# A method to automatically identify diachronic variation in collocations

**Marcos Garcia**
Universidade da Coruña, Grupo LyS
Dpto. de Letras, Facultade de Filoloxía
Campus da Zapateira, 15701, Coruña
Universidade da Coruña, CITIC
Campus de Elviña, 15701, Coruña
marcos.garcia.gonzalez@udc.gal

**Marcos García-Salido**
Universidade da Coruña, Grupo LyS
Dpto. de Letras, Facultade de Filoloxía
Campus da Zapateira, 15701, Coruña
marcos.garcias@udc.gal

## Abstract

This paper introduces a novel method to track collocational variations in diachronic corpora that can identify several changes undergone by these phraseological combinations and to propose alternative solutions found in later periods. The strategy consists of extracting syntactically-related candidates of collocations and ranking them using statistical association measures. Then, starting from the first period of the corpus, the system tracks each combination over time, verifying different types of historical variation such as the loss of one or both lemmas, the disappearance of the collocation, or its diachronic frequency trend. Using a distributional semantics strategy, it also suggests linguistic structures that convey meanings similar to those of extinct collocations. A case study on historical corpora of Portuguese and Spanish shows that the system speeds up and facilitates the finding of some diachronic changes and phraseological shifts that are harder to identify without using automated methods.

## 1 Introduction

One of the main characteristics of natural language is change, as there is no evidence of any language which does not show different types of variation. Change seems to affect all the strata of natural languages: phonology, morphology, syntax, and semantics. Besides this language-internal perspective, the study of language variation may also take into account the external causes of change: that is, geographical, social, or historical factors, among others (Chambers and Schilling, 2013).

Historical (*diachronic*) studies of language, carried out by philologists and historical linguists, have shown how language evolves over time, finding interesting cross-linguistic generalizations. In those cases where digitalized resources exist, several corpus linguistics and natural language processing (NLP) methods have been applied to automate the discovering of language change, thus alleviating the effort of searching for linguistic variation (Curzan, 2008; Dipper, 2008). In this regard, frequency-based strategies are useful to identify increases and decreases in the use of some linguistic phenomena (Hilpert and Gries, 2016). The rise of distributional semantics methods (both count-based and neural network approaches) also allowed researchers to track semantic change in different time periods (Sagi et al., 2009; Gulordava and Baroni, 2011; Kulkarni et al., 2015; Hamilton et al., 2016; Bamler and Mandt, 2017; Gamallo et al., 2018).

A particular case of diachronic variation is the evolution of lexical combinations over time. In this respect, research on the diachrony of complex predicates has provided useful knowledge for theoretical studies on language evolution (Anderson, 2006; Butt and Lahiri, 2013; Elenbaas, 2013). From a different perspective, historical analyses of collocational patterns have shown that some lexical restrictions vary diachronically, while some others seem to be more persistent. Thus, studies such as Alba-Salas (2007) or García-Salido (2017) explore how Spanish causative verbs such as *hacer* ('to make') or *poner* ('to put') were replaced by *dar* (literally, 'to give') to express causation with different nouns such as *miedo* ('fear') or *vergüenza* ('embarrasment'): *hacer vergüenza* ⇒ *dar vergüenza*; *poner miedo* ⇒ *dar miedo*. These examples show the asymmetry of collocations, understood as combinations where one of their lexical units (LUs) (the COLLOCATE: *hacer*, *poner*, or *dar*) is lexically selected by the other (the BASE: *miedo*, *vergüenza*) (Mel'čuk, 1998).

Understanding the properties of collocations and other multiword expressions, both in a specific period of time and diachronically, is crucial not only to understand how a particular lan-



*Proceedings of the 1st International Workshop on Computational Approaches to Historical Language Change*, pages 71–80
Florence, Italy, August 2, 2019. 

guage evolves, but also to develop computational methods for language processing (Sag et al., 2002; Ramisch and Villavicencio, 2018). However, this type of analyses has benefited less from computational approaches, whereby NLP systems could facilitate the automatic identification of variations in lexical combinations. Tools such as *DiaCollo* (Jurish, 2015) or JESEME (Hellrich and Hahn, 2017) are able to track changes in word associations and lexical semantics, but they are not specifically designed to analyze combinations of syntactically dependent lexical units like the ones exemplified above.

Taking the above into account, we present a new method to analyze, in historical corpora, the diachronic distribution of collocations and their internal components. Besides the period when certain collocations start to be used, the method identifies four variation types: (1) the disappearance of both lexical units of the collocation; the loss of (2) the base or (3) of the collocate, and (4) the loss of certain combinations whose constituent lemmas are still used. In each case, the system searches for other similar combinations and proposes possible replacements. Furthermore, it classifies the increase, decrease, or stability of collocations that continue to be used.

