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

Elena Álvarez Mellado[1], Constantine Lignos[2]
[1]NLP & IR Group, School of Computer Science, UNED
[2]Michtom School of Computer Science, Brandeis University
elena.alvarez@lsi.uned.es, lignos@brandeis.edu

Abstract

We present a new corpus of Twitter data annotated for codeswitching and borrowing between Spanish and English. The corpus contains 9,500 tweets annotated at the token level with codeswitches, borrowings, and named entities. This corpus differs from prior corpora of codeswitching in that we attempt to clearly define and annotate the boundary between codeswitching and borrowing and do not treat common "internet-speak" (*lol*, etc.) as a codeswitch when used in an otherwise monolingual context. The result is a corpus that enables the study and modeling of Spanish-English borrowing and codeswitching on Twitter in one dataset. We present baseline scores for modeling the labels of this corpus using Transformer-based language models. The annotation itself is released with a CC BY 4.0 license, while the text it applies to is distributed in compliance with the Twitter terms of service.

Keywords: codeswitching, borrowing, corpora

1. Challenges in Annotating Codeswitching and Borrowing

Over the last decade, several initiatives have drawn attention to codeswitching as a challenging NLP task, such as various shared tasks (Aguilar et al., 2018; Molina et al., 2016; Solorio et al., 2014) and a recent benchmark (Aguilar et al., 2020).

Prior work on creating datasets of codeswitching has framed the task of language identification for codeswitched utterances as token-level annotation, 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.

Further labels have been proposed and adopted to the repertoire of tags in codeswitching datasets, such as *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, and *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). However, this tagset does not take into account that not all other-language inclusions are necessarily a codeswitch. In the prior work we examined, 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. After all, language mixing can happen in cases that are not necessarily codeswitching. 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?

Lexical borrowing—the incorporation of single lexical units from one language into another language (Haugen, 1950)—is a process that resembles codeswitching in that it conveys the inclusion of tokens from another language into otherwise monolingual contexts, but it is different in nature (Poplack and Dion, 2012). Codeswitches are usually produced by multilingual individuals, and thus they comply with the grammatical structure of both languages *at the same time*. By definition, codeswitches are not integrated into the recipient language, unlike established borrowings.

On the other hand, lexical borrowings can be produced by monolingual speakers, and they are often accompanied by morphological and phonological modification. Lexical borrowings comply with the grammatical patterns of the recipient language and can eventually become assimilated into it until the perception of being an inclusion of "foreign" origin disappears (Lipski, 2005). With this in mind, it seems unlikely that in a sentence such as *Intentando comprar* online *uno de los nuevos discos duros que sacó Samsung, pero qué lata tener que rellenar tanto formulario*,[1] the word *online* could be considered a codeswitch. Our analysis of existing codeswitching corpora suggests that tokens like *online* have usually been annotated as codeswitches, although in the absence of formal annotation guidelines, it is unclear whether that was intentional or not. On the other

---

[1]"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."

hand, a sentence such as *I got it, but prefiero usar mi Dell para cosas sencillas* is clearly a codeswitched utterance.[2]

Given that lexical borrowing and codeswitching are two distinct linguistic phenomena and the prevalence of lexical borrowings in social media messages (Luján García, 2017; Sanou, 2018; Stewart et al., 2021), we believe that specific guidelines should be provided on how to deal with the codeswitching-borrowing distinction when annotating a dataset for codeswitching. If codeswitching is the phenomenon of interest, then only having a collection of tweets that are rich in other-language inclusions is not a sufficient solution. As Poplack and Dion (2012) state, distinguishing codeswitching and borrowing is "the thorniest issue in the field of contact linguistics today."

However, previous work on codeswitching dataset design has rarely attempted to establish an explicit difference between intrasentential codeswitching—codeswitching that happens inside one sentence—and lexical borrowing. Previous work in Hindi by Bali et al. (2014) proposed a statistical approach based on frequency and linguistic typology to capture the distinction between borrowing and mixing, while Patro et al. (2017) proposed different metrics to compute the likelihood of a word being borrowed being codeswitched. While many codeswitching datasets include Spanish data, the codeswitching/borrowing distinction has not been explicitly explored, and to the best of our knowledge, no annotated dataset has made this distinction.

In fact, a glance at the most frequent codeswitches in an English-Spanish codeswitched Twitter datasets (Maharjan et al., 2015) reveals that social media abbreviations and well-established internet terms (such as *lol*) are among the most frequent codeswitches annotated. It is debatable whether these words are in fact true codeswitches, as they can be considered examples of the internet dialect that monolinguals also use rather than evidence of bilinguals switching between languages, as Maharjan et al. (2015) point out:

> In the case of abbreviations, some of them such as "lol" and "lmao" have become social media lingo rather than abbreviations of English words and thus cross language barriers.

