How to cite: Castro, A., Nazar, R., & Renau, I. (2021). New verbs and dictionaries: A method for the automatic detection of neology in Spanish verbs. *International Journal of Lexicography*, *34*(3), 382-399. https://doi.org/10.1093/ijl/ecab009
https://academic.oup.com/ijl/article-abstract/34/3/382/6306811?redirectedFrom=fulltext

New verbs and dictionaries: a method for the automatic detection of neology in Spanish verbs

Abstract

The appearance of new verbs can be observed regularly, but verbs are not frequently investigated in neology, and they are difficult to detect automatically. In this study, a corpus-based method is proposed to detect Spanish verbs with a series of algorithms that analyse the morphology of regular verbs. The vocabulary was drawn from a large corpus and contrasted with a major dictionary of Spanish. Then, a series of filters were applied to distinguish between valid neologism candidates and spelling mistakes. Around 88% of the neologisms proposed by the method were correct and we estimate that the system detected 76% of the neologisms present in the corpus. This procedure can be included in the workflow of a lexicographic project as a regular part of the task, as a systematic way of collecting new verbs from the data and avoiding under-representation or bias.

Keywords: computational lexicography, corpus-based lexicography, neology, Spanish, verbs

1. Introduction

In this paper, we present a methodological proposal for the detection of new Spanish verbs in a general corpus, and for their incorporation in lexicographic tasks concerning the activity of updating dictionaries. Verbs are an open class of words and new units are regularly added to the vocabulary, as we are observing now due to the coronavirus crisis: verbs such as *cuarentenar* 'to be in quarantine' or *desconfinar* 'to deconfine' are new verbs created by combining existing morphological structures in an innovative way. That is why Cook (2010: 13) explains that 'POS tagging of unknown words, including neologisms, can benefit greatly from exploiting word structure'. However, morphological analysis is one of the main sources of error in automatic neology detection, as many authors have reported. Thus, Renouf (2016) states that POS taggers struggle with words not found in a dictionary. Janssen (2009) also explains that POS taggers can be less effective in neologisms and have a much higher error rate. Amore et al. (2018) report problems regarding POS tagging of Italian verbs, that, as the rest of Romance languages, have tens of possible inflected forms for each unit. Langemets et al. (2020: 79) also conclude that, 'in order to make neologism discovery more effective we need more advanced tools for automatic language processing. There were a large number of mistakes in tokenisation, lemmatisation, POS tagging, Name Entity Recognition (NER) etc.'.

In Spanish, as in many other languages, verb morphology is composed of the stem, carrying the semantic information, and the inflection, carrying the grammatical information (RAE and ASALE 2009). Thus, in new verbs, it is only the inflection that allows us to automatically identify these units as verbs in the corpus. By comparing the list of the verbs obtained this way with the headwords present in a dictionary, it is possible to detect those that are not yet recorded and, as a consequence, might be candidates for neologism. This allows us to obtain not only the neologisms themselves, but also corpus-based evidence of their use, such as their frequency and dispersion, which is important for taking informed decisions on whether or not to include them in the dictionary. The approach described, consisting basically of the automatic extraction of neologisms from the corpus and their contrast against a dictionary, is well-established (see Section 2). We propose methods to reduce the noise that such contrast usually produces. Moreover, our method for verb detection, which is rule-based and simple both conceptually and technically, is not present in current POS-taggers for Spanish, for which verbs are still a challenge, particularly those with enclitics (Parra Escartín and Martínez Alonso 2015). For that reason, we consider this study a contribution to the field of natural language processing, computational lexicography and automatic detection of neologisms.

This paper is structured as follows: in the next section we offer a brief account of previous research on neology detection and classification. In Section 3 we present our methodology. In Section 4 we describe the results and, finally, in Section 5 we offer our conclusions.

2. The siege of neology

2.1. *Theoretical and methodological considerations about neology*

A neologism can be broadly defined as a new formal or semantic linguistic unit (Rey 1976), and it can manifest in various forms: as Klosa-Kückelhaus and Kernerman (2020: 2) state, 'the general understanding of neologisms includes new words, new multiword units, new elements of word formation, and new meanings of any of them'. For example, cases such as the already mentioned verbs *cuarentenar* and *desconfinar* are classified as formal neologisms, as they are new lexical units that are added to the vocabulary. Formal neologisms can originate using already available linguistic resources (internal neology), or can be acquired from other languages as loans (external neology). Furthermore, a new meaning can be added to an already existing word. An example of this, called semantic neology, would be the case of the verb *rastrear* 'to follow the trace, to investigate', which is currently being used with the meaning of 'tracking the contacts of a person who has been infected with covid-19 in order to put them in quarantine'.

Different researchers have dealt with the theoretical and methodological considerations on how to discriminate between a true neologism and *hapax legomena,* nonce word formations (occasionalisms), or simply spelling mistakes. A new word may appear a few times in the corpus as a creative resource in a specific situation, but that is not enough to qualify it as a neologism (Cabré 2015, Abel and Stemle 2018). Cabré (1999) established some non-mutually exclusive criteria to determine whether a word is a neologism: a) if the unit has arisen in recent years; b) if it is not listed in dictionaries; c) if it exhibits formal instability and d) if it is perceived as new by native speakers. Later, however, Cabré and

Nazar (2012) acknowledged how problematic these criteria can be, since a word can fulfil some of them and still not be considered a neologism. Firstly, a term may have a long history of use among a reduced circle of experts and remain unknown to a wider audience. Secondly, being in the dictionary is also not a reliable criterion, especially in the case of semantic neology. And thirdly, someone may not be familiar with a particular term due to age, origin or for cultural or sociological reasons.

One cannot find an entirely satisfactory answer to this problem, as the process of lexicalisation of a neologism is a continuum and there is no sharp distinction between a word that is used for the first time and a neologism that becomes a part of the vocabulary of a language (Kerremans and Prokić 2018). From the methodological point of view, this discussion is often solved at the last step of the process by human analysis of neologism candidates extracted from the corpus by some automatic system.

