Borrowing or Codeswitching?
Annotating for Finer-Grained Distinctions in Language Mixing

Elena Álvarez-Mellado\textsuperscript{1}  Constantine Lignos\textsuperscript{2}

\textsuperscript{1}NLP & IR group, UNED
\textsuperscript{2}Michtom School of Computer Science, Brandeis University

LREC 2022
Table of Contents

1 What is codeswitching

2 What is lexical borrowing

3 The task
   - The dataset
   - Annotation guidelines

4 Modeling

5 Conclusions
Table of Contents

1. What is codeswitching
2. What is lexical borrowing
3. The task
   - The dataset
   - Annotation guidelines
4. Modeling
5. Conclusions
What is codeswitching?

Alternating between two or more languages in the same discourse. For ex., a speaker that is bilingual in English and Spanish may alternate between both languages in the same sentence: *Sometimes I’ll start a sentence in Spanish y termino en español* (Poplack, 1980)

- Codeswitches are fluent multiword interferences that combine more than one language
- Codeswitching frequently takes place in conversation among multilingual speakers
Codeswitching has become a frequent NLP task over the last decade (Aguilar et al., 2020, 2018; Molina et al., 2016; Solorio et al., 2014). The task has consisted in identifying the language of each token in codeswitched utterances (for example, in social media messages). Token-level annotation task, where every token receives a language identification tag (Maharjan et al., 2015, among others). For example, in a collection of English-Spanish codeswitched tweets, tokens in Spanish will be labeled with a language identification tag for Spanish, and tokens in English will be labeled with a language identification tag for English.
Codeswitching in NLP

Besides $\text{lang1}$ and $\text{lang2}$, additional labels have been proposed for use in codeswitching datasets:

- **ambiguous** for words whose language is difficult to determine even in context
- **other** for tokens in languages other than the main languages under study
- **mixed** for intralexical codeswitching (words that combine morphemes from different languages)
- **NE** for named entities
- **none** for punctuation marks, emoji, Twitter mentions, etc.

This repertoire of labels has become the usual tagset in codeswitching shared tasks (Aguilar et al., 2018; Molina et al., 2016; Solorio et al., 2014)
Codeswitching example

I ENG got ENG it ENG, N but ENG prefiero SPA usar SPA mi SPA Dell NE para SPA cosas SPA sencillas SPA. N

Ay SPA dios SPA, N I ENG' N m ENG tired ENG . N
Our research question

Given an utterance that is mostly monolingual, should we assume that any token from another language is a codeswitch, or could it be something else?

After all, language mixing can happen in cases that are not necessarily codeswitching.
What is lexical borrowing?

Lexical borrowing is the incorporation of words from one language into another language.
For ex., using in Spanish words that come from English: podcast, app, online, crowdfunding, spin-off, big data, fake news...

- Lexical borrowing is a type of linguistic borrowing.
  - Linguistic borrowing is the process of reproducing in one language the patterns of other languages Haugen (1950)

- Borrowing and code-switching are related and have frequently been described as a continuum Clyne et al. (2003)
## Lexical borrowing vs Code-switching

<table>
<thead>
<tr>
<th></th>
<th>Code-switching</th>
<th>Lexical Borrowing</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>Speaker</strong></td>
<td>bilinguals</td>
<td>monolinguals</td>
</tr>
<tr>
<td><strong>Grammar compliance</strong></td>
<td>both languages</td>
<td>recipient language</td>
</tr>
<tr>
<td><strong>Level of integration</strong></td>
<td>not integrated</td>
<td>can be integrated</td>
</tr>
</tbody>
</table>
Are these two examples of codeswitching?

✓ I got it, but prefiero usar mi Dell para cosas sencillas.¹
× Intentando comprar online uno de los nuevos discos duros que sacó Samsung, pero qué lata tener que rellenar tanto formulario.².

Most frequent codeswitches in an English-Spanish codeswitched Twitter dataset were social media abbreviations and well-established internet terms (such as lol) (Maharjan et al., 2015).

¹“I got it, but I’d rather use my Dell for simple things.”
²“Trying to buy one of the new hard disks released by Samsung online, but what a pain it is to have to fill in so many forms.”
If codeswitching is the phenomenon of interest, then having a collection of tweets that are rich in other-language inclusions is not sufficient.

As Poplack and Dion (2012) state, distinguishing codeswitching and borrowing is “the thorniest issue in the field of contact linguistics today.”
Table of Contents

1. What is codeswitching
2. What is lexical borrowing
3. The task
   - The dataset
   - Annotation guidelines
4. Modeling
5. Conclusions
The task

- The difference between codeswitching and borrowing has been explored and discussed in the Linguistics literature (Poplack and Dion, 2012)
- This distinction has not been implemented in NLP codeswitching datasets
  - In the prior work, we were not able to identify explicit, published definitions or guidelines on what should constitute a codeswitch—or perhaps more crucially, what is not a codeswitch—and what should constitute a borrowing.
- Our goal with this task was to:
  - create an annotated dataset that implements the codeswitching/borrowing distinction
  - develop annotation guidelines that assist annotators distinguish between true codeswitches and lexical borrowings
  - explore the performance of Transformer-based models for the task of predicting labels in a rich language-mix setting
The dataset was:

- An existing corpus already annotated for codeswitching
- We selected the codeswitching-dense corpus by Lignos and Marcus (2013)
  - 9,500 tweets
  - 198,706 tokens
  - Primarily a Spanish dataset, mainly composed of Spanish tweets that may have an English inclusion (a codeswitch or a borrowing)
  - Annotated for codeswitching (SPA/ENG/OTH/NE/N)
- We reannotated it implementing the borrowing/codeswitching distinction
Labels

Our reannotation considered the following labels:

- **SPA**: for tokens in Spanish
- **ENG**: for tokens in English
- **OTH**: for tokens in languages other than ENG or SPA
- **BOR**: for recent borrowings (in English or other languages)
- **ENT**: for named entities
- **N**: for punctuation marks, Twitter symbols (such as hashtags and mentions), URLs, etc.
Annotation guidelines

1 English words related to Twitter terminology: *tweet*, *follower*.
Annotation guidelines

1. English words related to Twitter terminology: tweet, follower.
2. Technology words: server, hosting, user, post, blog, app, online, chat.
Annotation guidelines

1. English words related to Twitter terminology: *tweet, follower*.
2. Technology words: *server, hosting, user, post, blog, app, online, chat*.
3. English words registered in *Diccionario de la lengua española* (RAE, 2021), the general dictionary of standard Spanish: *look, marketing*.

Álvarez-Mellado E., Lignos C.  
Borrowing or Codeswitching?  
LREC 2022 18 / 27
Annotation guidelines

1. English words related to Twitter terminology: tweet, follower.
2. Technology words: server, hosting, user, post, blog, app, online, chat.
3. English words registered in Diccionario de la lengua española (RAE, 2021), the general dictionary of standard Spanish: look, marketing.
4. English words that are already registered in Diccionario de Americanismos (ASALE, 2010), a specialized dictionary that covers the vocabulary spoken in American Spanish and that has a rich representation of well-established lexical borrowings from English used in Latin America: man nice, party.
Annotation guidelines

1. English words related to Twitter terminology: tweet, follower.
2. Technology words: server, hosting, user, post, blog, app, online, chat.
3. English words registered in *Diccionario de la lengua española* (RAE, 2021), the general dictionary of standard Spanish: look, marketing.
4. English words that are already registered in *Diccionario de Americanismos* (ASALE, 2010), a specialized dictionary that covers the vocabulary spoken in American Spanish and that has a rich representation of well-established lexical borrowings from English used in Latin America: man nice, party.
5. English words that are the headword of an entry in Spanish Wikipedia: hip hop, whisky.
Annotation guidelines

1. English words related to Twitter terminology: tweet, follower.
2. Technology words: server, hosting, user, post, blog, app, online, chat.
3. English words registered in Diccionario de la lengua española (RAE, 2021), the general dictionary of standard Spanish: look, marketing.
4. English words that are already registered in Diccionario de Americanismos (ASALE, 2010), a specialized dictionary that covers the vocabulary spoken in American Spanish and that has a rich representation of well-established lexical borrowings from English used in Latin America: man nice, party.
5. English words that are the headword of an entry in Spanish Wikipedia: hip hop, whisky.
6. Words that have English origin but are used following Spanish grammatical structure, such as noun-adjective word order: mensajes offline, rating online.
Token counts by label

<table>
<thead>
<tr>
<th>Tag</th>
<th>Tokens</th>
</tr>
</thead>
<tbody>
<tr>
<td>SPA</td>
<td>134,110</td>
</tr>
<tr>
<td>N</td>
<td>39,280</td>
</tr>
<tr>
<td>ENT</td>
<td>15,373</td>
</tr>
<tr>
<td>ENG</td>
<td>6,819</td>
</tr>
<tr>
<td>BOR</td>
<td>2,857</td>
</tr>
<tr>
<td>OTH</td>
<td>267</td>
</tr>
<tr>
<td>Total</td>
<td>198,706</td>
</tr>
</tbody>
</table>
### 10 most frequent tokens per label

<table>
<thead>
<tr>
<th>Spanish</th>
<th>Count</th>
<th>English</th>
<th>Count</th>
<th>Borrowing</th>
<th>Count</th>
</tr>
</thead>
<tbody>
<tr>
<td>de</td>
<td>6,175</td>
<td>the</td>
<td>119</td>
<td>blog</td>
<td>231</td>
</tr>
<tr>
<td>que</td>
<td>4,298</td>
<td>I</td>
<td>105</td>
<td>post</td>
<td>130</td>
</tr>
<tr>
<td>y</td>
<td>3,281</td>
<td>you</td>
<td>84</td>
<td>web</td>
<td>107</td>
</tr>
<tr>
<td>el</td>
<td>3,257</td>
<td>to</td>
<td>81</td>
<td>internet</td>
<td>106</td>
</tr>
<tr>
<td>a</td>
<td>3,178</td>
<td>my</td>
<td>69</td>
<td>followers</td>
<td>55</td>
</tr>
<tr>
<td>la</td>
<td>3,175</td>
<td>a</td>
<td>69</td>
<td>online</td>
<td>41</td>
</tr>
<tr>
<td>en</td>
<td>3,120</td>
<td>it</td>
<td>62</td>
<td>blogs</td>
<td>41</td>
</tr>
<tr>
<td>no</td>
<td>2,410</td>
<td>is</td>
<td>60</td>
<td>software</td>
<td>33</td>
</tr>
<tr>
<td>me</td>
<td>2,086</td>
<td>in</td>
<td>56</td>
<td>Internet</td>
<td>33</td>
</tr>
<tr>
<td>es</td>
<td>1,764</td>
<td>on</td>
<td>49</td>
<td>timeline</td>
<td>32</td>
</tr>
</tbody>
</table>

