Proceedings of the LREC 2020 – 4th Workshop on Computational Approaches to Code Switching, pages 1–8
Language Resources and Evaluation Conference (LREC 2020), Marseille, 11–16 May 2020
European Language Resources Association (ELRA), licensed under CC-BY-NC

An Annotated Corpus of Emerging Anglicisms in Spanish Newspaper Headlines

Elena Álvarez-Mellado
Department of Computer Science, Brandeis University
415 South St, Waltham, MA 02453
ealvarezmellado@brandeis.edu

Abstract

The extraction of anglicisms (lexical borrowings from English) is relevant both for lexicographic purposes and for NLP downstream tasks. We introduce a corpus of European Spanish newspaper headlines annotated with anglicisms and a baseline model for anglicism extraction. In this paper we present: (1) a corpus of 21,570 newspaper headlines written in European Spanish annotated with emergent anglicisms and (2) a conditional random field baseline model with handcrafted features for anglicism extraction. We present the newspaper headlines corpus, describe the annotation tagset and guidelines and introduce a CRF model that can serve as baseline for the task of detecting anglicisms. The presented work is a first step towards the creation of an anglicism extractor for Spanish newswire.

Keywords: borrowing extraction, anglicism, newspaper corpus

1. Introduction

The study of English influence in the Spanish language has been a hot topic in Hispanic linguistics for decades, particularly concerning lexical borrowing or anglicisms (Gómez Capuz, 2004; Lorenzo, 1996; Medina López, 1998; Menéndez et al., 2003; Núñez Nogueroles, 2017a; Pratt, 1980; Rodríguez González, 1999).

Lexical borrowing is a phenomenon that affects all languages and constitutes a productive mechanism for word-formation, especially in the press. Chesley and Baayen (2010) estimated that a reader of French newspapers encountered a new lexical borrowing for every 1,000 words. In Chilean newspapers, lexical borrowings account for approximately 30% of neologisms, 80% of those corresponding to English loanwords (Gerding et al., 2014).

Detecting lexical borrowings is relevant both for lexicographic purposes and for NLP downstream tasks (Alex et al., 2007; Tsvetkov and Dyer, 2016). However, strategies to track and register lexical borrowings have traditionally relied on manual review of corpora.

In this paper we present: (1) a corpus of newspaper headlines in European Spanish annotated with emerging anglicisms and (2) a CRF baseline model for anglicism automatic extraction in Spanish newswire.

2. Related Work

Corpus-based studies of English borrowings in Spanish media have traditionally relied on manual evaluation of either previously compiled general corpora such as CREA[1] (Balteiro, 2011; Núñez Nogueroles, 2016; Núñez Nogueroles, 2018b; Oncíns Martínez, 2012), either new tailor-made corpora designed to analyze specific genres, varieties or phenomena (De la Cruz Cabanillas and Martínez, 2012; Diéguez, 2004; Gerding Salas et al., 2018; Núñez Nogueroles, 2017b; Patzelt, 2011; Rodríguez Medina, 2002; Vélez Barreiro, 2003).

In terms of automatic detection of anglicisms, previous approaches in different languages have mostly depended on resource lookup (lexicon or corpus frequencies), character n-grams and pattern matching. Alex (2008b) combined lexicon lookup and a search engine module that used the web as a corpus to detect English inclusions in a corpus of German texts and compared her results with a maxent Markov model. Furiassi and Hofland (2007) explored corpora lookup and character n-grams to extract false anglicisms from a corpus of Italian newspapers. Andersen (2012) used dictionary lookup, regular expressions and lexicon-derived frequencies of character n-grams to detect anglicism candidates in the Norwegian Newspaper Corpus (NNC) (Hofland, 2000), while Losnegaard and Lyse (2012) explored a Machine Learning approach to anglicism detection in Norwegian by using TiMBL (Tilburg Memory-Based Learner, an implementation of a k-nearest neighbor classifier) with character trigrams as features. Garley and Hockenmaier (2012) trained a maxent classifier with character n-gram and morphological features to identify anglicisms in German online communities.

