

One Text / Two Languages

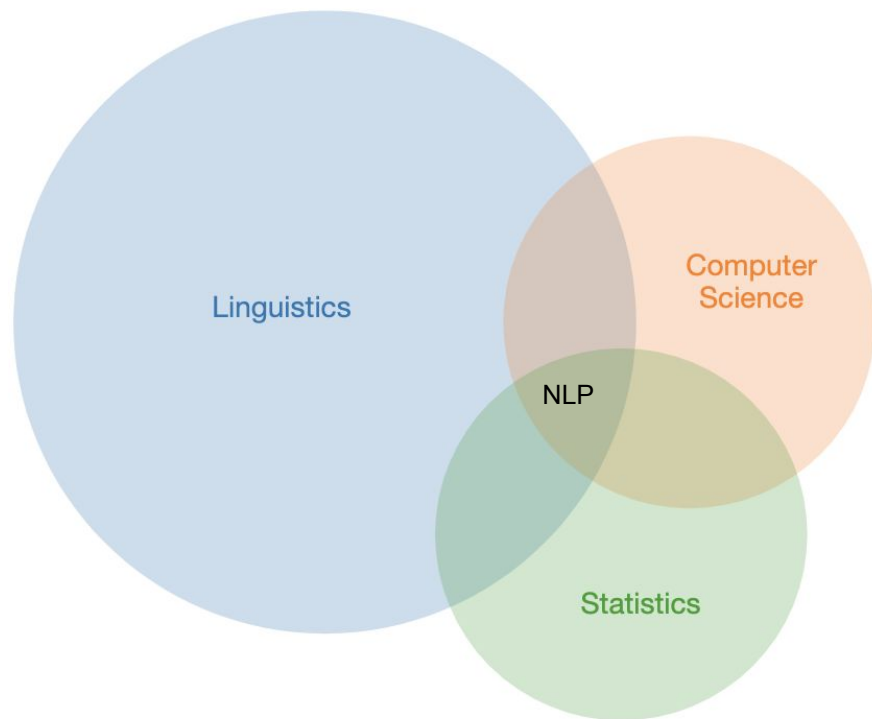
How linguists and programmers
approach code-switching

Dr. Jacqueline Serigos from George Mason
University

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

Intro

About me and this workshop



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 हैं 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

La última vez que volví a Guadalajara, in 1997, para celebrar mi ternura, as I call it ... –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?

Task 1 - What is code-switching?



For each text,

1. Underline all English tokens
2. Identify dominant language and embedded language
3. Classify words and phrase from the embedded language in terms of
 - Content words or grammatical words
 - Part of Speech: Nouns, verbs, adjectives, prepositions, articles
4. Define code-switching

**How do the two texts differ?

Text 1

Hindi - English

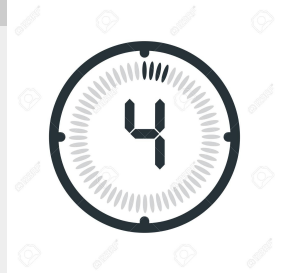
यहाँ पे मुझे कई बार थोड़ा अकेलापन लगते हैं क्योंकि दोस्त ही यहाँ family है, दोस्त ही परिवार है। जितने दोस्त हैं उन्ही के साथ आप कुछ समय बिता सकते हो और अपनी बातें share कर सकते हो, क्योंकि कोई immediate family या कोई direct family यहाँ तो है नहीं, तो मुझे ये लगता है कई बारी कि थोड़ा अकेलापन है, थोड़ा 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

La última vez que volví a Guadalajara, in 1997, para celebrar mi ternura, as I call it ... -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?

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

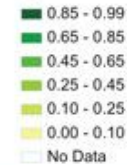
Why do people code-switch?

- More than half of the world's population is multilingual...
- So why not?

Linguistic Diversity Around the World

The index of linguistic diversity is a number ranging between zero (0) and one (1) and reflects the amount of linguistic diversity in each country. A linguistically diverse country is characterized by the presence of a number of linguistic groups. In a linguistically homogeneous country a great majority of population speak a single language. In the extreme case of diversity (index = 1), everybody speaks a different language; in the extreme case of homogeneity (index = 0), all people speak one language.

