Detecting Unassimilated Borrowings in Spanish: An Annotated Corpus and Approaches to Modeling

Anonymous ACL submission

Abstract

This work presents a new resource for borrowing identification and analyzes the performance and errors of several models on this task. We introduce a new annotated corpus of Spanish newswire rich in unassimilated lexical borrowings—words from one language that are introduced into another without orthographic adaptation—and use it to evaluate how several sequence labeling models (CRF, BiLSTM-CRF, and Transformer-based models) perform. The corpus contains 370,000 tokens and is larger, more borrowing-dense, OOV-rich, and topic-varied than previous corpora available for this task. Our results show that a BiLSTM-CRF model fed with either Transformer-based embeddings pretrained on codeswitched data or a combination of contextualized word embeddings (along with character embeddings and Spanish and English subword embeddings) outperforms results obtained by a multilingual BERT-based model.

1 Introduction and related work

Lexical borrowing is the process of bringing words from one language into another (Onysko, 2007; Poplack et al., 1988). Borrowings are a common source of out-of-vocabulary (OOV) words, and the task of detecting borrowings has proven to be useful both for lexicographic purposes and for NLP downstream tasks such as parsing (Alex, 2008a), text-to-speech synthesis (Leidig et al., 2014) and machine translation (Tsvetkov and Dyer, 2016).

Recent work has approached the problem of extracting lexical borrowings in European languages such as German (Alex, 2008b; Garley and Hockenmaier, 2012; Leidig et al., 2014), Italian (Furiassi and Hofland, 2007), French (Alex, 2008a; Chesley, 2010), Finnish (Mansikkaniemi and Kurimo, 2012), Norwegian (Andersen, 2012; Losnegaard and Lyse, 2012), and Spanish (Serigos, 2017), with a particular focus on English lexical borrowings (often called *anglicisms*).

Computational approaches to mixed-language data have traditionally framed the task of identifying the language of a word as a tagging problem, where every word in the sequence receives a language tag (Lignos and Marcus, 2013; Molina et al., 2016; Solorio et al., 2014). As lexical borrowings can be single (e.g. *app*, *online*, *smartphone*) or multi-token (e.g. *machine learning*), they are a natural fit for chunking-style approaches. Álvarez Mellado (2020b) introduced chunking-based models for borrowing detection in Spanish media which were later improved (Álvarez Mellado, 2020a), producing an F1 score of 86.41.

However, both the dataset and modeling approach used by Álvarez Mellado (2020a) had significant limitations. The dataset focused exclusively on a single source of news and consisted only of headlines. The number and variety of borrowings were limited, and there was a significant overlap in borrowings between the training set and the test set, which prevented assessment of whether the modeling approach was actually capable of generalizing to previously unseen borrowings. Additionally, the best results were obtained by a CRF model, and more sophisticated approaches were not explored.

The contributions of this paper are a new corpus of Spanish annotated with unassimilated lexical borrowings and a detailed analysis of the performance of several sequence-labeling models trained on this corpus. The models include a CRF, Transformer-based models, and BiLSTM-CRF with different word, subword and character embeddings (including contextualized embeddings pretrained on codeswitched data). The corpus contains 370,000 tokens and is larger and more topic-varied than previous resources. The test set was designed to be as difficult as possible; it covers sources and dates not seen in the training set, includes a high number of OOV words (92% of the borrowings in the test set are OOV) and is very borrowing-dense (20 borrowings per 1,000 tokens).

Media	Topics	Set(s)
ElDiario.es	General newspaper	Train, Dev.
El orden mundial	Politics	Test
Cuarto poder	Politics	Test
Politikon	Politics	Test
El salto	Politics	Test
La Marea	Politics	Test
Píkara	Feminism	Test
El blog salmón	Economy	Test
Pop rosa	Gossip	Test
Vida extra	Videogames	Test
Espinof	Cinema & TV	Test
Xataka	Technology	Test
Xataka Ciencia	Technology	Test
Xataka Android	Technology	Test
Genbeta	Technology	Test
Microsiervos	Technology	Test
Agencia Sinc	Science	Test
Diario del viajero	Travel	Test
Bebe y más	Parenthood	Test
Vitónica	Lifestyle & sports	Test
Los otros 18	Sports	Test
Foro atletismo	Sports	Test
Motor pasión	Automobiles	Test

Table 1: Sources included in each dataset split (URLs provided in the appendix)

Set	Tokens	ENG	OTHER	Unique
Training	231,126	1,493	28	380
Development	82,578	306	49	316
Test	58,997	1,239	46	987
Total	372,701	3,038	123	1,683

Table 2: Corpus splits with counts

2 Data collection and annotation

2.1 Contrasting lexical borrowing with codeswitching

Linguistic borrowing can be defined as the transference of linguistic elements between two languages. Borrowing and code-switching have been described as a continuum (Clyne et al., 2003).

Lexical borrowing involves the incorporation of single lexical units from one language into another language and is usually accompanied by morphological and phonological modification to conform with the patterns of the recipient language (Haugen, 1950; Onysko, 2007; Poplack et al., 1988).

On the other hand, code-switches are by definition not integrated into a recipient language, unlike established loanwords (Poplack, 2012). While code-switches require a substantial level of fluency, comply with grammatical restrictions in both languages, and are produced by bilingual speakers in bilingual discourses, lexical borrowings are words used by monolingual individuals that eventually become lexicalized and assimilated as part of the recipient language lexicon until the knowledge of "foreign" disappears (Lipski, 2005).

2.2 Data selection

Our dataset consists of Spanish newswire annotated for unassimilated lexical borrowings. All of the sources used are European Spanish online publications (newspapers, blogs, and news sites) published in Spain and written in European Spanish.

Data was collected separately for the training, development, and test sets to ensure minimal overlap in borrowings, topics, and time periods. The training set consists of a collection of articles appearing between August and December 2020 in *elDiario.es*, a progressive online newspaper based in Spain. The development set contains sentences in articles from January 2021 from the same source.

The data in the test set consisted of annotated sentences extracted in February and March 2021 from a diverse collection of online Spanish media that covers specialized topics rich in lexical borrowings and usually not covered by *elDiario.es*, such as sports, gossip or videogames (see Table 8).

To focus annotation efforts for the training set on articles likely to contain unassimilated borrowings, the articles to be annotated were selected by first using a baseline model and were then human-annotated. To detect potential borrowings, the CRF model and data from Álvarez Mellado (2020b) was used along with a dictionary look-up pipeline. Articles that contained more than 5 borrowing candidates were selected for annotation.

The main goal of data selection for the development and test sets was to create borrowing-dense, OOV-rich datasets, allowing for better assessment of generalization. To that end, the annotation was based on sentences instead of full articles. If a sentence contained a word either flagged as a borrowing by the CRF model, contained in a wordlist of English, or simply not present in the training set, it was selected for annotation. This data selection approach ensured a high number of borrowings and OOV words, both borrowings and non-borrowings. While the training set contains 6 borrowings per 1,000 tokens, the test set contains 20 borrowings per 1,000 tokens. Additionally, 90% of the unique borrowings in the development set were OOV (not present in training). 92% of the borrowings in the test set did not appear in training (see Table 2).

2.3 Annotation process

The corpus was annotated with BIO encoding using Doccano (Nakayama et al., 2018) by a native speaker of Spanish with a background in linguistic annotation. The annotation guidelines (which we provide in the appendix) were based on those of Álvarez Mellado (2020a) but were expanded to account for a wider diversity of topics. English lexical borrowings were labeled `ENG`, other borrowings were labeled `OTHER`. Here is an example from the training set:[1]

> **Benching** [ENG], estar en el banquillo de tu **crush** [ENG] mientras otro juega de titular.