In order to evaluate the usefulness of the proposed method, we carry out a case study on several historical corpora of Spanish and Portuguese. The analyses, both quantitative and qualitative, indicate that the presented approach allows historical linguists to rapidly analyze the diachronic evolution of collocations, showing some interesting changes in lexical combinations of the two languages. The system is freely available and can be applied to any historical corpus parsed in a CoNLL-like format.[1]

The remainder of this paper is organized as follows. Section 2 presents some related work on computational approaches to language change, and Section 3 briefly discusses the theoretical properties of collocations. In Section 4 we describe our method to identify diachronic variation of these expressions. Then, Section 5 shows the results of both quantitative and qualitative evaluations of the system as well as an error analysis, and finally, the conclusions and further work are addressed in Section 6.

[1]The annotated corpora and the software used in this paper are released under open-source licences at http://www.grupolys.org/~marcos/pub/diachronic_collocations.zip

## 2 Related Work

Besides historical linguistic approaches adopted by the philological tradition, the availability of diachronic corpora in digital formats allowed researchers from different areas to implement computational approaches to explore historical language change. In this regard, Lieberman et al. (2007) analyzed the past tense of English verbs over $1,200$ years, showing that the rate of regularization (i.e., the emergence of an *-ed* past form) is directly related to frequency.

Using distributional semantic methods, Sagi et al. (2009) and Cook and Stevenson (2010) found examples of meaning shift by working with historical corpora combining quantitative and qualitative analyses. The former study identified the probability of semantic change by measuring the density of a vector space. The latter concentrated on amelioration and pejoration cases, that is, words that change from negative to positive opinions (e.g., the meaning of *nice* was 'foolish'), or from positive to negative ones (e.g., *vulgar* meant 'common').

More recently, several works have taken advantage of the Google Books Ngrams to train English distributional models of different periods in order to find semantic change over time (Gulordava and Baroni, 2011; Wijaya and Yeniterzi, 2011; Kim et al., 2014; Kulkarni et al., 2015). Similarly, Hamilton et al. (2016) defined a methodology to quantify semantic change using four languages (Chinese, English, German, and French). The results of this article suggest that polysemous words are those with higher rates of semantic change, and that the meaning of frequent words is more stable over time. The Google Books Ngrams were also used to implement dependency-based distributional semantics methods to track the semantic change in Spanish (Gamallo et al., 2017, 2018). To avoid the alignment problem between the vector space of each time period, studies such as Bamler and Mandt (2017) and Rudolph and Blei (2018) learn a joint time-aware semantic space by means of dynamic embeddings.

Designed specifically to explore the diachronic contexts of words, *DiaCollo* allows historical linguists to analyze the typical collocates of a given word over time, providing useful information to identify potential semantic shifts (Jurish, 2015). JESEME also takes advantage of historical distributional semantics models to create diachronic

charts for tracking semantic variation and word emotion over time (Hellrich and Hahn, 2017).[2]

Inspired by several of these works, our method uses natural language processing techniques and distributional semantics methods to support historical linguists to find diachronic changes of collocations in different languages.

## 3 Collocations

There are at least two main views of the concept of *collocation*. In the Firthian tradition, collocations are arbitrary and recurrent co-occurrences of two or more words within a short space of each other in a text (Benson, 1990; Sinclair, 1991). From this point of view, collocations are word combinations occurring together in a given span with greater frequency than randomly expected (e.g., "night, dark").

Along with this statistical or empirical approach, in the field of phraseology, authors such as Hausmann (1989) or Mel'čuk (1998) conceive collocations as directional combinations of two syntactically related lexical units. According to this approach, one of the LUs that form the collocation (the BASE) is often defined as *autosemantic*, because it is chosen by the speaker due to its meaning. The base, in turn, restricts the selection of the other LU (the COLLOCATE), which conveys a particular meaning depending on a given base (e.g., "take$_{\text{Collocate}}$ (a) picture$_{\text{Base}}$", "black$_{\text{C}}$ coffee$_{\text{B}}$") and is therefore said to be *synsemantic*. This conception of collocations encompasses quite an ample range of compositional lexical combinations (Mel'čuk, 1998), ranging from support verb constructions—in which verbs provide a tenuous lexical meaning (e.g. *Peter took a walk* $\sim$ *Peter's walk*)—to other types of idiosyncratic couplings, where collocates express full meanings, but are not freely interchangeable with theoretical synonyms (see the case of Pt. *arrenegar* with the meaning 'abjure' used in some sections of the corpus almost exclusively in company of *demónio* 'devil' or *diabrura* 'deviltry' in Section 5).

In spite of the differences between the two approaches, there have been recent attempts at using statistical measures to automatically identify phraseological collocations. For instance, Pecina (2010) investigates the performance of a large set of statistical association measures in identifying phraseological combinations. The target collocations of Pecina are only partially coincident with the definition given above, as, along with collocations such as *make a decision*, they also include non-compositional combinations. More recently, Evert et al. (2017) and Uhrig et al. (2018) undertook a research with similar purposes, but, in contrast to Pecina (2010), who started from bigrams, they used dependency parsing to identify collocation candidates and, instead of manual identification of phraseological combinations, they used collocation dictionaries as gold standards.

