Twenty Years of Learner Corpus Research

Looking Back, Moving Ahead

Proceedings of the
First Learner Corpus Research Conference (LCR 2011)

Sylviane Granger, Gaëtanelle Gilquin
and Fanny Meunier (eds)

UCL PRESSES
UNIVERSITAIRES
DE LOUVAIN

Registration of copyright: D/2013/9964/14
ISBN: 978-2-87558-199-0
ISBN PDF version: 978-2-87558-200-3
ISSN: 2034-6417
Printed in Belgium

Cover design: Marie-Hélène Grégoire

Distribution: www.i6doc.com, online university publishers
Available on order from bookshops or at

Diffusion universitaire CIACO (University Distributors)
Grand-Rue, 2/14
1348 Louvain-la-Neuve, Belgium
Tel: +32 10 47 33 78
Fax: +32 10 45 73 50
duc@ciaco.com

Distributor in France:
Librairie Wallonie-Bruxelles
46 rue Quincampoix
75004 Paris, France
Tel: +33 1 42 71 58 03
Fax: +33 1 42 71 58 09
librairie.wb@orange.fr

Corpora and Language in Use

Corpora and Language in Use is a series aimed at publishing research monographs and conference proceedings in the area of corpus linguistics and language in use. The main focus is on corpus data, but research that compares corpus data to other kinds of empirical data, such as experimental or questionnaire data, is also of interest, as well as studies focusing on the design and use of new methods and tools for processing language texts.

The series also welcomes volumes that show the relevance of corpus analysis to application fields such as lexicography, language learning and teaching, or natural language processing.

Editorial Board

Series editors

Editorial Management

Université catholique de Louvain
Contact: cluse@uclouvain.be
http://www.uclouvain.be/cluse.html

Published volume

Granger, Sylviane, Gilquin, Gaëtanelle & Meunier, Fanny (eds). (2013). *Twenty Years of Learner Corpus Research: Looking Back, Moving Ahead.* Corpora and Language in Use – Proceedings 1. Louvain-la-Neuve: Presses universitaires de Louvain.

Forthcoming volumes

Sarda, Laure, Carter-Thomas, Shirley, Fagard, Benjamin & Charolles, Michel (eds). *Adverbials in Use: From Predicative to Discourse Functions.* Corpora and Language in Use – Monograph 1. Louvain-la-Neuve: Presses universitaires de Louvain.

Bolly, Catherine & Degand, Liesbeth (eds). *Text Structuring. Across the Line of Speech and Writing Variation.* Proceedings of LPTS2011, Louvain-la-Neuve, 16-18 November 2011. Corpora and Language in Use – Proceedings 2. Louvain-la-Neuve: Presses universitaires de Louvain.

Table of contents

Acknowledgements

We would like to thank our academic partners and sponsors for their support of the conference.

Academic partners

- Fonds National de la Recherche Scientifique
- Institut Langage et Communication, UCLouvain

Main sponsors

- Cambridge University Press
- CIACO/i6doc.com
- John Benjamins Publishing Company

Supporting sponsors

- Edinburgh University Press
- Oxford University Press
- Rodopi
- Routledge - Francis & Taylor Group

Wanner, L., Alonso Ramos, M., Vincze, O., Nazar, R., Ferraro, G., Mosqueira, E. & Prieto, S. (2013). Annotation of collocations in a learner corpus for building a learning environment. In S. Granger, G. Gilquin & F. Meunier (eds) *Twenty Years of Learner Corpus Research: Looking back, Moving ahead*. Corpora and Language in Use – Proceedings 1, Louvain-la-Neuve: Presses universitaires de Louvain, 493-503.