Compare these English tokens with the most frequent codeswitched tokens in prior codeswitching datasets, such as *lol*, *lmao* or *idk* (Maharjan et al., 2015).
Modeling

We evaluated four Transformer-based models on the dataset:

1. mBERT: multilingual BERT, trained on Wikipedia in 104 languages (Devlin et al., 2019)
Modeling

We evaluated four Transformer-based models on the dataset:

1. **mBERT**: multilingual BERT, trained on Wikipedia in 104 languages (Devlin et al., 2019)

2. **BETO**: a BERT-based model trained on a diverse set of international Spanish texts from different origins: OpenSubtitles, Global Voices, the United Nations (3 billion tokens) (Cañete, 2019; Cañete et al., 2020)

3. **RoBERTa BNE**: a RoBERTa-based model trained exclusively on data crawled from .es websites—those using the top-level domain for Spain—by the National Library of Spain (135 billion tokens) (Gutiérrez-Fandiño et al., 2021)

4. **RoBERTa Twitter**: a RoBERTa based model trained on English Twitter data (Barbieri et al., 2020)
Modeling

We evaluated four Transformer-based models on the dataset:

1. mBERT: multilingual BERT, trained on Wikipedia in 104 languages (Devlin et al., 2019)

2. BETO: a BERT-based model trained on a diverse set of international Spanish texts from different origins: OpenSubtitles, Global Voices, the United Nations (3 billion tokens) (Cañete, 2019; Cañete et al., 2020)

3. RoBERTa BNE: a RoBERTa-based model trained exclusively on data crawled from .es websites—those using the top-level domain for Spain—by the National Library of Spain (135 billion tokens) (Gutiérrez-Fandiño et al., 2021)

4. RoBERTa Twitter: a RoBERTa based model trained on English Twitter data (Barbieri et al., 2020)
Modeling

We evaluated four Transformer-based models on the dataset:

1. **mBERT**: multilingual BERT, trained on Wikipedia in 104 languages (Devlin et al., 2019)

2. **BETO**: a BERT-based model trained on a diverse set of international Spanish texts from different origins: OpenSubtitles, Global Voices, the United Nations (3 billion tokens) (Cañete, 2019; Cañete et al., 2020)

3. **RoBERTa BNE**: a RoBERTa-based model trained exclusively on data crawled from .es websites—those using the top-level domain for Spain—by the National Library of Spain (135 billion tokens) (Gutiérrez-Fandiño et al., 2021)

4. **RoBERTa Twitter**: a RoBERTa based model trained on English Twitter data (Barbieri et al., 2020)
Modeling results

<table>
<thead>
<tr>
<th>Model</th>
<th>Accuracy</th>
<th>Precision</th>
<th>Recall</th>
<th>F1</th>
</tr>
</thead>
<tbody>
<tr>
<td>mBERT</td>
<td>96.88</td>
<td>96.69</td>
<td>96.61</td>
<td>96.65</td>
</tr>
<tr>
<td>BETO</td>
<td>96.91</td>
<td>96.69</td>
<td>96.60</td>
<td>96.64</td>
</tr>
<tr>
<td>RoBERTa-BNE</td>
<td>93.73</td>
<td>93.19</td>
<td>93.23</td>
<td>93.21</td>
</tr>
<tr>
<td>RoBERTa Twitter</td>
<td>93.39</td>
<td>92.82</td>
<td>92.86</td>
<td>92.84</td>
</tr>
</tbody>
</table>

Table: Accuracy and micro-averaged precision, recall, and F1 score for baseline models (results from a single run)
# Table of Contents

1. What is codeswitching

2. What is lexical borrowing

3. The task
   - The dataset
   - Annotation guidelines

4. Modeling

5. Conclusions
Conclusions

- We have introduced a new dataset of tweets annotated both with lexical borrowings and Spanish-English codeswitches.
- The annotation builds on previous approaches to codeswitching dataset creation, but distinguishes lexical borrowing from true codeswitching.
- This distinction has been previously pointed out as crucial in the contact linguistics literature, but has not been made in previous codeswitching datasets.
- We have experimented with different Transformer-based models for the task of language identification and compared results in our dataset to previous work on other codeswitching datasets.
References


