One Text / Two Languages

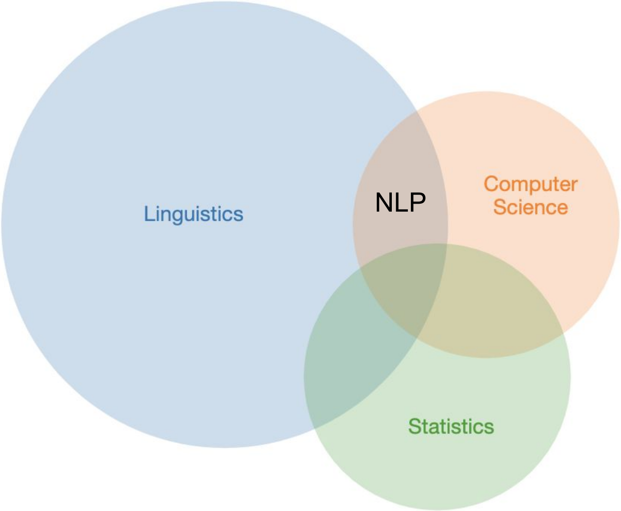
How linguists and programmers approach code-switching

Dr. Jacqueline Serigos

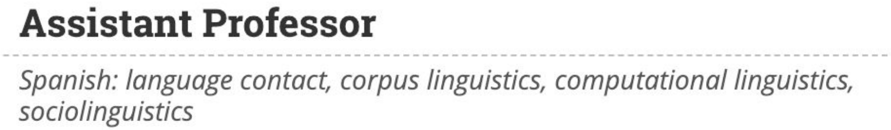
https://github.com/jserigos/One-Text-Two-Languages

Intro

About Me



GEORGE MASON UNIVERSITY



Assistant Professor

Spanish: language contact, corpus linguistics, computational linguistics, sociolinguistics

Goals of this workshop

1. Code-switching
2. NLP pipeline
3. Language Identification

No sticky note: "I'm happily working on it"

Blue sticky note: "I'm all done and ready to move on"

Orange sticky note: "I'm stuck, can someone help me?"

Alternatively, flag one of us down

Your background in Natural Language Processing

No experience

Some experience

Lots of experience

Code switching

What is it?

Why does it occur?

Why does it matter?

Text 1

Hindi - English

यहाँ पे मुझे कई बार थोड़ा अकेलापन लगते है क्योंकि दोस्त ही यहाँ family है, दोस्त ही परिवार है। जितने दोस्त हैं उन्ही के साथ आप कुछ समय बिता सकते हो और अपनी बातें share कर सकते हो, क्योंकि कोई immediate family या कोई direct family यहाँ तो है नहीं, तो मुझे ये लगता है कई बारी कि थोड़ा अकेलापन है, थोड़ा emotional support थोड़ी family support कम है यहाँ पे, जो कि India में थोड़ी easily available रहती है क्योंकि सब साथ में रहते हैं एक बड़े घर में, सब, पूरी family इकट्ठे रहेंगी। तो वो थोड़ा-सा missing लगता है।

Text 2

Spanish - English

Pues, el más feliz fue cuando hice find out, que estaba en el top ten percent. Yeah.. en lunch cuando nos juntamos todas las amigas, eran bien suave, los football games en los Fridays, yeah it was nice.

Task 1 - What is code-switching?

For each text,

1. Underline the English tokens
2. Identify dominant language and embedded language
3. Consider what types of words from the embedded language are being used
    - Content words or gramatical words
    - Part of Speech: Nouns, verbs, adjectives, prepositions, articles
4. What is code-switching?
5. Hypothesize why the speaker uses words from the embedded language

Text 1

Hindi - English

यहाँ पे मुझे कई बार थोड़ा अकेलापन लगते है क्योंकि दोस्त ही यहाँ family है, दोस्त ही परिवार है। जितने दोस्त हैं उन्ही के साथ आप कुछ समय बिता सकते हो और अपनी बातें share कर सकते हो, क्योंकि कोई immediate family या कोई direct family यहाँ तो है नहीं, तो मुझे ये लगता है कई बारी कि थोड़ा अकेलापन है, थोड़ा emotional support थोड़ी family support कम है यहाँ पे, जो कि India में थोड़ी easily available रहती है क्योंकि सब साथ में रहते हैं एक बड़े घर में, सब, पूरी family इकट्ठे रहेंगी। तो वो थोड़ा-सा missing लगता है।

