Simple Tools for Exploring Variation in Code-Switching for Linguists

Gualberto A. Guzman, Jacqueline Serigos, Barbara E. Bullock and Almeida Jacqueline Toribio

University of Texas at Austin
Empirical Methods in Natural Language Processing

1 November 2016

Introduction

- The study of code-switching (CS) is stymied by a paucity of data and by current methods
  - Decontextualized, isolated examples recruited to support or refute hypotheses about CS
  - Disagreement about what constitutes CS (nonce borrowing vs. single word switching)
  - Theoretical proposals about CS structure and usage cannot be fully tested
  - Local corpora are guarded
- NLP methods hold great promise for exploiting increasingly accessible multilingual corpora

Our Contributions

1. Make a linguistic case for classifying CS types according to how integrated the languages are
2. Improve on existing language identification systems
3. Introduce an Integration-index (I-index) derived from HMM transition probabilities
4. Employ methods for visualizing integration via a language signature (or switching profile)
5. Illustrate the utility of our simple metrics for linguists as applied to Spanish-English texts of different switching profiles

Related Work

Mixed Texts

- Multilingual documents can represent different types of mixing (King & Abney 2013)
  - Translations
  - Change of author/speaker
  - ‘Classic’ or intrasentential code-switching (Myers-Scotton 1993)

Related Work

Mixing Typology

- Muysken (2000, 2014) presents a typology of switching processes, each reflecting different levels of contributions from two (or more) languages and each associated with different historical and cultural embedding

Insertional Switching

- Example 1, English/Punjabi (Rampton et al. 2006)
  I don't mix with <kɑ[e:> ('black boys')
  - The Matrix Language (Myers-Scotton, 1993) supplies the morphosyntactic frame into which chunks of the other language are introduced (e.g., borrowing and small constituent insertion)
  - Argued to be prevalent in postcolonial and immigrant settings where there is asymmetry in speakers' competence of both languages

Alternational Switching

- Example 2, Moroccan Arabic/Dutch (Nortier 1990)
  <Maar 't hoeft niet> li-?anna ida seft ana
  ('But it need not be, for when I see, I . . . ')
  - Speakers are said to draw on 'universal combinatory' principles in building equivalence between discrete language systems while maintaining the integrity of each (MacSwan, 2000; Sebba, 2009)
  - Argued to be most common among proficient bilinguals in situations of stable bilingualism

Congruent Lexicalization

- Example 3, English/Afrikaans (Van Dulm, 2007)
  You've got no idea how <vinnig> I've been <slaan-ing> this <by mekaar>

  - Syntax of the languages is aligned and speakers produce a common structure using words from both languages
  - Argued to be attested among bilinguals who are fluent in typologically similar languages of equal prestige as well as in dialect/standard and post-creole/lexifier mixing

Back-Flagging

- Example 4, English/French (DuBois & Horvath, 2002)
  <Ça va>. Why don't you rewire this place and get some regular light switches? ('It's okay.')
  - Grammatical and lexical properties of the majority language serve as the base language into which emblematic minority elements are inserted
  - Said to signal ethnic identities once speakers have shifted to the majority language

Related Work

Multilingual Indices

- M-index or Mixing Index (Barnett et al., 2000)

  Indicates the degree to which various languages are represented in a text

  Limitation: Does not show how the languages are integrated

- CMI or Code-Mixing Index (Gambäck & Das, 2014, 2016)

  Ratio of language tokens that are from the majority language of the text

  Limitation: Segments the corpus into utterances, assumes a matrix language, and requires computing weights

Language Model

- Two tiers of annotation
  1. Language: Spanish, English, Punctuation, or Number
  2. Named Entity: yes or no
- Character n-gram (5-gram) and first order word-level Hidden Markov Model (HMM) model (Solorio & Liu, 2008)
- Two versions of the character n-gram model were tested
  1. One is trained on the CALLHOME transcripts
  2. The second is trained on the $\text{SUBTLEX}_{\text{US}}$ and ACTIV-ES subtitle corpora