Regarding the problem of automatic neology detection, the comparison of corpus data against a list of words obtained from dictionaries and other sources is a widely used methodology for automatic detection of lexical neology (Janssen 2009, Abel and Stemle 2018). Generally speaking, this approach consists of identifying lexical units that are not present in a reference list. There are several systems for neology detection that use this approach: Neoveille (Cartier 2016) for Chinese, French, Greek, Polish, Portuguese, Russian and Czech; NeoTrack (Janssen 2005) for Portuguese; Sextan (Vivaldi 2000) and BuscaNeo (Cabré and Estopà 2009) for Spanish and Catalan; AVIATOR (Renouf 1993) and NeoCrawler (Kerremans and Prokić 2018) for English; the Korean Neologism Investigation Project (Nam et al. 2020) for Korean; two online lexicographic resources for German linked to a neologism detector: the *Wortwarte* (Lemnitzer 2010) and the *Neologismenwörterbuch* (Klosa-Kückelhaus and Lüngen 2018), among others. Costin-Gabriel and Rebedea (2014) use the Google Books N-Gram corpus for neologism detection in English. Cartier (2016) and Kerremans and Prokić (2018) seem to work at the token level, i.e., without tagging. However, in general in this approach one must deal with the problem of POS-tagging when tokens have many different inflections, as lemmatisation is a key consideration if one wants to compare two lists of words. For example, in the case of Spanish verbs, it would not be accurate to compare only tokens, because some of the verb inflections have the same form as an adjective or a noun (e.g., *amo* 'I love' has the same form as the masculine noun *amo* 'master').

As we have observed, dictionaries are both receivers of neologisms and material for neology detection, working as stop lists. Concerning lexicographic projects, the fact that a word is missing from the lemma list does not necessarily mean that it is a neologism, as the criteria for the creation of the macrostructure of the dictionary involve many complex factors, including the type of user and their needs, the size and type of dictionary, the prescriptive approach of the dictionary, the low frequency and dispersion of the word, etc. (Alvar Ezquerra 2007, and Atkins and Rundell 2008, for dictionary projects in general). Indeed, Baayen and Renouf (1996: 69) state that dictionaries 'are not a reliable source for studying morphological productivity', because lexical productivity is a wider phenomenon. Cartier (2008) also argues that there are technical limitations related to the list of headwords, which are not displayed in all their forms in the dictionary. However, as Cañete et al. (2019) state, the inclusion of a new word in a dictionary can be considered as the last step of the lexicalisation of a neologism and its inclusion in the general vocabulary of a language, as it is sufficient evidence of the institutionalisation of the word. As Freixa and Torner (2020)

point out, this is not equivalent to saying that all institutionalised neologisms belong to the lemma list of a dictionary. The argument works the other way around: when conducting an accurate lexicographic procedure, there should not be unstable, very infrequent lexical units in a dictionary. In this study, our purpose is not to study language innovation but rather to contribute to the methods for updating dictionaries. For this reason, we consider that the approach of comparing a corpus against a lemma list is in this case appropriate.

2.2. *New verbs in Spanish*

Concerning, specifically, new verbs in Spanish, the most common mechanism is converting already existing nouns or verbs (either from Spanish or from other languages) into new verbs by affixation (Bohrn 2010, Fuentes et al. 2009). Regarding suffixation, very productive suffixes take part in the process, especially *-ificar, -ear* and *-izar* (Adelstein and Kuguel 2008, Bohrn 2010, RAE and ASALE 2009), e.g., *pesificar* 'to turn into pesos' (from *peso* and *-ificar*), *googlear* 'to google' (from *Google* and *-ear,* probably inspired by the English verb *to google*) and *oscarizar* 'to give an Oscar' (from *Oscar* and *-izar*). They are also often created by the addition of very productive prefixes to already existing verbs, such as *re-* (*revictimizar* 're-victimize', *reconfinar* 're-confine'), *des-* (*desconfinar* 'de-confine', *desescalar* 'de-escalate'), etc. Lavale-Ortiz (2016) also reports a substantial number of cases of parasynthetic new verbs in her corpus, such as *enturbantar* 'to put a turban' (made with *en-* and *-ar* added to the root) or *aproblemar* 'to cause problems' (made with *a-* and *-ar* added to the root).

Compared to nouns and adjectives, the creation of new verbs in Spanish is a less common phenomenon (Fuentes et al. 2009, Ortega Martín 2001, Sanmartín 2010). And yet, they contribute to a significant part of the processes of neology. Regarding affixation specifically, this mechanism has proven to be a current and constant way of creating new verbs (Lavale-Ortiz 2016). We described some studies on the topic, but the fact is that there is still plenty of room for research devoted specifically to Spanish new verbs. The situation does not seem to be very different in studies in other languages. Nevertheless, we would like to mention the work by Amore et al. (2018), who apply a semantic distributional model to conduct a corpus analysis of Italian new verbs. Also, Ginebra and Rull (2010) provide a quantitative analysis of new verbs in Catalan, and conclude that there is a tendency for these units to be transitive in a higher percentage of those verbs included in the dictionary —a tendency that, according to Bohrn (2010), can also be observed in Spanish verbs. Oliveira (2020) deals with the process of creation of new verbs in Portuguese, which take place mainly via suffixation, as it does in Spanish (and with equivalent suffixes: *-ificar, -ear, -izar,* etc.).

L'Homme (2015) has argued that, in the case of terminology, verbs have been often neglected, because studies are more focused on knowledge representation, in which nouns seem to be very well established. This argument may be applied to neology as well, as a way to explain the relative lack of interest in verbs. However, verbs are a key part of speech in lexicography, requiring complex corpus analyses and specific considerations on their lexicographic treatment that take into account their argument structure, polysemy, collocations, pedagogical aspects, etc. (Atkins et al. 1988, Battaner and Torner 2008, Boas

2001, El Maarouf et al. 2014, Hanks 2013, L'Homme 2003, Marello 2010, among many others). This complexity requires that studies in neology focus more on verbs and propose specific methods for detection and representation of new verbs in dictionaries. For example, Atkins et al. (1988), Hanks (2013) and Marello (2010) deal with the differences in the syntactic structure of verbs (e.g. the causative/inchoative alternation), which has to be taken into account when conducting automatic neology detection of verbs (see section 3.2.1 for how we dealt with the problem of pronouns used for creating inchoative structures in Spanish). Thus, a method for neology detection should not only detect the new verbs, but also the most frequent structures and other features in which the verbs are involved –e.g. alternations, prevalence of participle over other inflections, and collocations, among others. All these aspects could be of great help for lexicographers in their quest to create new verb entries.

3. Methodology

The general approach of our method was to extract verbs from corpora and compare them with the lemma list of a Spanish dictionary. The algorithm for the automatic extraction of Spanish verbs from corpora is based on a set of rules to detect verbal inflection. The basic idea was to hand-code a list of verb endings and to observe if the same root appeared in the corpus in combination with multiple endings. In the case that it did, we reconstructed an infinitive form and contrasted it against a reference dictionary. If no match was found, we submitted the candidate to a battery of tests to distinguish between true neologisms and spelling mistakes.