Text 2

Spanish - English

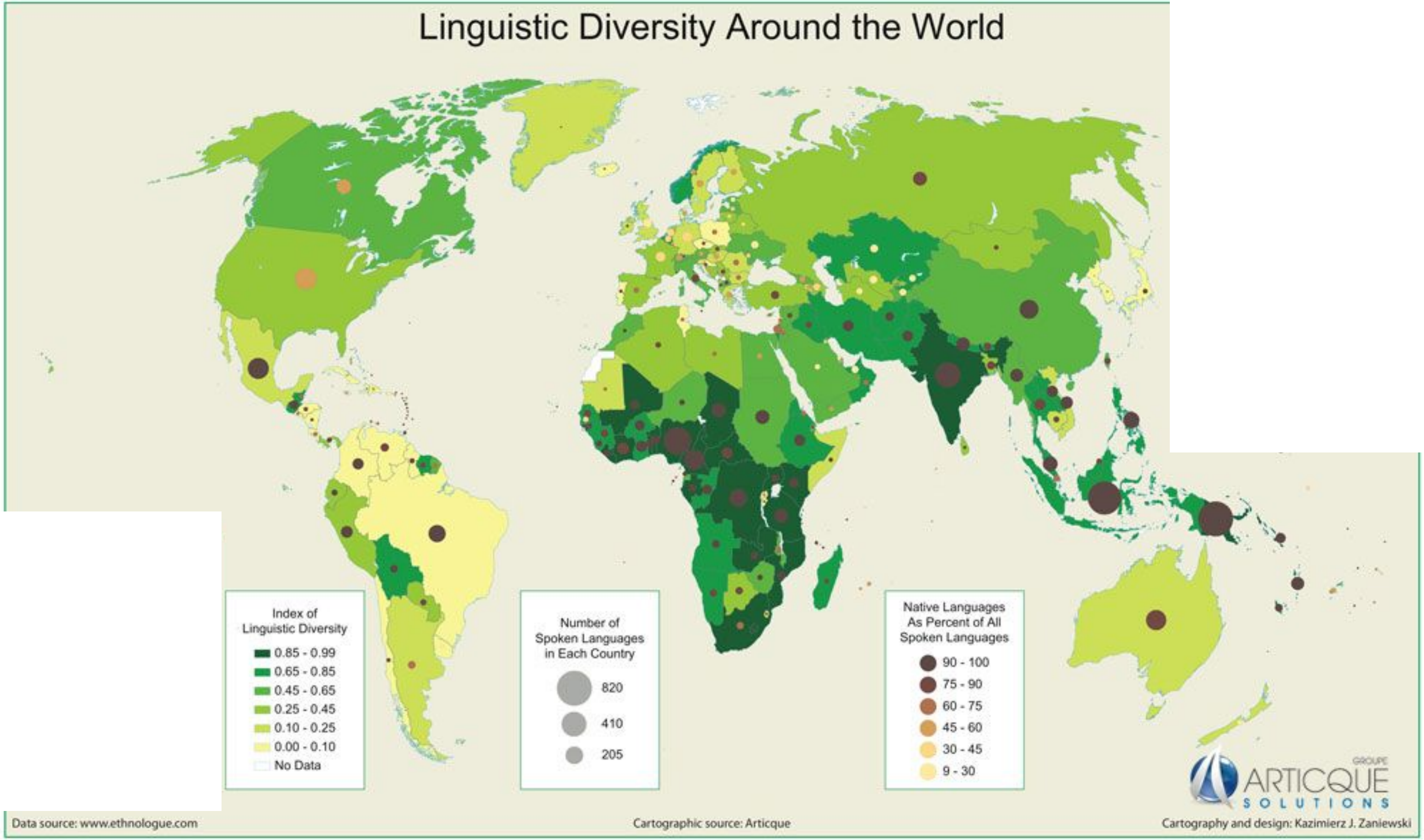
Pues, el más feliz fue cuando hice find out, que estaba en el top ten percent. Yeah.. en lunch cuando nos juntamos todas las amigas, eran bien suave, los football games en los Fridays, yeah it was nice.

Task 2 - How to identify languages

1. Outline a Pseudo Code to identify L1 & L2

Token	Language
Cuando	Spanish
hice	Spanish
find	English
out	English
que	Spanish
estaba	Spanish

Why?


Linguistic Diversity Around the World
Index of Linguistic Diversity
0.85 - 0.99
0.65 - 0.85
0.45 - 0.65
0.25 - 0.45
0.10 - 0.25
0.00 - 0.10
No Data
Number of Spoken Languages in Each Country
820
410
205
Native Languages As Percent of All Spoken Languages
90 - 100
75 - 90
60 - 75
45 - 60
30 - 45
9 - 30
GROUPE ARTICQUE SOLUTIONS
Data source: www.ethnologue.com
Cartographic source: Articque
Cartography and design: Kazimierz J. Zaniewski

Why does code-switching matter to linguists?

- Answer research questions
  - Who code-switches?
  - When do they code-switch?
  - Why do they code-switch?
- Debunk the myths about code-switching
  - X Children code-switch because they confuse the two languages
  - X People who code-switch do not know how to properly speak either language
  - X People code-switch because they are lazy

Why does code-switching matter?

What can I help you with?

Hey Siri, Can you play *Despacito*?

Turning text into data

Text 2

Spanish - English

Pues, el más feliz fue cuando hice find out, que estaba en el top ten percent. Yeah.. en lunch cuando nos juntamos todas las amigas, eran bien suave, los football games en los Fridays, yeah it was nice.

Token	Language	Part of Speech	Speaker
Cuando	Spanish	Adverb	Laura
hice	Spanish	Verb	Laura
find	English	Verb	Laura
out	English	Prep.	Laura
que	Spanish	Conjun.	Laura
estaba	Spanish	Verb	Laura