2. Dataset Description

With the previously mentioned challenges in mind, we developed a dataset consisting of tweets annotated to reflect codeswitching between Spanish and English as well as borrowing of English into Spanish. This section describes our annotation guidelines and the process of creating this dataset. In Section 2.1, we propose a set of guidelines to distinguish between codeswitching and borrowing that can be followed when annotating a corpus of English-Spanish codeswitched tweets. In Section 2.2, we describe how we reannotated an existing

codeswitching dataset to match our guidelines. Section 2.3 presents the counts of the resulting dataset and the most frequent tokens per label in the annotation. The stand-off annotation is released under a Creative Commons CC BY 4.0 license[3].

2.1. Annotation guidelines

Our annotation uses 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.

In general, tokens in Spanish were annotated as SPA, tokens in English were annotated as ENG. Internet acronyms (such as *lol*) were considered part of the internet jargon and annotated as Spanish whenever they occurred in monolingual Spanish contexts. Non-Spanish words that appear in contexts where no other codeswitching was happening and where there was a strong indication that the word was being borrowed (rather than codeswitched) was annotated as BOR.

Our goal with the borrowing tag was to address relatively recent borrowings into the Spanish language that might be spoken aloud in a normal conversational context, unlike the internet jargon referenced above.

The following were identified as borrowings, differentiating them from short codeswitches embedded in an otherwise monolingual segment:

1. English words related to Twitter terminology: such as *Follow Friday*, *tweet*, *follower*, etc.

2. Technology words: *server*, *hosting*, *user*, *post*, *blog*, *internet*, *template*, *app*, *online*, *chat*.

3. English words that are already registered in *Diccionario de la lengua española*[4] by Real Academia Española (Real Academia Española, 2021), the general dictionary of standard Spanish compiled by the Royal Spanish Academy, the main prescriptive institution on Spanish language.

4. English words that are already registered in *Diccionario de Americanismos*[5] by Asociación de Academias de la Lengua Española (Asociación de

---

[2]"I got it, but I'd rather use my Dell for simple things."

Academias de la Lengua Española, 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.

5. English words that are the headword of an entry in Spanish Wikipedia, such as music styles, genres and other cultural things (*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*).

The intuition behind guidelines #3–#5 was that if an English word either appears in some of the main prescriptive dictionaries of Spanish (which tend to be conservative in their approach to language change) or as the headword of a Wikipedia entry, then that would demonstrate that that English word is already well established among monolinguals at least in part of the Spanish-speaking world.

The criteria behind guideline #6 is that codeswitches by definition normally comply with the grammatical restrictions of both languages involved. On the other hand, when an item is borrowed from another language—as opposed to codeswitching into the other language—the borrowing will be produced with all the morphosyntactic requisites required by the recipient-language (Poplack and Dion, 2012). Consequently, when we encountered an English word that is used in an otherwise Spanish context and follows all the grammatical expectations of Spanish grammar rules, we annotated it as a borrowing, not a codeswitch.

Our guidelines build on prior work developed to assist annotators when identifying English borrowings in Spanish monolingual texts (Álvarez-Mellado, 2020a; Álvarez-Mellado, 2020b; Álvarez-Mellado and Lignos, 2022). That prior work on borrowing annotation was designed exclusively to identify borrowings in Spanish from Spain and used sources whose Spain-centric criteria has been previously pointed out (Blanch, 1995; Fernández Gordillo, 2014), which made them insufficient to identify well-established borrowings in a more geographically diverse collection of texts as our dataset. Consequently, our guidelines were adapted so that they could account for a more diverse representation of Spanish dialectal varieties than previous guidelines by considering additional lexicographic sources, such as Wikipedia or Diccionario de Americanismos. Our guidelines ensured that English words that are well-established borrowings in many parts of the Spanish speaking Americas (such as *man*,[6] *nice*[7] or *party*[8]) would not be mistaken for codeswitches.

We have tried to create guidelines for distinguishing borrowing and codeswitching that can be applied somewhat objectively and in the most reliable fashion possible. As we have stated previously, the distinction is challenging; no annotation guidelines will be perfect, and there is not one universal set of distinctions that all researchers or speakers of the languages involved will agree on. However, we believe these guidelines can at the very least produce reliable and replicable annotation, and that is an important step forward in this area.

2.2. Data Selection

To reduce the effort required to create our corpus and place it within existing work on codeswitching, we decided to select an existing corpus already annotated for codeswitching which we could reannotate to match our guidelines. We selected the codeswitching-dense corpus created by Lignos and Marcus (2013, hereafter LM2013) for several reasons.

