Borrowing or Codeswitching?

Annotating for Finer-Grained Distinctions in Language Mixing

Elena Álvarez-Mellado[1] Constantine Lignos[2]

[1]NLP & IR group, UNED

[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 in NLP

- 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 ex. 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 `lang1` and `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.

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 lexical borrowing?

Lexical borrowing is the incorporation of words form 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

	Code-switching	**Lexical Borrowing**
Speaker	bilinguals	monolinguals
Grammar compliance	both languages	recipient language
Level of integration	not integrated	can be integrated

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."

Borrowing or codeswitching?

- 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

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*.

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

Tag	Tokens
SPA	134,110
N	39,280
ENT	15,373
ENG	6,819
BOR	2,857
OTH	267
Total	198,706

10 most frequent tokens per label

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

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

Table of Contents

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

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)

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)

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

Model	Accuracy	Precision	Recall	F1
mBERT	96.88	**96.69**	**96.61**	**96.65**
BETO	**96.91**	**96.69**	96.60	96.64
RoBERTa-BNE	93.73	93.19	93.23	93.21
RoBERTa Twitter	93.39	92.82	92.86	92.84

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

Aguilar, G., AlGhamdi, F., Soto, V., Diab, M., Hirschberg, J., and Solorio, T. (2018). Named entity recognition on code-switched data: Overview of the CALCS 2018 shared task. In Proceedings of the Third Workshop on Computational Approaches to Linguistic Code-Switching, pages 138–147, Melbourne, Australia. Association for Computational Linguistics.

Aguilar, G., Kar, S., and Solorio, T. (2020). LinCE: A centralized benchmark for linguistic code-switching evaluation. In Proceedings of the 12th Language Resources and Evaluation Conference, pages 1803–1813, Marseille, France. European Language Resources Association.

ASALE (2010). Diccionario de americanismos. `https://lema.rae.es/damer/`.

Barbieri, F., Camacho-Collados, J., Espinosa Anke, L., and Neves, L. (2020). TweetEval: Unified benchmark and comparative evaluation for tweet classification. In Findings of the Association for Computational Linguistics: EMNLP 2020, pages 1644–1650, Online. Association for Computational Linguistics.

Cañete, J. (2019). Compilation of large spanish unannotated corpora.

Cañete, J., Chaperon, G., Fuentes, R., Ho, J.-H., Kang, H., and Pérez, J. (2020). Spanish pre-trained BERT model and evaluation data. In PML4DC at ICLR 2020.

Clyne, M., Clyne, M. G., and Michael, C. (2003). Dynamics of language contact: English and immigrant languages. Cambridge University Press.

Devlin, J., Chang, M.-W., Lee, K., and Toutanova, K. (2019). BERT: Pre-training of deep bidirectional transformers for language understanding. In Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers), pages 4171–4186, Minneapolis, Minnesota. Association for Computational Linguistics.

Gutiérrez-Fandiño, A., Armengol-Estapé, J., Pàmies, M., Llop-Palao, J., Silveira-Ocampo, J., Carrino, C. P., Gonzalez-Agirre, A., Armentano-Oller, C., Rodriguez-Penagos, C., and Villegas, M. (2021). Spanish language models.

Haugen, E. (1950). The analysis of linguistic borrowing. Language, 26(2):210–231.

Lignos, C. and Marcus, M. (2013). Toward web-scale analysis of codeswitching. In 87th Annual Meeting of the Linguistic Society of America, volume 90.

Maharjan, S., Blair, E., Bethard, S., and Solorio, T. (2015). Developing language-tagged corpora for code-switching tweets. In Proceedings of The 9th Linguistic Annotation Workshop, pages 72–84, Denver, Colorado, USA. Association for Computational Linguistics.

References (cont.)

Molina, G., AlGhamdi, F., Ghoneim, M., Hawwari, A., Rey-Villamizar, N., Diab, M., and Solorio, T. (2016). Overview for the second shared task on language identification in code-switched data. In Proceedings of the Second Workshop on Computational Approaches to Code Switching, pages 40–49, Austin, Texas. Association for Computational Linguistics.

Poplack, S. (1980). Sometimes i'll start a sentence in spanish y termino en espanol: Toward a typology of code-switching. Linguistics. An Interdisciplinary Journal of the Language Sciences La Haye, 18(7-8):581–618.

Poplack, S. and Dion, N. (2012). Myths and facts about loanword development. Language variation and change, 24(3):279–315.

RAE (2021). Diccionario de la lengua española, ed. 23.5. `http://dle.rae.es`.

Solorio, T., Blair, E., Maharjan, S., Bethard, S., Diab, M., Ghoneim, M., Hawwari, A., AlGhamdi, F., Hirschberg, J., Chang, A., and Fung, P. (2014). Overview for the first shared task on language identification in code-switched data. In Proceedings of the First Workshop on Computational Approaches to Code Switching, pages 62–72, Doha, Qatar. Association for Computational Linguistics.