3.1. *Materials*

Our materials comprised a corpus, for which we used the EsTenTen (Kilgarriff and Renau 2013) and a dictionary, used to contrast the verbs extracted from the corpus. As for the dictionary, we used the *Diccionario de la lengua española*, DLE (RAE 2014), mainly for technical reasons. The lemma list of this dictionary can be downloaded via Enclave RAE (https://enclave.rae.es), a website that offers an advanced interface for the DLE. The list of verbs from the dictionary comprised 11,893 verb lemmas.

The EsTenTen corpus consists of approximately 10^{10} running words from randomly downloaded web pages of Spanish-speaking countries. This corpus is already tokenised, but we ignored the tagging and used only the word forms. We only kept in the list those form types occurring at least 5 times in the corpus, a threshold we considered convenient to minimise accidental occurrences. We used 25% of this corpus to develop and test our method, and the rest to evaluate its performance. From the first part we obtained 1,054,411 different word forms, while from the rest of the corpus we obtained 2,954,973 different word forms. It should be clear that what we mean by this is form types and not form tokens, and also that form types are not to be confused with lemma types.

3.2. *Methods*

3.2.1. *Hand-coding a list of Spanish verb endings.* We obtained a list of verb endings from RAE and ASALE (2009), retaining only those with a minimum length of three letters. We also included a list of prefixes and enclitics, i.e., pronouns like *me, te, se, lo, los, la, las, le,*

les or *os,* that can be attached to verbs, as in *cómpraselo* 'buy him something' / 'buy it from him' or *peinarse* 'to comb oneself', etc.

According to RAE and ASALE (2009), verb endings are organised in paradigms or verb tenses and moods. Tenses are divided into simple and composites according to their lexical structure, with one or two elements respectively. We only used the simple tenses. Verbs are also classified in regular and irregular. Three regular conjugations exist, with infinitive forms ending in *-ar* as in *amar* 'to love', *-er,* like in *temer* 'to be afraid of' and *-ir*, like in *partir* 'to depart'. In the three cases, the root of the verb is kept intact in all inflected forms, with some exceptions due to accentuation. Irregular conjugations, in turn, correspond to all other verbs that do not conform to any of these three models, and their root is altered in their inflection. According to RAE and ASALE (2009), 90% of Spanish verbs correspond to those ending in *-ar*, and practically all cases of word formation with suffixes are also of this type (cases such as *-ear*, *-izar* and *-ificar,* as discussed in section 2).

We discarded word endings with fewer than three letters because they are error prone. For instance, a word ending like *-o* as in *yo amo* 'I love', corresponding to the first person singular of present tense indicative form of the verb *amar* 'to love', would coalesce with other word forms also ending in *-o* but pertaining to other grammatical categories, like adjectives (e.g. *atrevido* 'cheeky') or nouns (e.g. *sentimiento* 'feeling'). Table 1 shows the full list of 106 elements that were used.

Table 1. List of regular verb endings and enclitic pronouns.

Regular verb endings					
-aba	*-áramos*	*-aríamos*	*-eré*	*-iendo*	*-ieses*
-abais	*-aran*	*-arían*	*-eréis*	*-iera*	*-imos*
-ábamos	*-arán*	*-arías*	*-eremos*	*-ierais*	*-irá*
-aban	*-aras*	*-aron*	*-ería*	*-iéramos*	*-irán*
-abas	*-arás*	*-ase*	*-eríais*	*-ieran*	*-irás*
-ada	*-are*	*-aseis*	*-eríamos*	*-ieras*	*-iré*
-adas	*-aré*	*-ásemos*	*-erían*	*-iere*	*-iréis*
-ado	*-areis*	*-asen*	*-erías*	*-iereis*	*-iremos*
-ados	*-aréis*	*-ases*	*-íais*	*-iéremos*	*-iría*
-áis	*-aremos*	*-aste*	*-íamos*	*-ieren*	*-iríais*
-amos	*-áremos*	*-asteis*	*-ían*	*-ieres*	*-iríamos*
-ara	*-aren*	*-éis*	*-ías*	*-ieron*	*-irían*
-ará	*-ares*	*-emos*	*-ida*	*-iese*	*-irías*
-arais	*-aría*	*-erá*	*-idas*	*-ieseis*	*-iste*
	-aríais	*-erán*	*-ido*	*-iésemos*	*-isteis*
		-erás	*-idos*	*-iesen*	*-ndo*
Enclitics					
me	*te*	*se*	*nos*	*le*	
les	*la*	*lo*	*las*	*los*	

3.2.2. *Data processing.* As already mentioned, we used 25% of the corpus to develop and test our methodology, and left the rest of the corpus to evaluate the method's performance. This is important because one cannot evaluate a method with the same corpus that was used in its development.

With the list of word forms extracted from the corpus and the hand-coded list of verb endings, we implemented a script to extract all word forms matching a verb ending. Every time a match was found, the script separated the root from verb ending and stored them in a two-dimensional hash table. For illustration, consider Table 2, which shows some combinations of verb roots and endings.

Table 2. Examples of verb root and endings from verbs *rapear* 'to sing a rap', *retener* 'to retain' and *suprimir* 'to delete'.

Root	Endings
rape-	*-aba, -aban, -ada, -adas, -ado, -ados*
reten-	*-emos, -ida, -ido, -idos, -iendo, -éis, -ían*
suprim-	*-amos, -ida, -idas, -ido, -idos, -iendo, -iera, -ieran, -ieron, -iese, -iesen, -irán, -iré, -iría, -ían*

Here we depicted a root form like *rape-* as well as some typical verb endings, such as *-aba*. Together, they form the verb form *rapeaba* 'he/she was rapping', which was found in the corpus. If at least four different verb endings were observed attached to the same root, as in *rapeaba*, *rapeaban*, *rapeada* or *rapeadas*, then a new verb entry was created. The number of four combinations is certainly arbitrary, but is based on empirical testing.

Of course, what we needed to obtain was a list of verb lemmas and not a list of inflected forms. Therefore, once this part of the procedure was finished, the next step was to find the infinitive form of the verb, which in Spanish can only end in *-ar*, *-er* or *-ir*, as explained earlier. The way to find the infinitive form was to attach the newly found root to each of these endings and select the most frequent root-ending combination. For instance, in the case of a verb root like *reten-*, shown in Table 2, the algorithm would test lemma candidates with *-ar*: **retenar*; with *-er*: *retener*; and with *-ir*: **retenir*, but *retener*, the correct lemma, is the one that appears most frequently in the corpus. Table 3 shows an example of the result of this process, with the lemma, its frequency, root and endings. The DLE column states whether the lemma is found in the dictionary.