Annotation of collocations in a learner corpus for building a learning environment

Leo Wanner[a,b], Margarita Alonso Ramos[c], Orsolya Vincze[c], Rogelio Nazar[b], Gabriela Ferraro[b], Estela Mosqueira[c], Sabela Prieto[c]

[a]Institució Catalana de Recerca i Estudis Avançats,[b]Universitat Pompeu Fabra, [c]Universidade da Coruña

Abstract

Collocations in the sense of idiosyncratic lexical co-occurrences are one of the main barriers and challenges for any second language (L2) learner. In Computer Assisted Language Learning (CALL), a number of works deal with the automatic recognition of collocation errors and compilation of candidate lists for their correction. However, this is not sufficient. Firstly, to obtain a clear picture of the difficulties experienced by learners in order to be able to offer targeted aid to learners, a fine-grained linguistic analysis of collocation errors and their annotation in learner corpora is necessary. Secondly, programs must be developed that make concrete correction suggestions, besides providing correction candidate lists, and supply a learner with illustration and didactic material that is oriented towards the types of collocations with which this learner has difficulties. In our work, we attempt to push the state of the art one step further in both of these strands of research, focusing on Spanish as L2. Within the first strand, we carry out a detailed collocation-oriented annotation of a fragment of the corpus of learners of Spanish (CEDEL2). Within the second strand, we experiment with a number of strategies for choosing the most likely correction of a collocation error.

Keywords: L2, learner corpus, collocations, collocation errors, collocation error annotation, collocation error correction.

1. Introduction

The work described in this paper is carried out in the framework of a research project on the development of an active collocation learning environment for learners of Spanish as L2. Following Hausmann (1989), Mel'čuk (1998) and others, we assume that a collocation is a restricted binary co-occurrence of two lexical units (LUs) where one of them (the *base*) conditions the occurrence of the other (the *collocate*) and where between the two LUs a stable syntactic relation holds. Since Hausmann (1984), it has been repeatedly argued that collocations constitute one of the main barriers and challenges in second language learning (Granger 1998; Howarth 1998; Lewis 2000; Durrant & Schmitt 2009). Several quantitative and qualitative studies have been carried out since Granger's (1998) seminal work, comparing the use of collocations by native speakers and learners (see, for instance, Nesselhauf 2005; Gilquin 2007; Martelli 2007). However, so far, to the best of our knowledge, none of the studies focused on a detailed analysis of collocation errors and their annotation in learner corpora as needed for targeted support of learners. A similar state of affairs can be encountered in the field of collocation-oriented Computer Assisted Language Learning

(CALL). Although an increasing number of works deals with the identification of errors in collocation use (see, among others, Shei & Pain 2000; Chang & Chang 2004; Futagi *et al.* 2008; Chang *et al.* 2008; Park *et al.* 2008; Liu *et al.* 2009; Wu 2010; Wu *et al.* 2010; Chen 2011), hardly any attempts to go beyond the compilation of a list of possible corrections of an erroneous collocation from which then the learner has to choose. Often, this list consists of all collocations of the base in question identified in a reference corpus which possess the syntactic pattern of the erroneous sample.

In order to push the state of the art in theoretical collocation error analysis, collocation error annotation in learner corpora and collocation-oriented CALL a step further, we address the following two research questions for Spanish as L2: (1) Can the errors in collocation use by learners be systematized? (2) How can this systematization be exploited in CALL and, more specifically, in active CALL-based collocation learning, to offer the learner not only a list of possible corrections, but also concrete correction suggestions and didactic material targeted to the type of error. To address (1), we carried out an analysis and subsequent annotation of a fragment of the corpus of learners of Spanish CEDEL2 (Lozano 2009), distilling the results of the analysis into a collocation error typology. To address (2), we experimented so far with different strategies for selecting the most likely correction of an erroneous collocation.

In the next section, we present the analysis and annotation of CEDEL2 with respect to collocation errors. Section 3 summarizes the insights we gained from this analysis for CALL and outlines our preliminary experiments on automatic collocation recognition and correction. Section 4, finally, sketches how we plan to advance these experiments in order to benefit better from the results of our analysis.

2. Collocation error analysis and corpus annotation

As described in Alonso Ramos *et al.* (2010) and Vincze *et al.* (2011), we manually annotated 100 essays (amounting in total to 46420 words) from CEDEL2.[1] The annotation resulted in a fine-grained collocation error typology, but also revealed a number of challenges related to the annotation of collocations.

2.1. A closer look at collocation errors in learner corpora

Collocation error analysis can be carried out at least from two different angles: (1) seeking to identify the linguistic nature of the errors, and (2) seeking an explanation of the errors. Both angles are crucial in the context of second language learning / acquisition. In this section, we give a rough outline of the results of our analysis that tackles (1) and (2); for details, see Alonso Ramos *et al.* (2010).

