ADoBo — Automatic Detection of Borrowings

Annotation guidelines for lexical borrowings

With a focus on unassimilated anglicisms in Spanish newswire

ADoBo shared task at IberLEF 2021

Annotation guidelines for lexical borrowings

Author: Elena Álvarez Mellado

This publication is available from:

https://adobo-task.github.io/

February 2021

Table of Contents

List of Figures

1. Objective

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

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

These guidelines are the guidelines followed for the annotation of the dataset of ADoBo shared task at IberLEF 2021.

2. Tagset

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

3. What an unassimilated lexical borrowing is

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

3.1 Definition and scope

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

Lexical borrowing is the incorporation of single lexical units from one language (the donor language) into another language (the recipient language) and is usually accompanied by morphological and phonological modification to conform with the patterns of the recipient language [1–3].

Anglicisms are lexical borrowings that come from the English language [4–7]. For our annotation project, we will focus on direct, unassimilated, emerging anglicisms, i.e. lexical borrowings from the English language into Spanish that have recently been imported and that have still not been assimilated into Spanish, that is, words like *smartphone*, *influencer*, *hype*, *lawfare* or *reality show*.

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

```
... financiados a través de la plataforma de [crowdfunding](ENG) del
club [gourmet](OTHER) que tengas más cerca
```
[1]

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

3.2 Types of lexical borrowing

Lexical borrowings can be adapted (the spelling of the word is modified to comply with the phonological and orthographic patterns of the recipient language, as in *fútbol* or *tuit*) or unadapted (the word preserves its original spelling: *millennial*, *newsletter*, *like*). For this annotation project, we will be focusing on unassimilated lexical borrowings: this means that adapted borrowings will be ignored and only unadapted borrowings will be tagged (see section 4.1.1 for a full description on the differences between adapted and unadapted borrowings).

