Generation of a Spanish Artificial Collocation Error Corpus

Sara Rodríguez-Fernández[1], Roberto Carlini[1], Leo Wanner[1,2]
[1]NLP Group, Department of Information and Communication Technologies, Pompeu Fabra University
C/ Roc Boronat, 138, 08018 Barcelona (Spain)
[2]Catalan Institute for Research and Advanced Studies (ICREA)
sara.rodriguez.fernandez|roberto.carlini|leo.wanner@upf.edu

Abstract

Collocations such as *heavy rain* or *make* [*a*] *decision* are combinations of two elements where one (the *base*) is freely chosen, while the choice of the other (*collocate*) is restricted by the base. Research has consistently shown that collocations present difficulties even to the most advanced language learners, so that computational tools aimed at supporting them in the process of language learning can be of great value. However, in contrast to grammatical error detection and correction, collocation error marking and correction has not yet received the attention it deserves. This is unsurprising, considering the lack of existing collocation resources, in particular those that capture the different types of collocation errors, and the high cost of a manual creation of such resoures. In this paper, we present an algorithm for the automatic generation of an artificial collocation error corpus of American English learners of Spanish that includes 17 different types of collocation errors and that can be used for automatic detection and classification of collocation errors in the writings of Spanish language learners.

Keywords: artificial corpus, collocations, collocation errors, second language learning, computer assisted language learning

1. Introduction

Collocations, i.e., idiosyncratic word co-occurrences such as *ask* [*a*] *question*, *commit* [*a*] *murder*, *surmount* [*an*] *obstacle*, *faint suspicion*, *high expectation*, etc. are known to be one of the great challenges for language learners;[1] see, among others, Granger (1998), Lewis and Conzett (2000), Nesselhauf (2005) and Lesniewska (2006). According to Wible et al. (2003), "miscollocations" are the most frequent errors in the writings of students. Orol and Alonso Ramos (2013)'s study shows that the "collocation density" in learner corpora is nearly the same as in native corpora, while the collocation error rate in learner corpora is nearly 30% higher than in native corpora. Despite these palpable figures, collocation error identification and correction has not yet received in *Computer Assisted Language Learning* (CALL) the attention it deserves. Current collocation checkers focus mainly on collocation validation or identification of miscollocations (usually using mutual information- or distribution-based metrics) in the writings of learners and a display of lists of possible corrections, ordered in terms of the strength of their "collocationality" or similarity to the original miscollocation; see, e.g., (Chang et al., 2008; Liu et al., 2009; Wu et al., 2010; Ferraro et al., 2014). However, this is by far not sufficient. Ideally, learners should be given the same kind of feedback as given by language instructors when they mark students' essays: they use, as a rule, error type-specific symbols or acronyms during marking (e.g., 'SV' for "subject verb agreement", 'WO' for "wrong word order", 'WW' for "wrong word", etc.); see, e.g., (Nott, 2008). In other words, they classify the students' mistakes.

[1]In a collocation, one of the elements (the base) keeps the meaning it has in isolation, while the meaning of the other (the collocate) depends on the base. For instance, in *surmount* [*an*] *obstacle*, *obstacle* keeps its meaning, while the interpretation of the meaning of *surmount* depends on *obstacle*.

In order to be able to offer such advanced collocation checkers, sufficiently large collocation resources, and, in particular, learner corpora annotated with collocation error information, which could be used for training machine learning techniques, are needed. Unfortunately, in second language learning, corpora are usually too small. To remedy this bottleneck, artificial corpora have often been compiled in the context of automatic grammar error detection and correction; cf., e.g., Foster and Andersen (2009), Rozovskaya and Roth (2010), or Yuan and Felice (2013), among others. In our work, we adopt the same idea for automatic collocation error detection and correction. In what follows, we present an algorithm for the conversion of the Spanish GigaWord corpus into a collocation error corpus of American English learners of Spanish. As the blueprint of the error type occurrence and distribution, we use the Spanish learner corpus CEDEL2 (Lozano, 2009), which was annotated according to Alonso Ramos et al. (2010)'s three-dimensional fine-grained collocation error typology (see Section 2.). Section 3. presents the algorithm for the creation of the artificial corpus, and Section 4. provides a description of its characteristics. Finally, Section 5. concludes the paper.