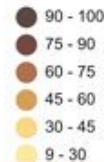
Index of Linguistic Diversity



Number of Spoken Languages in Each Country



Native Languages As Percent of All Spoken Languages



Countries with the largest number of spoken languages

Papua New Guinea	820
Indonesia	742
Nigeria	516
India	427
United States	311
Mexico	297
Cameroon	280
Australia	275
China	241

Countries with the highest index of linguistic diversity

Papua New Guinea	0.99
Vanuatu	0.97
Tanzania	0.96
Solomon Islands	0.96
Central African Republic	0.96
Chad	0.95
Dem. Rep. of Congo	0.95
Cameroon	0.94
India	0.93

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?



Hey Siri, play *Despacito*

Hey Siri, text my husband "Llego a casa en 10 minutos"

Turning text into data

Text 2

Spanish - English

La última vez que volví a
Guadalajara, in 1997, para
celebrar mi ternura, as I call it
... -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.

Chávez-Silverman, S. (2004). *Killer
crónicas: bilingual memories*. Univ of
Wisconsin Press.



Token	Lang.	POS	Named Entity	Speaker
a	Spanish	Prep	no	Susana
Guadalajara	Spanish	Noun	yes	Susana
,	NA	Punct	no	Susana
in	English	Prep	no	Susana
1997	NA	Num	no	Susana
para	Spanish	Prep	no	Susana

Natural Language Processing

Challenges of Code-switched Data

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

Monolingual Data

**Corpus of
Contemporary
American English**

450 million words

Google Books

American - 155 billion words

British - 34 billion words

Spanish - 45 billion words

Code-Switched Data

**Spanish in Texas
Corpus**

500,000 words

CESA

200,000
words

Set Up

- Download repo → <https://github.com/jserigos/One-Text-Two-Languages>
- Open jupyter notebook → [Step1-NLP Pipeline.ipynb](#)
- Install SpaCy → <https://spacy.io/usage/>
 - requires numpy >= 1.10. download will fail if older version of numpy

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

↑sudo

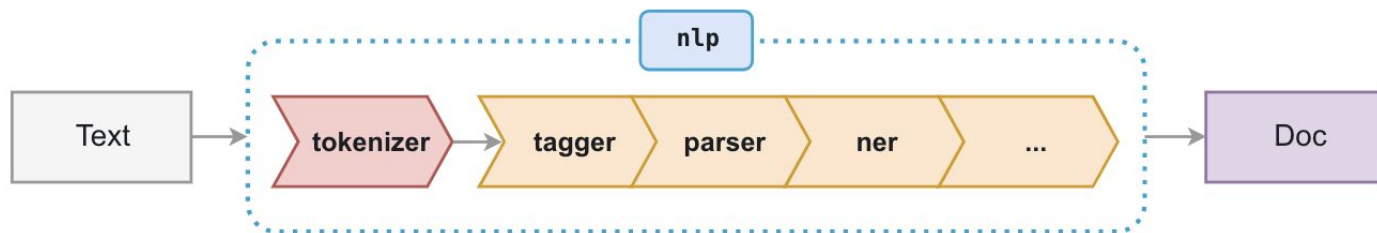
or

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

↑sudo

```
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)
```

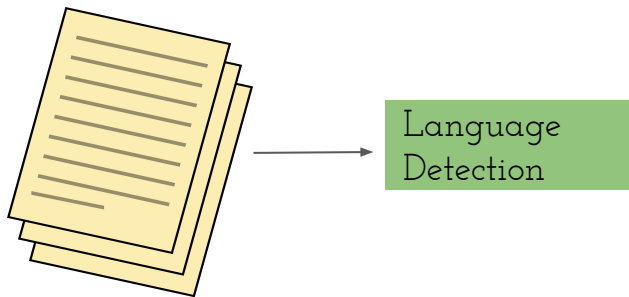
Spacy



```
Alice_text = '''Presently she began again. 'I wonder if I shall fall right THROUGH the earth! How funny it'll seem to c  
Alice_spacy = nlp_en(Alice_text)  
Alice_spacy[:100] # returns first 100 tokens  
len(Alice_spacy) # return the total number of tokens in the doc
```

```
Quixote_text = '''Con estas razones perdía el pobre caballero el juicio, y desvelábase por entenderlas y desentrañarles  
Quixote_spacy = nlp_sp(Quixote_text)  
Quixote_spacy[:100]
```