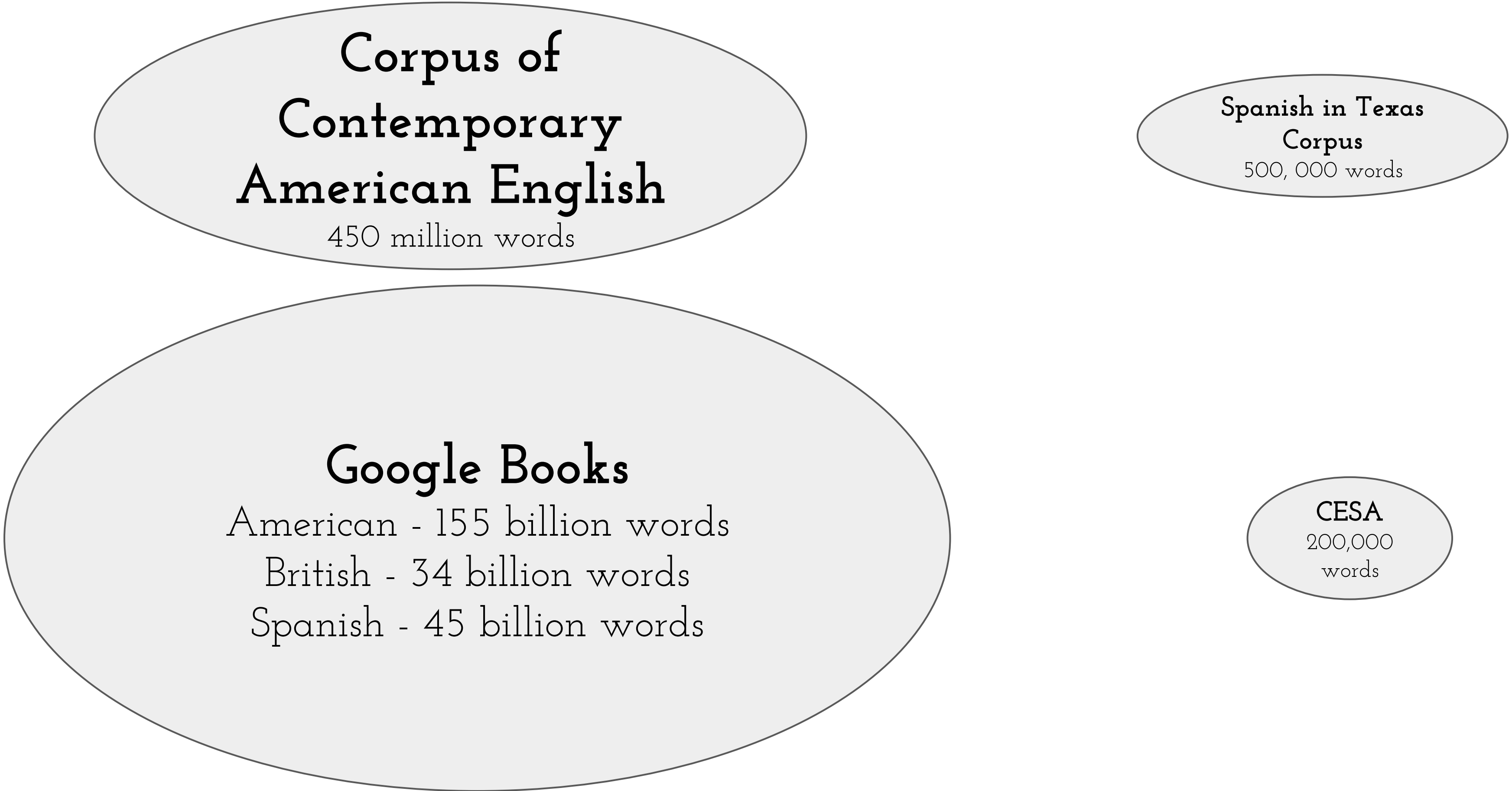
Natural Language Processing

Challenges of Code-switched Data

	Monolingual Texts	**Code-switched Texts**
Big Data	✔	~
NLP tools	✔ (NLTK, SpaCy)	~

Monolingual Data

Code-Switched Data



Corpus of Contemporary American English
450 million words

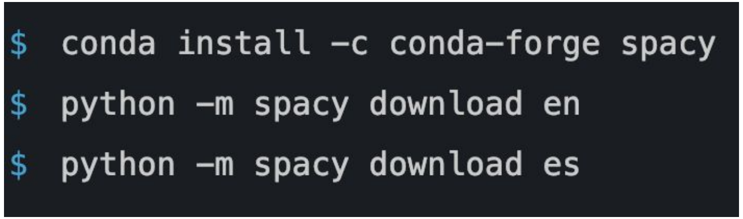
Google Books
American - 155 billion words
British - 34 billion words
Spanish - 45 billion words

Spanish in Texas Corpus
500, 000 words

CESA
200,000
words

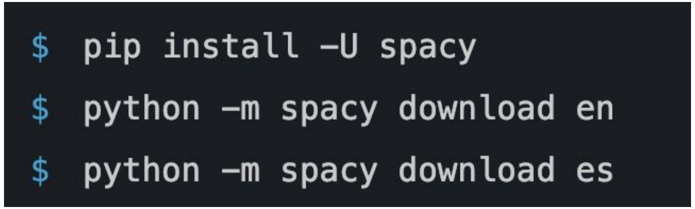
Install Libraries

- Download repo → https://github.com/jserigos/One-Text-Two-Languages
- Install SpaCy → https://spacy.io/usage/
  - This version of spaCy requires numpy >= 1.10. The download will fail if your version of numpy is too old.



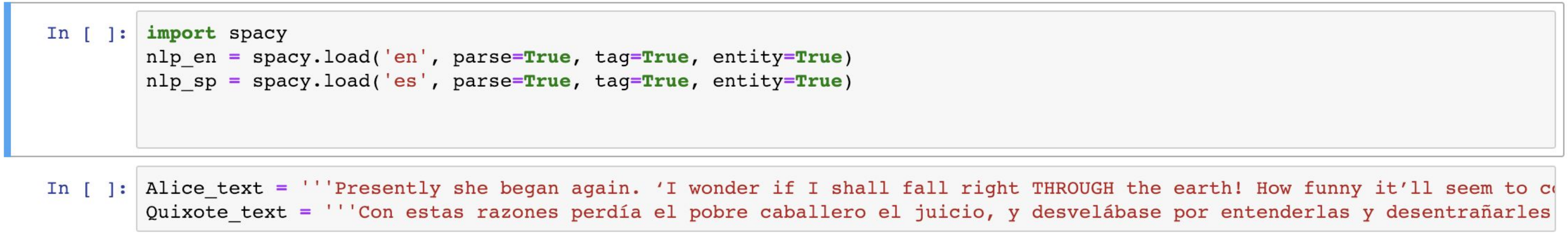
```
$  conda install -c conda-forge spacy
$  python -m spacy download en
$  python -m spacy download es
```