In Spanish, Serigos (2017a) extracted anglicisms from a corpus of Argentinian newspapers by combining dictionary lookup (aided by TreeTagger and the NLTK lemmatizer) with automatic filtering of capitalized words and manual inspection. In Serigos (2017b), a character n-gram module was added to estimate the probabilities of a word being English or Spanish. Moreno Fernández and Moreno Sandoval (2018) used different pattern-matching filters and lexicon lookup to extract anglicism cadidates from a corpus of tweets in US Spanish.

Work within the code-switching community has also dealt with language identification on multilingual corpora. Due to the nature of code-switching, these models have primarily focused on oral copora and social media datasets (Aguilar et al., 2018; Molina et al., 2016; Solorio et al., 2014). In the last shared task of language identification in code-switched data (Molina et al., 2016), approaches to English-Spanish included CRFs models (Al-Badrashiny and Diab, 2016; Shrestha, 2016; Sikdar and Gambäck, 2016; Xia, 2016), logistic regression (Shirvani et al., 2016) and LSTMs models (Jaech et al., 2016; Samih et al., 2016).

[1] `http://corpus.rae.es/creanet.html`

The scope and nature of lexical borrowing is, however, somewhat different to that of code-switching. In fact, applying code-switching models to lexical borrowing detection has previously proved to be unsuccessful, as they tend to overestimate the number of anglicisms (Serigos, 2017b). In the next section we address the differences between both phenomena and set the scope of this project.

3. Anglicism: Scope of the Phenomenon

Linguistic borrowing can be defined as the transference of linguistic elements between two languages. Borrowing and code-switching have frequently been described as a continuum (Clyne et al., 2003), with a fuzzy frontier between the two. As a result, a precise definition of what borrowing is remains elusive (Gómez Capuz, 1997) and some authors prefer to talk about code-mixing in general (Alex, 2008a) or "lone other-language incorporations" (Poplack and Dion, 2012).

Lexical borrowing in particular 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 (Onysko, 2007; Poplack et al., 1988). By definition, code-switches are not integrated into a recipient language, unlike established loanwords (Poplack, 2012). While code-switches are usually fluent multiword interferences that normally comply with grammatical restrictions in both languages and that 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" origin disappears (Lipski, 2005).

In terms of approaching the problem, automatic code-switching identification has been framed as a sequence modeling problem where every token receives a language ID label (as in a POS-tagging task). Borrowing detection, on the other hand, while it can also be transformed into a sequence labeling problem, is an extraction task, where only certain spans of texts will be labeled (in the fashion of a NER task).

Various typologies have been proposed that aim to classify borrowings according to different criteria, both with a cross-linguistic perspective and also specifically aimed to characterize English inclusions in Spanish (Gómez Capuz, 1997; Haspelmath, 2008; Núñez Nogueroles, 2018a; Pratt, 1980). In this work, we will be focusing on unassimilated lexical borrowings (sometimes called *foreignisms*), i.e. words from English origin that are introduced into Spanish without any morphological or orthographic adaptation.

4. Corpus description and annotation

4.1. Corpus description

In this subsection we describe the characteristics of the corpus. We first introduce the main corpus, with the usual train/development/test split that was used to train, tune and evaluate the model. We then present an additional test set that was designed to assess the performance of the model on more naturalistic data.

4.1.1. Main Corpus

The main corpus consists of a collection of monolingual newspaper headlines written in European Spanish. The corpus contains 16,553 headlines, which amounts to 244,114 tokens. Out of those 16,553 headlines, 1,109 contain at least one anglicism. The total number of anglicisms is 1,176 (most of them are a single word, although some of them were multiword expressions). The corpus was divided into training, development and test set. The proportions of headlines, tokens and anglicisms in each corpus split can be found in Table 1.

The headlines in this corpus come from the Spanish newspaper *eldiario.es*[2], a progressive online newspaper based in Spain. *eldiario.es* is one of the main national newspapers from Spain and, to the best of our knowledge, the only one that publishes its content under a Creative Commons license, which made it ideal for making the corpus publicly available[3].

Set	Headlines	Tokens	Headlines with anglicisms	Anglicisms	Other borrowings
Train	10,513	154,632	709	747	40
Dev	3,020	44,758	200	219	14
Test	3,020	44,724	202	212	13
Suppl. test	5,017	81,551	122	126	35

Table 1: Number of headlines, tokens and anglicisms per corpus subset.