In order to assess the quality of the guidelines and the annotation, a sample of 9,110 tokens from 450 sentences (60% from the test set, 20% from training, 20% from development) was divided among a group of 9 linguists for double annotation. The mean inter-annotation agreement computed by Cohen's kappa was 0.91, which is above the 0.8 threshold of reliable annotation (Artstein and Poesio, 2008).

2.4 Limitations

We believe it is best to be upfront about the potential limitations of this resource. The corpus consists exclusively of news published in Spain and written in European Spanish. This fact by no means implies the assumption that European Spanish represents the whole of the Spanish language.

The notion of assimilation is usage-based and community-dependant, and thus the dataset we present and the annotation guidelines that were followed were designed to capture a very specific phenomena at a given time and in a given place: unassimilated borrowings in the Spanish press.

In order to establish whether a given word has been assimilated or not, the annotation guidelines rely on lexicographic sources such as the prescriptivist *Diccionario de la Lengua Española* (Real Academia Española, 2020) by the Royal Spanish Academy, a dictionary that aims to cover world-wide Spanish but whose Spain-centric criteria has been previously pointed out (Blanch, 1995; Fernández Gordillo, 2014). In addition, prior work has suggested that Spanish from Spain may have a higher tendency of anglicism-usage than other Spanish dialects (McClelland, 2021). Consequently, we limit the scope of the dataset to European Spanish not because we consider that this variety represents the whole of the Spanish-speaking community, but because we consider that the approach we have taken here may not account adequately for the whole diversity in borrowing assimilation within the Spanish-speaking world.

[1] "Benching: being on your crush's bench while someone else plays in the starting lineup."

Set	Precision	Recall	F1
Development			
`ALL`	74.13	59.72	66.15
`ENG`	74.20	68.63	71.31
`OTHER`	66.67	4.08	7.69
Test			
`ALL`	77.89	43.04	55.44
`ENG`	78.09	44.31	56.54
`OTHER`	57.14	8.70	15.09

Table 3: CRF results on the development and test set

3 Modeling

The corpus was used to evaluate four types of models for borrowing extraction: (1) a CRF model, (2) several Transformer-based models, (3) a BiLSTM-CRF model with different types of unadapted embeddings (word, subword, and character embeddings) and (4) a BiLSTM-CRF model with previously fine-tuned Transformer-based embeddings pretrained on codeswitched data. By *unadapted* embeddings, we mean embeddings that have not been fine-tuned for the task of anglicism detection or a related task (e.g. codeswitching).

Evaluation for all models required extracted spans to match the annotation exactly in span and type to be correct. Evaluation was performed with SeqScore (Palen-Michel et al., 2021), using `conlleval`-style repair for invalid label sequences. All models were trained using an AMD 2990WX CPU and a single RTX 2080 Ti GPU.

3.1 Conditional random field model

As baseline model, we evaluated a CRF model with handcrafted features from Álvarez Mellado (2020b). The model was built using `pycrfsuite` (Korobov and Peng, 2014), a Python wrapper for `crfsuite` (Okazaki, 2007) that implements CRF for labeling sequential data. The model also uses the `Token` and `Span` utilities from `spaCy` library (Honnibal and Montani, 2017). The following handcrafted binary features from Álvarez Mellado (2020b) were used for the model:

– Bias: active on all tokens to set per-class bias
– Token: the string of the token

	Development			Test		
	Precision	Recall	F1	Precision	Recall	F1
BETO						
ALL	73.35	72.11	72.73	86.63	76.65	81.34
ENG	73.28	83.33	77.98	86.73	79.10	82.74
OTHER	100.00	2.04	4.00	71.43	10.87	18.87
mBERT						
ALL	**82.73**	**76.90**	**79.71**	**89.98**	**77.59**	**83.33**
ENG	82.97	87.58	85.21	90.45	80.23	85.03
OTHER	71.43	10.20	17.86	33.33	6.52	10.91

Table 4: Results on the development set and test set for Transformer-based models (BETO, mBERT)

– Uppercase: active if the token is all uppercase
– Titlecase: active if only the first character of the token is capitalized
– Character trigram: an active feature for every trigram contained in the token
– Quotation: active if the token is any type of quotation mark (‘ ’ " “ ” « »)
– Suffix: last three characters of the token
– POS tag: part-of-speech tag of the token provided by `spaCy` utilities
– Word shape: shape representation of the token provided by `spaCy` utilities
– Word embeddings: provided by Spanish word2vec 300 dimensional embeddings by Cardellino (2019)
– URL: active if the token could be validated as a URL according to `spaCy` utilities
– Email: active if the token could be validated as an email address by `spaCy` utilities
– Twitter: active if the token could be validated as a possible Twitter special token: `#hashtag` or `@username`

A window of two tokens in each direction was used for feature extraction. Optimization was performed using L-BFGS with the following hyperparameter values chosen following the best results from Álvarez Mellado (2020b) were set: c1 = 0.05, c2 = 0.01. As shown in Table 3, the CRF produced an overall F1 score of 66.15 on the development set (P: 74.13, R: 59.72) and an overall F1 of 55.44 (P: 77.89, R: 43.04) on the test set. The CRF results on our dataset are far below the F1 of 86.41 reported by Álvarez Mellado (2020b), showing the impact that a topically-diverse, OOV-rich dataset can have, especially on test set recall (43.04). These results demonstrate that we have created a more difficult task and motivate using more sophisticated models.

3.2 Transformer-based models

We evaluated two Transformer-based models:
– BETO base cased model: a monolingual BERT model trained for Spanish (Cañete et al., 2020)
– mBERT: multilingual BERT, trained on Wikipedia in 104 languages (Devlin et al., 2018)

Both models were run using the `Transformers` library by HuggingFace (Wolf et al., 2020). The same default hyperparameters were used for both models: 3 epochs, batch size 32, and maximum sequence length 256.

As shown in Table 4, the mBERT model performed best. Both models performed better on the test set that on the development set, despite the difference in topics between them, suggesting good generalization. This is a remarkable difference compared to the CRF results, where the CRF performed substantially worse on the test set than the development set.

3.3 BiLSTM-CRF

We explored several possibilities for a BiLSTM-CRF model fed with different types of word and subword embeddings. The purpose was to assess whether the combination of different embeddings that encode different linguistic information could outperform the best results obtained by the Transformer-based models in Section 3.2. All of our BiLSTM-CRF models were built using `Flair` (Akbik et al., 2018) with default hyperparameters (hidden size = 256, learning rate = 0.1, mini batch size = 32, maximum number of epochs = 150) and used embeddings provided by `Flair`.

3.3.1 Preliminary embedding experiments

We first ran exploratory experiments on the development set with different types of embeddings using `Flair` tuning functionalities. We explored the following embeddings: Transformer embeddings (mBERT and BETO), fastText embeddings (Bojanowski et al., 2017), one-hot embeddings, byte pair embeddings (Heinzerling and Strube, 2018), and character embeddings (Lample et al., 2016).

The best results were obtained by a combination of mBERT embeddings and character embeddings (F1: 74.00), followed by a combination of BETO embeddings and character embeddings (F1: 72.09). These results show that using contextualized embeddings unsurprisingly outperform non-contextualized embeddings for this task, and that subword representation is important for the task of extracting borrowings that have not been adapted orthographically. The finding regarding the importance of subwords is consistent with previous work; feature ablation experiments for borrowing detection have shown that character trigram features contributed the most to the results obtained by a CRF model (Álvarez Mellado, 2020b).

The worst result (F1: 39.21) was produced by a model fed with one-hot vectors, and the second-worst result was produced by a model fed exclusively with character embeddings. While it performed poorly (F1: 41.65), this fully unlexicalized model outperformed one-hot embeddings, reinforcing the importance of subword information for the task of unassimilated borrowing extraction.