Language Model

Named Entities

- We use the Stanford Named Entity recognizer with both the English and Spanish parameters

  If either Entity recognizer identified the token as a named entity, it was tagged as a named entity

- Unlike other taggers where named entities are viewed as language neutrals, our named entities retained their language identification from their first tier of annotation (Çetinoglu, 2016)

  We classify the language of Named Entities as they can trigger CS (Broersma & De Bot, 2006)

Integration Index

The I-index is calculated as follows:

$$\frac{1}{n-1} \sum_{1 \leq i \leq j \leq n} S(l_i, l_j),$$

where $S(l_i, l_j) = 1$ if $l_i \neq l_j$ and 0 otherwise

The factor of $1/(n-1)$ reflects the fact that there are $n-1$ possible switch sites in a corpus of size n

Integration Index

Examples

1. I don't mix with <kɑɭe:>
2. <Maar 't hoeft niet> li-?anna ida seft ana
3. You've got no idea how <vinnig> . . .
4. <Ça va>. Why don't you rewire this . . .
5. Anyway, al taxista right away le noté un acentito, not too specific.
6. Sí, ¿y lo otro no lo es? Scratch the knob and I'll kill you.

Table 1: Spans for Examples

Examples	1	2	3	4	5	6
Language 1	4	4	8	2	6	7
Language 2	1	5	4	12	6	7
Code-Switches	1	1	5	1	4	1
M-metric	0.47	0.98	0.8	0.32	1	1
I-index	0.25	0.125	0.45	0.08	0.36	0.08

Language Signature

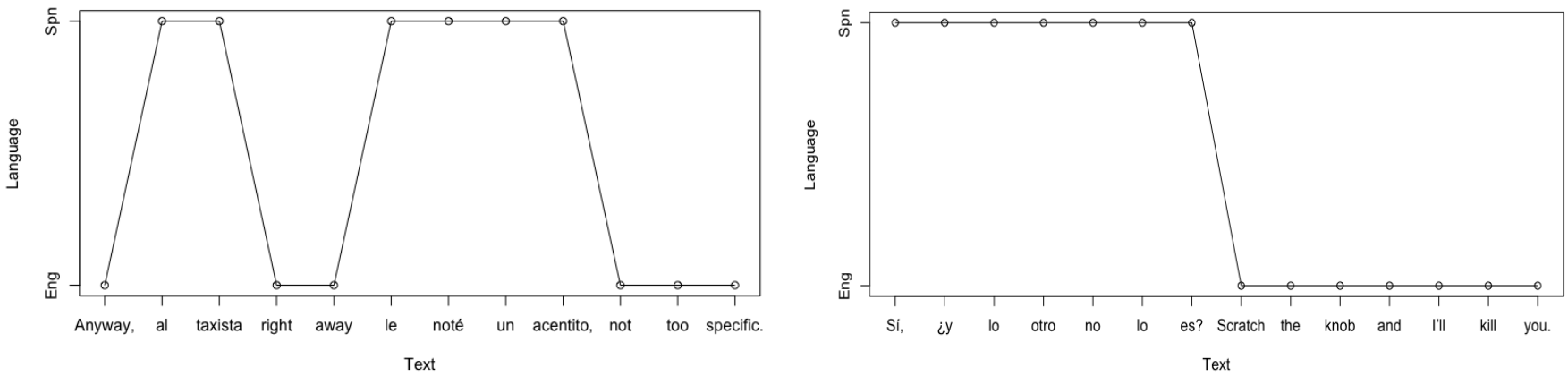
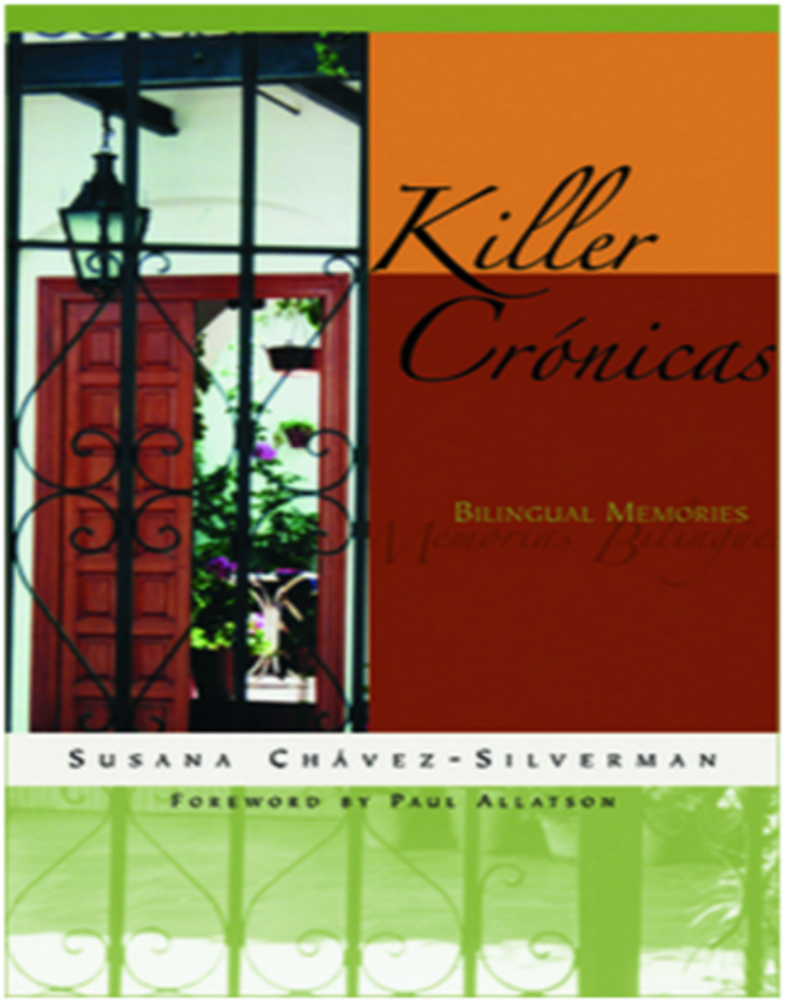