This paper also combines the statistical and phraseological approaches. Whereas phraseological collocations seem more interesting for diachronic investigations, statistical information can serve as a tool for identifying collocation candidates. The method proposed takes advantage of dependency parsing to identificate syntactically-related base–collocate candidate pairs, and uses statistical analysis in order to identify collocation candidates in each historical period.

## 4 Identification of diachronic changes on collocations

### 4.1 Method overview

The strategy for identifying historical variations on collocations consists of analyzing each of these combinations over time, starting from the first epoch when the collocation appears in the corpus. For each collocation, we identify whether it is still used in the following periods, and if it disappears, we verify what type of change it has undergone: loss of one or both LUs, or loss of the combination. As the collocation bases are those elements carrying the bulk of the lexical meaning, we check different candidates with the same base (or a very similar one) in those cases where only the collocate ceased to be used, with a view to finding examples such as the one referred above ($poner_C\ miedo_B \Rightarrow dar_C\ miedo_B$). As Section 4.4 will show, other alternatives (e.g., verbs with the same meaning of the collocations) can also be proposed.

### 4.2 Resources

In order to analyze the diachrony of collocations, our system needs historical corpora divided in different periods $p_1$, $p_2$, ..., $p_n$. Each corpus must have a CoNLL-like format containing lemmas, POS-tags and dependency labels. Also, the system uses word embeddings models to search for

[2] http://jeseme.org/

words with similar distributions. Optionally, it can take advantage of contemporary resources such as a dictionary of lemmas and a reference corpus (e.g., Wikipedia), used to reduce the noise present in historical corpora.

It is worth noting that in diachronic resources the *same word* can be written in different ways, due to variations in spelling, or because of morphological or phonological changes. For instance, the above mentioned Spanish word *vergüenza* can be found written as *berguensa*, *verguensa*, *berguenza*, or *verguença* (among others) in historical corpora (Vaamonde, 2015). As our objective is to find phraseological combinations of words, the system presented in this paper behaves better with normalized texts, where the lemmas have the same spelling across the different resources. Nevertheless, we take advantage of distributional models which encode subword information, so they can effectively tackle rare words present in historical resources (Bojanowski et al., 2017). In this regard, Section 5 includes experiments using normalized corpora (in Portuguese and Spanish) as well as a non-normalized historical corpus of Portuguese.

### 4.3 Extraction of collocation candidates

Once we have the analyzed corpora, we extract head–dependent pairs of the desired syntactic relations in order to identify candidates of collocations. For example, the *verb-object* dependency will extract instances such as 'eat, sausage' or 'take, shower'. These pairs are then ranked using statistical association measures to identify those candidates that are more likely to be phraseological collocations (Gries, 2013; Carlini et al., 2014; Evert et al., 2017).

### 4.4 Diachronic track of collocations

The process of tracking the diachronic evolution of collocations consists of the following steps:

- Starting from $p_1$, we select the *n* top collocations according the defined association measure and threshold. Optionally, in order to avoid possible noise in historical corpora, we select only those collocations whose internal elements are known (i.e., they appear in a contemporary dictionary), or have a very similar distribution (e.g., 0.9 of cosine similarity) to known present words.

- We calculate the ratio per period of each collocation dividing its frequency by the number of syntactic dependencies with the same relation (e.g., *subject*) in the same period.

- Then, for each collocation, we verify whether it appears in the next more recent period of the corpus (or ideally, in the reference one). If the collocation is not currently used:

  1. We traverse each period $p_{1+i}$ to identify when the collocation ceased to be used.
  2. Then, we analyze the type of change: (type 1) both the base and the collocate are not used anymore in the corpus; (2) the base, or (3) the collocate do not appear in further periods; (4) both LUs still occur, but the combination ceased to exist. In types 1 and 2 we use the distributional model to search for replacements for the base (for both types) and of the collocate (only for type 1). Using these candidates, we select further collocations whose base and collocate have cosine similarities greater than two given thresholds (*base_simil* and *collocate_simil*). In those cases where the base still appears in phraseological combinations (change 3, and eventually 4), we search for other combinations with the same base to find new collocates with the same lexical function.

     In *verb-object* collocations (e.g., *hacer venganza* or *tomar vingança*, 'take revenge' in Old Spanish and Portuguese) we also search (i) for verbs which convey the same meaning (e.g., *vingar*, 'revenge' in Portuguese), also using the word embeddings model, as well as (ii) for collocations with support verb constructions (*dar venganza*, 'take revenge' in Modern Spanish).

- If the collocation is still used in further historical periods, we obtain its frequency trend using the ratios of each period. This analysis classifies the trend of a collocation as *increase*, *decrease* or *stable* (types 5, 6, and 7, respectively).

Thus, the output of our system contains, for each collocation in the corpus (a) the period when it started to appear, (b) the type of change it undergone (if any), and the time when it happened,

as well as (c) the frequency trend of those collocations which have not suffered lexical variations. Additionally, for some combinations, it shows other expressions (collocations and eventually verbs) which could be replacements for those collocations which ceased to be used.