3.3.2 Optimal embedding combination

In light of the preliminary embedding experiments and our earlier experiments with Transformer-based models, we fed our BiLSTM-CRF model with different combinations of contextualized word embeddings (including BERT embeddings from Devlin et al.), byte-pair embeddings and character embeddings. Table 5 shows development set results from different combinations of embeddings. The best overall F1 on the development set (81.79) was obtained by the combination of BETO embeddings, BERT embeddings and byte-pair embeddings. The model fed with BETO embeddings BERT embeddings, byte-pair embeddings and character embeddings ranked second (F1=81.48).

Several things stand out from the results in Table 5. The BETO+BERT embedding combination consistently works better than mBERT embeddings, and BPE embeddings contribute to better results. Character embeddings, however, seem to produce little effect at first glance. Given the same model, adding character embeddings produced little changes in F1 or even slightly hurt the results (81.48 vs 81.79; 78.96 vs 79.01; 78.96 vs 78.64). Although character embeddings seem to make little difference in overall F1, recall was consistently higher in models that included character embeddings (77.46 vs 77.18; 76.62 vs 76.34; 77.18 vs 74.65), and in fact, the model with BETO+BERT embeddings, BPE embeddings and character embeddings produced the highest recall overall (77.46). This is an interesting finding, as our results from Sections 3.1 and 3.2 as well as prior work (Álvarez Mellado, 2020b) identified recall as weak for borrowing detection models.

The two best-performing models from Table 5 (BETO+BERT embeddings, BPE embeddings and optionally character embeddings) were evaluated on the test set. Table 6 gives results per type on the development and test sets for these two models. For both models, results on the test set were better (F1: 84.02, F1: 84.25) than on the development set (F1: 81.79, F1: 81.48). Although the best F1 score on the development set was obtained with no character embeddings, when run on the test set the model with character embeddings obtained the best score and the highest recall (R: 79.53), which again seems to corroborate the positive impact that character information can have in recall when dealing with previously unseen borrowings.

3.4 Borrowing detection as a transfer learning task from codeswitching

Finally, we decided to explore whether detecting unassimilated lexical borrowings could be framed as transfer learning from language identification in codeswitching. As before, we ran a BiLSTM-CRF model using `Flair`, but instead of using the unadapted Transformer embeddings, we used codeswitch embeddings[2], fine-tuned Transformer-based embeddings pretrained for language identification on the Spanish-English section of the codeswitching LinCE dataset (Aguilar et al., 2020).

Table 7 gives results for these models on the development and test sets. The two best-performing models on the development set where the BiLSTM-CRF with codeswitch and BPE embeddings (F1: 80.17) and the BiLSTM-CRF model with codeswitch, BPE and character embeddings (F1: 79.5). None of these outperformed the best development set results obtained by the two best performing BiLSTM models with unadapted embeddings from Section 3.3 (F1: 81.79, F1: 81.48).

However, these two models did outperform the BiLSTM-CRF models with unadapted embeddings when run on the test set; the model with codeswitch, BPE and character embeddings produced the best

[2] https://github.com/sagorbrur/codeswitch

Word embedding	BPE embedding	Char embedding	Precision	Recall	F1
mBERT	-	-	82.27	69.30	75.23
mBERT	-	✓	79.45	72.96	76.06
mBERT	multi	✓	81.37	73.80	77.40
mBERT	es, en	-	83.07	74.65	78.64
mBERT	es, en	✓	80.83	77.18	78.96
BETO, BERT	-	✓	81.44	76.62	78.96
BETO, BERT	-	-	81.87	76.34	79.01
BETO, BERT	es, en	✓	85.94	**77.46**	81.48
BETO, BERT	es, en	-	**86.98**	77.18	**81.79**

Table 5: Overall F1 score obtained on the development set with the BiLSTM-CRF model using different combinations of multilingual word embeddings, subword embeddings and character embeddings

	Development			**Test**		
	Precision	Recall	F1	Precision	Recall	F1
BETO+BERT and BPE						
ALL	**86.98**	77.18	**81.79**	**90.56**	78.37	84.02
ENG	86.90	88.89	87.88	90.68	80.87	85.49
OTHER	50.00	2.04	3.92	71.43	10.87	18.87
BETO+BERT, BPE and char						
ALL	85.94	**77.46**	81.48	89.57	**79.53**	**84.25**
ENG	86.58	88.56	87.56	89.92	82.08	85.82
OTHER	57.14	8.16	14.29	50.00	10.87	17.86

Table 6: Results on the development set and test set for BiLSTM-CRF model with BETO and BERT embeddings, BPE embeddings and optionally character embeddings

all around results on the test set (F1:85.49), a result that outperformed all results from all previous models both on ENG borrowings and OTHER borrowings. It should be noted that this transfer learning approach is indirectly using more data than just the training data from our initial corpus, as the codeswitch-based BiLSTM-CRF models benefit from the labeled data seen during pretraining for the language-identification task.

4 Error analysis

We compared the different results produced by the best performing model of each type on the test set: (1) the mBERT model, (2) the BiLSTM-CRF with BERT+BETO, BPE and character embeddings and (3) the BiLSTM-CRF model with codeswitch, BPE and character embeddings. We divide the error analysis into two sections. We first analyze errors that were made by all three models, with the aim of discovering which instances of the dataset were challenging for all models. We then analyze unique answers (both correct and incorrect) per model, with the aim of gaining insight on what are the unique characteristics of each model in comparison with other models.

4.1 Errors made by all models

4.1.1 Borrowings labeled as O

There were 137 tokens in the test set that were incorrectly labeled as O by all three models. 103 of these were of type ENG, 34 were of type OTHER. These errors can be classified as follows:

– Borrowings in upper case (12), which tend to be mistaken by models with proper nouns:

> Análisis de empresa basados en **Big Data** [ENG].[3]

– Borrowings in sentence-initial position (9), which were titlecased and therefore consistently mislabeled as O:

> **Youtuber** [ENG], mujer y afroamericana: Candace Owen podría ser la alternativa a Trump.[4]

Sentence-initial borrowings are particularly tricky, as models tend to confuse these with foreign named entities. In fact, prior work on anglicism detection based on dictionary lookup (Serigos, 2017) stated that borrowings in sentence-initial position were rare in Spanish and consequently chose to ignore all foreign words in sentence-initial position

[3] "Business analytics based on Big Data"

[4] "Youtuber, woman and African-American: Candace Owen could be the alternative to Trump"

	Development			Test		
	Precision	Recall	F1	Precision	Recall	F1
Codeswitch embeddings						
ALL	80.30	74.65	77.37	90.50	77.12	83.28
ENG	80.49	86.27	83.28	90.48	79.74	84.77
OTHER	50.00	2.04	3.92	100.00	6.52	12.24
Codeswitch embeddings + char embeddings						
ALL	77.84	75.21	76.50	89.83	80.39	84.85
ENG	78.01	86.93	82.23	89.80	83.13	86.34
OTHER	50.00	2.04	3.92	100.00	6.52	12.24
Codeswitch embeddings + BPE						
ALL	**83.08**	**77.46**	**80.17**	89.18	**81.48**	85.16
ENG	82.82	88.24	85.44	89.37	84.18	86.70
OTHER	100.00	10.20	18.52	57.14	8.70	15.09
Codeswitch embeddings + BPE + char embeddings						
ALL	82.48	76.90	79.59	**90.30**	81.17	**85.49**
ENG	82.21	87.58	84.81	90.42	83.78	86.97
OTHER	100.00	10.20	18.52	71.43	10.87	18.87

Table 7: Results on the development set and test set for LSTM-CRF model with different combinations of codeswitch embeddings, BPE embeddings and character embeddings

under the assumption that they could be considered named entities. However, these examples (and the difficulty they pose for models) prove that sentence-initial borrowings are not rare and therefore should not be overlooked.