NLP Pipeline

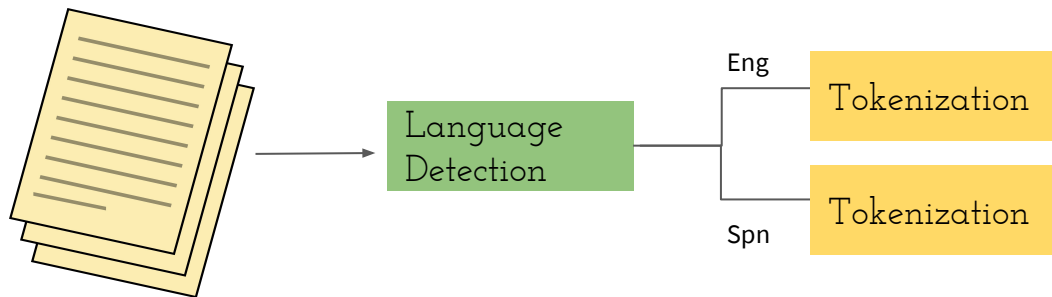


```
import spacy
from spacy_cld import LanguageDetector

nlp = spacy.load('en')
language_detector = LanguageDetector()
nlp.add_pipe(language_detector)
doc = nlp('This is some English text.')

doc._.languages # ['en']
doc._.language_scores['en'] # 0.96
```

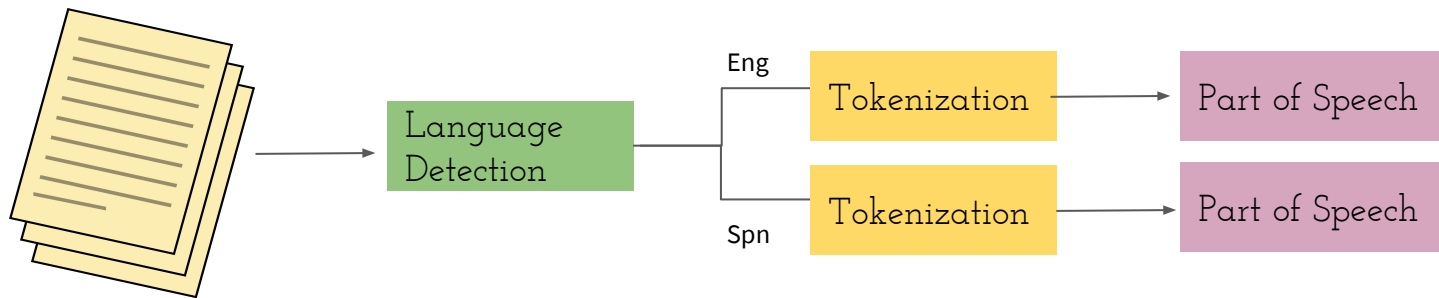
NLP Pipeline



```
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
```

