Writing assistants and automatic lexical error correction: word combinatorics

Leo Wanner[1], Serge Verlinde[2], Margarita Alonso Ramos[3]

[1]ICREA and Department of Information and Communication Technologies, Universitat Pompeu Fabra, Roc Boronat, 138, 08018 Barcelona (Spain)

[2]Leuven Language Institute, KU Leuven, Dekenstraat 6, B-3000 Leuven (Belgium)

[3]Faculty of Philology, Universidade da Coruña, Campus da Zapateira sn, 15071 A Coruña (Spain)

E-mail: leo.wanner@upf.edu, serge.verlinde@ilt.kuleuven.be, lxalonso@udc.es

Abstract

Genuine lexical writing assistants that attempt to detect lexical errors such as miscollocations are traditionally less common in Computer Assisted Language Learning than spell and grammar checkers. However, there is empirical evidence of the importance of capturing and correcting miscollocations in the writings of language learners, and therefore an increasing number of proposals deals with the detection of errors in collocations and the delivery of lists of correction suggestions. However, very few of these proposals take into account the varying ease with which learners can master different collocation and miscollocation types, or the fact that some collocation type errors might be more common than others, given that a writing assistant should be capable of handling at least the most common types of miscollocation. Furthermore, existing proposals explore collocation error-specific strategies, implicitly assuming that with one universal strategy all types of miscollocations can be detected and corrected. Our preliminary study, conducted on Spanish and French material, highlights one type of collocation in which learners err the most: support verb constructions (SVCs). To account for this, we explore a SVC-specific collocation error detection and correction strategy.

Keywords: CALL, collocation (error) typology, collocation identification, collocation error detection, collocation error correction

1. Introduction

Electronic lexicography supporters argue that e-lexicography needs to design new applications that take advantage of the potential offered by the electronic medium, while still drawing upon data from traditional lexicography (Gouws, 2011). From this perspective, lexical writing assistants seem to be an ideal solution. On the one hand, they imply the use of features offered only by the electronic medium (real time interaction with the user, on-the-fly error detection, correction suggestions, etc.), whereas on the other hand, they require the use of "traditional" data, i.e., lexical resources.

Genuine lexical writing assistants are much less common than spell and grammar checkers [1] (although they are increasing; see below) and are not as mature: performance tests show differences in quality between writing assistants that focus on lexical errors; many of them achieve only a limited rates of successful error recognition and/or correction. However, this is not only due to the immaturity of the technologies. In addition, lexical errors are very heterogeneous (including, e.g., preposition use, choice of synonyms, word combinatorics etc.) are thus more difficult to capture and, at the same time, very frequent in foreign language learners' text production (e.g., Granger, 2003 for French, Alonso Ramos et al., 2010b and Agustin Llach, 2011 for Spanish).

In this paper, we address the problem of writing assistants for collocations, where one of the elements of a combination (the *base*, *B*) is chosen freely and the other (the *collocate*, *C*) is chosen idiosyncratically depending on *B*; cf., e.g., *ask a question*, *poser une question*, *hacer una pregunta*, *eine Frage stellen*. Several studies show that collocations pose major problems to language learners (see, among others, Granger, 1998; Lewis, 2000; Nesselhauf, 2004, 2005; Lesniewska, 2006). Therefore, it is not surprising that over the last decade, proposals for collocation writing assistants have been put forward. Some of them allow for a verification of the correctness of a combination introduced via an interactive interface as a collocation and suggest, in the case of a presumed erroneous collocate, a list of possible corrections. Others provide (usually lists of) suggestions for the correction of detected collocation errors in the writing of learners (see, among others, Chang et al., 2008; Park et al., 2008; Yin et al., 2008; Liu et al., 2009; Wu et al., 2010 ; Ferraro et al., 2011, for a variety of different proposals).

Accuracy varies between proposals with respect to both recognition of collocation errors and provision of correction suggestions. However, most proposals attempt to cover all types of collocations (at least those with the same morpho-syntactic pattern: usually, V+N or N+V collocations), applying the same error detection and correction strategy. In light of the great variety of collocations, ranging from prototypical support verb constructions such as *take* [*a*] *walk* to combinations with semantically full verbs such as *fulfill* [*a*] *condition*, this might not be the best approach. Thus, on the one hand, learners might have more problems with one specific type of collocation than with others, whereas on the other hand, collocation type-specific detection and correction strategies might be more efficient than a universal strategy.