The headlines were extracted from the newspaper website through web scraping and range from September 2012 to January 2020. Only the following sections were included: economy, technology, lifestyle, music, TV and opinion. These sections were chosen as they were the most likely to contain anglicisms. The proportion of headlines with anglicisms per section can be found in Table 2.

Section	Percentage of anglicisms
Opinion	2.54%
Economy	3.70%
Lifestyle	6.48%
TV	8.83%
Music	9.25%
Technology	15.37%

Table 2: Percentage of headlines with anglicisms per section.

Using headlines (instead of full articles) was beneficial for several reasons. First of all, annotating a headline is faster and easier than annotating a full article; this helps ensure that a wider variety of topics will be covered in the corpus. Secondly, anglicisms are abundant in headlines, because they are frequently used as a way of calling the attention of the reader (Furiassi and Hofland, 2007). Finally, borrowings that make it to the headline are likely to be particularly salient or relevant, and therefore are good candidates for being extracted and tracked.

[2] http://www.eldiario.es/

[3] Both the corpus and the baseline model (Section 5) can be found at https://github.com/lirondos/lazaro.

4.1.2. Supplemental Test Set

In addition to the usual train/development/test split we have just presented, a supplemental test set of 5,017 headlines was collected. The headlines included in this additional test set also belong to *eldiario.es*. These headlines were retrieved daily through RSS during February 2020 and included all sections from the newspaper. The headlines in the supplemental corpus therefore do not overlap in time with the main corpus and include more sections. The number of headlines, tokens and anglicisms in the supplemental test set can be found in Table 1.

The motivation behind this supplemental test set is to assess the model performance on more naturalistic data, as the headlines in the supplemental corpus (1) belong to the future of the main corpus and (2) come from a less borrowing-dense sample. This supplemental test set better mimics the real scenario that an actual anglicism extractor would face and can be used to assess how well the model generalizes to detect anglicisms in any section of the daily news, which is ultimately the aim of this project.

4.2. Annotation guidelines

The term *anglicism* covers a wide range of linguistic phenomena. Following the typology proposed by Gómez Capuz (1997), we focused on direct, unadapted, 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. Other phenomena such as semantic calques, syntactic anglicisms, acronyms and proper names were considered beyond the scope of this annotation project.

Lexical borrowings can be adapted (the spelling of the word is modified to comply with the phonological and orthographic patterns of the recipient language) or unadapted (the word preserves its original spelling). For this annotation task, adapted borrowings were ignored and only unadapted borrowings were annotated. Therefore, Spanish adaptations of anglicisms like *fútbol* (from *football*), *mitin* (from *meeting*) and such were not annotated as borrowings. Similarly, words derived from foreign lexemes that do not comply with Spanish orthotactics but that have been morphologically derived following the Spanish paradigm (*hacktivista*, *hackear*, *shakespeariano*) were not annotated either. However, pseudo-anglicisms (words that are formed as if they were English, but do not exist in English, such as *footing* or *balconing*) were annotated.

Words that were not adapted but whose original spelling complies with graphophonological rules of Spanish (and are therefore unlikely to be ever adapted, such as *web*, *internet*, *fan*, *club*, *videoclip*) were annotated or not depending on how recent or emergent they were. After all, a word like *club*, that has been around in Spanish language for centuries, cannot be considered emergent anymore and, for this project, would not be as interesting to retrieve as real emerging anglicisms. The notion of *emergent* is, however, time-dependent and quite subjective: in order to determine which unadapted, graphophonologically acceptable borrowings were to be annotated, the online version of the *Diccionario de la lengua española*[4] (Real Academia Española, 2014) was consulted. This dictionary is compiled by the Royal Spanish Academy, a prescriptive institution on Spanish language. This decision was motivated by the fact that, if a borrowing was already registered by this dictionary (that has conservative approach to language change) and is considered assimilated (that is, the institution recommended no italics or quotation marks to write that word) then it could be inferred that the word was not emergent anymore.