## 5 Experiments

### 5.1 Data

To verify the usefulness of the proposed method for automatically finding changes on collocations, we carried out a case study on two historical corpora of Portuguese and Spanish (with 648k and 808k tokens of private letters, respectively) from the *P.S. Post Scriptum* project (CLUL, 2014; Vaamonde, 2015).

Both resources are divided into centuries, from the 16th to the 19th century, and include versions with normalized spelling. We used the provided tokens and lemmas, and applied two NLP pipelines to POS-tag (LinguaKit, Garcia and Gamallo (2015)) and parse (UDPipe, Straka and Straková (2017)) the corpora using Universal Dependencies 2.3 (Nivre et al., 2018). As contemporary resources of Portuguese and Spanish, we used the dictionaries included in LinguaKit, and recent versions of Portuguese and Spanish Wikipedia (November, 2018) processed using the same tools as the corpora.

For computing the distributional similarity we trained *fastText* embeddings (Bojanowski et al., 2017) with mixed historical and present corpora, of about $250M$ for each language. For Spanish, we used *cuentos españoles* and *romances españoles*;[3] for Portuguese, we combined the Colonia historical corpus (Zampieri and Becker, 2013) with a collection of novels from XIX century.[4] Apart from that, we randomly selected sentences containing about $200M$ tokens from the Wikipedia version of each language. These distributional models were also used as pre-trained word embeddings to train the UDPipe parsers which analyzed the corpora. Ideally, we could train different distributional models for each time period, but we decided to use a single model with data from different epochs due to the lack of large resources for historical Portuguese and Spanish.

[3] https://github.com/cligs/textbox/tree/master/spanish

[4] https://github.com/cligs/romancesportugueses

For both languages we restricted the analyses to *verb-object* collocations, and we used *log-likelihood* as the association measure (Uhrig et al., 2018). Moreover, as we deal with historical corpora, we defined a high-coverage approach by selecting candidates with a low *log-likelihood* value ($>=2.5$), and also other very frequent combinations (with a empirically defined ratio per century equal or greater than 0.18). The thresholds *base_simil* and *collocate_simil* were defined to 0.9 and 0.7, respectively.

It is worth mentioning that, since, to our knowledge, there is no gold-standard data on collocation diachronic variation, we cannot carry out a systematic analysis of our approach. Thus, we performed a preliminary evaluation aimed at having an overview of the precision of the system and knowing how it could help to automatize the work of historical linguists.

### 5.2 Results

First, we present some quantitative results obtained by evaluating a random set of the output in Portuguese and Spanish. Then, we discuss the outcome from a qualitative perspective, carrying out a brief analysis using a historical linguistics point of view. Finally, we also show some results of our system using a non-normalized diachronic corpus in Portuguese.

**Quantitative analysis:** Summing up the data of the five centuries, the system identified $1,932$ and $1,980$ changes of types 1 through 4 in Spanish and Portuguese, respectively. Most of these combinations (about 90%) were of type 4, due to the use of contemporary resources to restrict the analysis of unknown words. Besides, it extracted the historical trends (changes 5 to 7) of $3,129$ (Spanish) and $2,210$ (Portuguese) combinations.

To perform the quantitative evaluation we randomly selected the output of 100 collocations of types 1 to 4 for each language (we did not evaluated the results of types 5, 6, and 7, since they are obtained from the observed frequencies of the collocations). From this sample, we removed those combinations which were not proper collocation candidates due to parsing errors (e.g., the Spanish *llevar* plus *[el] alférez* –literally 'to take' plus 'the sub-lieutenant'— was incorrectly labeled as an object relation instead of subject), totaling 32% in Spanish, and 39% in Portuguese (see Table 1). Note that these values refer to parsing errors in

| *Evaluation* | Span. | Port. | Average |
|---|---|---|---|
| *Prec_Alt* | 47.1% | 56.8% | 52.1% |
| *Prec_Dia* | 62.5% | 73.8% | 68.8% |
| *Parsing errors* | 32.4% | 39.0% | 36.3% |

Table 1: Results of the quantitative evaluations in Spanish and Portuguese. *Prec_Alt* is the precision of the proposed alternatives, while *Prec_Dia* is the overall precision of the system. *Parsing errors* include those source combinations (not the target ones) which were wrongly analyzed by the parser. Average is micro-average.

the source combinations (those which suffered a change), not in the collocations proposed as alternatives for each variation type.

Then, we evaluated the output of each collocation as follows. For those collocations where the system did not give any alternative, we looked for other examples with the same meaning in the lists of collocations (false negatives). In those cases where the system provides alternatives, we checked whether these results have approximately the same meaning (e.g., *dar [um] alegrão* → *alegrar*, 'make happy' in Portuguese). We considered correct (i) the nonexistence of newer collocations with similar meanings (in the first case) as well as (ii) the identification of proper alternatives (in the second). Otherwise, the output was considered incorrect. Then, we carried out an error analysis aimed at knowing into more detail what types of error produced our method (see Section 5.3 below).