3.3 Multiword borrowings

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

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

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

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

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

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

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

4. What an unasimilated lexical borrowing is not

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

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

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

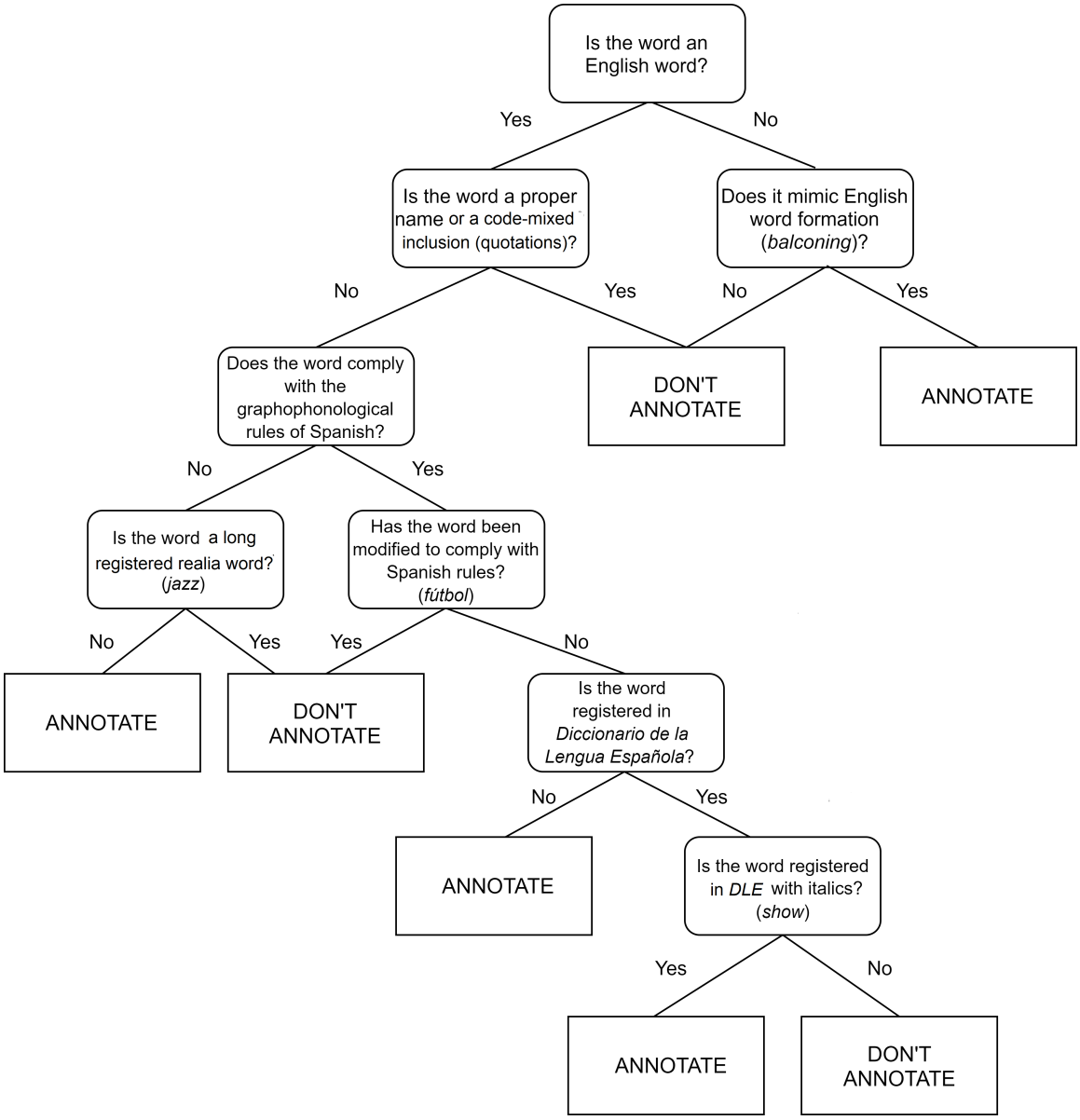

Fig. 1. Decision steps to follow during the annotation process to decide whether to annotate a word as an anglicism.

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

4.1 Assimilated vs unassimilated borrowings

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

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

4.1.1 Adapted borrowings

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

Borrowings that have not been adapted but whose original spelling complies with grapho-phonological rules of Spanish (and are therefore unlikely to be further adapted, such as *bar*, *fan*, *web*, *internet*, *club*, *set* or *videoclip*) will be tagged as a borrowing or not not depending on how recent or emergent they are. In order to determine which unadapted,

graphophonologically acceptable borrowings are to be annotated, the latest online version of the *Diccionario de la lengua española* [8] will be consulted (as of February 2021)[2]. If the DLE dictionary already registers the word with that meaning and with no italics or quotation marks, then it will be considered assimilated and therefore should not be tagged.

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

```
ganó el primer set
```

```
los tonos ‘[nude]’ (ENG)
```

It should be noted that this guideline only applies to lexical borrowings that comply with graphophonological rules of Spanish. Unadapted lexical borrowings that do not comply with graphophonological rules of Spanish (such as *show*, *look*, etc) will be tagged as borrowing, regardless of whether the word is included in the dictionary or not (although see section 4.1.4 for exceptions to this).

It is important to emphasize that, in order for an unadapted graphophonologically-compliant borrowing to be considered assimilated it should be registered in the dictionary both without italics and with the corresponding meaning. For instance, a word like *top* (that

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

is graphophonologically acceptable in Spanish) is registered in DLE with no italics, but it is only registered with the meaning of a piece of clothing. The word *top* as referring to the upper part of something (as in *top 5*) is not registered. Consequently, the borrowing *top* will be considered assimilated when referring to the piece of clothing, but unassimilated when used to talk about the best elements of a ranking or the upper part of something.

`un top estampado`

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

`la [top] desfiló (ENG)`

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

`el mundo post pandemia`

`un [post] de Facebook (ENG)`

4.1.2 Words derived from foreign lexemes

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

4.1.3 Pseudoanglicisms

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

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

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

4.1.4 Realia words

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

The purpose of this guideline is to account for cultural terms such as *pizza*, *whisky*, *jazz*, *blues*, *banjo* or *sheriff*. These are all borrowings that are reluctant to be adapted or translated, even when they have been around in the Spanish language for long. The reason is that they refer to cultural inventions (the name was imported along with the object it refers to), and, given their cultural significance, they never competed with a Spanish equivalent

and are seen as assimilated.

