Lázaro: An Extractor of Emergent Anglicisms in Spanish Newswire

A Master's Thesis

Presented to

The Faculty of the Graduate School of Arts and Sciences
Brandeis University

Graduate Program in Computational Linguistics
Department of Computer Science

Constantine Lignos, Advisor

In Partial Fulfillment
of the Requirements for the Degree
Master of Science

by

Elena Álvarez-Mellado

May 2020

Acknowledgments

What kind of times are these, when
To talk about trees is almost a crime
Because it implies silence about so many horrors?
"To Those Born After", Bertold Brecht

This thesis was written under the extraordinary circumstances that took place during the coronavirus pandemic of Spring 2020, a pandemic that was destructive all around the globe, but that was particularly devastating in Madrid, my hometown. Despite the uncertainty and collective anguish caused by this event, writing this thesis was still one of the most joyful and rewarding experiences I recall. There are several people I would like to thank for that.

Constantine Lignos has been a fantastic advisor and I would like to thank him for his guidance and enthusiastic support throughout the entire project.

I would also like to thank Kristen Sheets and Sam Haymann, my family in Boston. I find it impossible to imagine the last two years of my life without them.

My family has also been a major point of emotional support, not only during the writing of this thesis but throughout the entire experience of being abroad, especially my parents, Isabel and Amador.

Los Mellado, my noisy and fantastic extended family, were also a constant source of enjoyment and laughter in gloomy days thanks to their unbeatable sense of humour.

My beloved friends (many of whom work at hospitals) organized games, custom parties and even baking sessions to keep the spirits up during the quarantine. Seeing their faces on the screen was a real joy and the best source of motivation I could think of.

But, above all, I would like to thank Andrés Pajarón Lizcano, who turned his life upside down and followed me across the Atlantic. During the last two years, he has put up with my tiredness, my grumpiness and my long hours of study while taking care of us. I would like to thank him for his generosity, his unconditional support and for making life so much more enjoyable in all sorts of ways, even amidst of a global pandemic. *Mil gracias.*

ABSTRACT

Lázaro: An Extractor of Emergent Anglicisms in Spanish Newswire

A thesis presented to the Faculty of the
Graduate School of Arts and Sciences of Brandeis University
Waltham, Massachusetts

By Elena Álvarez-Mellado

The use of lexical borrowings from English (often called *anglicisms*) in the Spanish press evokes great interest, both in the Hispanic linguistics community and among the general public. Anglicism usage in Spanish language has been previously studied within the field of corpus linguistics. Prior work has traditionally relied on manual inspection of corpora, with the limitations that implies.

This thesis proposes a model for automatic extraction of unadapted anglicisms in Spanish newswire. This thesis introduces: (1) an annotated corpus of 21,570 newspaper headlines (325,665 tokens) written in European Spanish annotated with unadapted anglicisms and (2) two sequence-labeling models to perform automatic extraction of unadapted anglicisms: a conditional random field model with handcrafted features and a BiLSTM-CRF model with word and character embeddings. The best results are obtained by the CRF model, with an F1 score of 89.60 on the development set and 87.82 on the test set. Finally, a practical application of the CRF model is presented: an automatic pipeline that performs daily extraction of anglicisms from the main national newspapers of Spain.

Table of Contents

List of Tables

List of Figures

Chapter 1

Introduction

Lexical borrowing is a phenomenon that affects all languages and constitutes a productive mechanism of word formation. During the last decades, English has been a major influence on other European languages (Furiassi et al., 2012; Görlach, 2002) and, consequently, has produced numerous lexical borrowings (often called *anglicisms*), especially in the press. Chesley and Baayen (2010) estimated that a reader of French newspapers encounters a new lexical borrowing every 1,000 words, English borrowings outnumbering all other borrowings combined (Chesley, 2010). In Chilean newspapers, lexical borrowings account for approximately 30% of neologisms, 80% of those corresponding to English loanwords (Gerding et al., 2014). In European Spanish, Rodríguez González (2002) estimated that anglicisms could account for 2% of the vocabulary used in Spanish newspaper *El País* in 1991, a number that is likely to be higher today.

As a result, the influence of English on the Spanish language has attracted lots of attention, both in academia and among the general public, usually from a prescriptivist perspective (Balteiro, 2011a). The use of anglicisms is in fact a major concern for prescriptivist institutions such as *Real Academia Española* (RAE) or *Fundación del Español Urgente* (Fundéu), who advocate for limiting the usage of English borrowings.

Despite the interest that anglicism usage produces, the systematic study of novel angli-

cisms has mainly relied on manual inspection of limited corpora (where both the selection of anglicisms of interest and corpus lookup is done by hand), an approach that seems insufficient to account for an on-going phenomenon like anglicism incorporation. The study of anglicisms could benefit from applying computational techniques to monitor novel anglicism usage. The purpose of this thesis is to propose a computational model for the automatic extraction of unadapted anglicisms in Spanish newswire.

The structure of this thesis is as follows:

Chapter 2 sets the definition and scope of what an anglicism is and reviews previous approaches to the study of anglicisms from corpus linguistics, computational linguistics and code-switching methods.

Chapter 3 introduces a corpus of 21,570 newspaper headlines annotated with anglicisms, and presents the annotation guidelines and tagset followed during the annotation process.

Chapter 4 explore two sequence-labeling models for automatic extraction of anglicisms and applies them to the corpus presented in Chapter 3: a conditional random field model with handcrafted features and a BiLSTM-CRF model with character and word embeddings.

Chapter 5 introduces an automatic pipeline that uses the model introduced in Chapter 4 to perform daily extraction of anglicisms and to monitor anglicism usage in the main national newspapers of Spain.

Finally, Chapter 6 summarizes the conclusions of this work and offers ways in which this project could be improved and expanded.

The model, code and corpus presented in this thesis have been made publicly available[1]. Earlier portions of the content of this thesis appeared in the paper *An Annotated Corpus of Emerging Anglicisms in Spanish Newspaper Headlines* that was accepted at the Fourth Workshop on Computational Approaches to Linguistic Code-switching (Álvarez Mellado, 2020).

Finally, the name of this project, Lázaro, is an homage to Spanish philologist Fernando

[1] `https://lirondos.github.io/lazaro/`

Lázaro Carreter, whose prescriptive columns admonishing against the usage of anglicisms on the Spanish press became extremely popular during the decades of 1980s and 1990s.

Chapter 2

Literature review

This chapter sets the definition of what an anglicism is and reviews relevant prior work on anglicism usage and extraction from three different fields: Hispanic linguistics, computational linguistics and code-switching.

2.1 Anglicism: definition and scope

Linguistic borrowing is the process of reproducing in one language elements and patterns that come from another language (Haugen, 1950). Linguistic borrowing therefore involves the exchange between two languages and has been widely studied within the field of contact linguistics (Weinreich, 1963). Various typologies have been proposed to classify linguistic borrowings according to different criteria, such as typological characteristics, levels of language involved, integration of the borrowed element within the recipient language, etc. (Haspelmath and Tadmor, 2009; Matras and Sakel, 2007; Thomason and Kaufman, 1992).

The process of linguistic borrowing requires a language from which a certain element or pattern is taken (the donor language) and a second language into which that element in inserted (the recipient language). In that sense, borrowing and code-switching are similar and 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, 2008) 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). For example, all borrowed verbs in Spanish must undergo the process of adding the suffix *-ar* or *-ear* in order to comply with the morphological paradigm of Spanish verbs (*tuitear*, from *to tweet*). Similarly, unadapted borrowings that do not comply with phonological restrictions of Spanish will also be transformed. That phonological transformation might even end up showing on its spelling: the unadapted borrowing *spoiler* coexists with the adapted form *espóiler*, that better reflects the Spanish pronunciation.