2. Collocation Error Typology

Alonso Ramos et al. (2010) present a multidimensional collocation error typology, designed after carrying out an analysis of a fragment of a Spanish learner corpus, the *Corpus Escrito del Español L2* (CEDEL2) (Lozano, 2009). The first dimension, or *Location* dimension, describes where the error is produced, i.e., which element of the collocation is affected, namely the base or the collocate, or whether the error affects the collocation as a whole. The second, *Descriptive*, dimension accounts for the kind of error that has been produced (register, lexical or grammatical). Register errors capture context-inappropriate use of *per se* correct collocations. Lexical errors capture a mistake with re-

spect to one of the collocation elements (either wrong word or creation of a non-existing word) or the collocation as a whole (creation of an artificial single word instead of a collocation, creation of an artificial collocation, or use of a collocation with a different sense than intended). Grammatical errors concern the grammar of collocations (missing or superfluous determiner, wrong preposition, wrong subcategorization, etc.). Finally, the third dimension, the *Explanatory* dimension, models the cause of the errors, that is, whether they are caused by interlingual or intralingual reasons.

In previous works on error detection and correction, it has been common to divide errors according to the type of operation that needs to be carried out to make a particular error, that is a *substitution* operation, a *deletion* operation, or an *insertion* operation. Taking into account these operations in the context of learner collocation resources can be useful for developing more accurate strategies for error correction, and for providing better feedback to the learners. Given that the typology by Alonso Ramos et al. (2010) does not consider these operations, we have opted to include them and subdivide when possible the collocation error types into these three extra categories. We arrive, thus, at a fine-grained typology that takes into account, for each type of error: (1) the location of the error, i.e., base, collocate or collocation as a whole, (2) the error type that is produced, i.e., creation, government, order errors, etc., and (3) the type of operation that results in the particular error, i.e., substitution, deletion and insertion.[2] As a consequence, we obtain types of errors such as *Government Base Substitution*, where the preposition of the base is incorrectly chosen, *Pronoun Insertion*, where a reflexive pronoun is incorrectly inserted into the collocation, or *Collocate Creation*, where the collocate is an invented word, etc.

An analysis of the CEDEL2 corpus, annotated with collocation errors, reveals that some of the error types in the typology by Alonso Ramos et al. (2010) tend to occur very seldom. For this reason, we have opted to disregard them in our current work, arriving at the following classes of lexical and grammatical collocation errors (17 in total):

Lexical errors

• **SubB**. Erroneous choice of the base, as in **tener confidencia*, lit. 'have confidence [secret]'; corr.: *tener confianza*, lit. 'have confidence [trust]'.

• **SubC**. Erroneous choice of the collocate, as in **hacer una decisión*, lit. 'make a decision'; corr.: *tomar una decisión*, lit. 'take a decision'.

• **CrB**. Erroneous choice of a non-existing base, as in **hacer un llamo*, lit. 'make a *llamo* [non-existing word meaning *call*]'; corr.: *hacer una llamada*, lit. 'make a call'.

• **CrC**. Erroneous choice of a non-existing collocate, as in **serie televisual*, lit. '[non-existing word meaning *TV*] series'; corr.: *serie televisiva*, lit. 'TV series'.

Grammatical errors

• **DetD**. Erroneous omission of a determiner of the nominal base, as in **ir a _ escuela*, lit. 'go to school'; corr.: *ir a la escuela*, lit. 'go to the school'.

• **DetI**. Erroneous presence of a determiner of the nominal base, as in **hablar el inglés*, lit. 'speak the English'; corr.: *hablar inglés*, lit. 'speak English'.

• **GoBD**. Erroneous omission of a preposition governed by the base, as in **tener la oportunidad _ hacer algo*, lit. 'have the opportunity to do something'; corr.: *tener la oportunidad de hacer algo*, lit. 'tener la oportunidad of do something'.

• **GoBS**. Erroneous choice of the preposition governed by the base, as in **tener obligación a*, lit. 'have obligation to'; corr.: *tener obligación de*, lit. 'have the obligation of'.

• **GoCD**. Erroneous omission of a preposition governed by the collocate, as in **asistir _ una universidad*, lit. 'assist a university'; corr.: *assistir a una universidad*, lit. 'assist to a university'.

