Unsupervised machine learning for the accurate classification of the discourse marker *like* in code-switching utterances

Page Piccinini¹ & Eric Kramer²

¹Department of Linguistics, University of California, San Diego
²Medical Scientist Training Program, University of California, San Diego

I. Introduction

**BACKGROUND**
- Bilingual speakers often engage in code-switching, sometimes using both languages in a single utterance (e.g., Pfaff, 1979; Poplack, 1980; Myers-Scotton, 2008).
- This makes processing incoming speech more complicated, as speakers have to anticipate a possible change in language.
- It would be easier if there were cues in the signal warning a listener of an upcoming switch, which listeners could use to prepare for an upcoming change in languages.

**PAST WORK ON CODE-SWITCHING**
- In spontaneous speech Spanish-English bilinguals use the discourse marker *like* in English, Spanish, and code-switching utterances.
- An acoustic analysis found that the [l] and diphthong in *like* is produced differently depending on the type of utterance in which it occurs (Piccinini, 2011).
- Spanish *likes* begin more backed and end higher and more fronted than English.
- Code-switch *like* begins like the language preceding the switch, English or Spanish, and end somewhere in between both languages.

**RESEARCH QUESTIONS**
- What types of model best classify the tokens and teach us more about where differences lie between given categories?
- Is it the static differences at particular points in time that differentiate the tokens, or the general shape of the change over time?

**TOKENS**
- Tokens of *like* were collected from a corpus of spontaneous speech containing dyads of early Spanish-English bilinguals (8 female, 2 male).
- Each token was given one of four classifications:
  1. English
  2. Spanish
  3. Code-switch from English to Spanish (CS-ES)
  4. Code-switch from Spanish to English (CS-SE)

**STATISTICAL MODELS**
- Three logistic regression models were used to classify tokens using F1 and F2 values:
  1. Standard Logistic Regression
  2. Polynomial-Logistic Hybrid Regression
  3. An Independent Component Analysis (ICA) regression

**MODEL COMPARISON**
- The performance of each model was evaluated using receiver-operator curves (ROC), which plot the true positive rate (TPR) versus the false positive rate (FPR) for each model.
- TPR and FPR are defined below:
  \[
  TPR = \frac{TP}{TP + FN} \]
  \[
  FPR = \frac{FP}{FP + TN} \]
- The area under the curve (AUC) is the area underneath the ROC curve, and quantifies the overall performance of the model.
- AUC values range from 0.5 (random) to 1 (perfect).
- All ROC curves were built with the R package ROCR (Sing et al., 2005).

**RESULTS**
- There are significant differences in the number of tokens classified as English and Spanish tokens of the discourse marker *like* in code-switching utterances.
- *Like* is produced differently in English, Spanish, and code-switching utterances.
- Differences are significant in particular points in time.

**REFERENCES**