– Borrowings that also happen to be words in Spanish (8), such as the word *primer*, that is a borrowing found in makeup articles (*un primer hidratante*, "an hydrating primer") but also happens to be a fully Spanish adjective meaning "first" (*primer premio*, "first prize"). Borrowings like these are still treated as fully unassimilated borrowings by speakers, even when the form is exactly the same as an already-existing Spanish word and were a common source of mislabeling, especially partial mismatches in multitoken borrowings: *red* (which exists in Spanish meaning "net") in *red carpet*, *tractor* in *tractor pulling* or *total* in *total look*.

– Borrowings that could pass as Spanish words (58): most of the misslabeled borrowings were words that do not exist in Spanish but that could orthographically pass for a Spanish word. That is the case of words like *burpees* (hypothetically, a conjugated form of the non-existing verb *burpear*), *gimbal*, *mules*, *bromance* or *nude*.

– Other borrowings (50): a high number of mislabeled borrowings were borrowings that were othographically implausible in Spanish, such as *trenchs*, *multipads*, *hypes*, *riff*, *scrunchie* or *mint*. The fact that none of our models were able to correctly classify these orthographically implausible examples leaves the door open to further exploration of character-based models and investigating character-level perplexity as a source of information.

4.1.2 Non-borrowings labeled as borrowings

29 tokens were incorrectly labeled as borrowings by all three models. These errors can be classified in the following groups:

– Metalinguistic usage and reported speech: a foreign word or sentence that appears in the text to refer to something someone said or wrote.

> Escribir "**icon pack**" [ENG] en el buscador.[5]

– Lower-cased proper nouns: such as websites.

> Hay que acceder a la página **flywithkarolg** [ENG][6]

– Computer commands: the test set included blog posts about technology, which mentioned computer commands (such as *sudo apt-get update*) that were consistently mistaken by our models as borrowings. These may seem like an extreme case—after all, computer commands do contain English words—but they are a good example of the real data that a borrowing-detection system may encounter.

– Foreign words within proper nouns: lower-cased foreign words that were part of multitoken proper nouns.

> La serie "10.000 **ships** [ENG]" cuenta la odisea de la princesa Nymeria.[7]

– Acronyms and acronym expansions:

> El entrenamiento HITT (**high intensity interval training** [ENG])[8]

[5] "Type 'icon pack' on the search box"

[6] "You need to access the website flywithkarolg"

[7] "The series '10,000 ships' tells the story of princess Nymeria"

[8] "HITT training (High-intensity interval training)"

– Assimilated borrowings: certain borrowings that are already considered by RAE's dictionary as fully assimilated were labeled by all models as anglicisms.

Labios rojos, a juego con el **top** [ENG].[9]

4.1.3 Type confusion

Three tokens of type OTHER were marked by all models as ENG. There were no ENG borrowings that were labeled as OTHER by all three models.

Había **buffet** [ENG] libre.[10]

4.2 Unique answers per model

We now summarize the unique mistakes and correct answers made per model, with the aim of understanding what data points were handled uniquely well or badly by each model.

4.2.1 mBERT

There were 46 tokens that were incorrectly labeled as borrowings only by the mBERT model. These include foreign words used in reported speech or acronym expansion (21), proper names (11) and already assimilated borrowings (7).

There were 27 tokens that were correctly labeled only by the mBERT model. The mBERT model was particularly good at detecting the full span of multitoken borrowings as in *no knead bread*, *total white*, *wide leg* or *kettlebell swings* (which were only partially detected by other models) and at detecting borrowings that could pass for Spanish words (such as *fashionista*, *samples*, *vocoder*). In addition, the mBERT model also correctly labeled as O 12 tokens that the other two models mistook as borrowings, including morphologically adapted anglicisms, such as *craftear* (Spanish infinitive of the verb *to craft*) *crackear* (from *to crack*) or *lookazo* (augmentative of the noun *look*).

4.2.2 BiLSTM-CRF with unadapted embeddings

There were 23 tokens that were incorrectly labeled as borrowings solely by this model, the most common types being assimilated borrowings (such as *fan*, *clon*) and Spanish words (*fiestiones*) (9 each).

32 tokens were correctly labeled as borrowings only by this model. These include borrowings that could pass for Spanish words (*camel*, *canvas*). In addition, this model also correctly labeled as O 6 tokens that the other two mistook as borrowings, including old borrowings that are considered today as fully assimilated (such as *films* or *sake*) or the usage of *post* as a prefix of latin origin (as in *post-produccion*), which other models mistook with the English word *post*.

4.2.3 BiLSTM-CRF with codeswitch embeddings

The codeswitch-based system incorrectly labeled 18 tokens as borrowings, including proper nouns (7) such as *Baby Spice*, and fully asimilated borrowings (5), such as *jersey*, *relax* or *tutorial*.

This model correctly labeled 27 tokens that were mistakenly ignored by other models, including multitoken borrowings (*dark and gritty*, *red carpet*) and other borrowings that were non-compliant with Spanish orthograpich rules but that were however ignored by other models (*messy*, *athleisure*, *multitouch*, *workaholic*).

The codeswitch-based model also correctly labeled as O 16 tokens that the other two models labeled as borrowings, including acronym expansions, lower-cased proper names and orthographically unorthodox Spanish words, such as the ideophone *tiki-taka* or *shavales* (a non-standard writing form of the word *chavales*, "guys").

5 Conclusion

We have introduced a new corpus of Spanish newswire annotated with unassimilated lexical borrowings. The test set has a high number of OOV borrowings—92% of unique borrowings in the test set were not seen during training—and is more borrowing-dense and varied than resources previously available. We have used the dataset to explore several sequence labeling models (CRF, BiLSTM-CRF, and Transformer-based models) for the task of extracting lexical borrowings in an high-OOV setting. Results show that the BiLSTM-CRF model fed with character embeddings, Spanish and English subword embeddings and either Transformer-based embeddings pretrained on codeswitched data (F1: 85.49) or a combination of contextualized word embeddings (F1: 84.25) produced the best results and outperformed prior models for this task (CRF F1: 55.44) and multilingual Transformer-based models (mBERT F1: 83.33).

References

Fundéu. Fundación del Español Urgente. http://www.fundeu.es.

[9] "Red lips, matching top"

[10] "There was a free buffet"

Gustavo Aguilar, Sudipta Kar, and Thamar Solorio. 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.

Alan Akbik, Duncan Blythe, and Roland Vollgraf. 2018. Contextual string embeddings for sequence labeling. In *COLING 2018, 27th International Conference on Computational Linguistics*, pages 1638–1649.

Beatrice Alex. 2008a. *Automatic detection of English inclusions in mixed-lingual data with an application to parsing*. Ph.D. thesis, University of Edinburgh.

Beatrice Alex. 2008b. Comparing corpus-based to web-based lookup techniques for automatic English inclusion detection. In *Proceedings of the Sixth International Conference on Language Resources and Evaluation (LREC'08)*, Marrakech, Morocco. European Language Resources Association (ELRA).

Elena Álvarez Mellado. 2020a. An Annotated Corpus of Emerging Anglicisms in Spanish Newspaper Headlines. In *Proceedings of the Fourth Workshop on Computational Approaches to Code Switching*, pages 1–8, Marseille, France. European Language Resources Association.

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

Gisle Andersen. 2012. Semi-automatic approaches to Anglicism detection in Norwegian corpus data. In Cristiano Furiassi, Virginia Pulcini, and Félix Rodríguez González, editors, *The anglicization of European lexis*, pages 111–130. John Benjamins.

