Predicting variation in the frequency, dispersion, and the success of loanwords

Barbara E. Bullock, Jacqueline Serigos, Almeida Jacqueline Toribio
(The University of Texas at Austin)

What predicts the success of a loanword in a given community? We present a novel approach to this question by employing computational methods for automatic extraction of tokens, allowing for processing of a large corpus and providing an accountable basis by which to quantify the cultural dimensions of loanword dispersion. We aim to predict the potential success of English single noun loanwords relative to the success of their closest Spanish competitors. For example, in the Puerto Rican data examined here, the loanword *web* is highly successful compared to its Spanish equivalent *red*, whereas *cinnamon* competes less successfully with *canela*.

We draw on computational tools in the design of the corpus, created from the main pages of Puerto Rican news websites addressed to audiences of 3 distinct social classes: *Vocero* published for a working class population; *Nuevo Día*, directed at a broad readership; and *80 grados*, an online publication targeting a more intellectual readership. The corpus was downloaded automatically from archives using a hand-crafted Python script; 1.1 million words were extracted from each newspaper, resulting in a corpus of 3.3 million words. Once compiled, the corpus was processed using manual and automated steps to extract all English loanwords: (i) annotate the corpus for lemma using Tree Tagger (Schmid, 1995); (ii) identify tokens that are not recognized by the Spanish parameters; (iii) check if the non-Spanish tokens (e.g., French borrowings, proper names, onomatopeia) were recognized as English using a modified English dictionary; (iv) manually inspect the tokens to remove any noise.

First, the frequency of English loanwords is compared across the 3 subcorpora, using a Cramér-V test to estimate the effect size (Cohen 1988, Coe 2002, Kilgariff 2005). Then, following the analytical procedures of Zenner et al. (2012), we fit mixed models to each subcorpora separately with the dependent variable *SUCCESS RATE* of the anglicism (transformed in log odds) and the fixed predictors of orthographic *WORD LENGTH* and *LEXICAL FIELD* (levels: entertainment, technology, sports, business, social life) as fixed predictors. The *CONCEPT*, represented by the English lemma, serves as a random term. A separate model was fitted on the corpus as a whole with an additional fixed predictor, a scaled value for the *RELATIVE DISPERSION* of the loanword across subcorpora, which probes the degree to which a loanword is diffused across subcorpora. Our hypotheses maintain that: there will be significant differences in the frequency of borrowing between the subcorpora (*80Grados* > *NuevoDía* > *Vocero*); anglicisms of shorter length will have higher success across all corpora; within each subcorpora *LEXICAL FIELD* will be a significant predictor of success, but the levels of the factor will contribute differently toward accounting for the variation according to the subcorpus (e.g., *SPORTS* will have a stronger weight in *Vocero* than in *80Grados*).

This work affords new perspectives into the factors that influence borrowing behavior, by accounting for the differences in frequency, dispersion, and success of loanwords in terms of cultural and conceptual factors in addition to the well-studied linguistic ones (e.g., part of speech, transfer types (VanHout & Muysken 1994; Muysken 2000; Poplack & Sankoff 1984; Poplack et al. 1988)).

References:

Chesley, P., & Baayen, R. H. (2010). Predicting new words from newer words: Lexical borrowings in French. *Linguistics*, *48*(6), 1343–1374.

Coe, R. (2002, September 25). *It's the effect size, stupid: what effect size is and why it is important*. Retrieved July 7, 2014, from http://www.leeds.ac.uk/educol/documents/00002182.htm

Cohen, J. (2013). *Statistical Power Analysis for the Behavioral Sciences* (2nd ed.). Hoboken: Taylor and Francis.

Kilgarriff, A. (2005). Language is never, ever, ever, random. *Corpus Linguistics and Linguistic Theory*, *1*(2), 263–276. Muysken, P. (2000). *Bilingual speech: a typology of code-mixing*. Cambridge: Cambridge University Press.

Poplack, S., & Sankoff, D. (1984). Borrowing: the synchrony of integration. *Linguistics*, *22*(1), 99–136.

Poplack, S., Sankoff, D., & Miller, C. (1988). The social correlates and linguistic processes of lexical borrowing and assimilation. *Linguistics*, *26*(1), 47–104.

Schmid, H. (1994). Probabilistic part-of-speech tagging using decision trees. In *Proceedings of the international conference on new methods in language processing* (Vol. 12, pp. 44–49). Manchester, UK.

Van Hout, R., & Muysken, P. (1994). Modeling lexical borrowability. *Language Variation and Change*, *6*(01), 39–62.

Varra, R. M. (2013). *The social correlates of lexical borrowing in Spanish in New York City*. City University of New York.

Zenner, E., Speelman, D., & Geeraerts, D. (2012). Cognitive Sociolinguistics meets loanword research: Measuring variation in the success of anglicisms in Dutch. *Cognitive Linguistics*, *23*(4), 749–792.