First, all tweets in the original LM2013 corpus had been annotated by two annotators and the original decisions made by each annotator were readily available, allowing us to identify and revisit any disagreements between the two annotators.

Second, the LM2013 annotators had been instructed to annotate the data to attempt to differentiate borrowing and codeswitching. While the annotation does not have a borrowing tag, the annotators were instructed not to annotate borrowed words as codeswitches. In the LM2013 annotation process, this was referred to as the "monolingual grandparent" rule: if it is unclear whether a word is a borrowing or codeswitch, if one's monolingual grandparent would recognize it as part of their language, do not annotate it as a codeswitch. While this was intended to draw upon the life experiences of many bilinguals in that they are able to identify which words are understood by monolingual speakers, because it referred to an older relative it may have unintentionally had the effect of causing the annotators to annotate more recent borrowings as codeswitches.[9]

Third, the data used by LM2013 is primarily Spanish,[10] which gives more opportunities to annotate borrowings into Spanish, especially those from English which is responsible for a large number of emergent borrowings (Furiassi et al., 2012; Görlach, 2002).

The LM2013 dataset was reannotated following the guidelines from section 2.1 by a native speaker of Spanish that had a background in linguistics and had previ-

ous experience in linguistic annotation. Our reannotation differs from the original annotation of LM2013 in several ways. We completed adjudication of all 9,500 tweets annotated by LM2013. Due to limited adjudicator availability, they had only completed adjudication of 7,018 tweets, with the remainder having unresolved disagreements. The LM2013 dataset assigned language tags to named entities—names were annotated Spanish or English based on the surrounding language, not the source language of the name—but we found that this annotation was extremely inconsistent and we removed it, making our data more consistent with other codeswitching datasets that do not make this distinction.

LM2013's original annotation also distinguished between two types of entities: works of art (e.g., book titles) and other named entities. We removed this distinction as it was inconsistently annotated and most subsequent annotation either uses standard named entity types (PER, LOC, ORG, etc.) or a general "named entity" tag (Maharjan et al., 2015) like our annotation.

Finally, we explicitly annotated borrowings as such. The original LM2013 annotation did not do so and the LM2013 annotators were instructed to not treat borrowings as codeswitches, but the annotators followed this advice inconsistently. While the annotators were not consistent, the fact that they were instructed to make this distinction was useful, as it resulted in inconsistencies that we could easily identify in adjudication.

As discussed in Section 1, a major difference between our annotation and previous codeswitching corpora that we are aware of (including the original LM2013 annotation) is that we do not count "internet-speak" as a codeswitch into English, reflecting the fact that monolingual communities may use internet shorthand adapted from English even in purely monolingual non-English contexts.

In order to identify borrowing candidates that could fall under the definition established in Section 2.1, we manually looked for frequently used English borrowings, checked for inter-annotator disagreements (for example *Me mandas tu _mail_ y lo envío*,[11]) and automatically flagged inconsistencies in how certain tokens were sometimes labeled as SPA and sometimes annotated as ENG (such as *man*, *server* or *tweet*) across the whole corpus, even if annotators agreed on the individual annotations. This approach proved to be successful and revealed a high number of tokens that were actually borrowings and not true codeswitches (see Section 2.3).

The result of our reannotation process is a corpus that makes the distinction between borrowing and codeswitching and has a narrower definition of codeswitching than previous work, allowing for investigation of a finer-grained linguistic phenomenon.

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

Table 1: Token counts by label

2.3. Data counts

The corpus consists of 9,500 tweets that were annotated following the guidelines established in Section 2.1. Out of those 9,500 tweets, 2,017 contained at least one borrowing, 2,495 contained at least one codeswitch, and 403 contained both a borrowing and a codeswitch. In total there are 198,706 tokens in the corpus, and counts for each label are given in Table 1. The largest sources of tokens labeled OTH belong to tweets with codeswitches into languages that are neither Spanish nor English, typically Portuguese and Catalan.

Table 2 summarizes the most frequent tokens per label. As there are essentially zero purely-English tweets in this dataset, the counts of English tokens reflect the count of tokens of codeswitches into English. The most frequent tokens show a noticeable difference with the most frequent tokens from some other codeswitched data counts. While Maharjan et al. (2015) reported internet terms and abbreviations such as *lol*, *lmao* or *idk* among the most frequent codeswitched tokens, the most frequent English tokens in our dataset are exclusively function words distributed in an approximately Zipfian fashion. This matches what we would expect in a fully monolingual English corpus and is similar to the distribution that tokens labeled as Spanish follow.