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 that are often used by monolingual individuals without knowledge of the donor language. Therefore, code-switches are by definition not integrated into the recipient language, unlike established borrowings (Poplack, 2012). In fact, borrowings can eventually become assimilated into the recipient language lexicon and therefore lose the perception of being "foreign" (Lipski, 2005). Some authors establish the need of a borrowing being recognized as foreign by native speakers in order to be consider as such (Zenner et al., 2012). For instance, a word like *bar* was originally borrowed from English into Spanish, but it has been so assimilated that is now perceived as a native word by monolingual speakers of Spanish, and its English nature is only seem as etymological. On the other hand, a word like *whisky* (that has also been used in Spanish for some time) has never been fully assimilated and is perceived as a foreign word.

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 part-of-speech 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 extracted, as in a named entity recognition task. This thesis proposes approaching the task of lexical borrowing extraction in the fashion of a NER task.

This thesis is concerned with unadapted lexical borrowings from English (or *anglicisms*), i.e. words from English origin that are introduced into Spanish without any morphological or orthographic adaptation (Gómez Capuz, 1997; Núñez Nogueroles, 2018a; Pratt, 1980). The precise characteristics of these borrowings will be introduced in Section 3.2.

According to Thomason and Kaufman (1992), this type of lexical borrowing corresponds to level 1 on the borrowing scale, where borrowing is due to casual contact between languages and is motivated by cultural reasons. The borrowing scale described by Thomason and Kaufman distinguishes five levels of borrowing, according to the degree of contact between the donor language and the recepient language, and the typological effects these borrowings may have on the recipient language. The type of borrowing that this thesis is concerned with is level 1, a type of borrowing that affects only the lexicon and comprises the borrowing of content words that are not part of the basic vocabulary. Level 2 includes borrowing of function words (conjunctions and adverbs); level 3 involves adposition borrowing and the use of derivational affixes from borrowed words on native vocabulary; in level 4, borrowings will have typological effects on the recipient language, such as new syllable structure constraints or word order. In level 5, the pressure of the donor language causes significant typological disruption in the recepient language, such as changes in morphophonemic rules, word structure or concord rules.

2.2 Anglicisms in Hispanic linguistics

The study of English influence in the Spanish language has been a hot topic in Hispanic linguistics for decades, particularly concerning anglicism usage (Núñez Nogueroles, 2017a).

Seminal work on anglicism classification and analysis was done by Pratt (1980), where any word that entered Spanish via the English language was considered an anglicism, re-

gardless of its previous history or the original language the word was coined in. Unlike prior approaches to this issue, Pratt's work is purely descriptivist and is in fact openly critical towards the purist view that prior authors had on borrowing, an attitude that he considers to be unscientific and even xenophobic (Pratt, 1980, pg. 17). Many of the forecasts outlined in 1980 by Pratt concerning the influence of English on the Spanish language are in fact common practice in today's Spanish, like the use of the suffix *-s* in English-derived plural forms (*fans, tuits*) or the proliferation of words derived with *-ismo* or *-ización* (Pratt, 1980, pg. 239).

Later works on anglicisms have proposed different classifications (Lorenzo, 1996; Medina López, 1998; Núñez Nogueroles, 2018a; Rodríguez González, 1999) and have focused on various aspects of anglicism usage, such as orthographic integration (Rodríguez González, 2018), diachronic shifts (Gimeno Menéndez and Gimeno Menéndez, 2003), typological characteristics (Gómez Capuz, 1997), syntactic changes (Rodríguez Medina, 2002), sociocultural aspects (Gómez Capuz, 2004) or lexicographical coverage (Balteiro, 2011a).

More recently, corpus linguistics methods have also been applied to the study of English borrowings usage in Spanish media. These works, however, have mainly relied on manual inspection of either previously compiled general corpora such as CREA[1] (Balteiro, 2011b; Núñez Nogueroles, 2016, 2018b; Oncíns Martínez, 2012), either new tailor-made corpora designed to analyze specific genres, varieties or phenomena (De la Cruz Cabanillas and Tejedor Martínez, 2012; Diéguez, 2004; Gerding Salas et al., 2018; Gimeno Menéndez and Gimeno Menéndez, 2003; Núñez Nogueroles, 2017b; Patzelt, 2011; Rodríguez Medina, 2002; Vélez Barreiro, 2003).

2.3 Computational approaches to anglicism detection

In recent years, several automatic and semiautomatic approaches to anglicism detection have been proposed, aimed either towards improving how natural language processing systems

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

deal with out-of-vocabulary borrowings (Alex, 2008; Leidig et al., 2014; Mansikkaniemi and Kurimo, 2012), or towards enhancing corpus-based approaches to the study of linguistic borrowing (Andersen, 2012; Serigos, 2017a).

Previous approaches in different languages have mostly depended on resource lookup (lexicon or corpus frequencies), character n-gram probability and pattern matching. Alex (2008) 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. Also in German, Leidig et al. (2014) used language model perplexity and lexicon and corpora lookup to build an English inclusion classifier based on decision trees, support vector machines and a voting system. In French, Chesley (2010) extracted anglicisms from newswire using pattern-matching search. Furiassi and Hofland (2007) explored corpora lookup and character n-grams to extract pseudoanglicisms 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. Koo (2015) presented an unsupervised approach based on the expectation-maximization algorithm to detect loanwords in Korean. Mansikkaniemi and Kurimo (2012) proposed a method based on n-gram perplexity to detect foreign names in morphologically-rich languages like Finnish. Table 2.1 summarizes the approaches used in prior anglicism detection projects. Probabilities derived from character n-grams seem to be the most popular approach.

In Spanish, Serigos (2017a) built an extractor of anglicisms from a corpus of Argentinian newspapers and TV and film subtitles. This anglicism identifier calculated the probability of a word being English or Spanish based on character n-grams; when the difference between

Work	Pattern matching	Lexicon/corpus lookup	Char n-grams probability	Machine Learning model	Language
Alex (2008)		✓			German
Andersen (2012)	✓	✓	✓		Norwegian
Chesley (2010)	✓				French
Furiassi and Hofland (2007)		✓	✓		Italian
Garley and Hockenmaier (2012)		✓	✓	Maxent	German
Koo (2015)			✓	EM	Korean
Leidig et al. (2014)		✓	✓	SVM	German
Losnegaard and Lyse (2012)			✓	k-NN	Norwegian
Mansikkaniemi and Kurimo (2012)			✓		Finnish
Serigos (2017a)	✓	✓	✓		Spanish

Table 2.1: Previous approaches to anglicism detection.

both probabilities was smaller than a given threshold, a lexicon lookup module adjudicated whether the word was an anglicism or not. In order to avoid mistaking named entities with anglicisms, a capitalization module was set to ignore titlecased words. This system achieved an F1 score of 74.50 on unseen data.

Serigos's work is the first to show that computational methods could successfully be applied to perform automatic detection of anglicisms in Spanish newswire (Serigos, 2017a), and also the first to apply distributional techniques to analyze the semantic specificity of those extracted anglicisms (Serigos, 2017b). The work by Serigos is the closest both in nature and scope to the one presented in this project. Serigos, however, makes some assumptions that are addressed by this thesis. Here are some of the limitations of the anglicism identifier presented by Serigos (2017a):

1. Each word is analyzed in isolation, i.e. the previous word is considered irrelevant to determine whether the current word is an anglicism.

2. The way the capitalization module works (discarding any word in titlecase as a named entity, including at the beginning of the sentence) assumes that anglicisms can never appear on the first position of the sentence. Serigos argues that nouns in the first position of a sentence are extremely rare in Spanish, which is true in many contexts, but they are very normal in newspaper headlines.