• **GoCI**. Erroneous presence of the preposition governed by the collocate, as in **perder a clientes*, lit. 'lose to clients'; corr.: *perder clientes*, lit. 'lose clients'.

• **GoCS**. Erroneous choice of the preposition governed by the collocate, as in **ir por tren*, lit. 'go by train'; corr.: *ir en tren*, lit. 'go in train'.

• **PrD**. Erroneous use of a non-reflexive form of the verbal collocate (omission of the reflexive pronoun), as in **el hielo _ descongela*, lit. 'the ice melts'; corr.: *el hielo se descongela*, lit. 'el hielo melts itself'.

• **PrI**. Erroneous use of the reflexive form of the verbal collocate (insertion of the reflexive pronoun), as in **odio que uno se siente*, lit. 'hatred that one feels themselves'; corr.: *odio que uno siente*, lit. 'hatred that one feels'.

• **NumB**. Erroneous number of the base, as in **dar bienvenidas*, lit. 'give welcomes'; corr.: *dar la bienvenida*, lit. 'give the welcome'.

• **NumD**. Erroneous number of the base determiner, as in **buena notas*, lit. 'good[sing] marks'; corr.: *buenas notas*, lit. 'good[pl] marks'.

• **Gen**. Erroneous gender, as in **aumentar las precios*, lit. 'raise the[fem] prices'; corr.: *aumentar los precios*, 'lit. 'raise the[masc] prices'.

• **Ord**. Erroneous word order, as in **educación buena*, lit. 'education good'; corr.: *buena educación*, lit. 'good education'.

3. Generation of an Artificial Collocation Error Corpus

This section focuses on the methodology for the generation of the artificial corpus. In our work, errors are generated and introduced probabilistically, based on the collocation error distribution of the CEDEL2 corpus. In what follows, we first present a statistical analysis of the CEDEL2 corpus and then provide a detailed description of the error generation algorithm. Afterwards, the resources that are used for the creation of the artificial corpus are outlined.

3.1. Analysis of the learner corpus CEDEL2

In order to obtain relevant information about the error distribution in the learner corpus, we start from CEDEL2, carrying out a statistical analysis of the errors present in this corpus. The error distribution is shown in Table 1.[3] The

[2] In our work, we also consider the *Explanatory* dimension, used as source of information for the automatic generation of the errors

[3] Currently, we only consider error types whose raw frequencies are equal or above 5

Error type	Frequency	%
SuC	470	32.41
Gen	116	8.00
GoCD	98	6.76
SuB	96	6.62
DetD	87	6.00
DetI	78	5.38
CrB	72	4.96
GoBS	48	3.31
GoCI	48	3.31
GoCS	45	3.10
Ord	38	2.62
NumB	33	2.27
GoBD	32	2.21
PrI	27	1.86
CrC	25	1.72
PrD	23	1.59
NumD	10	0.69

Table 1: Frequency of collocation errors in CEDEL2

Error type	Frequency	%
GoCI + SuC	11	0.76
PrD + SuC	10	0.69
GoCD + SuC	9	0.62
DetI + NumB	6	0.41
DetI + GoBS	5	0.34
PrI + SuC	5	0.34
Ord + SuC	5	0.34

Table 2: Multiple error types

second column of the table refers to the number of times that each type of error occurs in the corpus, and the third column shows the percentage of the corresponding error type with respect to the total number of collocation errors found in the corpus.

We observed that a collocation can be often affected by several errors at the same time, for instance, containing an error in the base and another in the collocate, such as in **jugar tenis*, 'to play tennis', corr. *jugar al tenis*, lit. 'to play to the tennis', where there is an omission of the base determiner *el* 'the', and an omission of the collocate preposition *a* 'to'.[4] In the current state of our work, these cases are treated as separate occurrences of the errors, and the decision whether to insert two errors in a collocation is taken randomly by the system.

Furthermore, we found that a base or a collocate can be affected by several errors.[5] This occurs less often, but is nonetheless a phenomenon that needs to be reflected in the artificial corpus. Table 2 shows all combinations whose raw frequencies are equal or above 5, and presents their frequencies and percentages with respect to the total number of collocation errors in CEDEL2.