Ron Artstein and Massimo Poesio. 2008. Inter-coder agreement for computational linguistics. *Computational Linguistics*, 34(4):555–596.

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

Piotr Bojanowski, Edouard Grave, Armand Joulin, and Tomas Mikolov. 2017. Enriching word vectors with subword information. *Transactions of the Association for Computational Linguistics*, 5:135–146.

Cristian Cardellino. 2019. Spanish Billion Words Corpus and Embeddings. https://crscardellino.github.io/SBWCE/.

José Cañete, Gabriel Chaperon, Rodrigo Fuentes, Jou-Hui Ho, Hojin Kang, and Jorge Pérez. 2020. Spanish pre-trained bert model and evaluation data. In *PML4DC at ICLR 2020*.

Paula Chesley. 2010. Lexical borrowings in French: Anglicisms as a separate phenomenon. *Journal of French Language Studies*, 20(3):231–251.

Michael Clyne, Michael G Clyne, and Clyne Michael. 2003. *Dynamics of language contact: English and immigrant languages*. Cambridge University Press.

Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 2018. BERT: pre-training of deep bidirectional transformers for language understanding. *CoRR*, abs/1810.04805.

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

Cristiano Furiassi and Knut Hofland. 2007. The retrieval of false anglicisms in newspaper texts. In *Corpus Linguistics 25 Years On*, pages 347–363. Brill Rodopi.

Matt Garley and Julia Hockenmaier. 2012. Beefmoves: Dissemination, diversity, and dynamics of English borrowings in a German hip hop forum. In *Proceedings of the 50th Annual Meeting of the Association for Computational Linguistics (Volume 2: Short Papers)*, pages 135–139, Jeju Island, Korea. Association for Computational Linguistics.

Juan Gómez Capuz. 1997. Towards a typological classification of linguistic borrowing (illustrated with anglicisms in romance languages). *Revista alicantina de estudios ingleses*, 10:81–94.

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

Benjamin Heinzerling and Michael Strube. 2018. BPEmb: Tokenization-free pre-trained subword embeddings in 275 languages. In *Proceedings of the Eleventh International Conference on Language Resources and Evaluation (LREC 2018)*, Miyazaki, Japan. European Language Resources Association (ELRA).

Matthew Honnibal and Ines Montani. 2017. spaCy 2: Natural language understanding with bloom embeddings, convolutional neural networks and incremental parsing. https://spacy.io/.

Mikhail Korobov and Terry Peng. 2014. Python-crfsuite. https://github.com/scrapinghub/python-crfsuite.

Guillaume Lample, Miguel Ballesteros, Sandeep Subramanian, Kazuya Kawakami, and Chris Dyer. 2016. Neural architectures for named entity recognition. In *Proceedings of the 2016 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 260–270, San Diego, California. Association for Computational Linguistics.

Sebastian Leidig, Tim Schlippe, and Tanja Schultz. 2014. Automatic detection of anglicisms for the pronunciation dictionary generation: a case study on our German IT corpus. In *Spoken Language Technologies for Under-Resourced Languages*.

Constantine Lignos and Mitch Marcus. 2013. Toward web-scale analysis of codeswitching. Presented at the 87th Annual Meeting of the Linguistic Society of America.

John M Lipski. 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.

Gyri Smordal Losnegaard and Gunn Inger Lyse. 2012. A data-driven approach to anglicism identification in Norwegian. In Gisle Andersen, editor, *Exploring Newspaper Language: Using the web to create and investigate a large corpus of modern Norwegian*, pages 131–154. John Benjamins Publishing.

André Mansikkaniemi and Mikko Kurimo. 2012. Unsupervised vocabulary adaptation for morph-based language models. In *Proceedings of the NAACL-HLT 2012 Workshop: Will We Ever Really Replace the N-gram Model? On the Future of Language Modeling for HLT*, pages 37–40. Association for Computational Linguistics.

Jacob McClelland. 2021. A brief survey of anglicisms among spanish dialects. *Colorado Research in Linguistics*.

Giovanni Molina, Fahad AlGhamdi, Mahmoud Ghoneim, Abdelati Hawwari, Nicolas Rey-Villamizar, Mona Diab, and Thamar Solorio. 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.

Hiroki Nakayama, Takahiro Kubo, Junya Kamura, Yasufumi Taniguchi, and Xu Liang. 2018. doccano: Text annotation tool for human. https://github.com/doccano/doccano.

Eugenia Esperanza Núñez Nogueroles. 2017. An up-to-date review of the literature on anglicisms in Spanish. *Diálogo de la Lengua, IX*, pages 1–54.

Naoaki Okazaki. 2007. Crfsuite: a fast implementation of Conditional Random Fields (CRFs). http://www.chokkan.org/software/crfsuite/.

Alexander Onysko. 2007. *Anglicisms in German: Borrowing, lexical productivity, and written codeswitching*, volume 23. Walter de Gruyter.

Chester Palen-Michel, Nolan Holley, and Constantine Lignos. 2021. SeqScore: Addressing barriers to reproducible named entity recognition evaluation. In *Proceedings of the 2nd Workshop on Evaluation and Comparison of NLP Systems*, pages 40–50, Punta Cana, Dominican Republic. Association for Computational Linguistics.

Barbara Plank. 2016. What to do about non-standard (or non-canonical) language in NLP. *arXiv preprint arXiv:1608.07836*.

Shana Poplack. 2012. What does the nonce borrowing hypothesis hypothesize? *Bilingualism: Language and Cognition*, 15(3):644–648.

Shana Poplack, David Sankoff, and Christopher Miller. 1988. The social correlates and linguistic processes of lexical borrowing and assimilation. *Linguistics*, 26(1):47–104.

Chris Pratt. 1980. *El anglicismo en el español peninsular contemporáneo*, volume 308. Gredos.

Real Academia Española. 2020. Diccionario de la lengua española, ed. 23.4. http://dle.rae.es.

Felix Rodríguez González. 1999. Anglicisms in contemporary Spanish. An overview. *Atlantis*, 21(1/2):103–139.

Jacqueline Rae Larsen Serigos. 2017. *Applying corpus and computational methods to loanword research: new approaches to Anglicisms in Spanish*. Ph.D. thesis, The University of Texas at Austin.

Thamar Solorio, Elizabeth Blair, Suraj Maharjan, Steven Bethard, Mona Diab, Mahmoud Ghoneim, Abdelati Hawwari, Fahad AlGhamdi, Julia Hirschberg, Alison Chang, and Pascale Fung. 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.

Yulia Tsvetkov and Chris Dyer. 2016. Cross-lingual bridges with models of lexical borrowing. *Journal of Artificial Intelligence Research*, 55:63–93.

Thomas Wolf, Lysandre Debut, Victor Sanh, Julien Chaumond, Clement Delangue, Anthony Moi, Pierric Cistac, Tim Rault, Remi Louf, Morgan Funtowicz, Joe Davison, Sam Shleifer, Patrick von Platen, Clara Ma, Yacine Jernite, Julien Plu, Canwen Xu, Teven Le Scao, Sylvain Gugger, Mariama Drame, Quentin Lhoest, and Alexander Rush. 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. Association for Computational Linguistics.

A Data sources

See Table 8 for the URLs of the sources used in the test set.

B Annotation guidelines

B.1 Objective

This document proposes a set of guidelines for annotating emergent unassimilated lexical borrowings, with a focus on English lexical borrowings (or anglicisms). The purpose of these annotation guidelines is to assist annotators to annotate unassimilated lexical borrowings from English that appear in Spanish newswire, i.e. words from English origin that are introduced into Spanish without any morphological or orthographic adaptation.