3. Although the system includes loan phrases (i.e. anglicisms that are formed by more

than one word, such as *prime time*), given that the system considers every word in isolation (without considering the previous word), it assumes that all anglicisms occurring sequentially will be a single loanword. This is definitely the most frequent case, but this assumption prevents the system from considering any other possibilities.

4. Only two language tags were considered: Spanish and English. No tags for any other languages were included and borrowings from other languages were labeled as Spanish during the annotation process. As Serigos points out on the results discussion, this decision hurt the final metrics, as the anglicism identifier tended to identify non-English borrowings as anglicisms.

5. The system was evaluated on token level exclusively, and no span/phrase level evaluation was given for loan phrases. Token level takes into account each tag individually and any tag that is correct will be counted as a true positive, regardless of whether the full phrase that the individual token belongs to was correctly labeled or not. For example, in token level evaluation, if only *night* in *late night* was detected, *night* would still count as a true positive. In span level evaluation, only full matches over the entire phrase count, i.e. the entire phrase *late night* would have to be correctly labeled in order to count as a true positive. Span level evaluation may seem too harsh, because no credit is given to partial matches. However, token level evaluation can overstate results; after all, a model that would only detect English function words could get away with just detecting *and* in *rock and roll* or *by* in *stand by* and still get a generous result.

This thesis seeks to address the shortcomings derived from the assumptions in Serigos's work.

2.4 Work within code-switching

Regarding the work within the code-switching community, language identification on multilingual corpora has been widely explored. Due to the nature of code-switching, these models have primarily focused on oral corpora 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 conditional random field (CRF) models (Al-Badrashiny and Diab, 2016; Shrestha, 2016; Sikdar and Gambäck, 2016; Xia, 2016), logistic regression (Shirvani et al., 2016) and LSTM models (Jaech et al., 2016; Samih et al., 2016).

As mentioned in 2.2, 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, 2017a).

Chapter 3

Corpus description and annotation process

No previously annotated corpus was found to be suitable for anglicism extraction from Spanish newswire: previous work on anglicism detection did not publicly release the annotated data, and other public newswire corpus in Spanish did not include annotation for anglicisms. As a result, a new corpus was specifically retrieved and annotated for the task.

3.1 Corpus description

The corpus consists of a collection of monolingual newspaper headlines written in European Spanish. Using headlines was preferred to using full articles 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. Second, anglicisms are abundant in headlines, because they are frequently used as a way of calling the attention of the reader (Furiassi and Hofland, 2007), as in *'Wearables', robots y coches del futuro: diez tecnologías del CES 2016 que debe conocer todo CIO*[1] ("Wearables, robots and futuristic cars: ten technologies from the CES 2016 that every CIO should know") or *Ugly food: fruta fea y*

[1] https://www.eldiario.es/hojaderouter/ntssolutions/Wearables-robots-coches-CES_2016-CIO_6_500860010.html

deformada, pero igual de buena[2] ("Ugly food: fruit that is ugly and deformed, but equally good"). 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.

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

3.2 Annotation process

The term *anglicism* covers a wide range of linguistic phenomena. Narrowing down when a lexical borrowing has been fully adapted into the recipient language is a complex task and, consequently, previous work on borrowing annotation have pointed out the difficulty of deciding what to annotate as a lexical borrowing (Andersen, 2012; Serigos, 2017a).

Different anglicism identification projects have used different definitions and scopes of what an English inclusion is. Alex (2008) followed the definition proposed by Onysko (2007) and included English borrowings, code-switching and pseudoanglicisms (words formed from English elements that do not exist in English[4]). Leidig et al. (2014) also considered proper names, along with borrowings, pseudoanglicisms and hybrid forms (compound words with a German and an English part). Serigos (2017a), on the other hand, set a more restrictive scope and followed the definition of loanword proposed by Haugen (1950) ("words whose phonemic shape and meaning have been imported into a recepient language without morphemic substitution"). Serigos's annotation focused on words identifiable as English and excluded proper names and code-switched data. This scope is the closest to the one proposed in this thesis. None of these projects included semantic calques (loan translations such as *rascacielos*, "skyscraper") within the scope of the annotation.

[2] `https://www.eldiario.es/consumoclaro/ahorrar_mejor/Ugly-food-fruta-deformada-igual-buena_0_689781139.html`

[3] `http://www.eldiario.es/`

[4] An example of pseudoanglicism in Spanish would be *footing* ("jogging") or *balconing*.

This annotation project follows the typology proposed by Gómez Capuz (1997) and focuses 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 not annotated as anglicisms and only unadapted borrowings were annotated as such. Therefore, Spanish adaptations of anglicisms like *fútbol* (from *football*), *mitin* (from *meeting*), etc. were not tagged 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 tagged as anglicism either. However, pseudoanglicisms (words that are formed as if they were English, but do not exist in English, such as *footing* or *balconing*) were tagged as anglicisms.

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*[5] (Real Academia Española, 2014) was consulted. This dictionary is compiled by the Royal Spanish Academy, a prescriptive institution. This decision was motivated by the fact that, if a borrowing was already registered by this dictionary (that has a conservative approach

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

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 above 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[6]) 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 cultural innovation that came via the English language (such as name of things related to food or music), that was not perceived as new anymore and that had never really competed with a Spanish lexical equivalent, then it was ignored. Following this criterion, words like *jazz*, *swing* or *banjo* were not tagged as anglicisms[7]. This criterion proved to be extremely useful to deal with old unadapted anglicisms in the fields of music and food. Figure 3.1 contains the decision steps followed during the annotation process.

The corpus was annotated by a native speaker of Spanish using Doccano[8] (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

[6]Winter-Froemel and Onysko do not use *catachresis* with its standard meaning of "misapplication of a word", but with its original rhetoric sense "metaphor caused by necessity" (Winter-Froemel and Onysko, 2012, pg. 47)

[7]Interestingly, according to *Diccionario de la Lengua Española* and *Oxford English Dictionary*, the word *banjo* might come from the Spanish word *bandurria*. If this origin is true, then the word *banjo* in English would be an adapted Spanish borrowing, and *banjo* in Spanish would be an unadapted anglicism.