Although the previous guidelines covered most cases, they proved insufficient. Some anglicisms were unadapted (they preserved their original spelling), unacceptable according to the Spanish graphophonological rules, and yet did not satisfy the condition of being emergent. That was the case of words like *jazz* or *whisky*, words that do not comply with Spanish graphophonological rules but that were imported decades ago, cannot be considered emergent anymore and are unlikely to ever be adapted into the Spanish spelling system. To adjudicate these examples on those cases, the criterion of pragmatic markedness proposed by Winter-Froemel and Onysko (2012) (that distinguishes between catachrestic and non-catachrestic borrowing) was applied: if a borrowing was not adapted (i.e. its form remained exactly as it came from English) but referred to a particular invention or innovation that came via the English language, that was not perceived as new anymore and that had never competed with a Spanish equivalent, then it was ignored. This criteria proved to be extremely useful to deal with old unadapted anglicisms in the fields of music and food. Figure 1 summarizes the decision steps followed during the annotation process.

The corpus was annotated by a native speaker of Spanish using Doccano[5] (Nakayama et al., 2018). The annotation tagset includes two labels: `ENG`, to annotate the English borrowings just described, and `OTHER`. This `OTHER` tag was used to tag lexical borrowings from languages other than English. After all, although English is today by far the most prevalent donor of borrowings, there are other languages that also provide new borrowings to Spanish. Furthermore, the tag `OTHER` allows to annotate borrowings such as *première* or *tempeh*, borrowings that etymologically do not come from English but that have entered the Spanish language via English influence, even when their spelling is very different to English borrowings. In general, we considered that having such a tag could also help assess how successful a classifier is detecting foreign borrowings in general in Spanish newswire (without having to create a label for every possible donor language, as the number of examples would be too sparse). In total, the training set contained 40 entities labeled as `OTHER`, the development set contained 14 and the test set contained 13. The supplemental test set contained 35 `OTHER` entities.

5. Baseline Model

A baseline model for automatic extraction of anglicisms was created using the annotated corpus we just presented as training material. As mentioned in Section 3, the task of detecting anglicisms can be approached as a sequence

[4] `https://dle.rae.es/`

[5] `https://github.com/chakki-works/doccano`

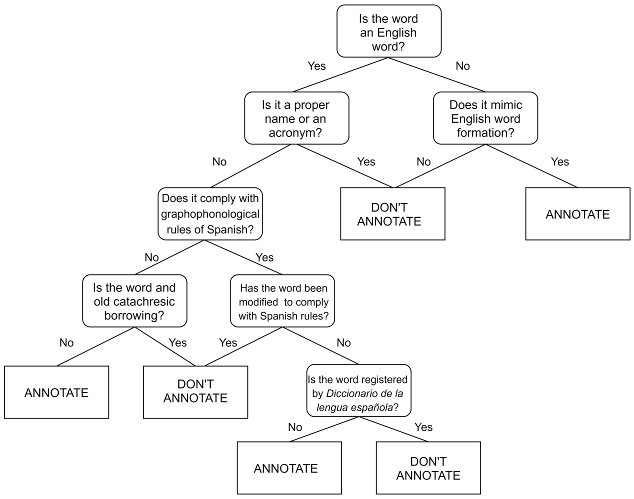
Is the word an English word?
Yes
No
Is it a proper name or an acronym?
Does it mimic English word formation?
No
Yes
No
Yes
Does it comply with graphophonological rules of Spanish?
DON'T ANNOTATE
ANNOTATE
No
Yes
Is the word and old catachresic borrowing?
Has the word been modified to comply with Spanish rules?
No
Yes
Yes
No
ANNOTATE
DON'T ANNOTATE
Is the word registered by *Diccionario de la lengua española*?
No
Yes
ANNOTATE
DON'T ANNOTATE

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

labeling problem where only certain spans of texts will be labeled as anglicism (in a similar way to an NER task). The chosen model was conditional random field model (CRF), which was also the most popular model in both Shared Tasks on Language Identification for Code-Switched Data (Molina et al., 2016; Solorio et al., 2014).

The model was built using `pycrfsuite`[6] (Korobov and Peng, 2014), the Python wrapper for `crfsuite`[7] (Okazaki, 2007) that implements CRF for labeling sequential data. It also used the `Token` and `Span` utilities from `spaCy`[8] library (Honnibal and Montani, 2017).