This project approaches the phenomenon of lexical borrowing from a synchronic point of view, which means that we will not be annotating all words that have been borrowed at some point of the history of the Spanish language (like arabisms), but only those that have been recently imported and have not been integrated into the recipient language (in this case, Spanish).

B.2 Tagset

We will consider two possible tags for our annotation: `ENG`, for borrowings that come from the English language (or anglicisms), and `OTHER` for other borrowings that comply with the following guidelines but that come from languages other than English.

B.3 What an unassimilated lexical borrowing is

In this section we provide an overview of what words will be considered as unassimilated lexical borrowings for the sake of our annotation project.

B.3.1 Definition and scope

The concept of *linguistic borrowing* covers a wide range of linguistic phenomena. We will first provide a general overview of what lexical borrowing is and what will be understood as an anglicism within the scope of this project.

Lexical borrowing is the incorporation of single lexical units from one language (the donor language) into another language (the recipient language) and is usually accompanied by morphological and phonological modification to conform with the patterns of the recipient language (Haugen, 1950; Onysko, 2007; Poplack et al., 1988).

Anglicisms are lexical borrowings that come from the English language (Gómez Capuz, 1997; Pratt, 1980; Rodríguez González, 1999; Núñez Nogueroles, 2017). For our annotation project, we will focus on direct, unassimilated, emerging anglicisms, i.e. lexical borrowings from the English language into Spanish that have recently been imported and that have still not been assimilated into Spanish, that is, words like *smartphone*, *influencer*, *hype*, *lawfare* or *reality show*.

Although this project focuses on lexical borrowings from English, we will also consider borrowings from other languages that comply with these guidelines. Borrowings from the English language will be annotated with the tag `ENG`, while borrowings from other languages shall be annotated with the tag `OTHER`:

```
... financiados a
través de la plataforma de
[crowdfunding](ENG) del club
[gourmet](OTHER) que tengas más
cerca
```
[11]

Other types of borrowings, such as semantic calques, syntactic anglicisms or literal translations will be considered beyond the scope of these annotation project and will not be covered in these guidelines.

B.3.2 Types of lexical borrowing

Lexical borrowings can be adapted (the spelling of the word is modified to comply with the phonological and orthographic patterns of the recipient language, as in *fútbol* or *tuit*) or unadapted (the word preserves its original spelling: *millennial*, *newsletter*, *like*). For this annotation project, we will be focusing on unassimilated lexical borrowings: this means that adapted borrowings will be ignored and only unadapted borrowings will be tagged (see section B.4.2 for a full description on the differences between adapted and unadapted borrowings).

B.3.3 Multiword borrowings

Lexical borrowings can be both single-token units (*online*, *impeachment*), as well as multiword expressions (*reality show*, *best seller*). Multitoken borrowings will be labeled as one entity.

[11] Examples in these guidelines will display the lexical borrowing that should be labeled between square brackets, with the the corresponding tag in parentheses. Examples with no words marked with brackets will illustrate cases where no lexical borrowing should be tagged.

Media	URL
El orden mundial	https://elordenmundial.com/
Cuarto poder	https://www.cuartopoder.es/
Politikon	https://www.politikon.es/
El salto	https://www.elsaltodiario.com/
La Marea	https://www.lamarea.com/
Píkara	https://www.pikaramagazine.com/
El blog salmón	https://www.elblogsalmon.com/
Pop rosa	https://www.poprosa.com/
Vida extra	https://www.vidaextra.com/
Espinof	https://www.espinof.com/
Xataka	https://www.xataka.com/
Xataka Ciencia	https://www.xatakaciencia.com/
Xataka Android	https://www.xatakandroid.com/
Genbeta	https://www.genbeta.com/
Microsiervos	https://www.microsiervos.com/
Agencia Sinc	https://www.agenciasinc.es/
Diario del viajero	https://www.diariodelviajero.com/
Bebe y más	https://www.bebesymas.com/
Vitónica	https://www.vitonica.com/
Los otros 18	https://www.losotros18.com/
Foro atletismo	https://www.foroatletismo.com/
Motor pasión	https://www.motorpasion.com/

Table 8: Media included in the test set

```
imagina ser un '[tech bro]' con millones de dólares (ENG)
```

The annotation should however distinguish between a multitoken borrowing and adjacent borrowings. A phrase like *signature look* is a multiword borrowing (the full phrase has been borrowed as a single unit) and should be annotated as such.

```
para recrear su [total look] (ENG)
```

However, a phrase like *look sporty* follows the NAdj order that is typical of Spanish grammar (but impossible in English): these are in fact two separate borrowings (*look* and *sporty*) that have been borrowed independently and happen to be colocated in a phrase. The annotation should capture these nuances:

```
un [look] (ENG) [sporty] (ENG) perfecto
```

B.3.4 Origin of the borrowings

Establishing the origin of a certain borrowings can sometimes be tricky, as the language of origin can sometimes be disputed. Additionally, certain borrowings might have originated in a certain language, but may have reached the recipient language through another language.

In order to establish the origin of borrowings, the origin attributed by reference dictionaries and institutions (Real Academia Española, 2020; fun) will be followed.

This means that words like *junior* and *senior* (whose frequency and perhaps even their pronunciation may have changed due to the influence of English) will still be considered as latinisms, as DLE registers their adaptated versions (*júnior* and *sénior*) as such (and mentions no English influence). Similarly, the word *barista* might have entered the Spanish language via English, but RAE's Observatorio de Palabras considers it of Italian origin (and should therefore be annotated with OTHER label).

B.4 What an unasimilated lexical borrowing is not

In the previous section we provided an overview of what words will be considered as an unassimilated lexical borrowing for the sake of our annotation project. In this section we will cover what an unassimilated lexical borrowing is *not*.

There are several phenomena that are close enough to unassimilated borrowing and that can sometimes be mistaken with. In this section we will list what phenomena will not be considered as unassimilated lexical borrowings (and are therefore beyond the scope of our annotation project), as well as provide guidelines in order to distinguish these cases and adjudicate them.

We will focus on three main phenomena: assimilated borrowings, proper names and code-mixed

inclusions.

Figure 1 summarizes the decision steps that can be followed when deciding if a certain word should be labeled or not as a lexical borrowing.

B.4.1 Assimilated vs unassimilated borrowings

This annotation project aims to capture unassimilated lexical borrowings. As a general rule, all unadapted lexical borrowings should be tagged. This means that direct borrowings that have not gone through any morphological or orthographic modification process should be labeled.

Lexical adaptation, however, is a diachronic process and, as a result, what constitutes an unadapted borrowing is not clear-cut. The following guidelines define what borrowings will be considered as unassimilated (and therefore should be tagged) versus those that have already been integrated into the recipient language (and therefore should not be tagged).

B.4.2 Adapted borrowings

Words that have already gone through orthographical or morphological adaptation (such as *fútbol*, *líder*, *tuit* or *espóiler*) will be considered assimilated and therefore should not be labeled. Partial adaptations (such as *márketing*, where an accent has been added) will also be excluded.

Borrowings that have not been adapted but whose original spelling complies with graphophonological rules of Spanish (and are therefore unlikely to be further adapted, such as *bar*, *fan*, *web*, *internet*, *club*, *set* or *videoclip*) will be tagged as a borrowing or not not depending on how recent or emergent they are. In order to determine which unadapted, graphophonologically acceptable borrowings are to be annotated, the latest online version of the *Diccionario de la lengua española* (Real Academia Española, 2020) will be consulted (as of February 2021)[12]. If the DLE dictionary already registers the word with that meaning and with no italics or quotation marks, then it will be considered assimilated and therefore should not be tagged.