2.1.1. Analysis of the linguistic nature of collocation errors

In contrast to the impression given by previous studies, errors in the use of collocations do not always have the same scope. In particular automatic collocation error detection and correction proposals seem to implicitly assume that (a) the base is always correct and (b) the collocation with the intended meaning exists in L2, such that the learner can only err in the choice of the collocate. The material in CEDEL2

[1] For the setup of the annotation and its detailed evaluation, see Vincze *et al.* (2011).

shows that the learner can err on both elements of the collocation (*cf.* (1) for wrong collocate uses and (2) for wrong base uses) or on the collocation as a whole (*cf.* (3)):[2]

1. wrong collocate uses: *empezar una familia*, lit. 'begin a family' (instead of *formar una familia*, lit. 'form a family'), *tomar una siesta*, lit. 'take a siesta' (instead of *echar una siesta*, lit. 'throw a siesta'), *hacer errores*, lit. 'make errors' (instead of *cometer errores* 'commit errors').
2. wrong base uses: *tener *limitades*, lit. 'have *limitades' (instead of *tener límites*, lit. 'have limits'), *conduzco breve*, lit. 'brief *conduzco' (instead of *trayecto corto*, lit. 'short distance'), *tener un relacionamento*, lit 'have *relacionamento' (instead of *una relación*, lit. 'have a relationship'), *policía exterior*, lit. 'external police' (instead of *política exterior*, lit. 'external politics'), *hablar francia*, lit. 'speak France' (instead of *hablar francés*, lit. 'speak French').
3. wrong collocation uses: *fábrica de carne*, lit. 'meat factory' (instead of *matadero* 'slaughterhouse'), *sitio de acampar*, lit. 'location to camp' (instead of *camping*), **escaparatear* (instead of *ir de escaparates* 'go window- shopping'), *hacer pinturas*, lit. 'make pictures' (instead of *pintar* 'to paint'), *dar respeto* 'command respect' (instead of *tener respeto*, 'have respect').

An erroneous use of a collocation can be rooted in the lexicon or in the grammar.[3] A lexicon error that concerns the collocate or the base consists either in an incorrect replacement of the element by an existing word in Spanish or in the use of a non-existing word (see examples in (1) and (2) above). When the error concerns the collocation as a whole, it may consist in (i) creation of a new LU instead of using a collocation (*e.g.* **escaparatear*); (ii) creation of a new expression with the structure of a collocation instead of using a single LU (*e.g. hacer pinturas*); (iii) use of a correct Spanish collocation with a different meaning than the intended one (*e.g. dar respeto*).

Grammatical errors also concern the base, the collocate or the collocation as a whole and consist mainly in the erroneous absence or presence of a determiner, wrong number use, or wrong government; *cf*:

4. determiner: *viene a mente*, lit. 'comes to mind' (instead of *viene a la mente*), *dar el amor*, lit. 'give the love' (instead of *dar amor*)
5. number: *tienen prejuicio*, lit. '[they] have prejudice' (instead of *tienen prejuicios*)
6. government: *jugamos con tarjetas*, lit. '[we] play with cards…' (instead of *jugamos a las cartas*)

For a few other rare grammatical collocate errors, see Alonso Ramos *et al.* (2010).

[2] When the erroneous use contains a non-existing word in Spanish (*limitad*, *conduzco*, and *relacionamento*), we reproduce it as it is in the literal translation. The last two examples in (3) are examples for the wrong use of *per se* correct collocations.

[3] Apart from lexical and grammatical collocation errors, we also identified register errors; *cf.* *#Yo tengo el deseo personal de ser bilingüe*, lit. 'I have the personal wish to be bilingual'. *Tener* [*un*] *deseo*, lit. 'have [a] wish' is a correct collocation in Spanish. However, it is used in a formal, emotionally charged context, which is not given in the setting of the student. Due to the very limited number of such errors, we ignore them for the time being.

2.1.2. Analysis of the sources of collocation errors

In general, we can distinguish between 'interlingual' (or 'L1-L2 transfer') and 'intralingual L2' errors. For grammatical errors, this distinction suffices: the error can either be traced to English as L1 (as in *han ganado control sobre* '[they] have gained control on', instead of *han tomado control de*, lit. '[they] have taken control of) or not (as in *estaba vacaciones*, lit. 'was vacations', instead of *estaba de vacaciones*, lit. 'was of vacacions').