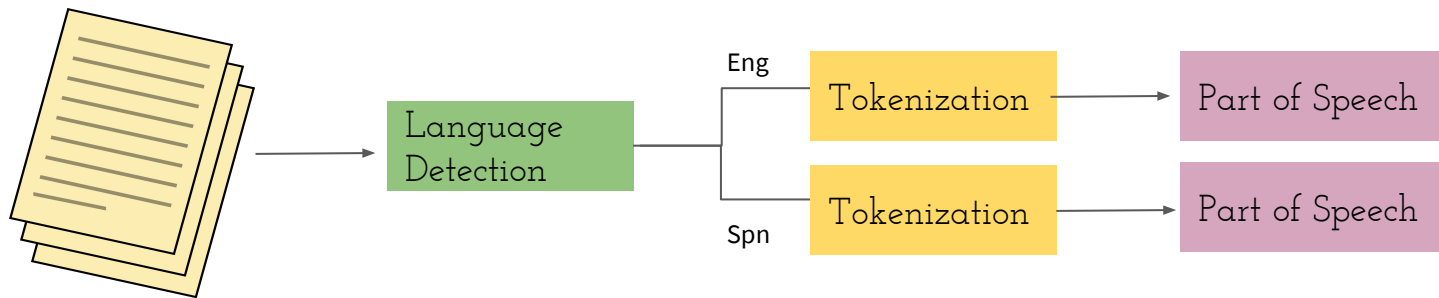
NLP Pipeline



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

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

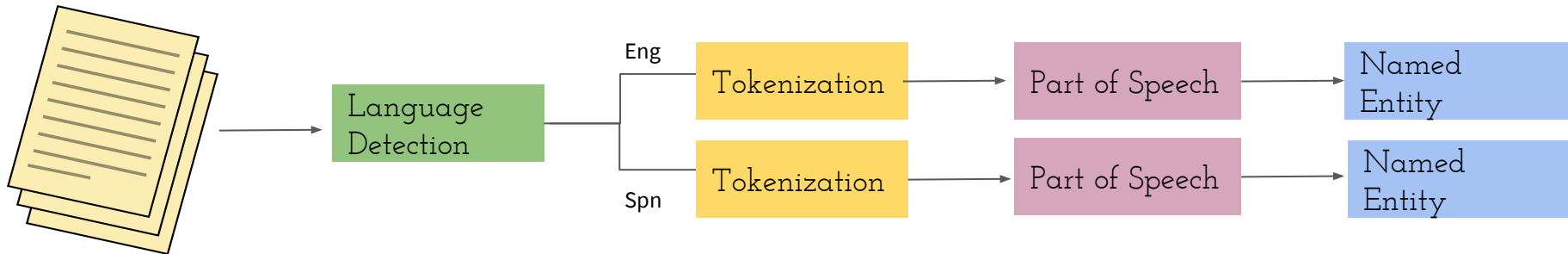

NLP Pipeline



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

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

NLP Pipeline



```
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_)
```

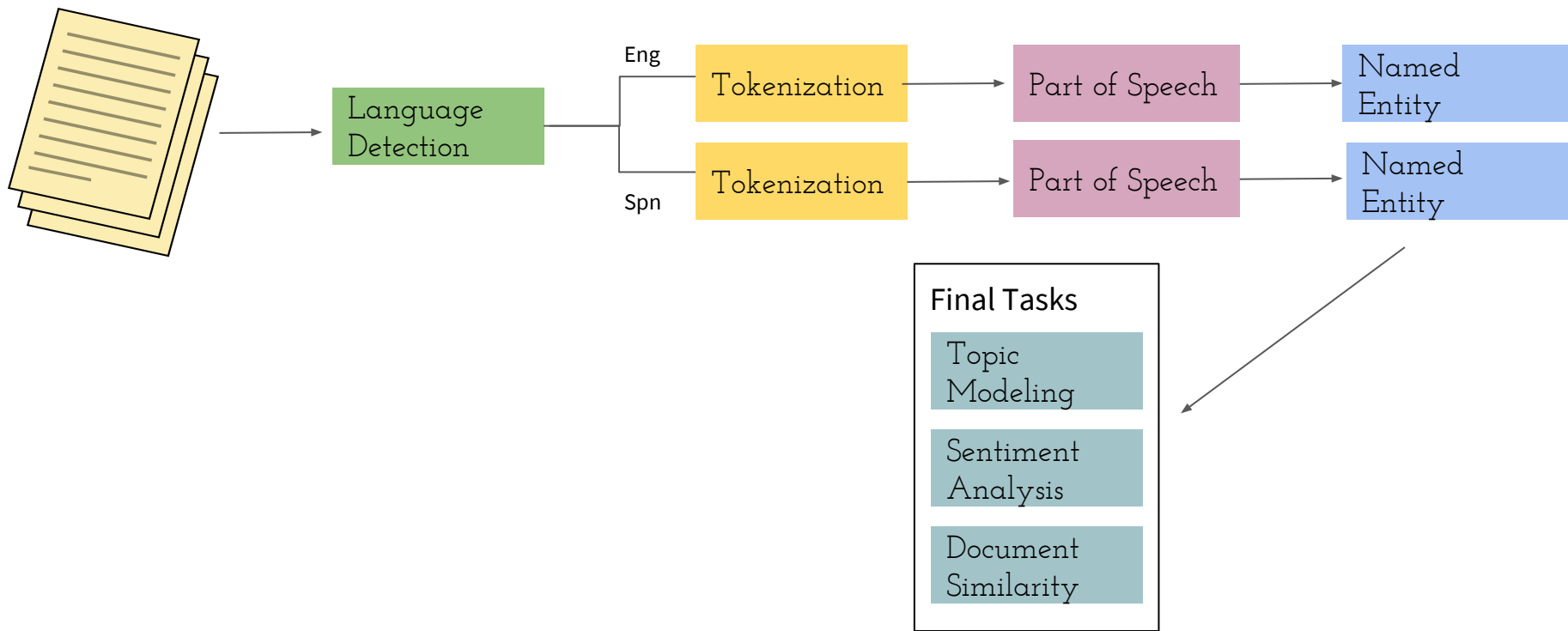
```
[(token, token.ent_iob_) for token in Alice_spacy]
```

```
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. Characters
2. Stop words
3. Character Ngrams

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 है। फुटबॉल को सामान्यतः 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.

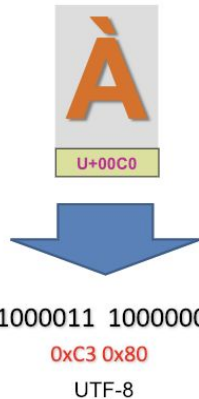
Text 4

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 다룰 수 있다.

Characters

- Character Encoding
 - UTF-8
 - ASCII - latin alphabet only
- Unicode
 - Industry standard
 - Contains a repertoire of 137,439 characters covering 146 modern and historic scripts
 - Each abstract character has a “code point”
 - Different scripts use unique range of code points
 - Code points U+0000 to U+007F (0-127) were the same as ASCII
 - <http://www.unicode.org/charts/>



<http://www.unicode.org/charts/>

1100

Hangul Jamo

11FF

	110	111	112	113	114	115	116	117	118	119	11A	11B	11C	11D	11E	11F