Figure 1: Chronological CS Plot for Examples 5 and 6

We offer the concept of a language signature that can visualize span length over time

Datasets

Killer Crónicas

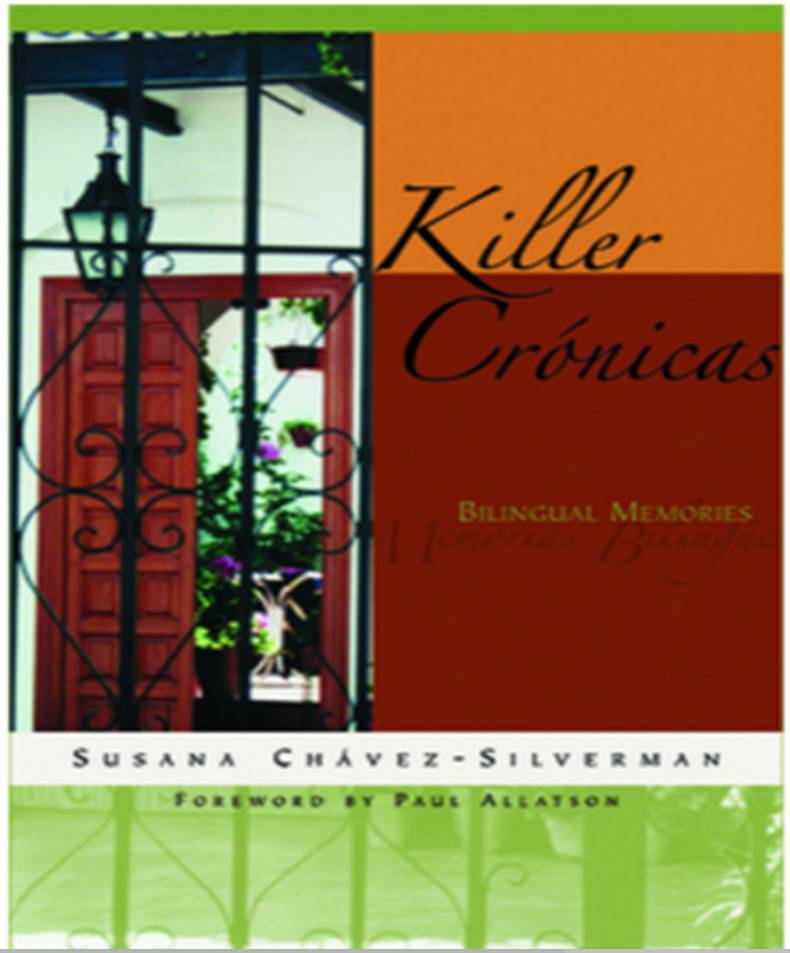


So different from when I lived in Spain, en la secundaria. De teenager, me regocijaba when my foreignness was apparent. Angry at my parents for uprooting me en la cúspide of what would be, alas, una short-lived y sólo semi-popularidad, I turned upon the foreign country toda la rabia y el veneno de mi terca (in)diferencia. Pero en Buenos Aires (y OJITO: eso que before I moved there, casi los únicos argentinos who'd impressed me favorably were either in books or dead or both . . .) I realized que nunca me había sentido más . . . qué sé sho—and I know que es medio cursi y trillado, pero—más yo misma. Y . . . (pausa porteña) ob-vio que había—hay—enormes, pero gaping differences between me and the average porteña. Pero an odd, opiate centeredness, hasta orgullo washed over me, more and more as the months passed y me daba cuenta de que la gente me pasaba in the streets not exactly like I was invisible sino como si fuera . . . one of them. Eso nunca, pero *nunca* me había pasado in any city, in any country before. Not even at "home." Y esa extraña comodidad o aceptación de mí misma, in my skin,