Lexical errors call for a more detailed identification of their source. In the case of lexical interlingual errors, we can distinguish between cases where the learner creates in L2 an LU from an LU in L1 or from another language (*cf.*, *e.g.*, *recibir un llamo* 'receive a call' or *ir de hiking*, lit. 'go of hiking') and cases where the learner extends the meaning of an existing LU in L2. In many cases of such an extension, the LU in L2 is a valid translation of an LU in L1, but with a different meaning than the intended one. Consider, *e.g.*, *juego de fútbol* 'game of football' instead of *partido de fútbol*, where *juego* is chosen because of a possible translation of *game* as *partido*. An error of extension is also often produced because an L2 LU is used due to its phonetic similarity with the equivalent form in L1; *cf. maternal* in *lengua maternal* 'mother tongue' instead of *materna*, or when the use of an L2 LU is avoided precisely because it seems formally too similar to its L1 equivalent – what can be considered a case of hypercorrection (*cf. atender el teléfono* 'attend the phone', which is discarded by the learner in favour of *acudir*, lit. 'come': *acudir el teléfono* because it appears too similar to the English *to attend*).

The lexical intralingual errors may consist in (a) the creation of an inexistent form in L2 as a result of a process of erroneous derivation by analogy with another form in L2 (*cf.* in *enseñanza *segundaria*, lit. 'secondary education', instead of the adjective *secundaria* 'secondary', the learner derives erroneously **segundaria* from the ordinal adjective *segundo* 'second'); (b) selection of a vaguer or a (more) generic LU than required (*cf. hacer citas*, lit. 'make appointments' instead of *concertar citas* 'arrange appointments'); or (c) selection of a wrong LU without a clear reason and without intervention of L1 (*cf. escribir el examen* 'write the exam', instead of *hacer el examen* 'make the exam').

2.2. Challenges in collocation error analysis and annotation

Collocation is known to be a notoriously difficult language phenomenon. Thus, already the decision of what is a collocation and what is a free word co-occurrence is problematic. The acceptance of a co-occurrence as a correct collocation by speakers also varies from one speaker community to another[4] and so does the interpretation of collocation errors by linguists. Let us discuss the major challenges.

2.2.1. Challenge to recognize collocations

The problem of recognizing collocations in learner texts can be ascribed to the difficulty of establishing clear and, most importantly, operational criteria for delimiting the notion of collocation. In practice, this results in the annotators having difficulty in telling collocations apart from free combinations, on the one hand, and from idioms, on the other hand. For instance, it is quite straightforward to agree on the fact that *buena nota* 'good grade' is a collocation, given that the semantic characteristics of a noun like *nota* 'grade' call for a qualification adjective. This is not so in the case of the combination *buena comida* 'good food', where the meaning of the noun *comida* 'food' does not necessarily require qualification. Consider, however, the combination *comida rica* 'delicious food', where the adjective *rico* 'delicious' has a rather restricted use; it is the adjective prototypically chosen to speak about good food. From our point of view, combinations such as *comida rica* should be considered collocations, and, consequently, other less idiomatic combinations, containing less restricted adjectives appearing with the same noun, such as *buena comida* 'good food' or even *comida fantástica* 'fantastic food' will be considered collocations as well.

[4] In the case of Spanish, the stock of collocations in Latin American Spanish differs considerably from that in Peninsular Spanish.

An example for the difficulty of distinguishing collocations from idioms[5] is the case of *darse cuenta* 'realize', which should be treated as a non-compositional expression, given its frozen syntactic structure. It was mistaken for a collocation by the annotators due to the fact that the verb *dar* 'give' is often used in light verb constructions, as in *dar un paseo* 'take a walk', *dar consejos* 'give advice', *etc.* We also noticed that correct collocations often passed unnoticed by annotators until an incorrect counterpart of the same combination was found. An example for this is the case of *país de origen* 'country of origin', which was not annotated as a collocation until the erroneous combination *países maternos*, lit. 'mother(ly) countries' was found in the corpus. At the same time, any error was bound to be perceived as a collocation error by the annotators. For instance, the free combination *recorrimos *por la isla* '[we] travelled all over the island' was annotated in the first iteration of the annotation process, probably because the preposition error made it more salient.