Table 3. Examples of the reconstruction of the lemma forms of the verbs and their comparison against the dictionary lemma list. Only *rapear* 'to sing a rap' was not present in the dictionary.

Lemma	Frequency	Root	DLE	Endings
rapear	206	*rape-*	0	*-aba, -aban, -ada, -adas, -ado, -ados*
retener	26390	*reten-*	1	*-emos, -ida, -ido, -idos, -iendo, -éis, -ían*

suprimir	32436	*suprim-*	1	*-amos, -ida, -idas, -ido, -idos, -iendo, -iera, -ieran, -ieron, -iese, -iesen, -irán, -iré, -iría, -ían*

If the newly found verb lemma matched the list obtained from the reference dictionary (the DLE), then there was no doubt we were looking at a legitimate Spanish verb. Otherwise, it could be a neologism candidate, an orthographic error or a word form pertaining to a different grammatical category. The next step of the procedure was then to analyse those candidates with value 0 in the DLE column in order to separate true neologisms from erroneous forms.

3.2.3. *Error reducing strategies.* In this part of the procedure we submitted all the verb candidates that did not match the reference dictionary to a series of tests to further separate true neologism candidates from misspellings and errors of other nature. Only if a candidate survived the full battery of tests was it considered a neologism candidate.

Prefixes. We used a list of prefixes in this process. If a prefix was detected in an infinitive, and the verb was listed in the dictionary without the prefix, then the verb was considered a neologism candidate, and then no other test was necessary. We adopted this strategy because one way for creating derivative neologisms is by adding prefixes to an existing word, such as in *recalendarizar* 'to re-calendarise', from the existing verb *calendarizar* 'calendarise' and the prefix *re-*. We extracted the list of prefixes from DLE, also by downloading it from Enclave RAE. The full list is shown in Table 4.

Table 4. List of prefixes used in the prefix detection process.

ana-, anti-, auto-, cata-, cis-, co-, con-, contra-, cuasi-, de-, des-, di-, dia-, dis-, em-, en-, entre-, es-, ex-, extra-, geo-, hiper-, hipo-, in-, infra-, inter-, intra-, para-, per-, peri-, pluri-, pos-, post-, pre-, pro-, re-, requete-, res-, rete-, semi-, sin-, so-, sobre-, son-, sub-, super-, tele-, tera-, trans-, tras- ,ultra-

Table 5, in turn, shows some examples of verbs not listed in the dictionary but bearing one of the prefixes. As can be seen in the Frequency columns, the lemma without the prefix, which does appear in the dictionary, is always far more frequent than the form with the prefix.

Table 5. Examples of verbs with their prefixes.

Candidate	Frequency	Prefix	Base verb	Frequency
autoevaluar	476	*auto-*	*evaluar*	350,844
desjerarquizar	66	*des-*	*jerarquizar*	9,668
recalendarizar	188	*re-*	*calendarizar*	655

Minimum length. Those candidates that did not match any prefix were subjected to the rest of the battery tests. If a given candidate did not pass all these tests, it was discarded. The first of such eliminating tests was to measure the length of the lemma in characters. As we considered word endings of at least three letters, we discarded in this step all lemmas with less than four characters. As explained by RAE and ASALE (2009), Spanish verbs tend not to be too short because they carry at least four morphological segments: the root, the thematic vowel, the tense/mood and the person/number. Moreover, verbs with affixation and enclitic pronouns are not infrequent, thus it is difficult to find verbs that can compress all this morphological information in less than four letters. Examples of verb lemmas that were eliminated in this step are **bar*, **kir* and **oir*. None of these appear as verb lemmas in the dictionary. The first form coincides with the noun *bar* 'bar', the second is not a word and the last one resembles the Spanish verb *oír* 'to hear' but lacking the diacritical mark due to misspelling.

Reinsertion of first letter. Our second attempt to separate neologism candidates from misspelling mistakes was to reinsert a first letter, as we observed this is a very common occurrence. Table 6 shows a few examples of false verb candidates that were obtained from the corpus, the forms **aber*, **econocer* and **guantar.* They certainly meet all the morphological criteria described so far (e.g. they present a root which combines with different word endings) and yet, they are not neologisms but spelling errors.

The idea was thus to try to reinsert an initial letter (such as *a*, *e*, *i*, *o*, *u*, *h*, *b*, *c*, *d*, *f*, *g*, *j*, *k*, *l*, *m*, *n*, *p*, *q*, *qu*, *r*, *s*, *t*, *v*) and then check again against the reference dictionary. If a match was found this time, then the candidate was discarded. Otherwise, it was submitted to the following tests. For example, in the cases of Table 6, by adding an initial letter, the system could match **aber,* a mistake, with the real verb *haber* 'to have', which is not a neologism candidate and was thus discarded.

Table 6. Examples of false candidates that were eliminated by inserting a first letter.

Candidate	Reinserted letter	Correct lemma
***aber**	*h-*	*haber*
***econocer**	*r-*	*reconocer*
***guantar**	*a-*	*aguantar*

Letter substitution. The remaining candidates were then screened for the detection of other forms of typographic errors. A typical case of misspelling is the substitution of one letter for another. For instance, a very common orthographic mistake in Spanish is the confusion between *b* and *v* (e.g. it was common to find the verb *avanzar* 'to move forward' misspelled as **abanzar*). In order to detect and discard this type of error, we applied a special case of a spell-checking algorithm consisting of a combination of letter substitution rules based on Spanish orthography (RAE and ASALE 2010). Table 7 shows the entire catalogue of rules. It should be read from left to right: for example, the letter or sequence of letters on the left is replaced by the letter or sequence on the right.

Table 7. List of pairs of characters used by the substitution rules.

s → c	n → m	nb → mb	ps → s	n → pn	ó → o
c → s	s → z	mb → nb	s → ps	pt → t	o → ó
b → v	z → s	mb → nv	ii → i	t → pt	ú → u
v → b	z → c	nv → mb	i → ii	bs → s	u → ú
j → g	y → ll	ep → pe	de → des	s → bs	u → ü
g → j	ll → l	pe → ep	di → dis	ns → s	ü → u
c → k	l → ll	p → pe	il → ll	s → ns	i → e
k → c	ll → y	pe → p	li → ll	st → s	e → i
k → qu	r → rr	nf → mf	in → ins	s → st	a → e
qu → k	rr → r	x → j	in → inc	á → a	e → a
ñ → n	h → -	j → x	oo → o	a → á	
n → ñ	y → i	ee → e	o → oo	é → e	
Ñ → ñ	i → y	e → ee	gn → n	e → é	
ny → ñ	mp → np	x → s	n → gn	í → i	
m → n	np → mp	s → x	pn → n	i → í	

If a letter substitution could be made in some verb candidate, the new form was then compared with those in the dictionary. If a match was found, then the candidate was discarded. Table 8 shows some examples of cases in which this happens.