This means that a word like *set* (when used to refer to a collection of things, a television studio or a part of a tennis match) will be considered assimilated because it is already registered in DLE dictionary with no italics, and therefore should not be labeled as `ENG`. On the other hand, a word like *nude*, although its spelling also complies with Spanish graphophonological rules, will be considered an unassimilated borrowing because it has not been registered yet in the dictionary, and should therefore be tagged as such.

```
ganó el primer set

los tonos 'nude' (ENG)
```

It should be noted that this guideline only applies to lexical borrowings that comply with graphophonological rules of Spanish. Unadapted lexical borrowings that do not comply with graphophonological rules of Spanish (such as *show*, *look*, etc) will be tagged as borrowing, regardless of whether the word is included in the dictionary or not (although see section B.4.6 for exceptions to this).

It is important to emphasize that, in order for an unadapted graphophonologically-compliant borrowing to be considered assimilated it should be registered in the dictionary both without italics and with the corresponding meaning. For instance, a word like *top* (that is graphophonologically acceptable in Spanish) is registered in DLE with no italics, but it is only registered with the meaning of a piece of clothing. The word *top* as referring to the upper part of something (as in *top 5*) is not registered. Consequently, the borrowing *top* will be considered assimilated when referring to the piece

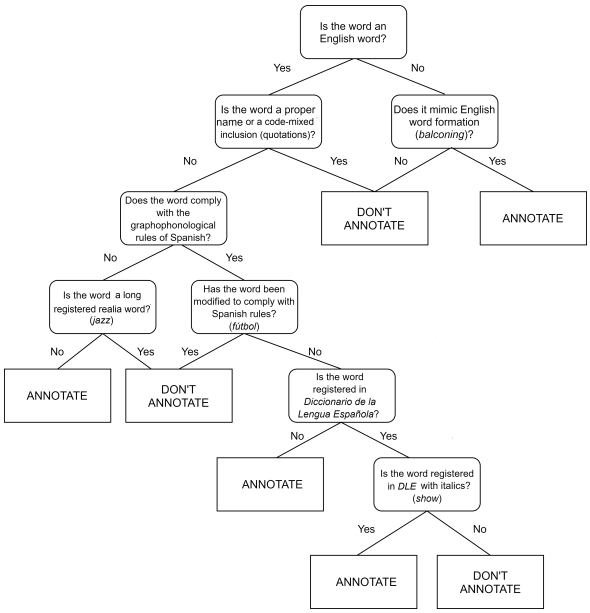

Figure 1: Decision steps to follow during the annotation process to decide whether to annotate a word as an anglicism.

[12] https://dle.rae.es/

of clothing, but unassimilated when used to talk about the best elements of a ranking or the upper part of something.

```
un top estampado
```

```
el [top] cinco de artistas (ENG)
```

```
la [top] desfiló (ENG)
```

Similarly, the word *post* will not be considered a borrowing when used as a prefix of Latin origin, but will be labeled with ENG when used to refer to something that is published on a social media platform.

```
el mundo post pandemia
```

```
un [post] de Facebook (ENG)
```

Additionally, assimilated borrowings can still be part of new unassimilated borrowings, in which case they will be labeled as such:

```
un [boys club] (ENG)
```

B.4.3 Words derived from foreign lexemes

Words derived from foreign lexemes that do not comply with Spanish orthotactics but that have been morphologically derived following the Spanish paradigm (such as *hacktivista*, *randomizar*, *shakespeariano*) will be considered assimilated and should therefore not be labeled as a borrowing.

Compound names where one of the lexemes is a borrowing will be labeled as a borrowing or not according to the degree of independence among the lexemes. A verb+noun compound (as *caza-clicks*) will not be labeled as a borrowing, because the elements are not independent from one another. However, noun-noun compounds where each of the lexemes work can work independent from one can be labeled as borrowings:

```
una casa-[loft] (ENG)
```

Similarly, prefixed borrowings will be labeled as a borrowing, as long as the borrowing keeps independence from the prefix:

```
la ex [influencer] (ENG)
```

For prefixed borrowings, it should be checked whether the prefix can also be considered part of the borrowing:

```
los [nano influencers] (ENG)
```

B.4.4 Number inflection

Unassimilated borrowings may be incorporated as invariable in number *los master*, with the same plural inflection that they had in the donor language (*los pappardelle*) or may form a new plural that is non-existant in the donor language (*los pappardelles*). For number inflection, we follow the same criteria that DLE (Real Academia Española, 2020) follows: a non-Italian plural like *pizzas* is still regarded as unadapted (and therefore should be written italicized even when the true Italian plural would be *pizze*). Consequently, non assimilated borrowings that have a non-cannonical plural inflection form will still be considered as an unassimilated borrowing and labeled as such.

```
una serie de animación de [mechas] (OTHER)
```

B.4.5 Pseudoanglicisms

Words that do not exist in English (or exist with a different meaning) but were coined following English morphological paradigm to imitate English words (such as *footing* or *balconing*) will be annotated as anglicisms.

```
la imagen del ‘[balconing]’ y las excursiones etílicas (ENG)
```

```
practicaba [footing] por la calle (ENG)
```

B.4.6 Realia words

Borrowings that refer to culture-specific elements (often called *realia words*) that were imported long ago but that have remained unadapted will not be tagged as borrowing. This means that if a borrowing is not adapted (i.e. its form remained exactly as it came from the donor language) but refers to a particular cultural object that came via the original language, that has been registered for a while in Spanish dictionaries and is not perceived as new anymore, then it will not be tagged as a borrowing, even if does not comply with graphophonologic rules of Spanish.

The purpose of this guideline is to account for cultural terms such as *pizza*, *whisky*, *jazz*, *blues*, *banjo* or *sheriff*. These are all borrowings that are reluctant to be adapted or translated, even when they have been around in the Spanish language for long. The reason is that they refer to cultural inventions (the name was imported along with the object

it refers to), and, given their cultural significance, they never competed with a Spanish equivalent and are seen as assimilated.

Therefore, unadapted borrowings that refer to cultural innovations (such as music, cooking, sport names etc) and that have been registered for long in the Spanish language[13] will not be tagged as emergent borrowings.

It should be noted that this only applies to borrowings that have been around enough time to be registered in dictionaries. A word like *hip hop* is a realia word, but it is still recent enough and has not been registered in the dictionary. In that case, it should be considered as unassimilated and tagged as such.

B.4.7 Latinisms

Borrowings that were introduced directly from Latin language (such as *deficit*, *curriculum*, etc) will not be considered emergent and therefore will not be tagged as a borrowing. However, it should be noted that unassimilated borrowings from other languages that happen to have a Latin etymology and and that are introduced with a distinct meaning (such as *adlib* or *premium* etc) will still be tagged as borrowings.

B.5 Borrowings vs names

B.5.1 Proper nouns

Non-Spanish proper nouns will not be tagged as borrowings. These include:

- person names: *Bernie Sanders*.
- organization names: *WikiLeaks*.
- product names: *Slack*.
- location names: *Times Square*.
- dates and celebrations: *St. Patrick's Day*, *Black Friday*.
- event names: *Brexit*, *procés*.
- social and political movements: *Black Lives Matter*, *MeToo*.
- treaties and documents: *New Deal*, *Privacy Shield*, *French Tech Visa*.
- titles of cultural productions: *Stranger Things*.

B.5.2 Borrowings in proper nouns

Borrowings that appear as part of proper nouns or named entities (such as book titles or organization names, as in *Los Hermanos Podcast*) will not be labeled as borrowings.