In order to generate errors that simulate "real" errors produced by learners, it is not sufficient to copy the error distribution observed in a learner corpus; an analysis of the most

[4]In Spanish, when the preposition *a* 'to' is followed by the determiner *el* 'the', the contracted form *al* is used

[5]We include here cases where an error that affects the collocation as a whole, i.e., **Ord**, and an error affecting either the base or the collocate is produced in the same collocation

Correct	Incorrect	#	%
None	*a* 'at'	16	33.33
	con 'with'	14	29.17
	de 'of'	13	27.08
	en 'in'	2	4.17
	por 'by', 'for'	2	4.17
	para 'to', 'for'	1	2.08

Table 3: Frequently confused prepositions (GoCI)

Correct	Incorrect	#	%
a 'at'	None	83	84.69
en 'in'		8	8.16
de 'of'		3	3.06
con 'with'		2	2.04
por 'by', 'for'		1	1.02
sobre 'over'		1	1.02

Table 4: Frequently confused prepositions (GoCD)

frequently confused words is also needed for the cases in which errors are produced through word replacements. In our case, we perform this analysis only for government errors, since, on the one hand, in lexical errors the number of possible options is infinite and thus the usefulness for our work very limited and, on the other hand, the only type of grammatical error where the incorrect choice of a word is considered an error are government errors.[6] The statistics concerning the wrong use of prepositions are presented in Tables 3, 4, 5, 6 and 7.

3.2. Algorithm for Creation of the Error Corpus

The algorithm for the generation of the collocation error corpus passes through three main stages: (1) collocation extraction, (2) collocation classification, and (3) error generation and injection. Firstly, all the N–V, N–Adj and V–Adj dependencies that occur in the corpus where the errors are to be inserted, are retrieved and classified, according to their POS pattern, into three groups: N–V, N–Adj and V–Adj. A statistical check is performed to reject non-collocations: we choose the asymmetrical normalized *Pointwise Mutual Information* (PMI) by Carlini et al. (2014) and consider as collocations only those dependencies whose PMI is higher than 0. Collocations are stored with their prepositions, determiners and pronouns, along with relevant information that will be used at later stages, such as their position in the sentence, lemmas, POS-tags, morphological information, and their sentential context.

Secondly, collocations are classified according to the types of errors that they can contain. For instance, N–Adj collo-

[6]Recall that the incorrect choice of determiner and pronoun are not seen as collocation errors

Correct	Incorrect	#	%
de 'of'	None	27	84.37
en 'in'		3	9.37
para 'to', 'for'		1	3.12
sobre 'over'		1	3.12

Table 5: Frequently confused prepositions (GoBD)

Correct	Incorrect	#	%
en 'in'	*por* 'by', 'for'	16	69.56
	a 'at'	4	17.39
	de 'of'	2	8.69
	con 'with'	1	4.35
de 'of'	*en* 'in'	4	57.14
	a 'at'	3	42.86
para 'to', 'for'	*en* 'in'	1	100
a 'at'	*en* 'in'	7	77.78
	de 'of'	1	11.11
	por 'by', 'for'	1	11.11
por 'by', 'for'	*en* 'in'	2	66.66
	a 'at'	1	33.33
contra 'against'	*con* 'with'	1	100

Table 6: Frequently confused prepositions (GoCS)

Correct	Incorrect	#	%
en 'in'	*de* 'of'	2	50
	sobre 'over'	1	25
	**in* 'in'	1	25
de 'of'	*para* 'to', 'for'	9	42.86
	a 'at'	8	38.09
	en 'in'	2	9.52
	que 'that'	1	4.76
	como 'as'	1	4.76
para 'to', 'for'	*por* 'by', 'for'	2	50
	a 'at'	1	25
	de 'of'	1	25
a 'at'	*de* 'of'	6	85.71
	en 'in'	1	14.28
por 'by', 'for'	*para* 'to', 'for'	6	54.54
	a 'at'	3	27.27
	de 'of'	2	18.18
sobre 'over'	*de* 'of'	2	100

Table 7: Frequently confused prepositions (GoBS)

cations cannot be affected by pronoun errors, V–Adv collocations cannot contain gender errors, collocations that do not contain a determiner cannot be affected by a determiner omission error, etc. A list of candidates is thus created for each type of error.