2.2.2. Challenge to interpret errors

Three kinds of problems constituted a challenge when labelling errors with specific error categories. Firstly, given that the error type labels reflect to some extent how the erroneous expression relates to its correction, error-type annotation in cases when more than one correction was possible was problematic. Thus, in *el viaje no *nos hizo gorditas,* lit. 'the trip didn't make us fatty', the combination *hizo gorditas* can be corrected either as the collocation *ponerse gordas*, lit. 'put oneselves fat' or as single verb *engordar* 'gain weight'. In the first case, the error should be described as the use of an incorrect collocate (*hacer* instead of *ponerse*), while in the second case, it should be described as the use of an erroneous analytical form (*hacer gorditas*) instead of a single lexical item (*engordar*). Secondly, some incorrect collocation-like combinations produced by the learners turned out to be literal translations of combinations in the native language that have no collocation equivalent in Spanish. For instance, the erroneous form **humo de segunda mano* corresponds to the English collocation *second hand smoke*, which can only be translated into Spanish by a complex phrase expressing the same meaning without constituting a phraseological expression: *humo del tabaco de otras personas* 'smoke from other people's tobacco'. In contrast, some expressions used by the learners do not constitute collocations, while the correct form to be used should be a collocation in Spanish. An example for this case is *tengo curiosidad* lit. '[I] have curiosity': **estoy curiosa* conocerlo, lit. '[I] am curious to get to know it', where the expression using the copulative verb and the adjective *curioso* 'curious' should be corrected as a collocation. Thirdly, two coexisting category labels

[5] According to Mel'čuk (2012), idioms represent a major subclass of phraseological multiword expressions: non-compositional multiword expressions.

had to be allowed in the cases where the source of the error could not be determined unambiguously. For instance, in the case of the incorrect collocation **hice citas*, lit. '[I] made appointments', the annotators found it feasible to treat the error both as a direct translation from English and as a generalization error, whereby the generic verb *hacer* 'make/do' is used instead of the correct and more restricted *concertar* 'arrange'.

3. Automatic collocation error correction

In Section 2, we saw that learners make a variety of different collocation errors and that the origin of these errors may be rather diverse – although L1 transfer is very prominent. In what follows, we study the insights gained from the corpus analysis with respect to automation of collocation error identification and correction and outline the experiments we carried out so far.

3.1. Insights from the collocation error analysis for CALL

The insights we gained from the detailed investigation of erroneous use of collocations in a fragment of a learner corpus can be summarized along the following three major lines:

I. A collocation error may concern any of the elements of the collocation (the base or the collocate) or the collocation as a whole. Thus, out of 266 lexical collocation errors identified in the annotated fragment of CEDEL2, 174 (61%) concerned the collocate, 61 (21%) the base, and 50 (18%) the collocation as a whole. These numbers suggest that a collocation correction program cannot be limited to the consideration of errors of the collocate, as most of the current programs do. New proposals are needed that equally treat collocation errors concerning the base and the collocation as such.

II. Learners may err in a collocation with any syntactic pattern, be it a verb+object, subject+verb, noun+modifier, or verb+modifier collocation. This means that the focus on verb+object collocations by most of the current CALL programs is not justified, although it is plausible given the early stage of the work in this area and the fact that verb+object collocation errors are very prominent in English as L2.

III. Most often, the source of a collocation error lies in a genuine L1-transfer. In our study, these were 67% of the errors. This figure corroborates the conclusions of other authors in this respect (see, *e.g.*, Nesselhauf 2005). It also means that during automatic error correction, the correct element can often be found among the synonyms of the erroneous element or among the translations of the L1 counterpart of the erroneous element – as exploited, for instance by Liu *et al.* (2009), Chang *et al.* (2008) and Futagi (2010). However, this is only half of the story: 33% of the errors have other origins, although in most of them the influence of L1 is still detectable. Thus, in a number of cases, the erroneous element is a 'false friend', *i.e.*, a formally similar but semantically different item, of an L1 word; other errors concern the direct use of an L1 lexical item or the adaptation of an L1 form to L2 morphology. For automatic collocation error detection and correction, this means that apart from bilingual dictionaries, the use of monolingual L1 and L2 dictionaries and morphological derivation models of L1 and L2 would be beneficial.