The following handcrafted features were used for the model:

- Bias feature
- Token feature
- Uppercase feature (y/n)
- Titlecase feature (y/n)
- Character trigram feature
- Quotation feature (y/n)
- Word suffix feature (last three characters)
- POS tag (provided by `spaCy` utilities)
- Word shape (provided by `spaCy` utilities)
- Word embedding (see Table 3)

Given that anglicisms can be multiword expressions (such as *best seller, big data*) and that those units should be treated as one borrowing and not as two independent borrowings, we used multi-token BIO encoding to denote the boundaries of each span (Ramshaw and Marcus, 1999). A window of two tokens in each direction was set for the feature extractor. The algorithm used was gradient descent with the L-BFGS method.

The model was tuned on the development set doing grid search; the hyperparameters considered were c1 (L1 regularization coefficient: 0.01, 0.05, 0.1, 0.5, 1.0), c2 (L2 regularization coefficient: 0.01, 0.05, 0.1, 0.5, 1.0), embedding scaling (0.5, 1.0, 2.0, 4.0), and embedding type (Bojanowski et al., 2017; Cañete, 2019; Cardellino, 2019; Grave et al., 2018; Honnibal and Montani, 2017; Pérez, 2017a; Pérez, 2017b) (see Table 3). The best results were obtained with c1 = 0.05, c2 = 0.01, scaling = 0.5 and word2vec Spanish embeddings by Cardellino (2019). The threshold for the stopping criterion delta was selected through observing the loss during preliminary experiments (delta = $1\text{e} - 3$).

Author	**Algorithm**	**# Vectors**	**Dimensions**
Bojanowski et al. (2017)	FastText	985,667	300
Cañete (2019)	FastText	1,313,423	300
Cardellino (2019)	word2vec	1,000,653	300
Grave et al. (2018)	FastText	2,000,001	300
Honnibal and Montani (2017)	word2vec	534,000	50
Pérez (2017a)	FastText	855,380	300
Pérez (2017b)	GloVe	855,380	300

Table 3: Types of embeddings tried.

In order to assess the significance of the the handcrafted features, a feature ablation study was done on the tuned model, ablating one feature at a time and testing on the development set. Due to the scarcity of spans labeled with the `OTHER` tag on the development set (only 14) and given that the main purpose of the model is to detect anglicisms, the baseline model was run ignoring the `OTHER` tag both during tuning and the feature ablation experiments. Table 4 displays the results on the development set with all features and for the different feature ablation runs. The results show that all features proposed for the baseline model contribute to the results, with the character trigram feature being the one that has the biggest impact on the feature ablation study.

Features	**Precision**	**Recall**	**F1 score**	**F1 change**
All features	97.84	**82.65**	**89.60**	
− Bias	96.76	81.74	88.61	−0.99
− Token	95.16	80.82	87.41	−2.19
− Uppercase	97.30	82.19	89.11	−0.49
− Titlecase	96.79	**82.65**	89.16	−0.44
− Char trigram	96.05	77.63	85.86	−3.74
− Quotation	97.31	**82.65**	89.38	−0.22
− Suffix	97.30	82.19	89.11	−0.49
− POS tag	**98.35**	81.74	89.28	−0.32
− Word shape	96.79	**82.65**	89.16	−0.44
− Word embedding	95.68	80.82	87.62	−1.98

Table 4: Ablation study results on the development test.

6. Results

The baseline model was then run on the test set and the supplemental test set with the set of features and hyperpa-

[6] `https://github.com/scrapinghub/python-crfsuite`

[7] `https://github.com/chokkan/crfsuite`

[8] `https://spacy.io/`

rameters mentioned on Section 5. Table 5 displays the results obtained. The model was run both with and without the OTHER tag. The metrics for ENG display the results obtained only for the spans labeled as anglicisms; the metrics for OTHER display the results obtained for any borrowing other than anglicisms. The metrics for BORROWING discard the type of label and consider correct any labeled span that has correct boundaries, regardless of the label type (so any type of borrowing, regardless if it is ENG or OTHER). In all cases, only full matches were considered correct and no credit was given to partial matching, i.e. if only *fake* in *fake news* was retrieved, it was considered wrong and no partial score was given.

Results on all sets show an important difference between precision and recall, precision being significantly higher than recall. There is also a significant difference between the results obtained on development and test set (F1 = 89.60, F1 = 87.82) and the results on the supplemental test set (F1 = 71.49). The time difference between the supplemental test set and the development and test set (the headlines from the the supplemental test set being from a different time period to the training set) can probably explain these differences.