0	ㄱ 1100	ㅋ 1110	ㆁ 1120	ㄴ 1130	ㄷ 1140	ㄹ 1150	ㅁ 1160	ㅂ 1170	ㅅ 1180	ㅈ 1190	ㅊ 11A0	ㅌ 11B0	ㅍ 11C0	ㅊ 11D0	ㅊ 11E0	ㅇ 11F0
1	ㄱ 1101	ㅋ 1111	ㆁ 1121	ㄴ 1131	ㄷ 1141	ㄹ 1151	ㅁ 1161	ㅂ 1171	ㅅ 1181	ㅈ 1191	ㅊ 11A1	ㅌ 11B1	ㅍ 11C1	ㅊ 11D1	ㅊ 11E1	ㅇ 11F1
2	ㄱ 1102	ㅋ 1112	ㆁ 1122	ㄴ 1132	ㄷ 1142	ㄹ 1152	ㅁ 1162	ㅂ 1172	ㅅ 1182	ㅈ 1192	ㅊ 11A2	ㅌ 11B2	ㅍ 11C2	ㅊ 11D2	ㅊ 11E2	ㅇ 11F2

Stop Words

- Short function words
 - She, is, who, what, are, in, before...
- <https://www.mostlymaths.net/2012/06/language-detection-in-python-with-nltk.html>

Character N-grams

- What's the probability that a given word is English? Spanish?
 - *queche* ?
 - *thamin* ?

Spanish training data:

"Esa es la razón por **que** he
querido salir".

English training data:

"**There** was a long **queue**
waiting for **the** train."

Unknown sequences:

"**que**" -> Spanish

"**the**" -> English

Character Ngram Model

1. Construct ngram model from two monolingual corpora
2. Calculate ratio of probability of a word in each

$$\frac{p(\text{word} | \text{Spanish})}{p(\text{word} | \text{English})}$$

Character Ngram Model

3. Label each word with most probable language except if ratio is near 1.0 \Rightarrow treat as ambiguous



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

4. Ambiguous words are assigned the language tag of the previous word
 - a. Functional Head Constraint (Belazi et al. 1994)
5. Performance of Codeswitchador = 96.9%

¡Muchas gracias!