Defining a Matrix Language in Language Mixing

Vivek Sharath, The Bilingual Annotations Tasks (BATs) Research Group, vivek.sharath@gmail.com

Introduction

- It's often assumed that one language, the 'matrix' language (ML) supplies the grammar into which words of the other language are inserted when bilinguals are mixing languages
- Identifying the ML is said to predict the distribution of types of elements in one language or the other (e.g., determiners, auxiliaries, complementizers, "do" verbs should always be in the ML)

Research Questions

- How is the ML identified?
  - Do different proposals for determining the ML converge?
- Is an ML always identifiable?
  - Does the type of language mixing matter?

Two types of language mixing

- Insertional: "Pero mi papá murió en nineteen thirty-two" (data = Spanish in Texas; Gloss: But my father died in nineteen thirty-two)
- Alternational: "Anyway, al taxista right away le noté un acentito, not too specific." (data = Killer Crónicas; Gloss: Anyway, I noticed the accent of the taxi driver right away, not too specific)

Methods

- Split data into sentences and identify the language and part-of-speech of every word in each sentence
- For each sentence, we generate the following METRICS:
  - M-Index: ratio of languages on scale of 0 (monolingual) to 1 (fully balanced)
  - I-Index: probability of switching languages from one word token to next on 0 to 1 scale
  - Burstiness (Goh & Barabási 2008): whether the switching is periodic (-1) or aperiodic (1)
  - Memory (Goh & Barabási 2008): whether sequences of monolingual spans are autocorrelated in terms of length(-1 to 1)
- Remove tokens without a language and sentences with no switching

Operationalizing the ML

- For each sentence, the ML is operationalized in 3 ways:
  1. Word count: is the sentence Spanish-dominant, English-dominant or Tie?
  2. Verb count: are most of the verbs in Spanish, English or Tie?
  3. Functional word count: are most of the function words in Spanish, English or Tie?
- Analysis
  - Dependent variable: Agreement between 3 versions of ML or not
  - Predictor variables: Metrics

ML Analysis Example

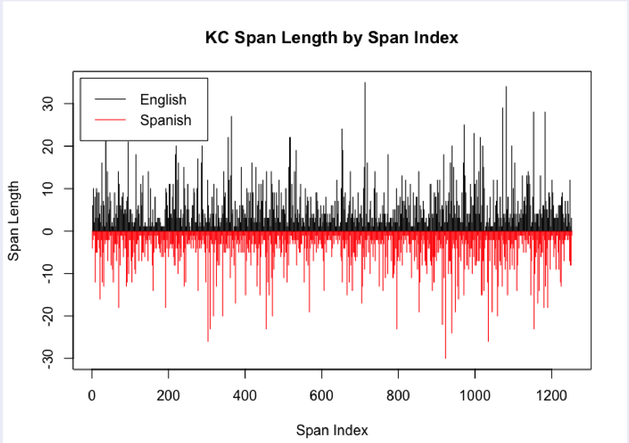
Anyway, al taxista right away le noté un acentito, not too specific.

ML Definition	English	Spanish	ML
Word Count	6	6	Tie
Verb	0	1	Spanish
Functional words	2	3	Spanish

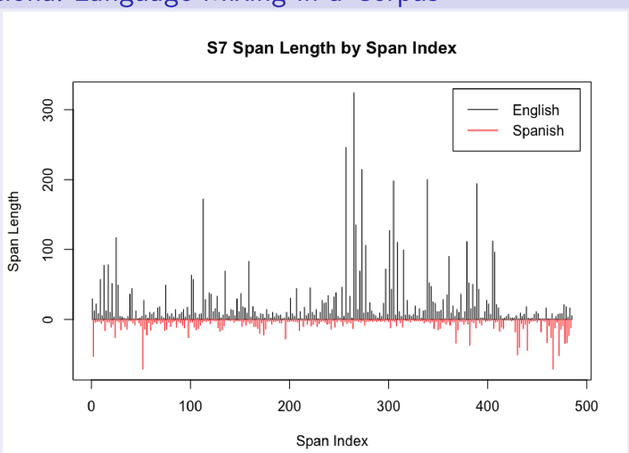
Experiments: Datasets

- Written texts: *Killer Crónicas* (KC), 40,469 words
- Spoken speech transcripts:
  - Spanish in Texas corpus (SpinTX), 500,000+ words
  - Solorio corpus (S7), 7,000 words
  - Miami corpus (Miami), 296,847 words