Table 8. Examples of misspelled verbs that were detected using substitution rules.

Candidate	Correct lemma	Frequency
***abanzar**	*avanzar*	319,185
***blokear**	*bloquear*	42,481
***habrir**	*abrir*	375,674

Notice a very frequent case which is the incorrect addition of the letter *h*, a very common spelling mistake in Spanish because *h* is silent. The system eliminated this letter, resulting in the verb *abrir* 'to open', which is not a neologism. The table also shows the frequency of the correctly spelled form in the corpus. If no match was found with these substitution rules, the candidate was then submitted to the rest of the tests.

Orthographic similarity coefficient. We placed the orthographic similarity measures after the previously described rules because they are more computationally expensive. We combined two different similarity coefficients in order to find the orthographic similarity between a verb candidate and each verb in the dictionary. The first one is the Jaccard index, commonly used in natural language processing (Equation 1). For this we used two-letter sequences as features.

$$J(A,B) = \frac{|A \cap B|}{|A \cup B|} \quad (1)$$

Again because of the computational cost that this comparison entails, we limited the application of such coefficient only to those verbs in the dictionary that began with the same letter as the candidate (except in the case of *h*, given that this is, as already explained, a very common occurrence), that were similar in length and which had a minimum frequency of 15 occurrences in the corpus. Table 9 shows some examples of verb candidates that were matched with verbs in the dictionary using the Jaccard similarity coefficient.

Table 9. Examples of verb candidates that matched and entry in the dictionary using an orthographic similarity coefficient.

Candidate	Similar lemma(s) and their frequency in the corpus
***abadonar**	*abandonar* (150,342)
***contatar**	*contar* (845,098); *contratar* (141,316); *contentar* (2,396); *contactar* (46,139); *constatar* (75,305)

There we can see cases like **abadonar* and **contatar,* each of which found a match with at least one entry in the dictionary using this coefficient: **abadonar* matched *abandonar* 'to abandon', and *contatar* matched five possible verbs in the DLE. As in previous cases, if at least one match with the dictionary was found, and such match was above the frequency threshold, then the candidate was discarded. Otherwise, it continued to be subjected to further tests.

To complement the previous similarity index, a new orthographic similarity coefficient was implemented. As in the previous case, this new index takes two words as arguments (the verb candidate and each of the entries in the dictionary) and proceeds to compare them letter by letter. In a first loop, it compares the first letter of one with the first letter of the other. If the letter was the same in both cases, a variable we call *match* was increased by one, and then the algorithm moved to the next letter in both words. If, however, the letter was not the same, then the first loop stopped and a second began, in a very similar fashion, only that now instead of the first letter, it compared the last letter in both words. Again, if the letter was the same, then the *match* variable was increased by one. Then, the pointer is moved to the previous letter and the comparison is made again. The process finished when another mismatch was found, now from the opposite direction. The result of the comparison was thus the quotient between the number of matches and the character length of the larger word (2). As with the previous case, comparisons resulting in a value over an empirically defined threshold were considered positive matches.

$$ortSim(A,B) = \frac{match(A,B)}{\max(\text{length}(A), \text{length}(B))} \quad (2)$$

Table 10 shows two examples of verb candidates, **abodar* and **colacar*, that found a match to some entries in the dictionary. The same criterion of frequency in the corpus was used to avoid comparisons with very low frequency verbs of the dictionary.

Table 10. Examples of matches using the second similarity coefficient.

Candidate	Matches with dictionary
abodar	*abocar* (1,289); *abogar* (4,875); *acodar* (113); *abordar* (165,173)
colacar	*colar* (7,177); *colocar* (234,393)

Distributional profiling. The last filter we applied consisted of a distributional similarity index, based on the intuition that a true verb will show a characteristic distributional pattern. This was by far the most computationally expensive of all, as it involved the extraction from the corpus of contexts of occurrence of the verbs. For this reason, it was left for the end, because only a reduced number of verb candidates passed the previous tests.

We defined what would be the typical profile of co-occurrence of a Spanish verb in its infinitive form by examining contexts of occurrence in the corpus of genuine verbs. We mainly resorted to verbal periphrases consisting of an auxiliary verb plus an infinitive form (Gómez Torrego 1999). We manually compiled a list of 50 mainly modal periphrases, such as *tengo que* 'I have to' or *hay que* 'it is necessary to'; aspectual periphrases like *comenzó a/comienza a* or *empiece a/empieza a* 'he/she start/started to', as well as other types, such as *se atrevió a* 'he/she dared to'; *sirve para* 'it is used for'. We also included other expressions that frequently precede infinitive forms, such as *el arte de* 'the art of'; *ganas de* 'lust for'; *máquinas de* 'machines for'; *un aparato para* 'a device for'; etc. Using this list, the last testing algorithm extracted contexts of occurrence of the candidate from the corpus and if it found that there were no matches of any of these expressions, then it eliminated the candidate. With this method, the algorithm was able to eliminate false candidates such as **acuer*, **brujer* or **shakespear*, that is, forms that do appear in the corpus with some frequency, but not with the predictors we defined for the category of verbs.

4. Results and evaluation

From the total of 2,954,973 form types with frequency > 5 in the corpus (i.e. the part that we used for evaluation), we obtained a list of 10,034 verb candidates. From these, 5,685 (56,65%) were already listed in the DLE and 4,349 (43,34%) were submitted to the battery of tests to discard orthographic errors. From this list, only 774 forms (17,79%) survived all the filters and were therefore considered neologism candidates.

Considering the list of verbs that did not match the dictionary, the vast majority was discarded by some filter, as shown in Table 11. Only the candidates with the categories of 'prefixes' and 'survivors' were considered true neologisms by the algorithm. The first category comprises those with a prefix and other verbs that passed all filters. The table also presents the results of manual evaluation of the performance of each filter or category.

Table 11. Results of the classification and application of filters.