Comparing the results with and without the OTHER tag, it seems that including it on the development and test set produces worse results (or they remain roughly the same, at best). However, the best precision result on the supplemental test was obtained when including the OTHER tag and considering both ENG and OTHER spans as BORROWING (precision = 87.62). This is caused by the fact that, while the development and test set were compiled from anglicism-rich newspaper sections (similar to the training set), the supplemental test set contained headlines from all the sections in the newspaper, and therefore included borrowings from other languages such as Catalan, Basque or French. When running the model without the OTHER tag on the supplemental test set, these non-English borrowings were labeled as anglicisms by the model (after all, their spelling does not resemble Spanish spelling), damaging the precision score. When the OTHER tag was included, these non-English borrowings got correctly labeled as OTHER, improving the precision score. This proves that, although the OTHER tag might be irrelevant or even damaging when testing on the development or test set, it can be useful when testing on more naturalistic data, such as the one in the supplemental test set.

Concerning errors, two types of errors were recurrent among all sets: long titles of songs, films or series written in English were a source of false positives, as the model tended to mistake some of the uncapitalized words in the title for anglicisms (for example, *it darker* in "‘You want it darker’, la oscura y brillante despedida de Leonard Cohen"). On the other hand, anglicisms that appear on the first position of the sentence (and were, therefore, capitalized) were consistently ignored (as the model probably assumed they were named entities) and produced a high number of false negatives (for example, *vamping* in "Vamping: la recurrente leyenda urbana de la luz azul ‘asesina’").

The results on Table 5 cannot, however, be compared to the ones reported by previous work: the metric that we report is span F-measure, as the evaluation was done on span level (instead of token level) and credit was only given to full matches. Secondly, there was no Spanish tag assigned to non-borrowings, that means that no credit was given if a Spanish token was identified as such.

Set	**Precision**	**Recall**	**F1 score**
Development set (− OTHER)	97.84	82.65	89.60
Development set (+ OTHER)			
ENG	96.79	82.65	89.16
OTHER	100.0	28.57	44.44
BORROWING	96.86	79.40	87.26
Test set (− OTHER)	95.05	81.60	87.82
Test set (+ OTHER)			
ENG	95.03	81.13	87.53
OTHER	100.0	46.15	63.16
BORROWING	95.19	79.11	86.41
Supplemental test set (− OTHER)	83.16	62.70	71.49
Supplemental test set (+ OTHER)			
ENG	82.65	64.29	72.32
OTHER	100.0	20.0	33.33
BORROWING	87.62	57.14	69.17

Table 5: Results on test set and supplemental test set.

7. Future Work

This is an on-going project. The corpus we have just presented is a first step towards the development of an extractor of emerging anglicisms in the Spanish press. Future work includes: assessing whether to keep the OTHER tag, improving the baseline model (particularly to improve recall), assessing the suitability and contribution of different sets of features and exploring different models. In terms of the corpus development, the training set is now closed and stable, but the test set could potentially be increased in order to have more and more diverse anglicisms.

8. Conclusions

In this paper we have presented a new corpus of 21,570 newspaper headlines written in European Spanish. The corpus is annotated with emergent anglicisms and, up to our very best knowledge, is the first corpus of this type to be released publicly. We have presented the annotation scope, tagset and guidelines, and we have introduced a CRF baseline model for anglicism extraction trained with the described corpus. The results obtained show that the the corpus and baseline model are appropriate for automatic anglicism extraction.

9. Acknowledgements

The author would like to thank Constantine Lignos for his feedback and advice on this project.

10. 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.

Al-Badrashiny, M. and Diab, M. (2016). The George Washington University System for the Code-Switching Workshop Shared Task 2016. In *Proceedings of the Second Workshop on Computational Approaches to Code Switching*, pages 108–111, Austin, Texas, November. Association for Computational Linguistics.

Alex, B., Dubey, A., and Keller, F. (2007). Using foreign inclusion detection to improve parsing performance. In *Proceedings of the 2007 Joint Conference on Empirical Methods in Natural Language Processing and Computational Natural Language Learning (EMNLP-CoNLL)*, pages 151–160.

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

Alex, B. (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, May. European Language Resources Association (ELRA).