We computed two precision values for each language (Table 1). On one hand, *Prec_Alt* evaluates the quality of the proposed alternatives by dividing the number of correct cases by the total number of collocations with alternatives (so this value ignores those cases where the system did not found expressions with similar meanings). On the other hand, *Prec_Dia* performs an overall evaluation of the system by taking into account these cases where it did not provide alternatives (correct cases divided by all the analyzed cases).

The results in Table 1 show that the performance of the system was better in Portuguese, even if this language had a large number of parsing errors. The two evaluation approaches had a similar behaviour in both languages (with differences of 15.4% in Spanish and of 17% in Portuguese).

It is worth mentioning that as our method is not a fully automatic system to identify the changes, but rather a tool for identifying potential variations to assist historical linguists, a qualitative evaluation is probably more appropriate than a quantitative one. Thus, qualitative analyses in both languages were carried out in order to know the usefulness of the system.

**Qualitative analysis:** As pointed out, changes of type 4 are the most frequently observed in both Spanish and Portuguese processing of the *P.S. Post Scriptum* corpora. In this regard, a manual revision is in order to evaluate the linguistic interest of these data. Thus, for instance, some of these results point to *bona fide* cases of collocational changes. That is the case of Portuguese *deitar missa* (lit. 'lay, mass', 'say a mass', lost in the 16th century) and *botar [uma] bênção* (lit. 'throw a blessing', 'give a blessing', until 18th c.) and Spanish *prestar paciencia* and *aprestar paciencia* (both meaning 'have patience').

In our setting, changes of type 1 are the less common, since we decided to analyze only those words which are present in further centuries or in present dictionaries. However, the system found some intriguing cases of type 1 (i.e., both words of the collocation do not appear in later periods of the corpus —but they still appear in current dictionaries), such as the Portuguese *obtundir acrimónia* ('lessen the curtness', lost in the 18th c.). Curiously enough, the historical *Corpus do Português* (Davies and Ferreira, 2006)[5] does not have any occurrence of the verb *obtundir*, and only 18 cases of *acrimónia*.

The Portuguese data also offers interesting cases of base loss such as *furtar [o] bisalho* ('steal a bag') and *perdoar [o] enfadamento* ('forgive an annoyance'). Regarding the latter, the system correctly proposes the alternative *perdoar [o] enfado*. Besides, it also identified the loss of the verb *arrenegar* ('to abjure', with a frequency of 3 in the *Corpus do Português*), present until the 18th in combination with bases such as *demónio* ('demon') or *diabrura* ('deviltry'), as examples of type 3 (collocate loss).

In Spanish, instances of change 2 (base loss) correspond to either very infrequent (*réprobo* 'reprobate', *requisitorio* 'requisition') or archaic nouns (*malhecho* 'misdeed'). An interesting case of collocate loss (type 3) is the verb *desenojar* 'to appease', which the system indicates that disappears in the 18th century. In the corpus of the *Nuevo diccionario histórico del español* (hence-

[5] https://www.corpusdoportugues.org/

forth CDH), a larger diachronic corpus of Spanish accessible only through a web interface (Instituto de Investigación Rafael Lapesa, 2013), this verb is mostly attested before 1700. Afterwards, in the 18th century its frequency decreases almost by a half (from 2.59 to 1.58 occurrences per million words, *opmw*), and continues to decrease steeply in later periods.

Amongst the changes of type 4, one finds the most relevant cases from a dichronic perspective. In the case of $aprestar|prestar$ *paciencia* the system identifies its loss around the 18th century and correctly predicts its substitution for the nowadays more common light-verb construction *tener paciencia* ('have patience'). In the larger CDH, *prestar paciencia* goes from 0.7 opmw in the 16th and 17th centuries to less than half (0.32) in the 18th c. and keeps decreasing. By the 20th century it seems almost extinct with only one occurrence in 1933.

A similar case is *meter paz* 'to put peace', the last attestations of which are dated by the system in the 16th century in favor of *poner paz*. The loss of this collocation, however, has greater implications, since a broader semantic change affecting the verb *meter* could be at play here. Corominas and Pascual (1996) (s.v. *meter*) point out that that the meanings of *meter* and *poner* ('to put') were more or less interchangeable in medieval Spanish. Nowadays, however, *poner* conveys the meaning 'change of position' and describes non-durative changes (achievements), whereas *meter* has a directional component and a durative interpretation (accomplishment), according to Cifuentes Honrubia (2004).

**Results in a non-normalized corpus:** Besides the previous experiments, we also carried out a test in a non-normalized and larger historical corpus of Portuguese, Colónia, with 6.2M tokens of essays from 16th to 20th centuries (Zampieri and Becker, 2013).[6] We analyzed combinations with a frequency equal or greater than 2 in the first time period in which it appeared, and used the same association measures and parameters as in the previous experiments.

In this case the system classified 2,622 changes of types 1 to 4, in which a brief analysis allowed us to identify interesting variations in historical collocational preferences in Portuguese. For instance, examples of type 1 such as *desafivelar gorjal* ('unfasten, gorjet', lost in the 19th century), of type 2 such as *fazer soído* ('make sound' or 'make noise', where *soído* is currently replaced by *som*) or *corromper [a] pudicícia* (lit. 'to corrupt the shyness') in the 18th and 19th centuries, respectively. Among the observed cases of type 3 there are interesting verb losses (or at least decreases in use) in cases such as *descantar [o] louvor* ('sing praises') or *manear [a] arma* ('handle a weapon'), currently less used than the collocates *cantar* and *manejar*, respectively.