To address these questions, we conducted research on French and Spanish texts, with the goals of (i) investigating whether language learners show any preference with respect to the use of a specific type of collocation and whether any peculiarities can be observed with respect to the distribution of miscollocations in learners' writings; and

[1] Some of the spell and grammar checkers also try to detect and correct lexical errors. However, we focus here on pure lexical "checkers".

(ii) assessing to what extent a universal strategy for the automatic detection and correction of miscollocations is feasible or whether collocation type-specific strategies are more promising. The outcome of our study was that learners make more errors on support verb constructions (SVCs), which they use relatively often, than on other V+N collocations, and that a collocation error detection and correction strategy that specifically targets SVC errors is required. We propose such a strategy, which we ultimately plan to integrate into the online application *Interactive Language Toolbox* (https://ilt.kuleuven.be/inlato/; Verlinde and Peeters, 2012) for French, and HaRenEs (http://patexpert-engine.upf.edu/HARenEs-devel/index.php)[2] for Spanish.

In the next section, we evaluate the collocation error type distribution in a fragment of a Spanish learner corpus (showing that SVCs play an extraordinary role in learners' writings) and explore the variety of strategies that can be applied to automatic collocation error detection and correction, assessing whether all of them are equally well-suited for all types of collocation. Section 3 elaborates on a possible strategy for the detection of SVC errors and their correction. In Section 4, we show how the collocation error detection/correction functionality can be integrated into a writing assistant environment. In section 5 we draw some conclusions from our presentation and outline the directions of our future work in the area of automatic collocation error detection and correction.

2. What should a collocation-oriented writing assistant focus on?

As aforementioned, collocations are of different types and levels of complexity. They are not all likely to be of the same relevance to learners (in the sense that some may be less common than others) or to pose equal difficulties to the learners. Their varying complexity has additional consequences for the prospect of successful automatic recognition and correction in the case of erroneous use or composition: some will be easier to recognize and more accurately corrected by the given state-of-the-art techniques than others; and some will require additional techniques.

Surprisingly, very few studies in Computer Assisted Language Learning (CALL) that deal with collocations address these questions: neither from the didactic nor the computational perspective. In cases where a proposal focuses on a distinct collocation type, this is nearly always performed *ad hoc*, with no theoretical justification, while strategies for automatic detection and correction of errors of collocations do not usually cope better with any particular type. In what follows, we attempt to shed some light on these questions.

[2] HaRenEs stands for "Herramienta de Ayuda a la Redacción en Español: Procesamiento de Colocaciones".

2.1 A closer look at the use of collocations

Collocations can be distinguished with respect to their syntactic patterns and semantic features. In the past, different types of typologies have been suggested; see, e.g., Hausmann (1985) and Heid (1996), who propose distinction between V+N, Adj+N, N+N, Adv+V, and Adv+Adj combinations; and Benson et al. (1997), who distinguishes between eight types of *grammatical collocations* and seven types of lexical collocations, some of them based on syntactic and some on semantic grounds. The most detailed and homogeneous typology is provided by *lexical functio*ns (LFs) as introduced in the Explanatory Combinatorial Lexicology (Mel'cuk, 1995). Each LF is characterized by a specific syntactic pattern and a semantic interpretation. The team led by M. Alonso Ramos of the University of La Coruña analyzed the use of collocations in a fragment of the Spanish learner corpus CEDEL2[3], with the LFs as reference typology. In total, 1948 LF instances have been identified;

Of them, 1491 were correct and 457 erroneous (i.e., about 23.5% of all used collocations were wrong). Of the correct collocations, 532 (35.7%) were LFs that capture SVCs.[4] The share of the other LFs was considerably lower: e.g., the LF 'intensity' (Magn) was used 97 times (6.55%) and the 'causation' (CausFunc) 87 times. From the 457 erroneous collocation instances, 110 (24%) were instances of the SVC-LFs; 83 were instances of the LFs 'realize' or 'fulfill' (Real); and 26 (5.69%) LFs CausFunc. The frequency of erroneous use of the other LFs oscillated between 1 and 9. It seems therefore clear that SVCs are both more used and more erred in by learners.[5] Previous studies also show that SVCs constitute a major challenge for learners. This is plausible because SVCs tend to be idiosyncratic, i.e., language-specific and unpredictable.[6] Thus, Nesselhauf (2004) reports an error rate of 32% of SVCs with the verb 'to make' produced by advanced learners of English with German as L1. Most mistakes were due to the inappropriate use of the verb. Bolly (2010:188) comes to a similar conclusion in a study on learners of French with Dutch and English as L1 with respect to the verb *faire* 'to make'.