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

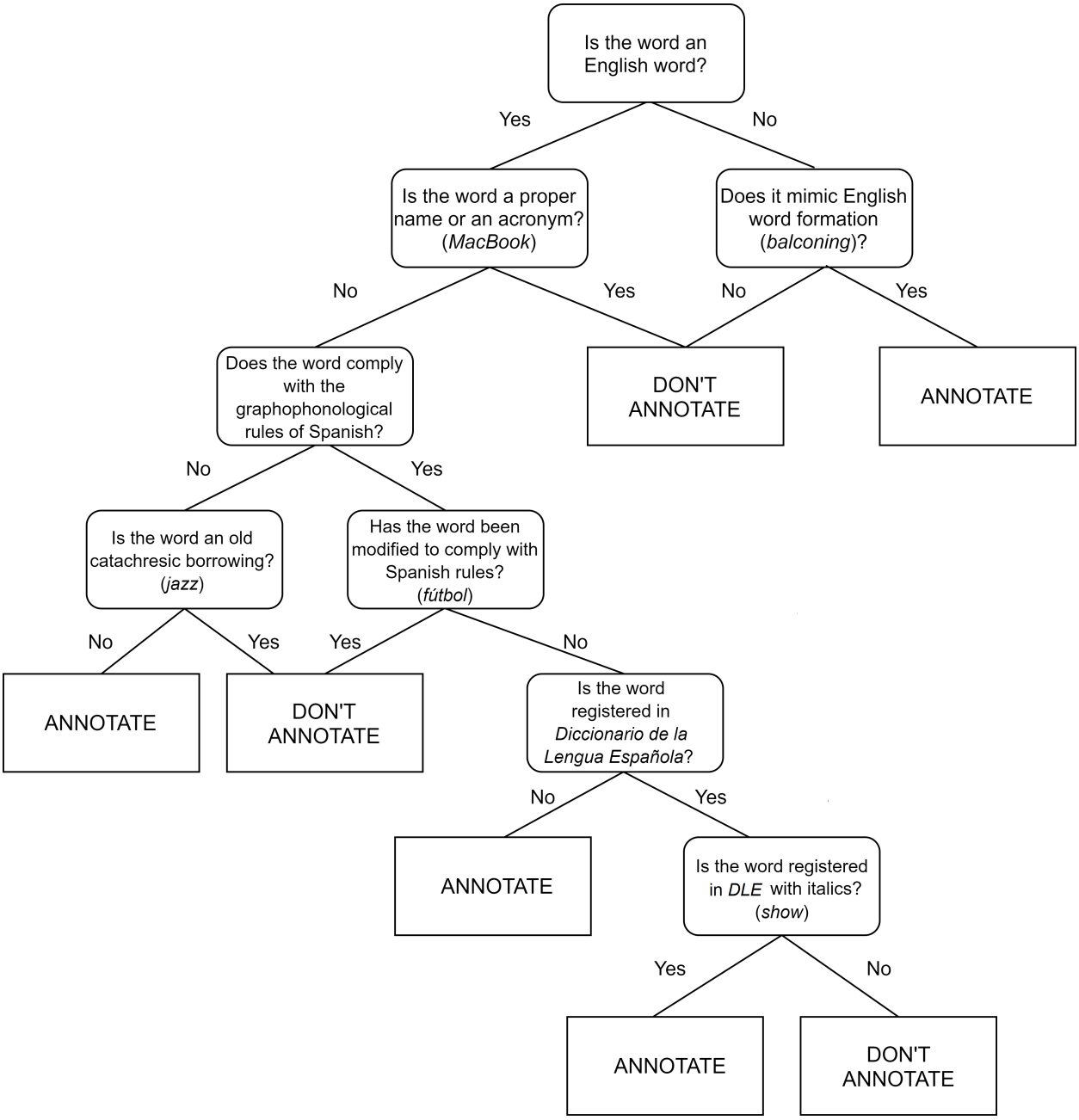

Figure 3.1: Decision steps to follow during the annotation process to decide whether to annotate a word as an anglicism.

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, I 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.

3.3 Corpus split and count

The corpus consists of two subcorpora: the main corpus (with the usual train/development/ test split that was used to train, tune and evaluate the model) and an additional test set that was designed to assess the performance of the model on more naturalistic data. The main corpus contains 16,553 headlines, while the supplemental test set contains 5,017 headlines. In total, the corpus contains 21,570 newspaper headlines, which amounts to 325,665 tokens.

The annotation guidelines and tagset presented in Chapter 3.2 were applied to both subcorpora of headlines. This section describes the distribution of anglicisms per corpus split and the difference between both subcorpora.

Main corpus

The headlines from the main corpus were extracted from *eldiario.es* 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, from manual inspection, they were the most likely to contain anglicisms. The proportion of headlines with anglicisms per section can be found in Table 3.1.

The main 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 spans of text labeled as anglicism (`ENG`) amounts to 1,176. Most of them (85%) were a single word, although some of them were multiword expressions (*prime time*) (see Table 3.2). 67 spans of text in the main corpus were labeled as `OTHER`.

The main corpus was divided into training, development and test set, following a pro-

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

Table 3.1: Percentage of headlines with anglicisms per section.

Anglicism	Count	Anglicism	Count
app	42	*prime time*	22
online	40	*black friday*	15
black	39	*big data*	11
apps	29	*low cost*	11
lobby	27	*fake news*	9
riders	25	*late night*	8
reality	24	*trending topic*	5
startups	23	*think tank*	5

Table 3.2: Most frequent single-token anglicisms (left) and multi-token anglicisms (right) in the main corpus.

portion of 60-20-20. This proportion split was done both in terms of total headlines and headlines with anglicisms to ensure balance across anglicism distribution (for example, 20% of all headlines and 20% of all headlines with anglicisms would be in the test set). The proportions of headlines, tokens and anglicisms in each corpus split can be found in Table 3.3.

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 supplemental test set contains 126 anglicisms and 35 spans of text labeled as other type of borrowing (see Table 3.3).

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
Total	21,570	325,665	1,233	1,304	102

Table 3.3: Number of headlines, tokens and anglicisms per corpus subset.

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.

Chapter 4

Models

Three models were explored for automatic extraction of unadapted anglicisms and applied to the corpus presented in Chapter 3: a simple baseline model with lexicon lookup, a conditional random field with handcrafted features and a BiLSTM-CRF model with word and character embeddings. The CRF and the BiLSTM-CRF are sequence-labeling models that have previously been applied to named-entity recognition (Lample et al., 2016; Sutton et al., 2012).

4.1 Baseline model with lexicon lookup

An intuitive approach to anglicism detection would be to label as anglicism any word appearing in the corpus of headlines that was also found in an English dictionary. Such a model was built using an English lexicon of 194,000 words[1]. For every word in each headline, the model checked whether it appeared in the lexicon. If the word being analyzed was the first word of the sentence, it was converted into lowercase; otherwise, the case was not changed (that means that most titlecased words would not be found in the lexicon). Words found in the lexicon were tagged as `ENG`, and no `OTHER` label was considered (all `OTHER` labels in the test set were converted to `O`). Every adjacent tags that were labeled as `ENG` were assumed

[1] `http://www.gwicks.net/dictionaries.htm`

to be a multiword anglicism. This very basic approach produced an F1 score of 1.32 on the development set (precision = 0.67, recall = 39.27) and an F1 score of 1.02 on the test set (precision = 0.52, recall = 30.66).

A second version of this model was also tried: this second model only labeled as anglicisms words that were found in the English lexicon and whose lemma (provided by `spaCy`) was not registered in the 23.3 edition of the *Diccionario de la Lengua Española* (Real Academia Española, 2014), according to the online lexicon[2] compiled by Rodríguez Alberich (2019) (91,347 lemmas). Again, any adjacent `ENG` tags were considered a multiword anglicism. This second model produced an F1 score of 25.40 on the development set (precision = 18.57, recall = 40.18) and an F1 score of 25.94 on the test set (precision = 18.67, recall = 42.45).

Although this second lookup model produced a better F1 score than the first one, the results were still very modest. This means that the simple lexicon lookup approach is insufficient for the task of extracting lexical anglicisms, which motivates the exploration of Machine Learning models.

4.2 Conditional random field model

CRF model with basic features

A CRF model for automatic extraction of anglicisms was created using the annotated corpus presented in Chapter 3 as training material. As mentioned in Chapter 2, the task of detecting anglicisms can be approached as a sequence labeling problem where only certain spans of texts will be labeled as anglicism (similar to an NER task). The chosen model was a 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`[3] (Korobov and Peng, 2014), a Python wrapper

[2] `https://dirae.es/static/lemario-23.3.txt`
[3] `https://github.com/scrapinghub/python-crfsuite`

for `crfsuite`[4] (Okazaki, 2007) which implements CRF for labeling sequential data. It also used the `Token` and `Span` classes from `spaCy`[5] library (Honnibal and Montani, 2017).

The following handcrafted features were used for the model. These are features that are commonly used in NER:

- Bias feature: a feature that is active on every single token to support setting per-class bias weights.
- Token feature: the string of the token.
- Uppercase feature (y/n): active if all characters in the token are uppercase.
- Titlecase feature (y/n): active if only the first character of the token is capitalized.
- Character trigram feature: a feature for every trigram in the token.
- Quotation feature (y/n): active if the token is any type of quotation mark (“”‘’«»).
- Word suffix feature: last three characters of the token.
- POS tag (provided by the `es_core_news_md` model from `spaCy`).
- Word shape (provided by the `es_core_news_md` model from `spaCy`).
- Word embedding (see Table 4.1).