B.5.3 Proper nouns in borrowings

Multiword borrowings and expressions can sometimes include proper nouns. Even when a proper noun in isolation cannot be considered a borrowing, proper nouns within a borrowed expression will be considered part of the borrowing, as long as the proper noun is part of the borrowing and is used following the grammar rules of the donor language (for example, in an English noun noun compound):

`Tecnología [made in Spain] (ENG)`

`[Google cooking] (ENG)`

B.5.4 Names of institutions and political roles

Non-Spanish names that refer to political institutions (such as *Parlament* or *Bundestag*) or to political roles and figures (*lehendakari*, *president*, *conseller*) will be excluded and will not be tagged as borrowings.

B.5.5 Words derived from proper nouns

Words derived from proper nouns (via metonymy or eponymy) will not be tagged as a borrowing, as long as the relation with the proper noun they come from is transparent to the speaker such as:

- products: *un iPhone*, *un whatsapp*, *un bizum*, *un Scalextric*, *el Satisfyer*.
- works of arts: *un monet*
- characters: *un frankestein*.

However, borrowings that originated from a proper noun in the donor language but entered the Spanish language as common nouns and are currently recognized as such, will be labeled as borrowings. In order to adjudicate which of these words are still used in Spanish as proper names and which are common nouns, dictionaries and other reference works can be consulted.

[13] RAE dictionary https://dle.rae.es/, Mapa de diccionarios https://webfrl.rae.es/ntllet/SrvltGUILoginNtlletPub and CREA http://corpus.rae.es/creanet.html and CORPES https://webfrl.rae.es/CORPES/view/inicioExterno.view can be consulted

B.5.6 Names of peoples or languages

Names of peoples or languages (such as *inuit*) will not be labeled as borrowings, even if the word is borrowed from another language and is not registered in Spanish dictionaries.

B.5.7 Ficticious creatures

Unadapted names of fictitious creatures (such as *hobbit* or *troll*) will be labeled as a borrowing.

```
   En un agujero en el suelo
vivía un [hobbit] (ENG)
```

B.5.8 Scientific units

Unadapted borrowings that refer to widespread scientific units (such as *hertz*, *newton*, *byte*, etc) will be considered assimilated and should not be tagged as a borrowing

B.5.9 Names of species

Scientific names of a species (such as Latin names) will not be tagged as a lexical borrowing (*anisakis*). Names of fruit, vegetable and plant varieties (such as *manzana golden*, *patatas Kennebec* or *aguacate Hass*) will also be excluded.

B.6 Borrowings vs other code-mixed inclusions

Borrowing (using units from one language in another language) and code-switching (intertwining segments of different languages in the same discourse) have frequently been described as a continuum (Clyne et al., 2003), with a fuzzy frontier between the two. As a result, it can be difficult to tell the difference between borrowing and other code-mixed inclusions. The following guidelines can assist annotators adjudicate edge cases.

When in doubt while dealing with code-mixed inclusion, the annotator may find it helpful to ask the following question as a rule of thumb: would it make sense to have this non-Spanish word registered in a dictionary of Spanish? If the answer is no (for instance, because the word reflects the literal quotation of what someone said or because the inclusions is metalinguistic usage rather than borrowing), then we are probably not in front of a borrowing but of another type of code-mixed inclusion (and should not be tagged as a borrowing).

B.6.1 Acronyms and acronym expansions

We consider acronyms to be a different phenomenon from borrowings. Consequently, acronyms will not be tagged as a borrowing, even if the acronym is of non-Spanish origin

```
un lector de CD
```

An acronym however may be tagged as a borrowing if it appears as part of a borrowed multiword expression, as in *CD player*, *peak TV*, *PC gaming*:

```
un [CD player] (ENG)
```

Acronym expansions, that is, the expansion of an acronym into the words that form the acronym (that is usually added in between brackets after an acronym has been introduced) will also not be considered a borrowing:

```
   La técnica de PCR (protein
chain reaction)
```

It is important to note that for a sequence to be considered as an acronym expansion it must appear after the acronym has been introduced and serve as a gloss to it (so that it expands what the letters in the acronym stand for). Usages where the full sequence is introduced in the text and later on acronymized for the sake of brevity can still be considered as borrowings.

```
   Utilizaron técnicas de
[Machine Learning] (también
conocido como ML) (ENG)
```

B.6.2 Digits

Similarly to proper nouns, digits in isolation cannot be considered borrowings. As a result, we cannot take for granted that digits within the surroundings of a borrowing will automatically be part of the borrowing.

```
[top ten] (ENG)
```

```
[top] 10 (ENG)
```

However, if the word order of the tokens makes it clear that the digit is part of a multitoken borrowing (because the order complies with the grammatical structure of an English noun-noun compound), we can label it as part of the borrowing:

```
los [10% banks] (ENG)
```

B.6.3 Metalinguistic usage

Non-Spanish words that appear to refer to the word itself in linguistic discourse and do not cover a lexical gap will not be tagged as a borrowing:

```
   El término viene de la palabra
'ghost', 'fantasma' en inglés
```

It should be noted that the newer, less adapted, less transparent a new word is, the more likely that the speaker will be aware of the decoding difficulty it may pose to the reader and will decide to add some sort of metalinguistic strategy or awareness around it, in the form of metacomments, word-pointers, meaning explanations, etc (*known as*, *so called*). Borrowings with these types of signals will still be considered borrowings, as long as they are covering a lexical gap.

True metalinguistic usage where the foreign word covers no lexical gap but exclusively provides linguistic information (such as etymological information) will not be considered a borrowing.

B.6.4 Literal quotations

Words or sequences in languages other than Spanish that are reflecting literally what someone said or wrote (as in a quotation, a statement or a slogan) will not be considered a borrowing.

`El eslogan 'Make America Great Again'`

`Es uno de los primeros resultados de Google cuando alguien busca "remote work in Spain" (trabajo en remoto en España).`

B.6.5 Expressions

In general terms, multiword borrowings will be tagged as borrowings. However, phrases and expressions that are not integrated into the sentence will be excluded. This means that autonomous expressions that are rather code switched sentences (rather than real borrowings) that work as a unit totally independently of the rest of the linguistic context (and that we would not expect to be registered in a dictionary) will not be considered or tagged as a borrowing.

`La innovación y la competencia tan escasas en la radiotelevisión o peor aún en Internet ("the winners takes all" o "most").`

B.7 Limitations of these guidelines

These guidelines are intended to assist annotators when labeling lexical borrowings. These guidelines, however, were created with a specific goal in mind (to capture unassimilated English lexical borrowings from a corpus of Spanish newswire) and may not be suitable if applied to a project with a different scope. These are some of the shortcomings and limitations that these guidelines may have.

B.7.1 Text genre

These guidelines were designed to specifically capture borrowings in a corpus of Spanish newswire. Newswire is a very specific genre of text that by no means represent the whole of a language (Plank, 2016).

B.7.2 Donor language

These guidelines were created with English lexical borrowings in mind, which are the most frequent source of borrowing today in the Spanish press. Although the criteria can be applied to other languages as well (and in fact the annotation tagset we propose includes the tag `OTHER` to account for borrowings from other languages other than English), a more fine-grained approach would require further guidelines.

B.7.3 Synchronic approach to borrowing

This project approaches emergent, unassimilated lexical borrowing in a synchronic fashion. The process of borrowing and the notion of assimilation is, however, time-dependent. A diachronic approach to lexical borrowing would require a wider scope, a different theoretical framework and an expanded set of criteria.

B.7.4 Geographic variety

The guidelines in this document were designed to capture borrowings used in Spanish newspapers, that is, written in the variety of Spanish that is spoken in Spain and may not be suitable to account for other dialects. For instance, according to the guidelines we have just introduced, a word like *living* (that is used heavily in some Latin American varieties to refer to the living room) would be considered unassimilated. It is arguable whether these criteria would be suitable for a project that tried to capture emergent lexical borrowings in Argentinian text, for example.