or



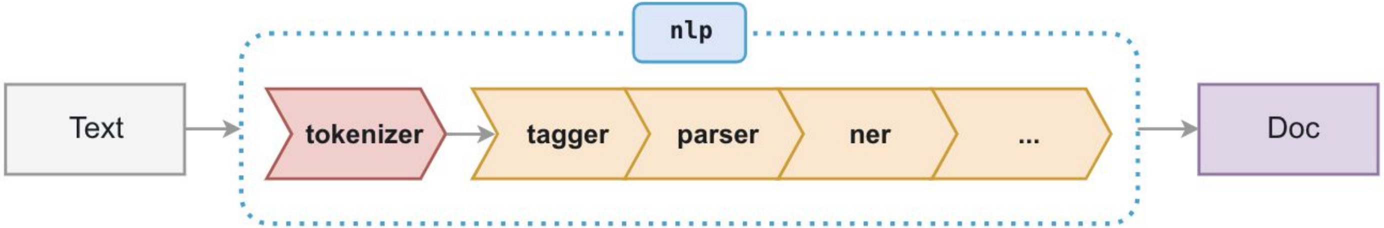
```
$  pip install -U spacy
$  python -m spacy download en
$  python -m spacy download es
```

Spacy

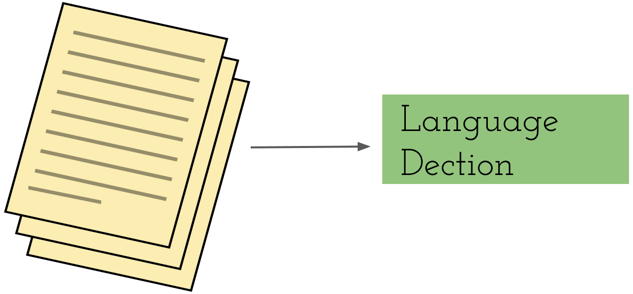


In [ ]:
```
import spacy
nlp_en = spacy.load('en', parse=True, tag=True, entity=True)
nlp_sp = spacy.load('es', parse=True, tag=True, entity=True)
```

In [ ]:
```
Alice_text = '''Presently she began again. ‘I wonder if I shall fall right THROUGH the earth! How funny it’ll seem to co
Quixote_text = '''Con estas razones perdía el pobre caballero el juicio, y desvelábase por entenderlas y desentrañarles
```