In this analysis, the system also proposed correct alternatives to changes of type 4, including the collocation *tomar aposento* ('to lodge') or the verb *carregar* ('to carry'), from *fazer aposento* and *fazer [a] carregação*, respectively.

In sum, this analysis allowed us to verify the usefulness of the proposed method to rapidly identify the target language changes also in non-normalized corpus such as the *Colónia*. It is worth recalling that, depending on the corpus properties and on the objectives of the research, the parameters of the system can be configured to suit the needs of the analysis.

### 5.3 Error analysis

In order to know in more detail the type of errors produced by our method we carried out an error analysis of each of the incorrect outputs of the quantitative evaluation. The errors were classified in the following three types, presented by their frequency (see Table 2 for the quantitative results):

1. Different sense of the collocates: the most common error type was the suggestion of a collocate with a different sense in those cases where the base still appears in the corpus, but in other combinations. For instance, the system proposed the combination *dar dilación* (literally 'to give a delay') as a replacement for the Spanish *sentir dilación* ('to regret a delay'). In Portuguese, *fazer recado* ('to do an errand') was suggested as a substitution of *esperar recado* ('to wait for an errand').

2. Different sense of the verbs: another frequent error, similar to the previous one, was the suggestion of single-word verb equivalents for collocations with different meanings. In Portuguese, *encomendar* ('to order') was proposed to replace *tomar encomenda* ('to take an order'), while *desacatar* ('to

[6] http://corporavm.uni-koeln.de/colonia/

| *Error type* | Span. | Port. | Average |
|---|---|---|---|
| *Collocate sense* | 50.0% | 75.0% | 61.8% |
| *Verb sense* | 38.9% | 18.8% | 29.4% |
| *Parsing* | 11.1% | 6.3% | 8.8% |

Table 2: Quantitative results of the error analysis per language. Average is micro average.

disobey') was the first suggestion for the Spanish *causar desacato* ('to cause disobedience').

3. Parsing: a less frequent error type was produced by incorrect annotations of the dependency parser. As an example, *ver auditor* ('to see an auditor', in Spanish) was analyzed as a *verb-object* relation instead of a *subject-verb* ('the auditor saw [...]').

Error types 1 and 2 were mainly produced due to our distributional semantics approach; as collocates have a particular meaning depending on the base they go with, standard distributional semantics models often fail to capture these specific senses. To avoid these problems, both non-compositional methods (e.g., representing the collocations as multiword units in the distributional models), or contextualized compositional strategies (which combine the vectors of the elements or their most prominent contexts) could be applied.

## 6 Conclusions and further work

In this paper we presented a system aimed at facilitating the diachronic detection of collocational variation. The method takes advantage of dependency parsing and of statistical association measures, together with a base–collocate approach, to find candidates of phraseological combinations. To the best of our knowledge, this is the first approach focused on the automatic identification of collocational changes in different languages.

For each collocation in the corpus, the system identifies the period it starts to appear and verifies whether it continues to be used. Those combinations which ceased to occur in later historical periods are analyzed in order to infer whether simple lexical substitutions have happened, or if the lexical restrictions of a collocation base have shifted. Also, the strategy takes advantage of distributional semantics methods to propose alternatives for those combinations which ceased to be used.

A case study on Portuguese and Spanish historical corpora shows that the system is useful both to speed up the finding of collocation changes as well as to detect phraseological and semantic variation in diachronic resources. In this regard, some interesting collocational and semantic changes have been pointed out based on a qualitative analysis of the results. It is worth mentioning that, even if the system is better suited for normalized historical corpora, the performed evaluations showed that it works reasonably well also in non-normalized resources. However, further research is needed to reduce the parsing errors in both normalized and non-normalized historical corpora.

Based on an error analysis, in future work we plan to improve the preprocessing with NLP tools adapted for non-normalized corpora as well as with more balanced word embeddings models trained on historical resources. Another future line of research could be the use of contextualized models of distributional semantics able to infer different senses of a word by the contexts where it appears. Finally, it would be interesting to embed the system in a visualization tool to support research in historical linguistics and in digital humanities.

## Acknowledgements

This research was supported by a 2017 Leonardo Grant for Researchers and Cultural Creators (BBVA Foundation) and by the Galician Government (Xunta de Galicia grant ED431B-2017/01). Marcos Garcia has been funded by a Juan de la Cierva-incorporación grant (IJCI-2016-29598), and Marcos García-Salido by a post-doctoral grant from Xunta de Galicia (ED481D 2017/009). We gratefully acknowledge the support of NVIDIA Corporation with the donation of the Titan Xp GPU used for this research.