Finally, errors are generated and inserted according to the error distribution presented in the CEDEL2 corpus. In each iteration, an error type is probabilistically chosen by the system; then a candidate from the list is taken, and an *error generator* produces an error, which is inserted into the sentence; otherwise, the candidate is ignored. In order to preserve the error distribution observed in the CEDEL2 corpus, the creation of the corpus ends when the number of candidates for any of the errors is equal to zero. The set of *Error Generators* that are used are presented below.

3.2.1. Error Generators

A total of six Generators is used to produce the 17 types of collocation errors that we target. 5 are developed for grammatical errors, and one generates all types of lexical errors.

1. Order Error Generator (OEG)

The OEG takes as input N–Adj and V–Adv collocations and swaps the order of the base and the collocate, generating order errors (**Ord**). In order to avoid the creation of uncontrolled grammatical errors, only collocations whose components appear in contiguous order are considered.

2. Gender Error Generator (GEG)

The GEG's role is to insert gender errors (**Gen**) into V–N and N–Adj collocations. In both types of collocations, gender errors are produced in the determiner of the base. In N–Adj collocations, the adjectival collocate is considered as a determiner itself, such that gender errors can be produced either in the base determiner or in the collocate. In the cases where a gender error can be inserted in both places, the GEG randomly chooses where to insert the error, i.e., in the determiner or in the adjective.

The GEG is made up of two main functions, one that changes the gender of the determiner, and one that changes the gender of the adjectival collocate. For determiners, the system first checks whether the input determiner is included in a list of irregular determiners, where both masculine and feminine forms are given. If so, the original determiner is replaced by its alternative form. Otherwise, common gender inflection rules are applied according to the determiner's last letters. For adjectives, a suffix map is used, where masculine suffixes are mapped to feminine ones, and vice versa. The system simply checks whether the adjective's last letters are included in the map, and replaces the original ending with the new one.

As a final step, the existence of the created form is guaranteed by checking its frequency in the reference corpus.

3. Number Error Generator (NEG)

The NEG inserts number errors into V–N and N–Adj collocations. As in the case of 'Gender' errors, 'Number' errors can be produced in the determiner or in the adjectival collocate. In contrast to 'Gender' errors, however, 'Number' errors can also affect the nominal base of the collocation. The NEG inserts, then, two types of errors: **NuBD** for errors produced in the determiner and adjectival collocate, and **NuBB** for errors produced in the base. In cases where the error can be inserted in more than one place, the NEG randomly chooses where to insert the error.

The NEG works as the GEG, i.e., two main functions are designed, one that deals with determiners and one that deals with adjectives and nouns. A list of irregular determiners together with number inflection rules is used for the former, while a suffix map is used for the latter.

4. Substitution Error Generator (SEG)

The SEG inserts replacement and deletion errors into N–V and N–Adj collocations. In the case of replacement errors, we only consider government replacement errors (**GoBS** and **GoCS**). The SEG takes as input collocations in which the target component (the base or the collocate) has a government preposition, and replaces it with another preposition, according to the statistics observed in the learner corpus.

Changing a preposition often results in an error, but in some occasions it can lead to a correct collocation that involves a change of meaning. In order to avoid the introduction of a false error, we developed an PMI-based association metric that calculates the association strength between the collo-

cate, the target preposition and the context of the collocation (a window of 2). Only when the contextual PMI of the original collocation is higher than the contextual PMI of the new collocation, is the error inserted.

Deletion errors can be produced in either prepositions, determiners or pronouns, giving rise to **GoBD**, **GoCD**, **DetD** and **PrD** errors. The mechanism of the SEG for deletion errors is the same as for replacement errors, the only difference being that while in replacement errors the replacement is a valid element, in deletion errors the replacement is void. Contextual PMI is also computed in deletion errors to check that the generated error is a true error.

5. Insertion Error Generator (IEG)

The IEG behaves as the SEG, with the difference that, in this case, none of the elements is changed nor removed, but rather a new element is inserted instead. The IEG generates government, determiner and pronoun insertion errors (**GoCI**,[7] **DetI** and **PrI**) in N–V and N–Adj collocations. As with the SEG, the IEG also uses contextual PMI scores to avoid the insertion of false errors.

