Regional stratification of *-in* in US states by multivariate analysis log-odds