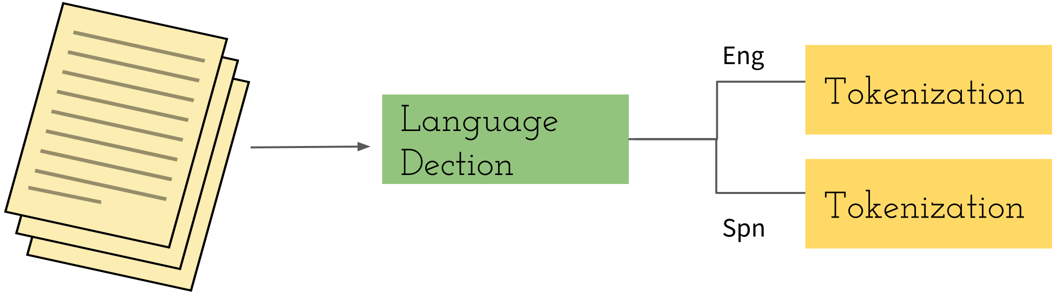


NLP Pipeline



Language Dection

NLP Pipeline



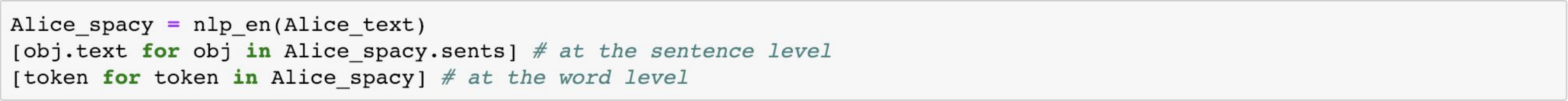
Language Dection

Eng

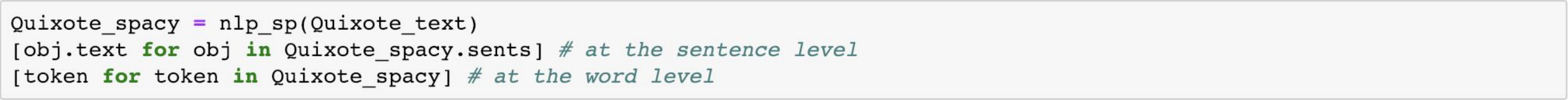
Tokenization

Spn

Tokenization

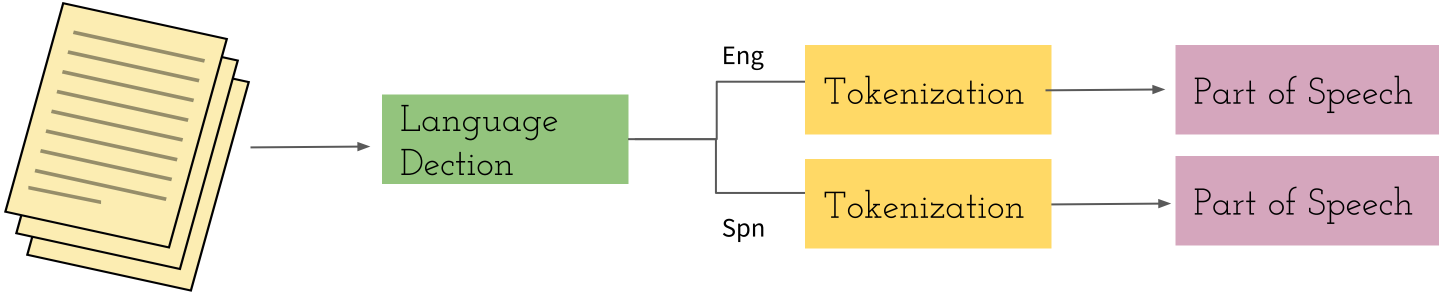


```
Alice_spacy = nlp_en(Alice_text)
[obj.text for obj in Alice_spacy.sents] # at the sentence level
[token for token in Alice_spacy] # at the word level
```



```
Quixote_spacy = nlp_sp(Quixote_text)
[obj.text for obj in Quixote_spacy.sents] # at the sentence level
[token for token in Quixote_spacy] # at the word level
```

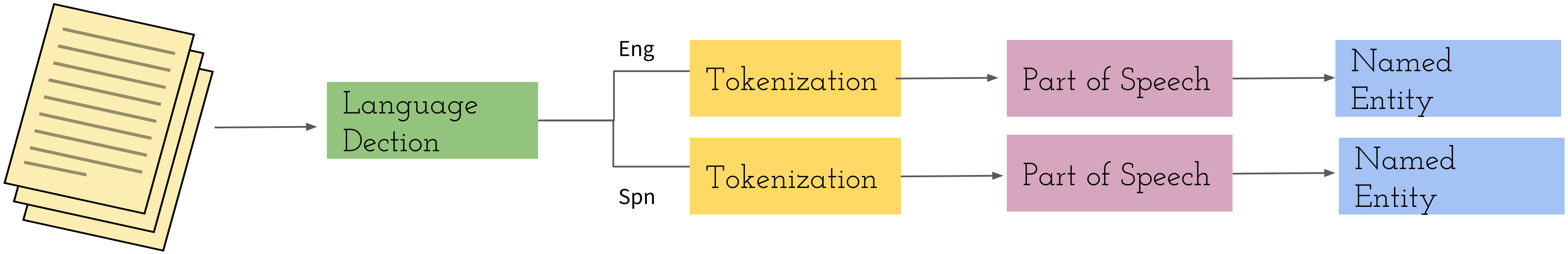
NLP Pipeline



```
[(token, token.pos_) for token in Alice_spacy]
```

```
[(token, token.pos_) for token in Quixote_spacy]
```