Filter/Selection	Tag	Total	% correct	% incorrect
Minimum length	false	21	100%	0%
Reinsertion of first letter	false	231	98,2%	1,7%
Letter substitution	false	873	99,5%	0,4%
Orthographic similarity 1	false	950	89,5%	10,4%
Orthographic similarity 2	false	568	97,3%	2,6%
Distributional profiling	false	932	90,4%	9,5%
Prefixes	true	398	96,7%	3,2%
Survivors	true	376	79,5%	20,47%

The best performing filter was the 'Minimum length'. This is also the less computationally expensive one, as it only counts the number of letters of the candidates. The 'Reinsertion of first letter' and 'Letter substitution' filters also had good performance, with 98% and 99% precision respectively. The 'Orthographic similarity 1' filter was less reliable, with 89% precision. It resulted in errors such as discarding genuine neologisms like *bloguear* 'to blog' because of its similarity with the verb *bloquear* 'to block'. The 'Distributional profiling' filter was also less effective, with 90% precision. It incorrectly eliminated valid neologisms such as *flashear* 'to be shocked' because they do not show the typical distribution of normal verbs, or perhaps because we collected fewer periphrases than were needed in the case of these less frequent verbs. One way to mitigate this problem would be to focus on the evolution of the frequency curve in the timeline, as done by Nazar and Vidal (2010). That is, if a word like *flashear* or *bloguear* shows a sharp rise in its frequency of use, then this can be taken as a strong indication of neology. We leave this possibility, however, for future work.

Table 12 presents the results of the overall evaluation, which informs how probable it is that the correct true/false tag will be assigned to a given candidate.

Table 12. Proportion of correct / incorrect results in the detection of neologisms

Tag	Correct		Incorrect		Total
<TRUE>	684	88,37%	90	11,62%	774
<FALSE>	3,364	94,09%	211	5,90%	3,575

We manually evaluated the total 4,349 verbs submitted to the battery of tests. We define precision and recall in terms of true positives (*tp*), the case of a true neologism that is correctly detected by the algorithm; false positives (*fp*), the case of a candidate that is not a neologism but is anyway promoted by the algorithm; and finally false negatives (*fn*), the case of a neologism that was not detected by the algorithm (i.e., a true neologism that was discarded by some filter). Using these three variables we calculated precision (3), which measures how likely it is that a promoted candidate will be a true neologism, and recall (4), which measures the exhaustivity of the method, i.e., how likely it is that a true neologism will be detected.

$$\text{precision} = \frac{tp}{(tp + fp)} \tag{3}$$

$$\text{recall} = \frac{tp}{(tp + fn)} \tag{4}$$

The overall precision of the detection of new verbs was 0.88 and recall was 0.76. Thus, close to 88% of the final list of selected verbs are valid neologism candidates, resulting from different word formation mechanisms.

With respect to the errors made by the algorithm, we found that they can be classified in three groups: a) errors due to failure to detect missing accents (e.g. **sonreir* was labelled a neologism whereas it is a misspelling of *sonreír* 'to smile') and other typos (e.g. *cosntruir* for *construir* 'to build'); b) errors due to failure to detect verbs from other languages (e.g. *aproveitar,* a Galician verb, was considered a valid Spanish verb), and c) errors due to incorrect interpretation of a form as a verb, with the subsequent creation of a nonexistent verb (e.g. **universar, *umbilicar*).

Table 13 shows some examples of the list of neologism candidates, and the type of neology mechanism according to Cabré (2015).

Table 13. Examples of new verbs detected with the proposed method.

Type of neology mechanism (Cabré 2015)	**New verbs**	**English translation**
Prefixation (adding a prefix to an existing verb)	*coorganizar, desactualizar, intercomunicar, precalcular, reasignar,* etc.	'to co-organise', 'to make something out-of-date', 'to inter-communicate', 'to pre-calculate', 'to reassign', etc.
Suffixation (adding a verb suffix to an existing noun)	*aperturar, chocolatear, cooperativizar, ficcionar, gelificar,* etc.	'to open', 'to add chocolate', 'to create an association', 'to turn into fiction', 'to turn into gel', etc.
Parasynthesis (adding a prefix and suffix to an existing noun)	*adinerar, desvirtualizar, enrutar,* etc.	'to enrich', 'to devirtualise', 'to put in a route', etc.
Loans (adding a Spanish verb suffix to a loan)	*customizar, draftear, loguear, resetear, spoilear,* etc.	'to customise', 'to make a draft', 'to log', 'to reset', 'to spoil', etc.

Some of these verbs are frequently used in everyday language, and the fact that they have not been incorporated into the dictionary might be due to different factors, as discussed in section 2. Freixa and Torner (2020), among others, point out that there can be a mismatch between the real institutionalisation of a word and the fact that the word is present in the dictionary. If we observe the 100 most frequent candidates in our results, most of them could be included in the dictionary, as they are frequent enough (they appear from 169 to 40,409

times in the corpus), they are morphologically correct and they are widespread across Spanish-speaking countries. This is the case of *vivenciar* 'to experiment', *direccionar* 'to address', *googlear* 'to google', *suplementar* 'to supplement', *referenciar* 'to refer', *resetear* 'to reset', *customizar* 'to customise', etc.

It is not the goal of this study to try to find out why these verbs were not included in the dictionary used to test the method. However, we can say that the resulting list of candidates, with very few errors, would allow a lexicographic team to work with empirically-based data and obtain valid information for potential new verb entries for dictionary updating. Once the team had this information, it could make informed decisions about which of these lemmas should be included and why, instead of using non-systematic ways of including new words in dictionaries (e.g. following social trends or through personal findings in newspapers). As already observed, there are many reasons why a new word should or should not be included in a dictionary. Some of them may be based on the characteristics of the dictionary project, combined with the characteristics of the potential new lemma. For example, in the examples in Table 13, most of the verbs could be included in a general descriptive dictionary, while *aperturar, enrutar, customizar, draftear, loguear, resetear* or *spoilear*, being unnecessary loans, would not be adequate for a prescriptive dictionary.

Finally, an interesting and unexpected result was the great number of verb entries from the DLE that were not present in the large corpus we used in our experiments. From the total of 11,893 verb lemmas of the dictionary, we found 6,208 that did not appear with the minimum frequency. This represents a very large proportion of verbs listed in the dictionary that are very infrequent or non-existent, and our intuition is that this is because they have fallen into desuetude (Algeo 1993). For that reason, we believe that the study of desuetude is complementary to the study of neology for updating dictionary lemma lists, but we will leave the exploration of this topic for future work.

5. Conclusions

In this paper, we proposed a method for the detection of new verbs, taking into account the difficulty of part of speech tagging and the less-than-perfect performance of current taggers. The proposed method, which is relatively simple, performs well in the detection of verbs in Spanish, and differentiates verbs that are already present in a dictionary such as DLE from the ones that the dictionary did not register. We also highlight the effectiveness of the method at reducing the noise of spelling mistakes in the corpus. The typification of the most frequent errors offers much help in reducing the burden of the otherwise manual revision of the data. Another characteristic of the proposed method is that it is based mainly on morphological features to reduce the error rate, avoiding the use of semantic features, which are harder to systematise.