44

Datasets

Killer Crónicas

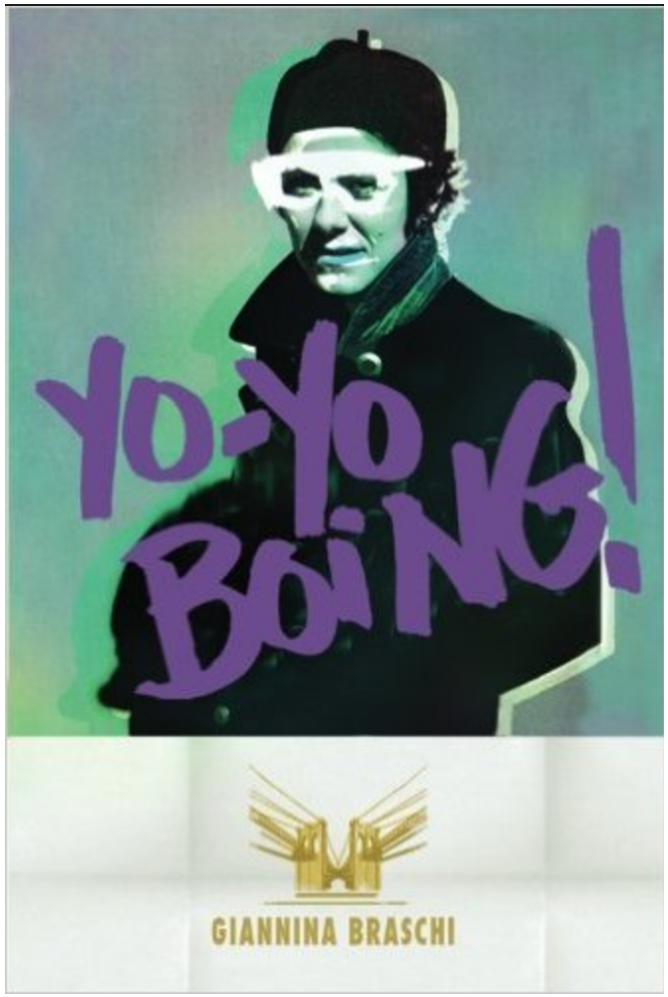


So different from when I lived in Spain, en la secundaria. De teenager, me regocijaba when my foreignness was apparent. Angry at my parents for uprooting me en la cúspide of what would be, alas, una short-lived y sólo semi-popularidad, I turned upon the foreign country toda la rabia y el veneno de mi terca (in)diferencia. Pero en Buenos Aires (y OJITO: eso que before I moved there, casi los únicos argentinos who'd impressed me favorably were either in books or dead or both . . .) I realized que nunca me había sentido más . . . qué sé sho—and I know que es medio cursi y trillado, pero—más yo misma. Y . . . (pausa porteña) ob-vio que había—hay—enormes, pero gaping differences between me and the average porteña. Pero an odd, opiate centeredness, hasta orgullo washed over me, more and more as the months passed y me daba cuenta de que la gente me pasaba in the streets not exactly like I was invisible sino como si fuera . . . one of them. Eso nunca, pero *nunca* me había pasado in any city, in any country before. Not even at "home." Y esa extraña comodidad o aceptación de mí misma, in my skin

Datasets

Yo-Yo Boing!

- Killer Crónicas: Bilingual Memoires (KC) is a 40,469-word work entirely in 'Spanglish'
- Yo-Yo Boing! (YYB) is a 58,494-word written in chapters of Spanish, English, and 'Spanglish'
- EMNLP 2014 (EN-ES)[1], 11,400 Spanish-English tweets with LangID



[1] ritual.uh.edu/code-switching/code-switching-resources

Experiments

Evaluation

Table 2: Language Accuracy on KC using different training corpora

	ACTIV-ES & $SUBTLEX_{US}$			CALLHOME		
Language	Accuracy	Precision	Recall	Accuracy	Precision	Recall
English	0.9507	0.9332	0.9729	0.9343	0.8931	0.9893
Spanish	0.9479	0.9021	0.9853	0.9442	0.9286	0.9422

Changing to equal-size corpora of 3.5M words (ACTIV-ES and $SUBTLEX_{US}$) resulted in a quantitative increase of 1% in language accuracy for both languages and better tagging of “ti”, “me” and “a” in mixed contexts

Results

M-metric and I-index

Table 3: Language Integration and Mixing

Corpus	M-index	I-index
Killer Crónicas	0.96	0.197
Yo-Yo Boing!	0.95	0.034
EN-ES	0.72	0.067

- Similar M-metrics for the two novel corpora
- Distinct levels of integration
- Twitter data is less bilingual than both, but more integrated than YYB