This suggests that what we are actually seeing labeled as English are true English sub-utterances—true codeswitches—and not the well-established borrowings that we find under the most frequent BOR column, or the "internet-speak" that we did not annotate as either a borrowing or codeswitch. However, it should be noted that the comparison of the amount and complexity of codeswitching in the data with other datasets is difficult because a quantitative codeswitching measure like that of Gambäck and Das (2016) cannot be meaningfully applied across corpora with different annotation guidelines.

2.4. Limitations

Our goal for creating this corpus is to study a narrowly-scoped slice of multilingual language usage on Twitter—how Spanish speakers codeswitch into English, how they use borrowings and how these two phenomena intertwine—with the cleanest and most care-

---

[11]"Send me your email address and I will sent it to you."

3198

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

Table 2: The ten most frequent tokens in the dataset for each label: Spanish, English, and borrowing

fully annotated dataset possible. This corpus is not designed to be an evenly-balanced English-Spanish codeswitching dataset. It is primarily a Spanish dataset, mainly composed of Spanish tweets that may have an English codeswitch or a borrowing.

In addition, although our guidelines seek to establish a division as clearly as possible, the distinction between codeswitching and borrowing is fuzzy and is far from being solved. Many gray areas exist between social media jargon, nonce borrowing, and true codeswitching.

Another limitation is that as the original data selection process used generic Spanish language identification—Burger et al. (2011) do not provide any details—we do not know the geographic distribution of the speakers in the dataset and what dialects of Spanish and English they used. While we do believe this is a useful dataset for studying codeswitching and borrowing, caution should be exercised before drawing any global conclusions about the usage and mixing of either language.

3. Baseline Models

We evaluated how well popular models perform in predicting the annotated labels of our dataset. The data was divided into a standard training, development, and test set split in 80/10/10 proportions.

We evaluated four Transformer-based models on the dataset:

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

- BETO base cased model: a BERT-based model trained on Spanish (Cañete et al., 2020)

- RoBERTa BNE: a RoBERTa based model trained on Spanish (Gutiérrez-Fandiño et al., 2021)

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

The aim behind this selection of models was to assess how Transformer-based models that had been trained on different types of text would perform on codeswitched data.

The only multilingual model that we included was mBERT. BETO is a BERT-based model trained on a diverse set of international Spanish texts from different origins, such as OpenSubtitles, Global Voices, and the United Nations (Cañete, 2019). RoBERTa BNE is a RoBERTa-based model that has been trained exclusively on data crawled from .es websites—those using the top-level domain for Spain—by the National Library of Spain. This means that while BETO training data is smaller (3 billion tokens vs BNE's 135 billion tokens), the model has probably been exposed to a more varied representation of the different varieties of Spanish spoken around the globe, which includes the area of Spanish where codeswitching may be more prevalent. Finally, we added a RoBERTa-based model trained on Twitter data under the assumption that an English monolingual model trained on social media text could potentially do better than other models trained on other genres.

All models were run using the Transformers library by HuggingFace (Wolf et al., 2020) with the same default (untuned) hyperparameters: 3 epochs, batch size 32, and a maximum sequence length of 256.[12]

Table 3 shows that both mBERT and BETO were the best performing models and achieved similar scores, with an F1 of 96.6. While we have bolded the highest number in each column, we have no reason to believe that any differences between the performance of mBERT and BETO are meaningful. Interestingly, the RoBERTa-based model trained on the crawled data by the National Library of Spain performed worse, despite being a larger model. This seems to suggest the importance that a diverse set of training material may have when modeling certain linguistic phenomena that are impacted by geographical and dialectal variation, such as codeswitching and borrowing.

---

[12]https://github.com/huggingface/transformers/tree/master/examples/pytorch/token-classification

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

While we cannot directly compare our scores to other results as they are the first results on this dataset, these results are similar to those obtained on other datasets. For example, in the LinCE benchmark baseline (Aguilar et al., 2020),[13] for the language identification task on Spanish-English codeswitching data the F1 scores were 94.16 with a BiLSTM, 98.12 with ELMo, and 98.53 with mBERT. The mBERT baseline F1 for our dataset is slightly lower at 96.65. One possible explanation is that our task is slightly more difficult for the model as it must make the borrowing-codeswitching distinction, but further experimentation and tuning would be required to support that claim.

The token-level accuracy of approximately 96.9 for mBERT and BETO baseline matches the accuracy of 96.9 reported by Lignos and Marcus (2013) on their data with a heuristic-based tagging approach. However, their evaluation excluded any tokens with named entity tags and their data was not annotated for borrowing, so it is likely their approach would get lower accuracy if evaluated on our reannotation.