## References

Josep Alba-Salas. 2007. On the life and death of a collocation: A corpus-based diachronic study of dar miedo/hacer miedo-type structures in Spanish. *Diachronica*, 24(2):207–252.

Gregory D. S. Anderson. 2006. The Origins of Patterns of Inflection in Auxiliary Verb Constructions. In *Auxiliary Verb Constructions*, Oxford Studies in Typology and Linguistic Theory, chapter 7. Oxford University Press.

Robert Bamler and Stephan Mandt. 2017. Dynamic word embeddings. In *Proceedings of the 34th International Conference on Machine Learning (ICML 2017)*, volume PMLR 70, pages 380–389.

Morton Benson. 1990. Collocations and general-purpose dictionaries. *International Journal of Lexicography*, 3(1):23–34.

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

Miriam Butt and Aditi Lahiri. 2013. Diachronic pertinacity of light verbs. *Lingua*, 135:7–29.

Roberto Carlini, Joan Codina-Filba, and Leo Wanner. 2014. Improving collocation correction by ranking suggestions using linguistic knowledge. In *Proceedings of the third workshop on NLP for computer-assisted language learning*, pages 1–12, Uppsala. LiU Electronic Press.

Jack K. Chambers and Natalie Schilling, editors. 2013. *The Handbook of Language Variation and Change*, 2nd edition. John Wiley & Sons, Inc, New Jersey.

José Luis Cifuentes Honrubia. 2004. Verbos locales estativos en español. *Estudios De Lingüística*, Anexo 2:73–118.

CLUL. 2014. P.S. Post Scriptum. Arquivo Digital de Escrita Quotidiana em Portugal e Espanha na Época Moderna. May 2018. http://ps.clul.ul.pt.

Paul Cook and Suzanne Stevenson. 2010. Automatically identifying changes in the semantic orientation of words. In *Proceedings of the 7th International Conference on Language Resources and Evaluation (LREC 2010)*, pages 28–34, Valletta, Malta. European Language Resources Association.

Joan Corominas and José A. Pascual. 1996. *Diccionario crítico etimológico castellano e hispánico*. 5. Gredos, Madrid.

Anne Curzan. 2008. Historical corpus linguistics and evidence of language change. In Anke Lüdeling and Merja Kytö, editors, *Corpus Linguistics. An international handbook*, volume 2, pages 1091–1108. Mouton de Gruyter, Berlin.

Mark Davies and Michael Ferreira. 2006. Corpus do Português (45 milhões de palavras, sécs. XIV-XX). http://www.corpusdoportugues.org.

Stefanie Dipper. 2008. Theory-driven and corpus-driven computational linguistics, and the use of corpora. In Anke Lüdeling and Merja Kytö, editors, *Corpus Linguistics. An international handbook*, volume 1, pages 68–96. Mouton de Gruyter, Berlin.

Marion Elenbaas. 2013. The synchronic and diachronic status of English light verbs. *Linguistic Variation*, 13(1):48–80.

Stefan Evert, Peter Uhrig, Sabine Bartsch, and Thomas Proisl. 2017. E-VIEW-alation – a Large-scale Evaluation Study of Association Measures for Collocation Identification. In Iztok Kosem, Carole Tiberius, Milos Jakubícek, Jelena Kallas, Simon Krek, and Vít Baisa, editors, *Electronic lexicography in the 21st century. Proceedings of eLex 2017 conference.*, pages 531–549. Lexical Computing CZ, Brno.

Pablo Gamallo, Iván Rodríguez-Torres, and Marcos Garcia. 2017. A web interface for diachronic semantic search in Spanish. In *Proceedings of the Software Demonstrations of the 15th Conference of the European Chapter of the Association for Computational Linguistics*, pages 45–48, Valencia, Spain. Association for Computational Linguistics.

Pablo Gamallo, Iván Rodríguez-Torres, and Marcos Garcia. 2018. Distributional semantics for diachronic search. *Computers & Electrical Engineering*, 65:438–448.

Marcos Garcia and Pablo Gamallo. 2015. Yet another suite of multilingual NLP tools. In *Languages, applications and technologies. Communications in Computer and Information Science*, International Symposium on Languages, Applications and Technologies (SLATE 2015), pages 65–75. Springer.

Marcos García-Salido. 2017. On causative *dar* and its alternatives in the history of Spanish. *Folia Linguistica*, 51(s38):91–124.

Stefan Th. Gries. 2013. 50-something years of work on collocations. *International Journal of Corpus Linguistics*, 18(1):137–165.

Kristina Gulordava and Marco Baroni. 2011. A Distributional Similarity Approach to the Detection of Semantic Change in the Google Books Ngram Corpus. In *Proceedings of the GEMS 2011 Workshop on GEometrical Models of Natural Language Semantics (GEMS'11)*, pages 67–71, Stroudsburg, PA, USA. Association for Computational Linguistics.

William L. Hamilton, Jure Leskovec, and Dan Jurafsky. 2016. Diachronic word embeddings reveal statistical laws of semantic change. In *Proceedings of the 54th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 1489–1501, Berlin, Germany. Association for Computational Linguistics.