The manner in which the element to be inserted is chosen depends on whether the target element is a preposition, a determiner or a pronoun. Prepositions are probabilistically chosen, according to the error statistics observed in the learner corpus, and inserted after the collocate. For determiners, the IEG inserts an indefinite article before the noun. Since neither the definite/indefinite confusion, nor the confusion of any determiner is considered as a collocation error in Alonso Ramos et al. (2010)'s typology, any determiner could be inserted in any case. For simplicity, we opted to always insert indefinite articles, choosing among the different forms depending on the noun number and gender. Finally, pronouns are inserted in two ways, following the rules of the Spanish grammar. For conjugated verbs, the correct pronoun that corresponds to the verb person and number is inserted before the verb. For infinitive forms, the reflexive pronoun *se* is added to the infinitive.

6. Lexical Error Generator (LEG)

The LEG inserts lexical substitution and creation errors in N–V, N–Adj and V–Adv collocations, in both the base and the collocate. The error types covered by the LEG are, therefore, **SuB**, **SuC**, **CrB** and **CrC**. The LEG finds or creates a replacement base or collocate and changes the original base or collocate by the replacement, an existing word in substitution errors, and a non-existing word in creation errors.[8]

Replacement words can be generated in different ways, i.e., (1) **transfer**, where the target base or collocate is translated into L1, (2) **affix change**, where a suffix (including gender inflection) is applied to the target element, (3) **transfer + affix change**, (4) **synonymy** (only for substitution errors), where the target element is replaced by one of its synonyms, and (5) **literal translation**, (exclusively for substitution collocation errors), where the base is translated into L1, and the verb that most often co-occurs with the base in the L1 is retrieved, translated into Spanish and used to replace the original verb. The choice of the method for generating the replacement is random. When unable to generate an error by means of the chosen option, the system selects another option until a valid replacement is found or until the options are finished.

3.3. Resources

The following resources have been used for the generation of the artificial corpus:

- **Base Corpus**. Spanish GigaWord corpus `https://catalog.ldc.upenn.edu/ldc2011t12`.
- **Learner corpus**. We use a learner corpus in order to obtain relevant information regarding the collocation errors that Spanish L2 learners make in their writings. As mentioned above, we use for this purpose CEDEL2. CEDEL2 is a Spanish L2 learner corpus (Lozano, 2009), which includes essays on different topics written by US learners of Spanish of different levels. Our working corpus is formed by 517 essays of levels ranging from pre-intermediate to advanced.
- **Reference corpus**. We use reference corpora (RC) to check word frequencies and co-occurrences. In particular, the algorithm makes use of two RCs, a Spanish RC and an English RC. The Spanish RC consists of 7 million sentences from newspaper material. For English, we use the British National Corpus (BNC), which contains 100 million words from texts of a variety of genres. In order to obtain syntactic dependency information, both corpora were processed with Bohnet (2010)'s dependency parser.
- **Spanish WordNet**. The algorithm also makes use of the Spanish WordNet, from the Multilingual Central Repository 3.0 (Gonzalez-Agirre et al., 2012) as a source of synonymy information. The NLTK library is used to access its contents.
- **Google Translate**. Google Translate is used as bidirectional translation engine, both to translate from Spanish to English, and from English to Spanish. Access to it is provided by the TextBlob Python library.
- **Morphological inflection tool**. Finally, the algorithm uses the morphological inflection tool by Faruqui et al. (2016). This tool allows for the generation of morphologically inflected forms of a word according to given morphological attributes. In our case, we use it for the generation of lexical errors, to inflect the words that are automatically created by the algorithm as replacement for bases and collocates.

4. The Artificially Generated Corpus

In order to check to what extent our artificial corpus simulates our learner corpus, we carried out an analysis of both of them. For this purpose, we took a sample of 50 sentences from each corpus and paid attention to three main aspects: (1) collocation errors, (2) non-collocation errors, and (3)

[7] In the CEDEL2 corpus the frequency of GoBI errors was rather small, so we opted for disregarding this type of error

[8] As in 'Gender' errors, the existence of the replacement words is checked in the RC.

sentence complexity. This is, on the one hand, because the analysis of the generated errors and their comparison to the "real" learners' errors is crucial for a qualitative evaluation of the resource. On the other hand, a comparison of the non-collocation errors and the sentence complexity between the "real" and the synthetic corpora might shed some extra light regarding the similarity of the two corpora. The analysis is presented below.