Andersen, G. (2012). Semi-automatic approaches to Anglicism detection in Norwegian corpus data. In Cristiano Furiassi, et al., editors, *The anglicization of European lexis*, pages 111–130.

Balteiro, I. (2011). A reassessment of traditional lexicographical tools in the light of new corpora: sports anglicisms in Spanish. *International Journal of English Studies*, 11(2):23–52.

Bojanowski, P., Grave, E., Joulin, A., and Mikolov, T. (2017). Enriching word vectors with subword information. *Transactions of the Association for Computational Linguistics*, 5:135–146.

Chesley, P. and Baayen, R. H. (2010). Predicting new words from newer words: Lexical borrowings in French. *Linguistics*, 48(6):1343.

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

De la Cruz Cabanillas, I. and Martínez, C. T. (2012). Email or correo electrónico? Anglicisms in Spanish. *Revista española de lingüística aplicada*, (1):95–118.

Diéguez, M. I. (2004). El anglicismo léxico en el discurso económico de divulgación científica del español de Chile. *Onomázein*, 2(10):117–141.

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

Garley, M. and Hockenmaier, J. (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, July. Association for Computational Linguistics.

Gerding, C., Fuentes, M., Gómez, L., and Kotz, G. (2014). Anglicism: An active word-formation mechanism in Spanish. *Colombian Applied Linguistics Journal*, 16(1):40–54.

Gerding Salas, C., Cañete González, P., and Adam, C. (2018). Neología sintagmática anglicada en español: Calcos y préstamos. *Revista signos*, 51(97):175–192.

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

Grave, E., Bojanowski, P., Gupta, P., Joulin, A., and Mikolov, T. (2018). Learning word vectors for 157 languages. In *Proceedings of the International Conference on Language Resources and Evaluation (LREC 2018)*.

Gómez Capuz, J. (2004). *Los préstamos del español: lengua y sociedad*. Cuadernos de Lengua Española. Arco Libros.

Haspelmath, M. (2008). Loanword typology: Steps toward a systematic cross-linguistic study of lexical borrowability. *Empirical Approaches to Language Typology*, 35:43.

Hofland, K. (2000). A self-expanding corpus based on newspapers on the web. In *Proceedings of the Second International Conference on Language Resources and Evaluation (LREC'00)*, Athens, Greece, May. European Language Resources Association (ELRA).

Jaech, A., Mulcaire, G., Ostendorf, M., and Smith, N. A. (2016). A neural model for language identification in code-switched tweets. In *Proceedings of the Second Workshop on Computational Approaches to Code Switching*, pages 60–64, Austin, Texas, November. Association for Computational Linguistics.

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.

Lorenzo, E. (1996). *Anglicismos hispánicos*. Biblioteca románica hispánica: Estudios y ensayos. Gredos.

Losnegaard, G. S. and Lyse, G. I. (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.

Medina López, J. (1998). *El anglicismo en el español actual*. Cuadernos de lengua española. Arco Libros.

Menéndez, F., Menéndez, M., and Morales, H. (2003). *El desplazamiento lingüístico del español por el inglés*. Cátedra lingüística. Cátedra.

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.

Moreno Fernández, F. and Moreno Sandoval, A. (2018). Configuración lingüística de anglicismos procedentes de Twitter en el español estadounidense. *Revista signos*, 51(98):382–409.

Núñez Nogueroles, E. E. (2016). Anglicisms in CREA: a quantitative analysis in Spanish newspapers. *Language* *design: journal of theoretical and experimental linguistics*, 18:0215–242.

Núñez Nogueroles, E. (2017a). An up-to-date review of the literature on anglicisms in spanish. *Diálogo de la Lengua, IX*, pages 1–54.

Núñez Nogueroles, E. E. (2017b). Typographical, orthographic and morphological variation of anglicisms in a corpus of Spanish newspaper texts. *Revista Canaria de Estudios Ingleses*, (75):175–190.

Núñez Nogueroles, E. E. (2018a). A comprehensive definition and typology of anglicisms in present-day Spanish. *Epos: Revista de filología*, (34):211–237.

Núñez Nogueroles, E. E. (2018b). A corpus-based study of anglicisms in the 21st century spanish press. *Analecta Malacitana (AnMal electrónica)*, (44):123–159.