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, multi-token BIO encoding adapted from from Ramshaw and Marcus (1999) was used to denote the boundaries of each span. A window of two tokens in each direction was set for the feature extractor. Optimization was performed using L-BFGS.

The model was tuned on the development set using grid search; the hyperparameters explored were c1 (L1 regularization coefficient: 0.01, 0.05, 0.1, 0.5, 1.0), c2 (L2 regularization

[4] `https://github.com/chokkan/crfsuite`
[5] `https://spacy.io/`

Author	System	# 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 4.1: Embeddings used in experiments.

coefficient: 0.01, 0.05, 0.1, 0.5, 1.0), embedding scaling[6] (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,b) (see Table 4.1). The best results were obtained with c1 = 0.05, c2 = 0.01, scaling = 0.5 and word2vec Spanish embeddings from the Spanish Billion Words Corpus[7] by Cardellino (2019). The threshold for the stopping criterion delta was selected through observing the loss during preliminary experiments (delta = 1e − 3).

Feature ablation

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 model was run ignoring the `OTHER` tag (`OTHER` labels were converted to `O`), both during tuning and the feature ablation experiments. Table 4.2 displays the results on the development set with all features and for the different feature ablation runs.

The best results were obtained with all features (F1 score = 89.60), which demonstrates that all the features proposed for the CRF model contribute positively to the model's performance. The character trigram feature seems to be the one that has the biggest impact on the feature ablation study (F1 change = −3.74), which confirms that character trigrams are

[6]Multiplicative scaling applied to each dimension of the embedding.

[7]`https://crscardellino.github.io/SBWCE/`

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 4.2: Ablation study results on the development test.

in fact more relevant than the token itself or its embedding representation when determining whether a given word is an anglicism or not.

Additional features

Several additional features were explored in order to improve the results of the model. However, none of the additional features tried produced better results than those obtained with the basic features. The following additional features were tried:

- Character bigram feature.
- Character 4-gram feature.
- Digit feature (y/n): active if the token contains a number (*2002*, *salu2*).
- Lemma feature: the lemma of the token (provided by the `es_core_news_md` model from `spaCy`).
- Punctuation feature (y/n): active if the token is a punctuation symbol (*¿?¡!,:;.*).
- First position in the sentence feature (y/n): active if the token is the first element of the sentence.

- Graphotactic feature: this feature aims to capture the graphotactic shape of the word (i.e. the combination of vowels and consonants) in order to assess whether the given word complies with the graphotactic rules of Spanish. A set of rules transforms every letter in the word into either vowel (v) or consonant (c). Combinations of consonants that are permitted by Spanish rules of spelling (*ch, bl, br, fl, fr, cl, cr, pr, pl, tr, dr, rr, ll, cc, nn...*) (Real Academia Española, 2010) are represented differently to those not permitted (*wh, ph, th...*). Similarly, letters allowed at the end of the word (*a, e, i, o, u, r, s, l, n, d, z*) are marked differently to those letters not expected at the end of Spanish words. This shape representation of the word is then splitted into trigrams which are finally fed as features to the model.

- Spanish lexicon feature (y/n): active if the lemma of the word (provided by `spaCy`) was registered in the 23.3 edition of the *Diccionario de la Lengua Española* (Real Academia Española, 2014), according to the online lexicon[8] compiled by Rodríguez Alberich (2019) (91,347 lemmas).

- English lexicon feature (y/n): active if the word was found in an English lexicon[9] (370,103 words).

- Word probability in Spanish feature: this feature calculated the probability of the word being Spanish by creating a probability distribution of trigrams over a Spanish lexicon[10].

- Word probability in English feature: this feature calculated the probability of the word being English by creating a probability distribution of trigrams over an English lexicon[9].

- Higher probability of being English feature (y/n): active if the probability of the word being English was higher than the probability of being Spanish.

[8] `https://dirae.es/static/lemario-23.3.txt`
[9] `https://github.com/dwyl/english-words/blob/master/words_alpha.txt`
[10] `https://github.com/julox/spanish_lexicon/blob/master/spanish_lexicon.csv`

Features	**Precision**	**Recall**	**F1 score**	**F1 change**
Basic features	97.84	82.65	89.60	
Basic features + Bigram	95.16	80.82	87.41	−2.19
Basic features + Quatrigram	97.28	81.74	88.83	−0.77
Basic features + Digit	**97.85**	**83.11**	**89.88**	**+0.28**
Basic features + Lemma	97.81	81.74	89.05	−0.55
Basic features + Punctuation	96.26	82.19	88.67	−0.93
Basic features + Sentence position	96.76	81.74	88.61	−0.99
Basic features + Graphotactic shape	94.27	82.65	88.08	−1.52
Basic features + Lexicon (ES)	94.76	82.65	88.29	−1.31
Basic features + Lexicon (EN)	96.76	81.74	88.61	−0.99
Basic features + Probability (ES)	97.84	82.65	89.60	0.00
Basic features + Probability (EN)	97.84	82.65	89.60	0.00
Basic features + Probability EN > ES	96.22	81.28	88.12	−1.48
Basic features + Perplexity threshold	**97.86**	**83.56**	**90.15**	**+0.55**

Table 4.3: Additional features tried. Results on the development set.

- Perplexity threshold feature: the perplexity of the word was calculated from the probability distribution over trigrams based on a lexicon of Spanish[10]. If the word was in percentile 80 (i.e. the perplexity of the word was higher than 80% of the lexicon), then the high perplexity feature was activated. This feature was inspired by the work of Mansikkaniemi and Kurimo (2012), where a similar technique was applied to detect foreign names in Finnish. In their work, the authors set percentile 70 as a good threshold to detect non-Finnish inclusions; in these experiments with anglicisms in Spanish language, percentile 80 was found to produce the best results on the development set.

The results produced by these additional features on the development set can be found on Table 4.3. Only two of these additional features produced better results on the development set: the digit feature (F1 score = 89.88, F1 change = +0.28) and the perplexity threshold feature (F1 score = 90.15, F1 change = +0.55). However, both features produced worse results than the CRF model with basic features when tried on the test set: −0.06 on F1 for the digit feature and −2.24 for the perplexity threshold feature. As a result, no real improvement could be obtained by incorporating additional features.

Results and discussion

The CRF model was trained on the training set, tuned on the development set and then run on the test set and the supplemental test set with the set of features and hyperparameters discussed on Chapter 4.2. Table 4.4 displays the results obtained on the development set, test set and supplemental test set. The model produced an F1 score of 89.60 on the development set, a result that greatly outperforms the very modest result of the lexicon lookup baseline introduced in Section 4.1 (F1 score = 25.40).

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`[11]). 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 substantial 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

[11] This is similar to a common scoring variant for NER, where no entity types are considered, only whether the span of text is just name or not name.

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 4.4: Results on test set and supplemental test set.

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 demonstrates 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.

In order to assess the ability of the model to detect previously unseen anglicisms, I checked how many of the anglicisms that were correctly labeled by the model were not present in the training set. After all, it would not be of much value to have a model that memorized whatever anglicisms were present in the training set and that could exclusively detect anglicisms

that had been previously seen during training. I inspected what proportion of true positives, false negatives and false positives on the development set, test set and supplemental test set had never been seen during training. The model seems to be generalizing reasonably well: out of 104 unique true positives on the development set, 36 were not present in the training set. In other words, 35% of the unique anglicisms that were correctly identified as such on the development set had never been seen before by the model. On the other hand, 31 out of 38 unique false negatives (over 80%) had not been seen during training, i.e. 31 of the 38 anglicisms that were incorrectly ignored by the model had not been seen during training. In total, out of 219 anglicisms in the test set that were not seen during training, 39 were correctly identified (17.80%). Concerning false positives, all of the 4 false positives found on the development set had not been seen during training. These numbers show that the model is capable of detecting new anglicisms that had never been seen before. However, there is still room for improvement in terms of generalization, as the great majority of false negatives are, in fact, previously unseen anglicisms.