3.2. Experiments on the automatic collocation error correction

An operational collocation learning environment needs to take all of the above findings into account. However, the work we present here is still work in progress; that is why some findings have not yet been considered. So far, we focused on lexical errors, carrying out experiments on the detection and correction of errors of the collocate, experimenting with verb+noun, noun+verb noun+adjective and adjective+noun collocations, and taking into account that a considerable part of the errors are motivated by L1-transfer.

Our program for the detection and correction of collocations (henceforth, *Collidentificator*) takes as input a binary *X+N* or *N+X* combination (with *X* being a verb or an adjective and *N* the base).[6] The combination can come either from the CEDEL2 corpus or typed in for verification by the learner. For two of the correction selection metrics which have been developed to correct the collocations in the writings of learners (see below), *Collidentificator* furthermore requires the sentence in which the combination occurs.

Collidentificator operates in three stages and uses a number of auxiliary resources: (1) a Spanish native reference corpus which consists of about 5GB of newspaper material; (2) the Open Office thesaurus of Spanish; (3) the Spanish WordNet and a bilingual Spanish-English dictionary compiled from Wikipedia.

In the first stage, *X+N* is checked for its status as collocation in the reference corpus. So far, the collocation check uses a frequency-based metric. If *X+N* qualifies as a collocation, positive feedback is given. If not, in the second stage, a number of checks are performed:

a. is *X'+N* (with *X'* as synonym of *X* encountered in the Open Office thesaurus or in the Spanish WordNet) a valid collocation?

b. Is *X'+N* (with *X'* as one of the Spanish translations of *X*'s English translation equivalents identified in the bilingual Spanish-English dictionary) a valid collocation?

If *X'+N* is a valid collocation (as, *e.g.*, *contar cuentos*, lit. 'tell fairy stories' for the learner's **decir cuentos*), it is suggested to the learner as a correction of *X+N*. If this is not the case (as, *e.g.*, **terminar* [*un*] *problema*, lit. 'terminate a problem' for the learner's **concluir* [*un*] *problema*), in the third stage, all collocations with *N* as base and with the same syntactic pattern as *X+N* are retrieved from the reference corpus. The obtained collocations are ordered with respect to their prominence and with respect to their probability to be the correction of the erroneous collocation according to one of several correction selection metrics. The first collocation in the list is then offered as the most likely correction; all other collocations are equally displayed for consideration.

We experimented with three different correction selection metrics to assess the probability of a candidate collocation *C+N* to be the correction of *X+N*:[7] 1) affinity metric, 2) lexical context metric, and 3) context feature metric.

[6] For the sake of simplification, we use henceforth 'X+N' for both 'X+N' and 'N+X'.

[7] See Ferraro *et al.* (2011) for formal details of the metrics.

The *affinity metric* is a local metric that takes into account the co-occurrence (or association) strength of C with N, the graphic similarity of C with X, and synonymy of C with X. As association strength measure, we use log-likelihood. The graphic similarity (which we calculate as Dice coefficient) captures mistyped collocates or erroneously chosen collocates due to their graphic similarity to the intended one (as, *e.g.*, *rise* instead of *raise* in English). The *lexical context metric* takes into account the context in which $X+N$ occurs (in our case, the corresponding sentence in CEDEL2). It is thus grounded in the assumption that the semantics of a collocation can be approximately deduced from the sentential context in which this collocation appears. More precisely, we assume that given the sentential context $c_1, c_2,\ldots, c_n$ of X in the original sentence of the learner, the candidate C with the highest affinity to $c_1, c_2,\ldots, c_n$ is the most adequate correction of X—with "affinity" meaning here the highest co-occurrence frequency. The *context feature metric* is similar to the lexical context metric in that it draws upon the context of $X+N$ in the original sentence of the student. However, there are also two significant differences: Firstly, it may take into account not only lexical tokens (although this is what we tested it with so far; see below), but any kind of contextual features (POS tags, grammatical functions, punctuation, *etc.*). Secondly, its interpretation of these features is very different: Given the sentential context $c_1, c_2,\ldots, c_n$ of X in the original sentence of the learner and the candidate collocate C, the idea is to assess whether any of the contextual features c of X speaks for the preference of C. For this purpose, we find the maximal probability of each feature c, given a candidate $C+N$. The suggestions of the three metrics tend to diverge, although common suggestions are also observed. Thus, for the learner's erroneous *concluir* 'conclude' in **concluir un problema*, lit. 'conclude a problem' the affinity metric suggests *resolver* 'resolve', the lexical context metric *solucionar* 'solve', and the context feature metric *acabar* 'terminate' (all three suggestions are correct); for *realizar* in **realizar* [*una*] *meta*, lit. 'realize a goal', the affinity metric suggests an erroneous **hacer* 'make', while the other two propose the correct *alcanzar* 'reach'; and for *cambiar* in **cambiar al cristianismo*, lit. 'change to Christianity', all three metrics suggest the correct *convertir* 'convert'.