NLP Pipeline



Language Dection

Eng

Tokenization

Part of Speech

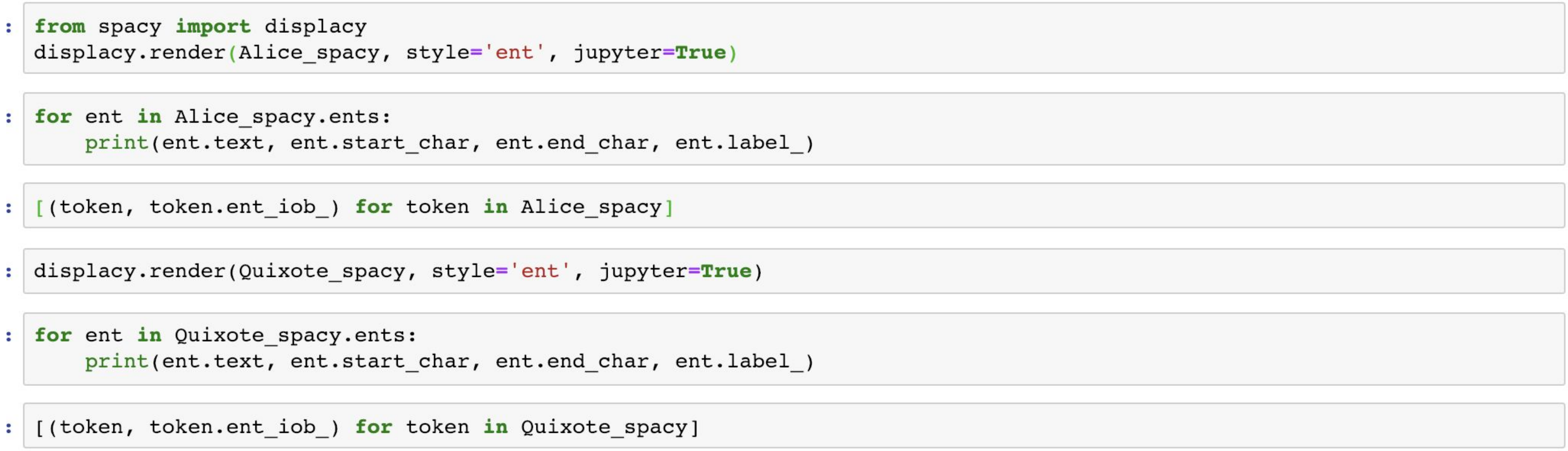
Named Entity

Spn

Tokenization

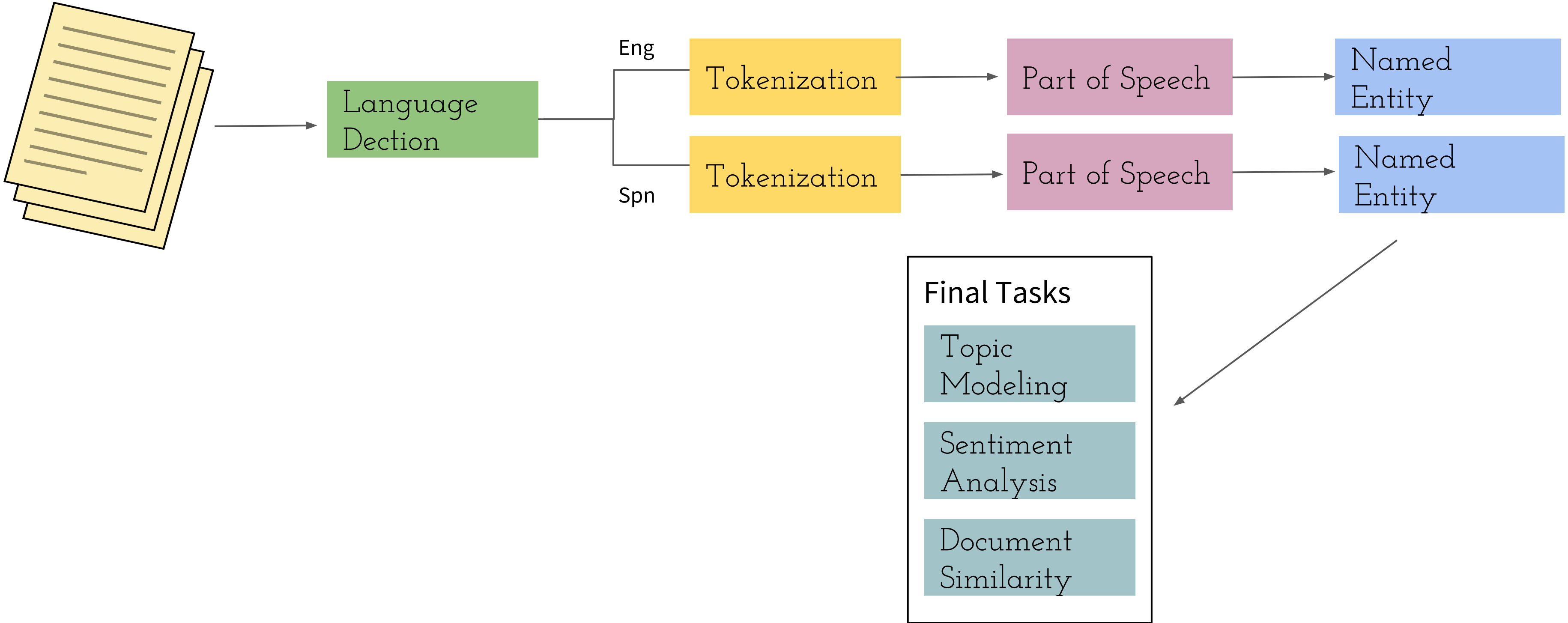
Part of Speech

Named Entity



```
: from spacy import displacy
  displacy.render(Alice_spacy, style='ent', jupyter=True)

: for ent in Alice_spacy.ents:
      print(ent.text, ent.start_char, ent.end_char, ent.label_)

: [(token, token.ent_iob_) for token in Alice_spacy]

: displacy.render(Quixote_spacy, style='ent', jupyter=True)

: for ent in Quixote_spacy.ents:
      print(ent.text, ent.start_char, ent.end_char, ent.label_)

: [(token, token.ent_iob_) for token in Quixote_spacy]
```

NLP Pipeline



Language Identification

1. By document
2. At the word level

At Document Level

- Ideas? How to identify languages?

Text 1

Hindi

एसोसिएशन फुटबॉल जिसे आमतौर पर सिर्फ फुटबॉल या सॉकर कहा जाता है, दुनिया के सबसे लोकप्रिय खेलों में से एक है । यह एक सामूहिक खेल है और इसे ग्यारह खिलाड़ियों के दो दलों के बीच खेला जाता हैं । फुटबॉल को सामान्यत: एक आयताकार घास या कृत्रिम घास के मैदान पर खेला जाता है जिसके दोनों छोरों पर एक एक गोल होता है। खिलाड़ियों द्‌वारा विरोधी दल के गोल में चालाकी से गेंद को डालना ही इस खेल का उद्‌देश्य है। खेल में गोलरक्षक ही एक मात्र ऐसा खिला\

Text 2

English

Association football, more commonly known as football or soccer, is a team sport played with a spherical ball between two teams of eleven players. It is played by 250 million players in over 200 countries and dependencies, making it the world's most popular sport. The game is played on a rectangular field called a pitch with a goal at each end. The object of the game is to score by moving the ball beyond the goal line into the opposing goal.

Text 3

Spanish

El fútbol es un deporte de equipo jugado entre dos conjuntos de once jugadores cada uno y algunos árbitros que se ocupan de que las normas se cumplan correctamente. Es ampliamente considerado el deporte más popular del mundo, pues lo practican unas 270 millones de personas.