Results

Span Distributions

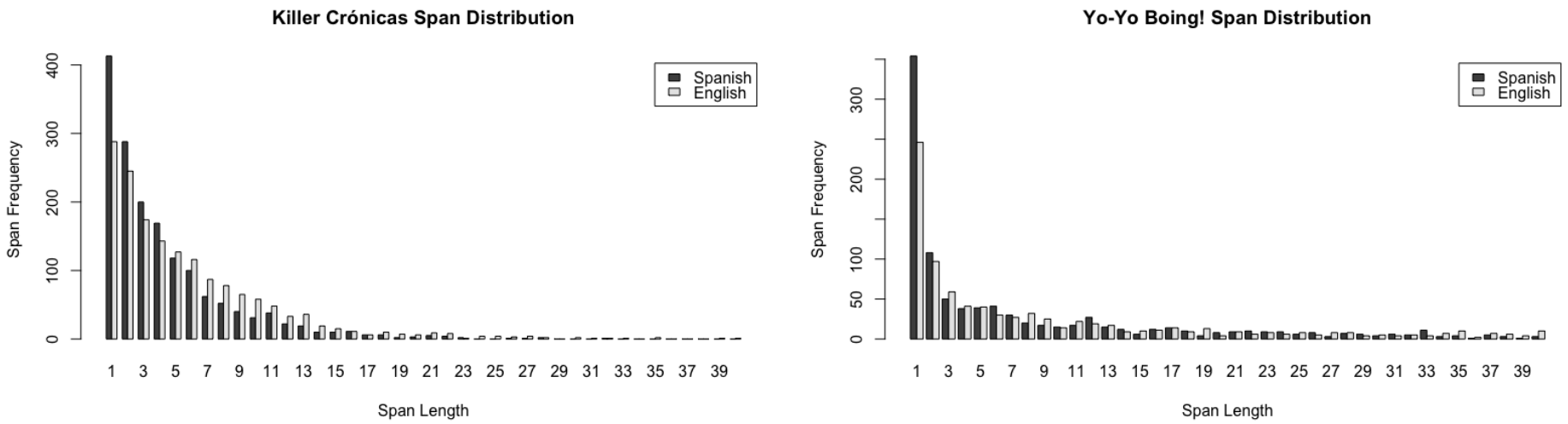


Figure 2: Span Distributions

- KC: Rapid exponential decay in span length vs. frequency
- YYB: Heavy tail, indicating a higher frequency of large spans compared to KC

Results

NER Performance

Table 4: KC NER Classification Performance

	Accuracy	Precision	Recall
Same Language	96.72%	79.19%	65.30%
Opposite Language	88.92%	33.24%	74.85%
English Only	96.65%	83.94%	58.08%
Spanish Only	89.00%	34.42%	82.06%

- Using only the English classifier yields the highest precision
- The Spanish classifier resulted in the highest recall rate
- These statistics are completely dependent on the dataset

Conclusion

What we've achieved

- Simple and easily calculated measure—the I-index—for quantifying language integration in multilingual texts
- Methods of visualizing the language profile of mixed-language documents
- Improved automatic language-identification system for classifying Spanish-English bilingual documents based on Solorio & Liu (2008)
- Increased accuracy by 1% in our model by experimenting with training corpora
- Reduced the greediness and latency of the Stanford Named Entity Recognizer by chunking text into spans of length 1000

Conclusion

What we've found

- KC and YYB have almost identical M-metrics, but different switching profiles
  - KC has a higher I-index, reflecting short, switched spans in each language relative to YYB
  - YYB illustrates a low I-index, exposing the alternation of monolingual with mixed-language chapters
- EN-ES showed a lower M-metric and an I-index that indicates much less language integration than KC, but more than YYB
- Implication: Some texts are more suitable for the study of intrasentential CS than others

Conclusion

Where we're going

- Our metrics are language independent, so we are free to study all kinds of mixed-language corpora
- Currently experimenting with quantifications of burstiness and memory because the I-index does not capture time course of CS (Goh & Barabási, 2008)
- Better visualizations of CS in mixed-language texts
- POS tagging

Thank you

Contact Information

- Gualberto A. Guzman, Jacqueline Serigos, Barbara E. Bullock, Almeida Jacqueline Toribio
  University of Texas at Austin

  {gualbertoguzman,jserigos}@utexas.edu
  {bbullock,toribio}@austin.utexas.edu

Supplementary Slides

Burstiness

Burstiness is defined by

$$B \equiv \frac{\sigma\tau/m_\tau - 1}{\sigma\tau/m_\tau + 1} = \frac{\sigma_\tau - m_\tau}{\sigma_\tau + m_\tau},$$