4. Conclusion

In this paper 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 that 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.

5. Bibliographical 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, July. 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, May. European Language Resources Association.

Álvarez-Mellado, E. and Lignos, C. (2022). Detecting unassimilated borrowings in Spanish: An annotated corpus and approaches to modeling. *arXiv preprint arXiv:2203.16169*.

Álvarez-Mellado, E. (2020a). An annotated corpus of emerging anglicisms in Spanish newspaper headlines. In *Proceedings of the The 4th Workshop on Computational Approaches to Code Switching*, pages 1–8, Marseille, France, May. European Language Resources Association.

Álvarez-Mellado, E. (2020b). Lázaro: An extractor of emergent anglicisms in Spanish newswire. Master's thesis, Brandeis University.

Asociación de Academias de la Lengua Española. (2010). Diccionario de americanismos. https://lema.rae.es/damer/.

Bali, K., Sharma, J., Choudhury, M., and Vyas, Y. (2014). "I am borrowing ya mixing ?" an analysis of English-Hindi code mixing in Facebook. In *Proceedings of the First Workshop on Computational Approaches to Code Switching*, pages 116–126, Doha, Qatar, October. Association for Computational Linguistics.

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, November. Association for Computational Linguistics.

Blanch, J. M. L. (1995). Americanismo frente a españolismo lingüísticos. *Nueva revista de filología hispánica*, 43(2):433–440.

Burger, J. D., Henderson, J., Kim, G., and Zarrella, G. (2011). Discriminating gender on Twitter. In *Proceedings of the 2011 Conference on Empirical Methods in Natural Language Processing*, pages 1301–1309, Edinburgh, Scotland, UK., July. Association for Computational Linguistics.

Cañete, J., Chaperon, G., Fuentes, R., Ho, J.-H., Kang, H., and Pérez, J. (2020). Spanish pre-trained BERT

---

[13]https://ritual.uh.edu/lince/

model and evaluation data. In *PML4DC at ICLR 2020*.

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

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, June. Association for Computational Linguistics.

Fernández Gordillo, L. (2014). La lexicografía del español y el español hispanoamericano. *Andamios*, 11(26):53–89.

Furiassi, C., Rodríguez González, F., and Pulcini, V. (2012). The anglicization of european lexis. *The Anglicization of European Lexis*, pages 1–366.

Gambäck, B. and Das, A. (2016). Comparing the level of code-switching in corpora. In *Proceedings of the Tenth International Conference on Language Resources and Evaluation (LREC'16)*, pages 1850–1855.

Görlach, M. (2002). *English in Europe*. OUP Oxford.

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.

Lipski, J. M. (2005). Code-switching or borrowing? No sé so no puedo decir, you know. In *Selected proceedings of the second workshop on Spanish sociolinguistics*, pages 1–15. Cascadilla Proceedings Project Somerville, MA.

Luján García, C. (2017). Analysis of the presence of Anglicisms in a Spanish internet forum: some terms from the fields of fashion, beauty and leisure. *Alicante Journal of English Studies*.

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, June. Association for Computational Linguistics.

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, November. Association for Computational Linguistics.

Patro, J., Samanta, B., Singh, S., Basu, A., Mukherjee, P., Choudhury, M., and Mukherjee, A. (2017). All that is English may be Hindi: Enhancing language identification through automatic ranking of the likeliness of word borrowing in social media. In *Proceedings of the 2017 Conference on Empirical Methods in Natural Language Processing*, pages 2264–2274, Copenhagen, Denmark, September. Association for Computational Linguistics.

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

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

Sanou, R. M. (2018). Anglicismos y redes sociales. *Cuadernos de la ALFAL*, 10:176–191.

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, October. Association for Computational Linguistics.

Stewart, I., Yang, D., and Eisenstein, J. (2021). Tuiteamos o pongamos un tuit? investigating the social constraints of loanword integration in Spanish social media. In *Proceedings of the Society for Computation in Linguistics 2021*, pages 286–297, Online, February. Association for Computational Linguistics.

Wolf, T., Debut, L., Sanh, V., Chaumond, J., Delangue, C., Moi, A., Cistac, P., Rault, T., Louf, R., Funtowicz, M., Davison, J., Shleifer, S., von Platen, P., Ma, C., Jernite, Y., Plu, J., Xu, C., Le Scao, T., Gugger, S., Drame, M., Lhoest, Q., and Rush, A. (2020). Transformers: State-of-the-art natural language processing. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing: System Demonstrations*, pages 38–45, Online, October. Association for Computational Linguistics.