The proposed algorithm can be easily incorporated into a workflow for creating or updating a lexicographic database, so it can be used regularly to detect new verbs and contribute to a more dynamic and updated dictionary. The vision of the electronic dictionary as 'up-to-date' and a 'dynamic repository of knowledge' was already presented by De Schryver (2003: 159), but almost 20 years later there is still plenty of work to do in this respect. The automatic interrogation of a corpus to find verbs not present in a list of headwords is a reliable method for tracing new verbs and adding them to a list of possible future verb entries to the dictionary, aiding the lexicographer in her/his decision-making

process. Information such as the first appearance in the corpus, as well as the frequency and dispersion of the word, can also be included in the entry, in order to orient the user to the neological nature of a specific unit. Depending on the type of dictionary and the conditions of the lexicographic projects (such as technical, methodological, economical, etc.), a dictionary for the Internet can be updated regularly with information about the frequency of the word, taking data from corpora of different periods of times, comparing them and showing the results in each entry. The general approach is not new (Renouf 1993), but it has not yet been fully incorporated into the lexicographic procedures of Spanish dictionaries, nor is it a standard procedure in other lexicographic traditions. As future work, we would like to add supplementary information about the new verbs to the method. For example, the method should be able to recognise whether a verb is only used in its participle form, or if it is used with a specific alternation, among other features that could help lexicographers to add accurate information to the dictionary entry and having first clues about its macrostructure.

References

A. Dictionaries

Real Academia Española. 2014. *Diccionario de la lengua española* (Twenty-third edition.). Madrid: Espasa.

Real Academia Española. Online. *Enclave RAE*. Accessed on 13 June 2020. https://enclave.rae.es.

B. Other literature

Abel, A. and E. Stemle. 2018. 'On the Detection of Neologism Candidates as Basis for Language Observation and Lexicographic Endeavours: The STyrLogism Project' In Čibej, J., V. Gorjanc, I. Kosem and S. Krek (eds), *Proceedings of the XVIII EURALEX International Congress*, *EURALEX 2018*, *Ljubljana*, *Slovenia*, *July 17-21, 2018*. Ljubljana: University of Ljubljana, 535-544.

Adelstein, A. and I. Kuguel. 2008. *De* salariazo *a* corralito*, de* carapintada *a* blog*. Nuevas palabras en veinticinco años de democracia.* Los Polvorines: Universidad Nacional de General Sarmiento.

Algeo, J. 1993. 'Desuetude among New English Words.' *International Journal of Lexicography* 6.4: 281-293.

Alvar Ezquerra, M. 2007. 'El neologismo español actual' In Luque Toro, L. (ed), *Léxico español actual. Actas del I Congreso Internacional de Léxico Español Actual, Venice-Treviso*, *March 14-15*, *2005*. Venezia: Libreria Editrice Cafoscarina, 11-35.

Amore, M., S. McGregor and E. Jezek. 2018. 'Distributional Analysis of Verbal Neologisms: Task Definition and Dataset Construction' In Cabrio, E., A. Mazzei and F. Tamburini (eds), *Proceedings of the Fifth Italian Conference on Computational Linguistics*, *CLiC-it 2018*, *Torino*, *Italy*. *December 10-12, 2018*. CEUR-WS.org, n. p.

Atkins, B. S., J. Kegl and B. Levin. 1988. 'Anatomy of a Verb Entry: From Linguistic Theory to Lexicographic Practice.' *International Journal of Lexicography* 1.2: 84-126.

Atkins, B. S., and M. Rundell. 2008. *The Oxford Guide to Practical Lexicography*. Oxford: Oxford University Press.

Baayen, R. H. and A. Renouf. 1996. 'Chronicling the Times: Productive Lexical Innovations in an English Newspaper.' *Language* 72.1: 69-96.

Battaner, P. and S. Torner. 2008. 'La polisemia verbal que muestra la lexicografía' In Azorín, D., B. Alvarado, J. Climent, M. I. Guardiola, R. Lavale Ortiz, C. Marimón Llorca, J. J. Joaquín Martínez, X. A. Padilla, H. Provencio, I. Santamaría-Pérez, L. Timofeeva and E. Toro (eds). *Actas del II Congreso Internacional de Lexicografía Hispánica: el diccionario como puente entre las lenguas y culturas del mundo.* Alicante: Universidad de Alicante, 204-246.

Boas, F. 2001. 'Frame Semantics as a Framework for Describing Polysemy and Syntactic Structures of English and German Motion Verbs in Contrastive Computational Lexicography.' In Rayson, P., A. Wilson, T. McEnery, A. Hardie and S. Khoja (eds), *Proceedings of the Corpus Linguistics 2001 Conference.* Lancaster: Lancaster University, 64-73.

Cabré, M. T. (1999). *Terminology. Theory, Methods and Applications.* Amsterdam: John Benjamins.

Cabré, M. T. and R. Estopà. 2009. 'Trabajar en neología con un entorno integrado en línea: la estación de trabajo OBNEO.' *Revista de Investigación Lingüística* 12: 17-38.

Cabré, M. T. and R. Nazar. 2012. 'Towards a New Approach to the Study of Neology.' *Neologica* 6: 63-80.

Cabré, M. T. 2015. 'Bases para una teoría de los neologismos léxicos: primeras reflexiones' In Alves, I. and E. Simões (eds), *Neologia das línguas românicas*. São Paulo: CAPES, Humanitas, 79-107.

Cañete, P., S. Fernández-Silva and B. Villena. 2019. 'Estudio de los neologismos terminológicos difundidos en el diario *El País* y su inclusión en el diccionario.' *Círculo de Lingüística Aplicada a la Comunicación* 80: 135-158.

Cartier, E. and J.-F. Sablayrolles. 2008. 'Néologismes, dictionnaires et informatique.' *Cahiers de Lexicologie* 2008-2.93: 175-192.

Cartier, E. 2016. 'Neoveille, système de repérage et de suivi des néologismes en sept langues.' *Neologica* 10: 101-131.

Cook, C. 2010. *Exploiting Linguistic Knowledge to Infer Properties of Neologisms*. Ph.D. Thesis, University of Toronto.

Costin-Gabriel, C. and T. E. Rebedea. 2014. 'Archaisms and Neologisms Identification in Texts' In *2014 RoEduNet Conference 13th Edition: Networking in Education and Research Joint Event RENAM 8th Conference*. IEEE, 1-6.

De Schryver, G.-M. 2003. 'Lexicographers' Dreams in the Electronic-Dictionary Age'. *International Journal of Lexicography* 16.2: 143-199.