4.1. Collocation errors

A look at the generated errors points to some important conclusions, mainly that, even when some of the generated errors resemble indeed learners' errors, in some cases, the algorithm fails to generate errors correctly. Thus, firstly, not all the combinations in which errors are inserted are real collocations. Some are free combinations; cf., e.g., *representantes _ las islas* 'islands' representatives'; orig. *representantes de las islas* and *llenaba el plaza* 'filled the square'; orig. *llenaba la plaza*.

Secondly, the injection of an "error" does not always produce a collocation error but, rather, results in a correct collocation involving a change of meaning. For instance, in *la depresión nerviosa que le causó la muerte a su mujer*, lit. 'the nervous depression that caused the death to his wife'; orig. *la depresión nerviosa que le causó la muerte de su mujer* 'the nervous depression that caused the death of his wife'. In other cases, the injection of the "error" results in a change of determination, such as in *consumir una droga* 'to use a drug' lit. 'to consume a drug'; orig. *consumir _ droga* 'to use drugs' lit. 'to consume drug'.

Finally, the injection of an error may result in the generation of unexpected errors. For example, the substitution of *instrumento* 'instrument' by its synonym *herramienta* 'tool' in *es un herramienta que manejaremos*, lit. 'it is a tool that we will use'; orig. *es un instrumento que manejaremos*, lit. 'it is an instrument that we will use', produces a determiner error, since there is no agreement between the changed base *herramienta* and the determiner.

4.2. Non-collocation errors

This section summarizes our findings regarding the production of errors outside the context of collocations. In particular, we consider orthographical, grammatical, lexical, punctuation and discourse marking errors. Our base corpus (the GigaWord) is assumed to be well written, and thus to be free of any error, apart from those collocation errors that were automatically generated. A closer look at it reveals that it contains indeed only very few spelling and grammatical mistakes. Some spelling errors are present, although their proportion and variety is much smaller than in the CEDEL2 corpus: only an unaccented word and 4 typos have been found. The only type of grammatical error observed in the GigaWord sample are agreement errors. Lexical, punctuation and discourse marker errors have not been observed.

4.3. Sentence complexity

In order to measure the sentence complexity, we select several features that can approximate the level of sentence complexity. These features and the values obtained for the two samples are presented in Table 8. In order to obtain the POS and syntactic features, the samples have been processed with (Bohnet, 2010)'s dependency parser.

Feature	CEDEL2	GigaWord
Total words	1,301	2,021
Average sentence length	26.02	40.42
Sentence noun ratio	5.14	10.10
Sentence adjective ratio	1.54	5.18
Sentence verb ratio	3.62	3.50
Sentence adverb ratio	1.56	0.90
Sentence punctuation ratio	2.24	3.38
Sentence coordination ratio	1.10	1.10
Sentence subordination ratio	0.94	0.50
Sentence relativization ratio	0.72	0.66
Sentence passivization ratio	0.18	0.18
Sentence apposition ratio	0.08	0.66

Table 8: Syntactic complexity features in the GigaWord and CEDEL2 samples

As can be observed in Table 8, the values for some of the features, such as the coordination of passivization ratios are rather similar in both corpora. However, each corpus also shows its own morpho-syntactic profile. For instance, the apposition ratio is 8 times higher in the GigaWord corpus than in the L2 corpus. Nouns and adjectives are also significantly more common in the GigaWord corpus, as is the use of punctuation marks. On the contrary, learners tend to use more adverbs and subordinate clauses. As expected, sentence length is substantially shorter in L2 writings.

5. Conclusions and Future Work

We presented an algorithm for the automatic generation of a collocation error corpus of Spanish. The algorithm is able to insert 17 types of errors in error-free data. Such a resource can prove useful for the development of computational collocation tools designed to provide valuable feedback to language learners regarding the types of errors they make. For our experiments, we use the Spanish GigaWord as base corpus.

Using this algorithm, we generated an artificial collocation error corpus, showing that between the CEDEL2 learner corpus and the artificial corpus there still are some differences, which affect both collocation and non-collocation errors, and sentence complexity (in addition to differences in domain and style). All these differences are likely to imply that an algorithm trained on artificial data may not perform as well on L2 data as it may on the artificial data.

To validate the generated error corpus, we carried out some preliminary experiments on collocation error recognition and classification, using LSTMs, in which we achieved an average precision of 0.95 and an average recall of 0.67. As expected, performance falls when the evaluation is carried out on L2 data: when experiments are run on the CEDEL2 corpus, an average precision of 0.58 and recall of 0.39 is achieved.