Table 4.5 displays the list of unseen anglicisms per set split divided according to whether they were correctly identified by the model (true positives) or not (false negatives). It should be noted that the same anglicism can appear both as a true positive and a false negative (even in the same set). This is due to the fact that context plays a major role on how the model detects anglicisms, and therefore, the same unseen anglicism could be detected when found in a given context and yet ignored when found in a different context. Similarly, the fact that a certain previously-unseen anglicism was not detected cannot be fully attributed to it not appearing in the training set: in fact, what several of these false negatives had in common is that they appeared on the first position of the sentence (and were, therefore, capitalized; for example, *vamping* in *Vamping: la recurrente leyenda urbana de la luz azul 'asesina"*[12]). These anglicisms tended to be consistently ignored (as the model probably assumed they were named entities) but perhaps could have been detected had they appeared in another

[12]https://www.eldiario.es/consumoclaro/cuidarse/Vamping-recurrente-leyenda-urbana-asesina_0_882262754.html

	True Positives	False Negatives
Dev set	*arcade*, *balconing*, *brain hacking*, *breaking of*, *call centers*, *communities*, *dating show*, *daytime*, *deep learning*, *docushows*, *edredoning*, *fast food*, *fitness*, *fracking*, *game jams*, *gintonic*, *girl-group*, *hardcore-punk*, *hip*, *hip hop*, *influencers*, *invent*, *machine learning*, *made in China*, *merchandising*, *microvlogging*, *morning show*, *networking*, *punky*, *redneck*, *roll-on*, *routers*, *skaters*, *stick*, *tickets*, *webcam*	*backstage*, *chat*, *deluxe*, *drone*, *ethiojazz*, *femtech*, *geek*, *golden visa*, *he*, *her*, *him*, *influencer*, *influencers*, *instagramer*, *made in China*, *mindfulness*, *noodles*, *off*, *ok*, *ok,* *boomer*, *packs*, *share*, *she*, *showman*, *social media*, *social trading*, *spoiler*, *spoilers*, *spray*, *trailer*, *vamping*
Test set	*ambient*, *box office*, *brit-pop*, *check*, *dating show*, *early adopter*, *eSports*, *geek*, *grooming*, *legaltech*, *look*, *low-cost*, *open source*, *panoselfie*, *sex symbol*, *spoiler*, *spots*, *stripper*, *take away*, *vamping*	*bluesmen*, *cool*, *copyright*, *cyber monday*, *cyborg*, *deluxe*, *docu-realities*, *ebook*, *email*, *email marketing*, *esports*, *excel*, *for the record*, *free jazz*, *free to play*, *green new deal*, *hip hop*, *india pale ale*, *lawfare*, *loop*, *me too*, *microneedling*, *mix*, *packs*, *proptech*, *queer*, *snapchat*, *stock*, *tour*, *wikis*
Suppl. test set	*best-seller*, *blockbusters*, *datings*, *factchecking*, *latin trap*, *loot boxes*, *match*, *podcast*, *renting*, *road movie*, *shock*, *star wars*, *talents*, *tipsters*, *tour*	*bulldog*, *caucus*, *celebrities*, *click*, *cool*, *delcygate*, *doodle*, *drag*, *dumping*, *electroshock*, *film*, *ghosting*, *hooligans*, *impeachment*, *k-pop*, *made in Chile*, *palmeroning*, *pause*, *performance*, *pulp*, *queer*, *rave*, *road movie*, *share*, *silver*, *sphynx*, *stand*, *westerns*

Table 4.5: Unique unseen anglicisms on development set, test set and supplemental set.

position in the sentence.

Concerning false positives (non-anglicisms that were incorrectly labeled as such), they can be classified according to the following types:

1. Neologisms in Spanish: *puntocom*, *pin parental*.

2. Non-English borrowings: *gourmet*, *kale-borroka*, *exconsellers*. These borrowings were the ones that got correctly labeled when the `OTHER` tag was included.

3. Proper names or entities: *lorazepam*.

4. Orthographically adapted borrowings: *láser*.

5. Titles from songs, films or series: 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. Examples: *it darker* in *'You want it darker', la oscura y brillante despedida de Leonard Cohen*[13].

6. Partial matches from multi-token anglicisms: *marketing* instead of *email marketing*; *trading* instead of *social trading*.

4.3 Neural model

The corpus presented in Chapter 3 was also used to train a neural model. For this task, the library `NCRF++`[14] (Yang et al., 2018) was used. `NCRF++` is a PyTorch-based framework that implements a neural sequence-labeling model in three layers: character sequence layer, word sequence layer and inference layer. For the anglicism detection model, the architecture chosen was a character CNN + word LSTM + CRF model, an architecture that has successfully been used before on other sequence-labeling tasks like named entity recognition (Lample et al., 2016). The advantage of this model over the CRF is its ability to learn more complex character patterns than the trigram character features in the CRF.

The model was trained including both `ENG` and `OTHER` tags and with `ENG` tags only (`OTHER` tags were replaced by `O`). Different learning rates were also tried. Table 4.6 displays the results obtained on the development set. The best results were obtained with $lr = 0.0075$ (F1 score = 84.62). In all cases, the word embeddings used were word2vec Spanish embeddings by Cardellino (2019) (as they were the ones that produced the best results on the CRF model), the character embedding dimension was set to 30, hidden dimensions were set to 200 and iterations were set to 100.

[13] `https://www.eldiario.es/cultura/musica/you-want-it-later-leonard-cohen_0_572193176.html`

[14] `https://github.com/jiesutd/NCRFpp`

The best performing hyperparameter settings were then run on the test set. The result obtained was F1 = 86.67, over two points more than the results obtained on the development set. Although the model performs reasonably well, these results are still worse than the F1 scores obtained with the non-neural CRF model, both on the development and the test set (F1 on dev = 89.60, F1 on test = 87.82).

Labels included	**Learning rate**	**F1 score**
`ENG` + `OTHER`	1.5e−2	80.89
`ENG` only	1.5e−2	83.55
`ENG` only	1.5e−3	84.29
`ENG` only	7.5e−3	**84.62**

Table 4.6: Results on the development set with a charCNN + wordLSTM + CRF model.

4.4 Discussion

In this chapter, different models have been explored for automatic detection of anglicisms. First, a simple lexicon lookup baseline was built (F1 score = 25.94 on the test set). Then, two sequence-labeling models were tried: a CRF model with handcrafted features, and a BiLSTM-CRF model with word and character embeddings.

The CRF model obtained an F1 score on development set of 89.60 and 87.82 on the test set. Although additional features were tried, it seems that the problem is resistant to being improved by feature engineering. Additionally, the feature ablation study showed that the character trigram was the feature that contributed the most for detecting anglicisms.

On the other hand, the neural model should be able to learn more complex character patterns than the trigram character features in the CRF, but it obtained worse results than the CRF model: F1 = 84.62 on the development set and F1 = 86.67 on the test set.

Chapter 5

A practical application: `@lazarobot`

A model like the one presented in Chapter 4.2 can be used to build a pipeline for automatic extraction of anglicisms in Spanish newswire. Such a system could detect novel anglicisms and assist monitoring anglicism frequencies and usages on the Spanish daily press. The conditional random field model presented on Chapter 4.2 was used to build (1) a pipeline that performs automatic extraction of anglicisms on a daily basis from the main national newspapers of Spain and (2) a Twitter bot that tweets the output of that pipeline. At the time of writing this thesis, five newspaper are included in this pipeline: *eldiario.es*[1], *El País*[2], *El Mundo*[3], *ABC*[4] and *El Confidencial*[5].