A preliminary evaluation demonstrated that our procedure is able to judge whether a combination is a correct or an incorrect collocation in Spanish with an accuracy of 0.90. When the procedure failed, it tended to judge a correct collocation as incorrect. This is due to our purely frequency-based collocation criteria, which need to be improved. For the error correction stage, we evaluated first the accuracy with which we are able to provide lists of collocations containing the right correction. This accuracy amounts to 0.73 for the context feature metric; the other two metrics have lower accuracies. The mean reciprocal rank (MRR) of the top five suggestions with the contextual feature metric is 0.72 (to compare: Wu *et al.* (2010) achieved on their experimental data an MRR of 0.518). Then, we evaluated the capacity of our procedure to offer the right correction using the context feature metrics, with features being simply words in the original sentence of the learner. The accuracy was 0.542.

4. Conclusion

The fine-grained annotation of a fragment of CEDEL2 with collocation error tags has been a very costly and challenging task. However, it provided us with valuable insights concerning the range and kind of problems learners of Spanish experience with the use of collocations, and resulted in a valuable resource for CALL. In our experiments, we already took some of these insights into account. However, the quality of the achieved results is still too low to be used in practical CALL – although it is better than other state-of-the-art collocation correction programs offer. We need to improve considerably. Furthermore, we need and plan to broaden our work in order to take into account the other insights of our corpus study; in particular, we plan to use the obtained rich corpus annotation for training a classifier cluster for automatic recognition and correction of collocation errors.

Acknowledgements

The presented work has been partially supported by national and European FEDER Funds (FFI2008-06479-C02-01/02, and FFI2011-30219-C02-01/02).

References

Alonso Ramos, M., Wanner, L., Vincze, O., Casamayor, G., Vázquez, N., Mosqueira, E. & Prieto, S. (2010). Towards a motivated annotation schema of collocation errors in learner corpora. In N. Calzolari, K. Choukri, B. Maegaard, J. Mariani, J. Odijk, S. Piperidis, M. Rosner & D. Tapias (eds) *Proceedings of the 7th conference on International Language Resources and Evaluation (LREC'10)*, Valletta, 17-23 May, 2010, 3209-3214.

Chang, J.S. & Chang, Y.C. (2004). Computer assisted language learning based on corpora and natural language processing: The experience of project CANDLE. In L. Anthony, S. Fujita & Y. Harada (eds) *Proceedings of IWLeL 2004: An Interactive Workshop on Language e-Learning*, Tokyo, 10 December, 2004, 15-23.

Chang, Y.C., Chang, J.S., Chen, H.J. & Liu, H.C. (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.

Chen, H.-J. (2011). Developing and evaluating a web-based collocation retrieval tool for EFL students and teachers. *Computer Assisted Language Learning* 24(1), 59-76.

Durrant, P. & Schmitt, N. (2009). To what extent do native and non-native writers make use of collocations? *IRAL* 47, 157-177.

Ferraro, G., Nazar, R. & Wanner, L. (2011). Collocations: A challenge in computer assisted language learning. In I. Boguslavsky & L. Wanner (eds) *Proceedings of the*

5th *International Conference on Meaning-Text Linguistics*, Barcelona, 8-9 September, 2011, 69-79.