Therefore, unadapted borrowings that refer to cultural innovations (such as music, cooking, sport names etc) and that have been registered for long in the Spanish language[3] will not be tagged as emergent borrowings.

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

4.1.5 Latinisms

Borrowings that were introduced directly from Latin language (such as *deficit*, *curriculum*, etc) will not be considered emergent and therefore will not be tagged as a borrowing. However, words of Latin origin that have been reintroduced via English (or other languages) with a distinct meaning and have not been assimilated yet (such as *adlib* or *premium* etc) will be tagged as a borrowing.

4.2 Borrowings vs names

4.2.1 Proper nouns

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

- person names: *Bernie Sanders*.

[3]RAE dictionary https://dle.rae.es/ and Mapa de diccionarios https://webfrl.rae.es/ntllet/SrvltGUILoginNtlletPub can be consulted

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

4.2.2 Names of institutions and political roles

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

4.2.3 Words derived from proper nouns

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

- products: *un iPhone*, *un whatsapp*, *un bizum*, *un Scalextric*, *el Satisfyer*.

- works of arts: *un monet*
- characters: *un frankestein*.

4.2.4 Names of peoples or languages

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

4.2.5 Ficticious creatures

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

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

4.2.6 Scientific units

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

4.2.7 Names of species

Scientific names of a species (such as Latin names) will not be tagged as a lexical borrowing (*anisakis*).

4.3 Borrowings vs other code-mixed inclusions

Borrowing (using units from one language in another language) and code-switching (intertwining segments of different languages in the same discourse) have frequently been described as a continuum [9], with a fuzzy frontier between the two. As a result, it can be difficult to tell the difference between borrowing and other code-mixed inclusions. The following guidelines can assist annotators adjudicate edge cases.

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

4.3.1 Acronyms and acronym expansions

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

```
un lector de CD
```

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

`un [CD player] (ENG)`

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

`La técnica de PCR (protein chain reaction)`

4.3.2 Metalinguistic usage

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

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

4.3.3 Literal quotations

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

`El eslogan ‘Make America Great Again’`

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

4.3.4 Expressions

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

```
Tecnología [made in Spain] (ENG)
```

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

5. Limitations of these guidelines

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

5.1 Text genre

These guidelines were designed to specifically capture borrowings in a corpus of Spanish newswire. Newswire is a very specific genre of text that by no means represent the whole of a language [10].

5.2 Donor language

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

5.3 Synchronic approach to borrowing

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

5.4 Geographic variety

The guidelines in these document were designed to capture borrowings used in Spanish newspapers, that is, written in the variety of Spanish that is spoken in Spain. The reason for

this limitation is by no means the assumption that European Spanish represents the whole of the Spanish language, but quite the opposite.

The notion of assimilation is usage-based and community-dependant. For instance, according to the guidelines we have just introduced, a word like *living* (that is used heavily in some Latin American varieties to refer to the living room) would be considered unassimilated. It is arguable whether these criteria would be suitable for a project that tried to capture emergent lexical borrowings in Argentinian text, for example.

Consequently, we limit the scope of these guidelines to European Spanish not because we consider that this variety represents the whole of the Spanish-speaking community, but because we consider that the approach we have taken here may not account adequately for the whole diversity in borrowing assimilation within the Spanish-speaking world.

References

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

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

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

[4] 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.

[5] Pratt C (1980) *El anglicismo en el español peninsular contemporáneo*. Vol. 308 (Gredos), .

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

[7] Núñez Nogueroles E (2017) An up-to-date review of the literature on anglicisms in Spanish. *Diálogo de la Lengua, IX* :1–54.

[8] Real Academia Española (2014) Diccionario de la lengua española, ed. 23.3.

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

[10] Plank B (2016) What to do about non-standard (or non-canonical) language in NLP. *arXiv preprint arXiv:160807836* .