Oncíns Martínez, J. L. (2012). Newly-coined anglicisms in contemporary Spanish. a corpus-based approach. In Cristiano Furiassi, et al., editors, *The anglicization of European lexis*, pages 217–238.

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

Patzelt, C. (2011). The impact of English on Spanish-language media in the usa. In *Multilingual Discourse Production: Diachronic and Synchronic Perspectives*, volume 12, page 257. John Benjamins Publishing.

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

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

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

Pratt, C. (1980). *El anglicismo en el español peninsular contemporáneo*, volume 308. Gredos.

Ramshaw, L. A. and Marcus, M. P. (1999). Text chunking using transformation-based learning. In *Natural language processing using very large corpora*, pages 157–176. Springer.

Rodríguez Medina, M. J. (2002). Los anglicismos de frecuencia sintácticos en español: estudio empírico. *RAEL. Revista electrónica de lingüística aplicada*.

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

Samih, Y., Maharjan, S., Attia, M., Kallmeyer, L., and Solorio, T. (2016). Multilingual code-switching identification via LSTM recurrent neural networks. In *Proceedings of the Second Workshop on Computational Approaches to Code Switching*, pages 50–59, Austin, Texas, November. Association for Computational Linguistics.

Serigos, J. (2017a). Using distributional semantics in loanword research: A concept-based approach to quantifying semantic specificity of anglicisms in Spanish. *International Journal of Bilingualism*, 21(5):521–540.

Serigos, J. R. L. (2017b). *Applying corpus and computational methods to loanword research: new approaches to Anglicisms in Spanish*. Ph.D. thesis, The University of Texas at Austin.

Shirvani, R., Piergallini, M., Gautam, G. S., and Chouikha, M. (2016). The Howard University System submission for the Shared Task in Language Identification in Spanish-English Codeswitching. In *Proceedings of the Second Workshop on Computational Approaches to Code Switching*, pages 116–120, Austin, Texas, November. Association for Computational Linguistics.

Shrestha, P. (2016). Codeswitching detection via lexical features in conditional random fields. In *Proceedings of the Second Workshop on Computational Approaches to Code Switching*, pages 121–126, Austin, Texas, November. Association for Computational Linguistics.

Sikdar, U. K. and Gambäck, B. (2016). Language identification in code-switched text using conditional random fields and Babelnet. In *Proceedings of the Second Workshop on Computational Approaches to Code Switching*, pages 127–131, Austin, Texas, November. Association for Computational Linguistics.

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.

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

Vélez Barreiro, M. (2003). *Anglicismos en la prensa económica española*. Ph.D. thesis, Universidade da Coruña.

Winter-Froemel, E. and Onysko, A. (2012). Proposing a pragmatic distinction for lexical anglicisms. In Cristiano Furiassi, et al., editors, *The anglicization of European lexis*, page 43.

Xia, M. X. (2016). Codeswitching language identification using subword information enriched word vectors. In *Proceedings of the Second Workshop on Computational Approaches to Code Switching*, pages 132–136, Austin, Texas, November. Association for Computational Linguistics.

11. Language Resource References

Cardellino, Cristian. (2019). *Spanish Billion Words Corpus and Embeddings*. `https://crscardellino.github.io/SBWCE/`.

Cañete, José. (2019). *Spanish Word Embeddings*. Zenodo, `https://doi.org/10.5281/zenodo.3255001`.

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

Korobov, M and Peng, T. (2014). *Python-crfsuite*. `https://github.com/scrapinghub/python-crfsuite`.

Hiroki Nakayama and Takahiro Kubo and Junya Kamura and Yasufumi Taniguchi and Xu Liang. (2018). *doccano: Text Annotation Tool for Human*. https://github.com/doccano/doccano.

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

Pérez, Jorge. (2017a). *FastText embeddings from SBWC*. Available at https://github.com/dccuchile/spanish-word-embeddings#fasttext-embeddings-from-sbwc.

Pérez, Jorge. (2017b). *GloVe embeddings from SBWC*. Available at https://github.com/dccuchile/spanish-word-embeddings#glove-embeddings-from-sbwc.

Real Academia Española. (2014). *Diccionario de la lengua española, ed. 23.3*. http://dle.rae.es.