5.1 Pipeline description

The pipeline connects on a daily basis to the RSS feeds of the newspapers mentioned above and extracts the entire articles (both headlines and article bodies) published within the last 24 hours. These articles are then preprocessed (for HTML tag removal, etc) and then sent to the CRF model. In this particular case and given that no evaluation was required in

[1] `https://www.eldiario.es/`
[2] `https://www.elpais.com/`
[3] `https://www.elmundo.es/`
[4] `https://www.abc.es/`
[5] `https://www.elconfidencial.com/`

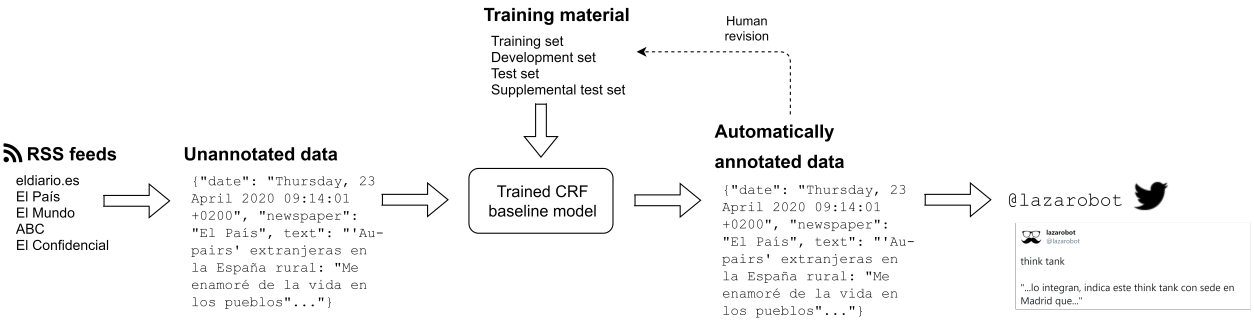

Figure 5.1: Extraction pipeline from RSS feeds to `@lazarobot`.

this scenario, the CRF model was trained using all four sets presented in Section 2: the training set, the development set, the test set and supplemental test set. The model outputs the lexical borrowings found by the model on the extracted articles (both `ENG` and `OTHER`). Each of these borrowings is then tweeted by the Twitter account `@lazarobot`[6], along with the context where the borrowing was found and the URL to the original newspaper article. Figure 5.1 displays the structure of the pipeline.

This bot allows us to see the model working on an environment quite different to the one used during training and testing. First of all, in this scenario the model deals with text coming not only from *eldiario.es* (the newspaper that was used for the training corpus), but from four additional newspapers. Furthermore, the text that is provided to the model as input is not only headlines anymore, but full articles, with the substantial difference in nature that that implies (longer texts, more sophisticated contexts, more abundant types of borrowings, presence of other code-mixed data, etc).

5.2 Pipeline output

The output produced by this pipeline confirms some of the system's shortcomings that were already observed during testing and can also help spot new ones that were never revealed when evaluating the model on a more limited scenario, like the test set and the supplemental

[6] `https://www.twitter.com/lazarobot`

test set.

English titles of films, songs, documents, etc. continued to be a source of false positives, particularly in very long titles. For example, the word *strategy* was incorrectly labeled as anglicism when found in the following article excerpt: "La contribución española al plan de reconstrucción económica de la UE tras la pandemia (*Spain's non paper on a European recovery strategy*, 19/4) es excelente." (*El País*, "España acierta cuando va de Europa"[7]).

Code-mixed data (such as English quotations) also poses a difficult challenge for the model, as it tends to mistake any code-mixed data with possible borrowings. In order to mitigate the effect of quotations and other code-mixed data in general, a post-processing filter was applied to the pipeline, so that multi-token spans consisting of four or more tokens that were labeled as anglicisms (a case that was very unlikely to be true borrowings and very likely to be code-mixed data) would be ignored.

The output of this pipeline also reveals the limitations of the CRF model when dealing with multi-token anglicisms: when dealing with an anglicism like *personal shopper*, the model was only capable of detecting *shopper* and ignored the word *personal*, probably due to the fact that *personal* happens to also be a word in Spanish. A perfect model would have been able to recognize that, although *personal* is an adjective in Spanish, given that Spanish is an NAdj language (the modifying adjective follows the noun), it would have had to follow the noun if it was in fact a Spanish usage. Similarly, only *sour* in *sour diesel* was detected as anglicism, probably influenced also by the overlap in trigrams between *diesel* and the Spanish word *diésel*.

A more complex case in terms of linguistic nuance was *rol playing* (*...las simulaciones (rol playing) aplicadas a la enseñanza...*[8]). The word *rol* is the Spanish adaptation of the English borrowing *role*. The model only identified *playing* as anglicism and ignored *rol*, which is not incorrect because *rol* has already been orthographically adapted into Spanish

[7] https://elpais.com/economia/2020-04-22/espana-acierta-cuando-va-de-europa.html

[8] https://www.eldiario.es/andalucia/UNIA-adaptacion-docencia-evaluacion-postgrados_0_1019448359.html

and no adaptations are considered by the model. However, even when part of the anglicism has been orthographically adapted into Spanish, the construction *rol playing* still follows the English word order of noun-noun compounds (which is foreign to Spanish). This case could be considered an example of both a syntactic anglicism and a partially adapted lexical anglicism, which illustrates how complex and multifaceted the process of adapting a borrowing can be.

In terms of multi-token anglicisms, the pipeline also produced some interesting examples that illustrate the difficulty of establishing whether two consecutive anglicisms are one multi-token anglicism or two independent anglicisms: *burger gourmet* was labeled as a single multi-token anglicism (*...disfrutar de una burger gourmet en nuestro restaurante favorito...*[9]). The phrase *burger gourmet*, however, follows the syntactic word order expected in Spanish (noun followed by adjective), which indicates that this is not a single multi-token anglicism (such as *fake news* or *big data*), but that these are in fact two independent anglicisms that are being collocated following Spanish grammar rules. The same conclusions can be applied to *show online* (*...un show online e interactivo desde su hogar...*[10]), which was also incorrectly labeled as a single multi-token anglicism. These examples prove that it cannot be assumed that any two consecutive anglicisms can be subsumed into a single multi-token anglicism.

Finally, in addition to demonstrating how the model deals with language in the wild, this pipeline can be seen as a first step towards an automatic system for tracking anglicism usage over time in the Spanish press, as well as for documenting the incorporation of new anglicisms and monitoring frequency shifts.

[9] `https://www.elmundo.es/metropoli/gastronomia/2020/04/25/5e9dc9ee21efa0b9728b45c3.html`

[10] `https://www.elmundo.es/metropoli/otros-planes/2020/04/28/5ea6f96821efa0be668b45d8.html`

Chapter 6

Conclusions and future work

The goal of this thesis has been to introduce and explore various resources for automatic extraction of emergent anglicisms in Spanish newswire.

First, this thesis has introduced a new corpus of Spanish newspaper headlines annotated with anglicisms. The corpus consists of 21,570 newspaper headlines written in European Spanish annotated with emergent anglicisms. The annotation scope, tagset and guidelines have been presented in Section 3.

Second, two sequence-labeling models have been explored for anglicism extraction in Spanish newswire and applied to the corpus mentioned above: a conditional random field model (CRF) with handcrafted features, and a BiLSTM-CRF model with word and character embeddings (Chapter 4). The best results were obtained by the CRF model, with an F1 score of 89.60 on the development set and 87.82 on the test set. The model shows some generalization ability (that is, the model is capable of detecting new anglicisms that were never seen during training). However, there is still room for improvement, as the majority of false negatives were previously unseen anglicisms. Anglicisms on the first position of the sentence are also challenging for this model, as they tend to be confused with named entities.