References

Alonso Ramos, M., Wanner, L., Vincze, O., del Bosque, G. C., Veiga, N. V., Suárez, E. M., and González, S. P.

(2010). Towards a motivated annotation schema of collocation errors in learner corpora. In *LREC*.

Bohnet, B. (2010). Very high accuracy and fast dependency parsing is not a contradiction. In *Proceedings of the 23rd international conference on computational linguistics*, pages 89–97. Association for Computational Linguistics.

Carlini, R., Codina-Filba, J., and Wanner, L. (2014). Improving collocation correction by ranking suggestions using linguistic knowledge. In *Proceedings of the third workshop on NLP for computer-assisted language learning at SLTC 2014, Uppsala University*, number 107. Linköping University Electronic Press.

Chang, Y., Chang, J., Chen, H., and Liou, H. (2008). An Automatic Collocation Writing Assistant for Taiwanese EFL learners. A case of Corpus Based NLP Technology. *Computer Assisted Language Learning*, 21(3):283–299.

Faruqui, M., Tsvetkov, Y., Neubig, G., and Dyer, C. (2016). Morphological inflection generation using character sequence to sequence learning. In *Proceedings of NAACL-HLT*, pages 634–643.

Ferraro, G., Nazar, R., Alonso Ramos, M., and Wanner, L. (2014). Towards advanced collocation error correction in Spanish learner corpora. *Language Resources and Evaluation*, 48(1):45–64.

Foster, J. and Andersen, Ø. E. (2009). Generrate: generating errors for use in grammatical error detection. In *Proceedings of the fourth workshop on innovative use of nlp for building educational applications*, pages 82–90. Association for Computational Linguistics.

Gonzalez-Agirre, A., Laparra, E., and Rigau, G. (2012). Multilingual central repository version 3.0. In *LREC*, pages 2525–2529.

Granger, S. (1998). Prefabricated patterns in advanced EFL writing: Collocations and Formulae. In A. Cowie, editor, *Phraseology: Theory, Analysis and Applications*, pages 145–160. Oxford University Press, Oxford.

Lesniewska, J. (2006). Collocations and second language use. *Studia Lingüística Universitatis lagellonicae Cracoviensis*, 123:95–105.

Lewis, M. and Conzett, J. (2000). *Teaching Collocation. Further Developments in the Lexical Approach*. LTP, London.

Liu, A. L.-E., Wible, D., and Tsao, N.-L. (2009). Automated suggestions for miscollocations. In *Proceedings of the NAACL HLT Workshop on Innovative Use of NLP for Building Educational Applications*, pages 47–50, Boulder, CO.

Lozano, C. (2009). Cedel2: Corpus escrito del español l2.

Nesselhauf, N. (2005). *Collocations in a Learner Corpus*. Benjamins Academic Publishers, Amsterdam.

Nott, D. (2008). Marking studentsâ written work: principles and practice, available from: www.llas.ac.uk/resources/gpg/2956, accessed on 10.07.2016. In *Good Practice Guide*. Subject Centre for Languages, Linguistics and Area Studies, Southampton.

Orol, A. and Alonso Ramos, M. (2013). A Comparative Study of Collocations in a Native Corpus and a Learner Corpus of Spanish. *Procedia–Social and Behavioural Sciences*, 96:563–570.

Rozovskaya, A. and Roth, D. (2010). Training paradigms for correcting errors in grammar and usage. In *Human language technologies: The 2010 annual conference of the north american chapter of the association for computational linguistics*, pages 154–162. Association for Computational Linguistics.

Wible, D., Kuo, C.-H., Tsao, N.-L., Liu, A. L.-E., and Lin, H.-L. (2003). Bootstrapping in a language learning environment. *Journal of Computer Assisted Learning*, 19(4):90–102.

Wu, J.-C., Chang, Y.-C., Mitamura, T., and Chang, J. (2010). Automatic collocation suggestion in academic writing. In *Proceedings of the ACL Conference, Short paper track*, Uppsala.

Yuan, Z. and Felice, M. (2013). Constrained grammatical error correction using statistical machine translation. In *CoNLL Shared Task*, pages 52–61.