El Maarouf, I., V. Baisa, J. Bradbury and P. Hanks. 2014. 'Disambiguating Verbs by Collocation: Corpus Lexicography Meets Natural Language Processing.' In Calzolari, N., K. Choukri, T. Declerck, H. Loftsson, B. Maegaard, J. Mariani, A. Moreno, J. Odijk and S. Piperidis (eds), *Proceedings of the Ninth International Conference on Language Resources and Evaluation (LREC'14).* Reykjavik: European Language Resources Association, 1001-1006.

Freixa, J. and S. Torner. 2020. 'Beyond Frequency: On the Dictionarization of New Words in Spanish.' *Dictionaries: Journal of the Dictionary Society of North America* 41.1: 131-153.

Fuentes, M., C. Gerding, A. C. Pecchi, G. A. Kotz G. and P. Cañete. 2009. 'Neología léxica: Reflejo de la vitalidad del español de Chile.' *Revista de Lingüística Teórica y Aplicada* 47.1: 103-124.

Ginebra, J. and X. Rull. 2010. 'Tendències en el règim sintàctic dels neologismes verbals: una aproximació' In Cabré, M. T., O. Domènech, R. Estopà, J. Freixa and M. Lorente (eds), *Actes del I Congrés Internacional de Neologia de les Llengües Romàniques.* Barcelona: IULA, 377-390.

Gómez Torrego, L. 1999. 'Los verbos auxiliares. Las perífrasis verbales de infinitivo' In Bosque, I. and V. Demonte (eds), *Gramática descriptiva de la lengua española*. Madrid: Espasa: 3323-3388.

Hanks, P. 2013. *Lexical Analysis: Norms and Exploitations.* Cambridge, MA: MIT Press.

Janssen, M. 2005. 'Neo Track: Semiautomatic Neologism Detection.' Communication presented in *XXI Encontro Nacional da Asociação Portuguesa de Lingüística*. Lisbon.

Janssen, M. 2009. 'Detección de neologismos: una perspectiva computacional.' *Debate Terminológico* 5: 68-75.

Kerremans, D. and J. Prokić. 2018. 'Mining the Web for New Words: Semi-Automatic Neologism Identification with the NeoCrawler.' *Anglia* 136.2: 239-268.

Kilgarriff, A. and I. Renau. 2013. 'esTenTen, a Vast Web Corpus of Peninsular and American Spanish.' *Procedia-Social and Behavioral Sciences* 95: 12-19.

Klosa-Kückelhaus, A. and H. Lüngen. 2018. 'New German Words: Detection and Description' In Čibej, J., V. Gorjanc, I. Kosem and S. Krek (eds), *Proceedings of the XVIII EURALEX International Congress*, *EURALEX 2018*, *Ljubljana*, *Slovenia*, *July 17-21, 2018*. Ljubljana: University of Ljubljana, 559-569.

Klosa-Kückelhaus, A. and I. Kernerman. 2020. 'Global Viewpoints on Lexicography and Neologisms: An Introduction.' *Dictionaries: Journal of the Dictionary Society of North America* 41.1: 1-9.

Langemets, M., J. Kallas, K. Norak and I. Hein. 2020. 'New Estonian Words and Senses: Detection and Description.' *Dictionaries: Journal of the Dictionary Society of North America* 41.1: 69-82.

Lavale-Ortiz, R. 2016. 'Hacia una revisión del concepto de neologismo aplicado a los verbos denominales aparecidos en la prensa española.' *Revista Española de Lingüística Aplicada* 29.1: 165-190.

Lemnitzer, L. 2010. *Die Wortwarte. Wörter von heute und morgen. Eine Sammlung von Neologismen.* Accessed on 11 June 2020. http://www.wortwarte.de.

L'Homme, M. C. 2003. 'Capturing the Lexical Structure in Special Subject Fields with Verbs and Verbal Derivatives. A Model for Specialized Lexicography.' *International Journal of Lexicography* 16.4, 403-422.

L'Homme, M. C. 2015. 'Predicative Lexical Units in Terminology.' In Gala, N., R. Rapp and G. Bel-Enguix (eds), *Language Production, Cognition, and the Lexicon.* Berlin: Springer, 75-93.

Marello, C. 2010. 'Verbos con construcciones tanto transitivas como intransitivas y/o pronominales en los diccionarios monolingües y bilingües italianos y españoles.' In

Castillo Carballo, M. A. and J. M. García Platero (coords), *La lexicografía en su dimensión teórica.* Málaga: Universidad de Málaga, 411-433.

Nam, K., S. Lee and H. Jung. 2020. 'The Korean Neologism Investigation Project: Current Status.' *Dictionaries: Journal of the Dictionary Society of North America* 41.1: 105-129.

Nazar, R. and V. Vidal. 2010. Aproximación Cuantitativa a la Neología. In Cabré, T., O. Domènech, R. Estopà, J. Freixa and M. Lorente (eds), *Actas del Congreso Internacional de Neología en las lenguas románicas (CINEO)*. Barcelona: IULA, 867-880.

Ortega Martín, M. P. 2001. 'Neología y prensa: un binomio eficaz'. *Espéculo. Revista de Estudios Literarios* 18.

Parra Escartín, C. and H. Martínez Alonso. 2015. Choosing a Spanish Part-of-Speech Tagger for a Lexically Sensitive Task. *Procesamiento del Lenguaje Natural.* 54: 29-36.

Real Academia Española (RAE) and Asociación de Academias de la Lengua Española (ASALE). 2009. *Nueva gramática de la lengua española*. Madrid: Espasa.

Real Academia Española (RAE) and Asociación de Academias de la Lengua Española (ASALE). 2010. *Ortografía de la lengua española*. Madrid: Espasa.

Renouf, A. 1993. 'Sticking to the Text: A Corpus Linguist's View of Language.' *ASLIB Proceedings* 45.5: 131-136.

Renouf, A. 2016. 'Big Data and its Consequences for Neology.' *Neologica* 10: 15-37.

Rey, A. 1976. 'Néologisme, un pseudo-concept?' *Cahiers de Lexicologie* 28.1: 3-17.

Sanmartín, J. 2010. 'El neologismo castellano en un corpus de prensa editada en la comunidad valenciana: ¿un hecho diferencial?' In Cabré, M. T., O. Domènech, R. Estopà, J. Freixa and M. Lorente (eds), *Actes del I Congrés Internacional de Neologia de les Llengües Romàniques.* Barcelona. IULA, 693-709.

Vivaldi, J. 2000. 'SEXTAN: prototip d'un sistema d'extracció de neologisms' In Cabré, M. T., J. Freixa and E. Solé (eds), *La Neologia en el tombant de segle*. Barcelona: IULA, 165-177.