Finally, the CRF model was used to build an automatic pipeline that performs daily extraction of anglicisms from the main national newspapers of Spain (Chapter 5), a pipeline

that can assist lexicographic work by detecting novel anglicisms and by tracking anglicism usage in the Spanish daily press. This pipeline allows us to see the model working on a more realistic and complex scenario than the one in the test set. In this new setting, code-mixed data is by far the most challenging issue for the model: titles of films or series as well as non-Spanish quotations tend be incorrectly labeled as anglicisms.

The corpus, the model and the extraction pipeline have been made publicly available[1], and the output of the extraction pipeline can also be seen on the Twitter account `@lazarobot`[2].

In terms of future work, there are several fronts in which this project could be improved or expanded.

The BiLSTM-CRF model could be further developed to improve performance, in particular by adding character model pretraining. After all, the CRF ablation study showed that the character trigram feature was the one that contributed the most to the CRF model. The neural model should be able to learn more complex character patterns than the CRF model and therefore obtain a better performance.

The output produced by the automatic pipeline could be used as training material after human careful revision. Using this data as additional training material could be beneficial in several ways. First of all, this data would contain the correct annotation of examples that were previously incorrectly labeled by the model, which should help improve the model performance. Second, this data would not be only headlines any more, but full articles, which should be informative for the model. Finally, the daily output produced by the automatic pipeline ensures a constant flow of potential new annotated data, which will be particularly useful to train future versions of the neural model.

The phenomenon of anglicisms is very wide and affects many other aspects of language besides the lexicon. However, the models presented in this thesis are only concerned with unadapted lexical borrowings and do not attempt to cover other forms of anglicisms, such as syntactic anglicisms or semantic calques. Developing computational models to address

[1] `https://lirondos.github.io/lazaro`
[2] `https://twitter.com/lazarobot`

these phenomena would be instrumental in order to have the full picture of the influence of English in Spanish language.

This thesis has dealt exclusively with detecting anglicisms in Spanish text. It remains open whether the presented approach could succesfully be implemented to detect anglicisms in other languages. In particular, the CRF model relies on a set of handcrafted features. It would be interesting to assess whether these features are also reliable in other languages. Perhaps some of the additional features that were discarded for the Spanish model (such as lexicon-derived features) could be more useful than character trigrams when detecting anglicisms in languages that are typologically and orthographically closer to English.

Similarly, this model aims to capture anglicisms in newswire and, consequently, has been trained and tested on newspaper articles, which is a very specific (and not necessarily representative) type of linguistic data (Plank, 2016). As a result, the model relies on the text input being normalized and standardized. It would be interesting to assess how a model like this performs on less standard data. For example, Twitter data is very prone to anglicism use and many of the informal and non-technical anglicisms that end up being used in the press first appear on Twitter conversations. It would be interesting to see if a model like the one introduced in this thesis (that relies heavily on character trigrams) could successfully detect anglicisms on less standarized data like Twitter data.

Finally, the output produced by the automatic pipeline can be a very interesting piece of data to analyze over time. The study of anglicism usage in the Spanish press has traditionally been limited to manual inspection of static corpora. In that sense, the pipeline introduced by this thesis facilitates the analysis of anglicisms on larger amounts of data and in real-time. This pipeline could assist lexicographic work to monitor anglicism usage over time and to study them in context. Analyzing what anglicisms survive and which ones do not would be particularly interesting and could help us understand more about language contact in general and the process of borrowing in particular.

References

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

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. Association for Computational Linguistics.

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

Álvarez Mellado, E. (2020). 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.

Andersen, G. (2012). Semi-automatic approaches to Anglicism detection in Norwegian corpus data. In Furiassi, C., Pulcini, V., and Rodríguez González, F., editors, *The anglicization of European lexis*, pages 111–130.

Balteiro, I. (2011a). Prescriptivism and descriptivism in the treatment of anglicisms in a series of bilingual Spanish-English dictionaries. *International Journal of Lexicography*, 24(3):277–305.

Balteiro, I. (2011b). 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.

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

Cañete, J. (2019). Spanish word embeddings. `https://doi.org/10.5281/zenodo.3255001`.

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

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 Tejedor Martínez, C. (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.

Furiassi, C., Pulcini, V., and Rodríguez González, F. (2012). *The anglicization of European lexis*. John Benjamins Publishing.

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. 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.

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

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.

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

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

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)*.

Haspelmath, M. and Tadmor, U. (2009). *Loanwords in the world's languages: a comparative handbook*. Walter de Gruyter.

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

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. European Language Resources Association (ELRA).

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

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. Association for Computational Linguistics.

Koo, H. (2015). An unsupervised method for identifying loanwords in Korean. *Language Resources and Evaluation*, 49(2):355–373.

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

Lample, G., Ballesteros, M., Subramanian, S., Kawakami, K., and Dyer, C. (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.

Leidig, S., Schlippe, T., and Schultz, T. (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*.

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 Andersen, G., editor, *Exploring Newspaper Language: Using the web to create and investigate a large corpus of modern Norwegian*, pages 131–154. John Benjamins Publishing.

Mansikkaniemi, A. and Kurimo, M. (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.

Matras, Y. and Sakel, J. (2007). *Grammatical borrowing in cross-linguistic perspective*, volume 38. Walter de Gruyter.

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

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

Nakayama, H., Kubo, T., Kamura, J., Taniguchi, Y., and Liang, X. (2018). doccano: Text annotation tool for human. `https://github.com/doccano/doccano`.

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

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.

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

Oncíns Martínez, J. L. (2012). Newly-coined anglicisms in contemporary Spanish. A corpus-based approach. In Furiassi, C., Pulcini, V., and Rodríguez González, F., 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.

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

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

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.

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

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

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

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.

Real Academia Española (2010). *Ortografía de la lengua española*. Espasa. `http://aplica.rae.es/orweb/cgi-bin/buscar.cgi`.

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

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

Rodríguez González, F. (2002). Spanish. In Görlach, M., editor, *English in Europe*, chapter 7, pages 128–150. Oxford University Press.

Rodríguez González, F. (2018). Aspectos ortográficos del anglicismo. *Lebende Sprachen*, 63(2):350–373.

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

Rodríguez Alberich, G. (2019). Lemario y palabras nuevas en la edición 23.3 del DLE. `https://blog.dirae.es/entradas/lemario-edicion-23-3/`.

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. Association for Computational Linguistics.

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

Serigos, J. R. L. (2017b). 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.

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. 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. 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. 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. Association for Computational Linguistics.

Sutton, C., McCallum, A., et al. (2012). An introduction to conditional random fields. *Foundations and Trends in Machine Learning*, 4(4):267–373.

Thomason, S. G. and Kaufman, T. (1992). *Language contact, creolization, and genetic linguistics*. Univ of California Press.

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

Weinreich, U. (1963). Languages in contact (1953). *The Hague: Mouton*.

Winter-Froemel, E. and Onysko, A. (2012). Proposing a pragmatic distinction for lexical anglicisms. In Furiassi, C., Pulcini, V., and Rodríguez González, F., 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. Association for Computational Linguistics.

Yang, J., Liang, S., and Zhang, Y. (2018). Design challenges and misconceptions in neural sequence labeling. In *Proceedings of the 27th International Conference on Computational Linguistics (COLING)*.

Zenner, E., Speelman, D., and Geeraerts, D. (2012). Cognitive sociolinguistics meets loanword research: Measuring variation in the success of anglicisms in Dutch. *Cognitive Linguistics*, 23(4):749–792.