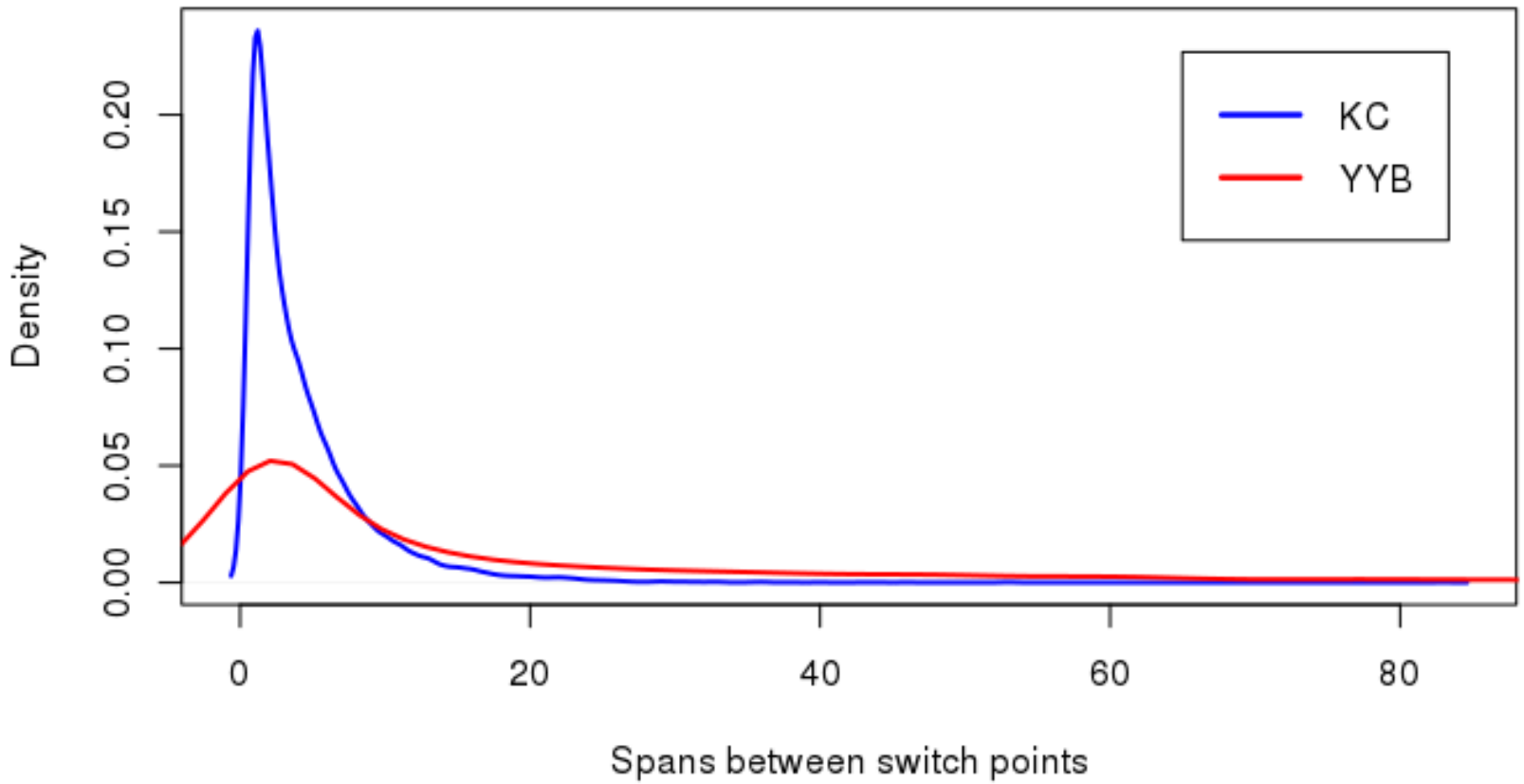
where σ_τ is the standard deviation of the span lengths and m_τ the mean.

- Value from -1 to 1, measuring how much a series of events differs from a Poisson Distribution

Supplementary Slides

Burstiness Visualization

Figure 3: Language Span Density By Corpus



Supplementary Slides

Memory

Memory is defined by

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

where n_r is the number of events, τ_i is the current span length, τ_{i+1} the next 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

- Value from -1 to 1, measuring how much a series of events "remembers" a short or long sequence of switching