Futagi, Y. (2010). The effects of learner errors on the development of a collocation detection tool. In R. Basili, D. Lopresti, C. Ringlstetter, K.U. Schulz & L.V. Subramaniam (eds) *Proceedings of the Fourth Workshop on Analytics for Noisy Unstructured Text Data,* Beijing, 17 September, 2011, 27-34.

Gilquin, G. (2007). To err is not all. What corpus and elicitation can reveal about the use of collocations by learners. *Zeitschrift für Anglistik und Amerikanistik* 55(3), 273-291.

Granger, S. (1998). Prefabricated patterns in advanced EFL writing: Collocations and formulae. In A.P. Cowie (ed.) *Phraseology. Theory, Analysis, and Applications*. Oxford: Clarendon Press, 145-160.

Hausmann, F.J. (1984). Wortschatzlernen ist Kollokationslernen. Zum Lehren u. Lernen französischer Wortverbindungen. *Praxis des neusprachlichen Unterrichts* 31(4), 395-407.

Hausmann, F.J. (1989). Le dictionnaire de collocations. In F.J Hausmann, O. Reichmann, H.E. Wiegand, L. Zgusta (eds) *Wörterbücher – Dictionaries – Dictionnaires, vol. 1*. Berlin: de Gruyter, 1010-1019.

Howarth, P. (1998). The phraseology of learners. In A.P. Cowie (ed.) *Phraseology. Theory, Analysis and Applications*. Oxford: Oxford University Press, 161-186.

Lewis, M. (2000). *Teaching Collocation. Further Developments in the Lexical Approach.* London: Language Teaching Publications.

Liu, A.L, Wible, D. & Tsao, N.L (2009). Automated suggestions for miscollocations. In J. Tetreault, J. Burstein & C. Leacock (eds) *Proceedings of the Fourth Workshop on Innovative Use of NLP for Building Educational Applications,* Boulder CO, 5 June, 2009, 47-50.

Lozano, C. (2009). *CEDEL2: Corpus Escrito del Español L2*. In C.M. Bretones Callejas, J.F. Fernández Sánchez, J.R. Ibáñez Ibáñez, M.E. García Sánchez, M.E. Cortés de los Ríos, S. Salaberri Ramiro, M.S. Cruz Martínez, N. Perdú Honeyman & B. Cantizano Márquez (eds) *Applied Linguistics Now: Understanding Language and Mind / La Lingüística Aplicada Hoy: Comprendiendo el Lenguaje y la Mente*. Almería: Universidad de Almería, 80-93.

Martelli, A. (2007). *Lexical Collocations in Learner English: A Corpus-based Approach*. Alessandria: Edizioni dell'Orso.

Mel'čuk, I. (1998). Collocations and lexical functions. In A.P. Cowie (ed.) *Phraseology. Theory, Analysis, and Applications*. Oxford: Clarendon Press, 23-53.

Mel'čuk, I. (2012). Phraseology in the language, in the dictionary and in the computer. In K. Kuiper (ed.) *The Yearbook of Phraseology* vol. 3, Berlin: Mouton de Gruyter.

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

Park, T., Lank, E., Poupart, P. & Terry, M. (2008). "Is the sky pure today?" AwkChecker: An assistive tool for detecting and correcting errors. In *UIST '08: Proceedings of the 21st ACM symposium on User interface software and technology*, Monterey CA, 19-22 October, 2008, 121-130.

Shei, C.C. & Pain, H. (2000). An ESL writer's collocational aid. *Computer Assisted Language Learning* 13(2), 167-182.

Vincze, O., Alonso Ramos, M., Mosqueira, S., & Prieto, S. (2011). Exploring a learner corpus for the development of a CALL environment for learning Spanish collocations. In I. Kosem & K. Kosem (eds) *Electronic lexicography in the 21st century: New applications for new users. Proceedings of eLex 2011*, Bled, Slovenia, 10-12 November 2011, 280-285.

Wu, J.-C., Chang, Y.-C., Mitamura, T. & Chang, J.S. (2010). Automatic collocation suggestion in academic writing. In *Proceedings of the ACL 2010 Conference Short Papers*, Uppsala, 11-16 July 2010, 115-119.

Wu, S. (2010). *Supporting collocations learning*. Doctoral dissertation, University of Waikato.