Text 4

Korean

축구는 출전 선수 11명씩 각각 한 팀을 이루어 두 팀이 겨루며, 세계적으로 최고 인기를 누리는 스포츠이다.[1] 경기장은 직사각형이며, 바닥은 천연잔디나 인조잔디, 흙 등으로 이뤄져 있다. 경기장 양 끝에 놓인 상대방 골대 사이로 공을 통과시키면 득점이 된다. 선수 중 골키퍼만 팔과 손으로도 공을 건드릴 수 있으며, 나머지 선수는 팔과 손을 제외한 신체 부위로만 공을 다룰수 있다.

wikipedia.com

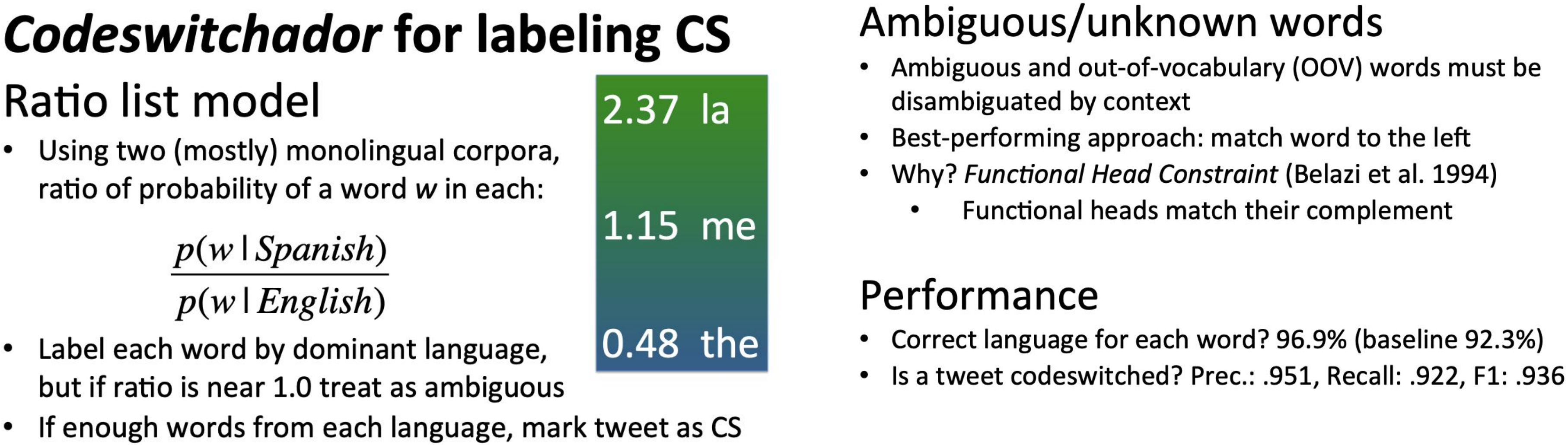
Character Encoding

- https://github.com/jserigos/codeswitch-annotation/blob/master/notebooks/HindiEnglish.ipynb

Stop Words

- https://www.mostlymaths.net/2012/06/language-detection-in-python-with-nltk.html

Ngrams



***Codeswitchador* for labeling CS**

Ratio list model

- Using two (mostly) monolingual corpora, ratio of probability of a word *w* in each:

$$\frac{p(w \mid Spanish)}{p(w \mid English)}$$

- Label each word by dominant language, but if ratio is near 1.0 treat as ambiguous
- If enough words from each language, mark tweet as CS

2.37	la
1.15	me
0.48	the

Ambiguous/unknown words

- Ambiguous and out-of-vocabulary (OOV) words must be disambiguated by context
- Best-performing approach: match word to the left
- Why? *Functional Head Constraint* (Belazi et al. 1994)
  - Functional heads match their complement

Performance

- Correct language for each word? 96.9% (baseline 92.3%)
- Is a tweet codeswitched? Prec.: .951, Recall: .922, F1: .936

From Constantine Lignos and Mitch Marcus (University of Pennsylvania, Johns Hopkins University)

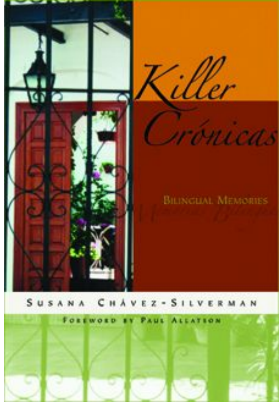
¡Muchas gracias!