Alternational Language Mixing in a Corpus



Insertional Language Mixing in a Corpus



Ratio: M-index (Barnett et al., 2000)

- Quantifies inequality of distribution of language tags

$$\text{M-Index} \equiv \frac{1 - \sum {p_j}^2}{(k-1) \cdot \sum p_j}. \quad (1)$$

where $k > 1$ is the total number of languages represented in the corpus, p_j is the total number of words in the language j over the total number of words in the corpus, and j ranges over the languages present in the corpus.

Ratio: I-index (Guzmán et al., 2017)

- Proportion of CS points relative to possible switch points

$$\text{I-Index} \equiv \frac{1}{n-1} \sum_{1 \leq i = j-1 \leq n-1} S(l_i, l_j) \quad (2)$$

where $n > 1$ is the total number of tokens in the corpus, l_i, l_j range over all tokens, and $S(l_i, l_j) = 1$ only if $l_i = l_j$.

Time-course: Burstiness (Goh & Barabási, 2008)

- Manner and extent of CS: bursts or periodic

$$\text{Burstiness} \equiv \frac{(\sigma_\tau / m_\tau - 1)}{(\sigma_\tau / m_\tau + 1)} = \frac{(\sigma_\tau - m_\tau)}{(\sigma_\tau + m_\tau)} \quad (3)$$

where σ_τ is the standard deviation of the span lengths and m_τ the mean.

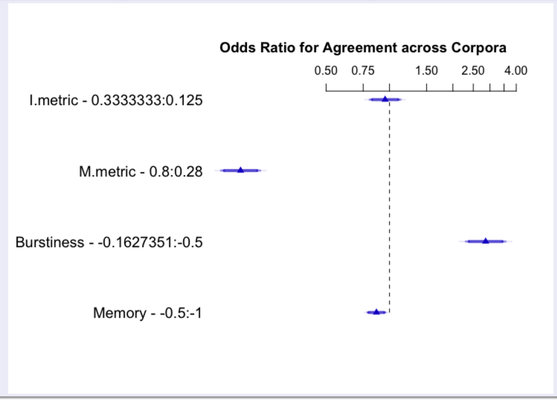
Time Course: Memory (Goh & Barabási 2008)

- Extent to which the length of language spans are autocorrelated

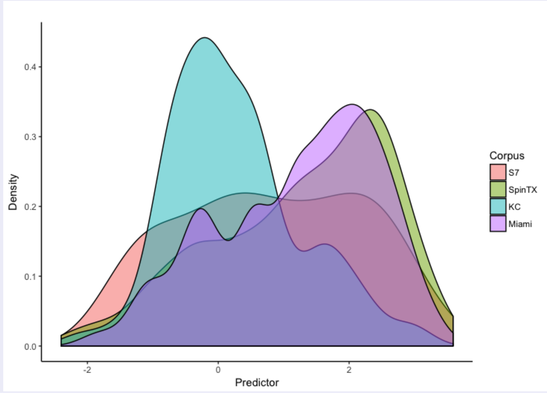
$$\text{Memory} \equiv \frac{1}{n_r - 1} \sum_{i=1}^{n_r - 1} \frac{(\tau_i - m_1)(\tau_{i+1} - m_2)}{\sigma_1 \sigma_2} \quad (4)$$

where n_r is the number of events, τ_i is the i-th span length, σ_1, m_1 the standard deviation and mean of all spans except the last, and σ_2, m_2 the standard deviation and mean

Odds Ratio for Agreement across Corpora



Predicting Agreement across Corpora



Results & Conclusions: Relevance for linguistics

- The three definitions of the ML agree over 44% of the time across all corpora for both types of language mixing
- Our metrics are effective in classifying whether the three definitions of the ML converge or not ($p < 0.001$)
  - However, some bilingual corpora do not show a clean separation in agreement, most notably in S7