Franz Josef Hausmann. 1989. Le dictionnaire de collocations. *Wörterbücher, Dictionaries, Dictionnaires*, 1:1010–1019.

Johannes Hellrich and Udo Hahn. 2017. Exploring diachronic lexical semantics with JeSemE. In *Proceedings of ACL 2017, System Demonstrations*, pages 31–36, Vancouver, Canada. Association for Computational Linguistics.

Martin Hilpert and Stefan Th. Gries. 2016. Quantitative approaches to diachronic corpus linguistics.

In *The Cambridge Handbook of English Historical Linguistics*, pages 36–53. Cambridge University Press.

Instituto de Investigación Rafael Lapesa. 2013. Corpus del Nuevo diccionario histórico (CDH). http://web.frl.es/CNDHE.

Bryan Jurish. 2015. DiaCollo: On the trail of diachronic collocations. In *Proceedings of the CLARIN Annual Conference*, pages 28–31, Wroclaw.

Yoon Kim, Yi-I Chiu, Kentaro Hanaki, Darshan Hegde, and Slav Petrov. 2014. Temporal analysis of language through neural language models. In *Proceedings of the ACL 2014 Workshop on Language Technologies and Computational Social Science*, pages 61–65, Baltimore, MD, USA. Association for Computational Linguistics.

Vivek Kulkarni, Rami Al-Rfou, Bryan Perozzi, and Steven Skiena. 2015. Statistically Significant Detection of Linguistic Change. In *Proceedings of the 24th International Conference on World Wide Web*, pages 625–635. International World Wide Web Conferences Steering Committee.

Erez Lieberman, Jean-Baptiste Michel, Joe Jackson, Tina Tang, and Martin A Nowak. 2007. Quantifying the evolutionary dynamics of language. *Nature*, 449(7163):713–716.

Igor Mel'čuk. 1998. Collocations and lexical functions. In Anthony Paul Cowie, editor, *Phraseology. Theory, analysis and applications*, pages 23–53. Clarendon Press, Oxford.

Joakim Nivre, Mitchell Abrams, Željko Agić, and Lars Ahrenberg *et al.* 2018. Universal dependencies 2.3. LINDAT/CLARIN digital library at the Institute of Formal and Applied Linguistics (ÚFAL), Faculty of Mathematics and Physics, Charles University.

Pavel Pecina. 2010. Lexical association measures and collocation extraction. *Language Resources and Evaluation*, 44(1-2):137–158.

Carlos Ramisch and Aline Villavicencio. 2018. Computational treatment of multiword expressions. In Ruslan Mitkov, editor, *Oxford Handbook on Computational Linguistics*, 2nd edition. Oxford University Press.

Maja Rudolph and David Blei. 2018. Dynamic Embeddings for Language Evolution. In *Proceedings of the 2018 World Wide Web Conference on World Wide Web (WWW 2018)*, pages 1003–1011. International World Wide Web Conference Committee (IW3C2).

Ivan A. Sag, Timothy Baldwin, Francis Bond, Ann A. Copestake, and Dan Flickinger. 2002. Multiword expressions: A pain in the neck for NLP. In *Proceedings of the 3rd International Conference on Computational Linguistics and Intelligent Text Processing*, volume 2276/2010 of *CICLing '02*, pages 1–15, London, UK. Springer-Verlag.

Eyal Sagi, Stefan Kaufmann, and Brady Clark. 2009. Semantic density analysis: Comparing word meaning across time and phonetic space. In *Proceedings of the Workshop on Geometrical Models of Natural Language Semantics*, pages 104–111, Athens, Greece. Association for Computational Linguistics.

John Sinclair. 1991. *Corpus, concordance, collocation*. Oxford University Press, Oxford.

Milan Straka and Jana Straková. 2017. Tokenizing, POS Tagging, Lemmatizing and Parsing UD 2.0 with UDPipe. In *Proceedings of the CoNLL 2017 Shared Task: Multilingual Parsing from Raw Text to Universal Dependencies*, pages 88–99, Vancouver, Canada. Association for Computational Linguistics.

Peter Uhrig, Stefan Evert, and Thomas Proisl. 2018. Collocation candidate extraction from dependency-annotated corpora: Exploring differences across parsers and dependency annotation schemes. In *Lexical Collocation Analysis*, pages 111–140. Springer.

Gael Vaamonde. 2015. PS Post Scriptum: Dos corpus diacrónicos de escritura cotidiana. *Procesamiento del Lenguaje Natural*, 55:57–64.

Derry Tanti Wijaya and Reyyan Yeniterzi. 2011. Understanding Semantic Change of Words over Centuries. In *Proceedings of the 2011 International Workshop on DETecting and Exploiting Cultural diversiTy on the Social Web (DETECT'11)*, pages 35–40, New York, NY, USA. Association for Computing Machinery.

Marcos Zampieri and Martin Becker. 2013. Colonia: Corpus of Historical Portuguese. In *ZSM Studien, Special Volume on Non-Standard Data Sources in Corpus-Based Research*, volume 5. Shaker.