More Resources

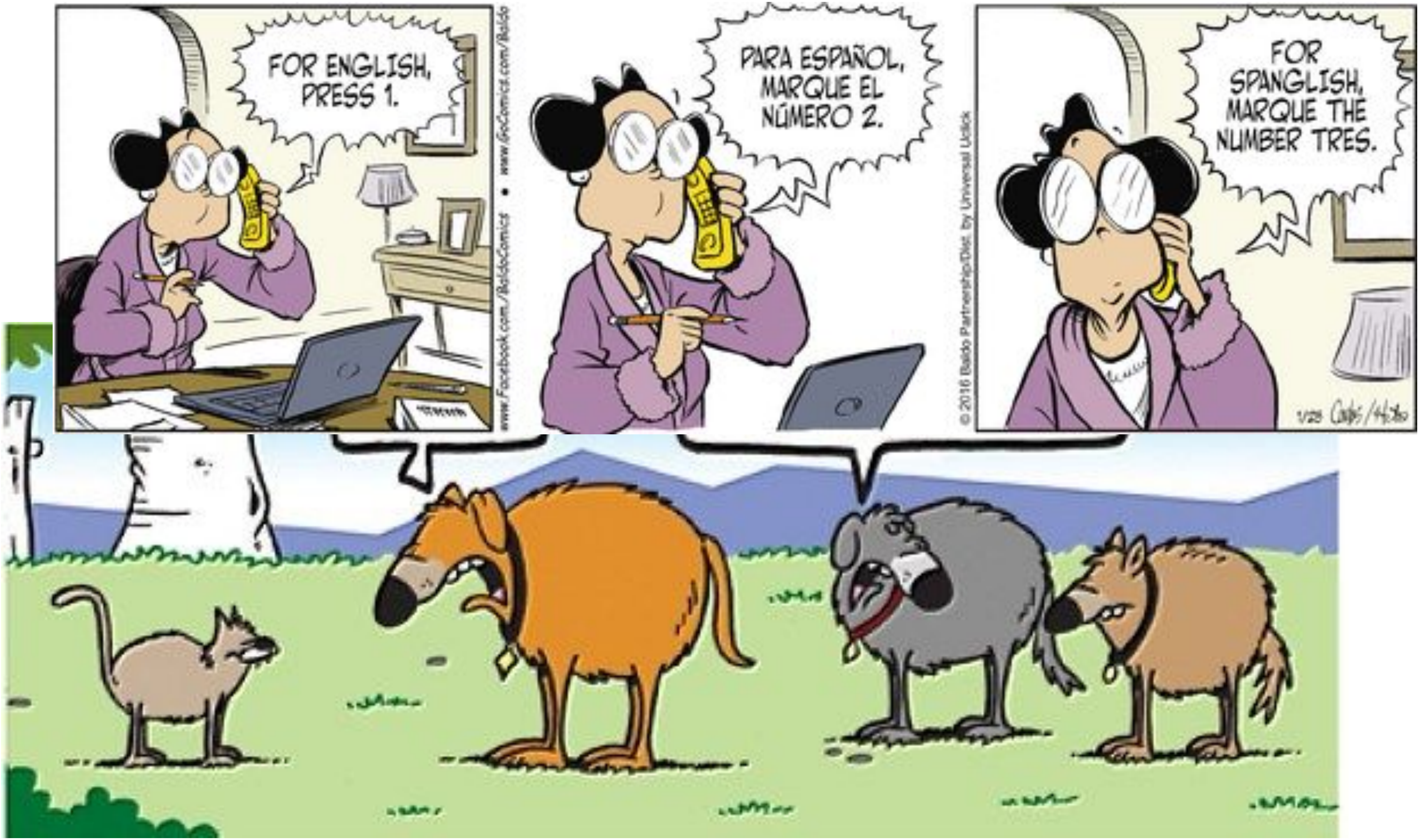
- Language Ideologies
- Code-switching Examples
  - Jamila Lyiscott’s Ted Talk *Three ways to speak English*

Killer Crónicas

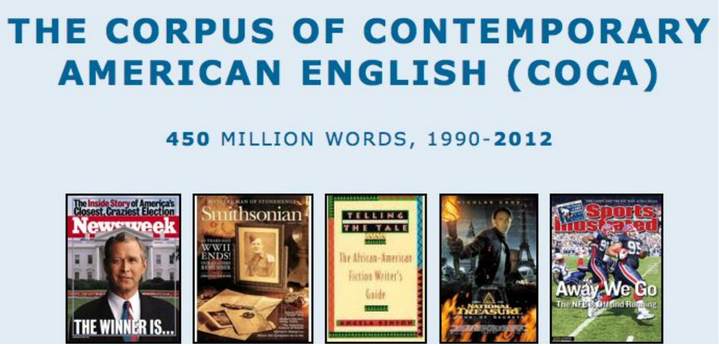
By Susana Chávez-Silverman



La última vez que volví a Guadalajara, in 1997, para celebrar mi ternura, as I call it (such a more gorgeous word for the thing than the real word for "tenure," la permanencia. No, I refuse the "real" translation. I prefer la ternura. Tenderness).—I know, ya sé, mamá. It's not the real word for it and I should speak right. La gente va a creer que I don't know right from wrong. Pero tú ya no estás para retarme, and I've always done my own thing anyway, que no?


FOR ENGLISH, PRESS 1.
PARA ESPAÑOL, MARQUE EL NÚMERO 2.
FOR SPANGLISH, MARQUE THE NUMBER TRES.
www.Facebook.com/BaldoComics • www.GoComics.com/Baldo
© 2016 Baldo Partnership/Dist. by Universal Uclick
1/23

Monolingual Data



THE CORPUS OF CONTEMPORARY AMERICAN ENGLISH (COCA)

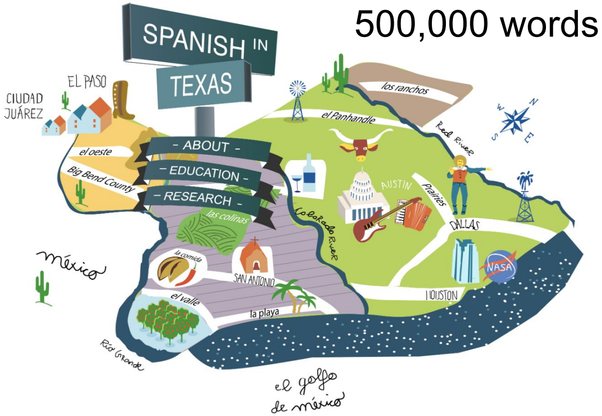
450 MILLION WORDS, 1990-2012

The Inside Story of America's Closest, Craziest Election
Newsweek
THE WINNER IS...
Smithsonian
WWII ENDS!
TELLING THE TALE
The African-American Fiction Writer's Guide
NATIONAL TREASURE
Sports Illustrated
Away We Go
The NFL Is Off and Running

Google books

Corpus	Size (words)
American	155 billion
British	34 billion
Spanish	45 billion

Code-Switched Data



500,000 words