[3] CEDEL2 is an L1 English-L2 Spanish learner corpus under construction by Cristóbal Lozano in the framework of a bigger corpus-oriented project directed by Amaya Medikoetxea at the Universidad Autónoma de Madrid. Currently, CEDEL2 contains about 730,000 words of essays in Spanish on a predefined range of topics by native speakers of English and (to a smaller extent, for contrastive studies) by native speakers of Spanish. The level of Spanish of the authors of the essays varies from "elementary", "lower intermediate", "intermediate", and "advanced" to "very advanced". For further information on CEDEL2 see http://www.uam.es/proyectoinv/woslac/cedel2.htm; cf. also Lozano (2009).

[4] For readers familiar with LFs: the LFs in question were Oper1/2/3; A detailed presentation of the LFs can be found in (Mel'čuk, 1996).

[5] Note, however, that this does not mean that the share of erroneous SVCs in all used SVCs is bigger than in the case of other collocation types. For instance, in our study, the share of erroneous SVCs oscillated around 17%, while the share of erroneous 'fulfill' collocations (Real) in the total of the used 'fulfill' collocations was about 25%.

[6] As shown in (Alonso Ramos et al., 2011), in a significant number of SVC errors, learners either literally translate L1 collocates into L2, or, on the contrary, attempt to avoid collocates that they perceive as a "too similar to" the L1 collocates.

The consequence we can draw from this abundance of SVCs is that the detection of SVC errors and their correction is a high priority task of collocation-oriented writing assistants across different L2s. The prominence of other types of miscollocations may depend more on L2; additional studies, such as that conducted by Alonso Ramos et al. (2010b), are needed to obtain a clearer picture in this respect.

2.2 A closer look at collocation error detection

Collocation error detection passes through collocation identification. In Computational Linguistics, collocation detection in corpora has been discussed and studied since the late eighties (cf. e.g., Choueka, 1988; Church and Hanks, 1989; Smadja, 1993). Mostly, word co-occurrence frequency-oriented metrics are used; see Pecina (2008) for an extensive list of such metrics. Wanner (2004) and Wanner et al. (2005) are among the few reports of semantic co-occurrence instead of word co-occurrence. In CALL, where the interest in collocations is considerably more recent, word co-occurrence metrics for the identification of miscollocations and collocations in a reference corpus are equally prominent (see, e.g., Yin et al., 2008; Chang et al., 2008; Liu et al., 2009; Dahlmeier and Ng, 2011; Ferraro et al., 2011). A co-occurrence (most often V+N co-occurrences) is considered a miscollocation if its frequency in a reference corpus is below a given threshold. Using this technique, Chang et al. (2008) report a precision of 97.5% for the recognition of English collocations and 90.7% for the recognition of English miscollocations from learners with Chinese as L1; Ferraro et al. (2011) report an accuracy of 90% for the recognition of Spanish miscollocations from learners with English as L1.

Obviously, this implies that correct rare (e.g., literary) collocations will be qualified as miscollocations. However, the results of collocation classification experiments suggest that this risk is likely to vary between collocation types. Thus, Moreno et al. (2013) report on a higher accuracy of the recognition of genuine SVCs (Oper1-LFs) than of other types of collocation by a Support Vector Machine (SVM) classifier when Vs are used as classification features. Our study in Section 2.1 also suggests that SVCs are common and thus the risk of interpretation of a rare SVC as a miscollocation by a frequency-based metric is reduced.[7] Furthermore, SVCs are a type of lexical co-occurrence that tends to be included in general purpose dictionaries, such that lists of SVCs to match with (as, e.g. Shei and Pain, 2000) during the collocation error detection procedure are more likely to be retrieved for SVCs.

2.3 A closer look at collocation error correction

State-of-the-art collocation error correction strategies are more diverse than (mis)collocation recognition strategies. Some focus on L1 interference in learners (see, e.g., Chang et al., 2008 and Dahlmeier and Ng, 2011). Chang et al. (2008) first extract

[7] This assumption requires further verification by broader empirical studies.

V+N co-occurrences from a given written text. Then, they check the extracted co-occurrences against a collocation list previously obtained from a reference corpus. Co-occurrences not found in the collocation list are variegated in that their verbal elements are substituted by all English translations of their L1 counterpart (Chinese, in this case) in an electronic dictionary. The variants are again matched against the collocation list. The resulting matching co-occurrences containing the noun of a non-matching co-occurrence are offered as correction suggestions. The Mutual Reciprocal Rank (MRR) of the correction list is reported to reach 0.66.

Dahlmeier and Ng (2011) work with *confusion sets* of semantically similar words. Given an input text in L2, they generate L1 paraphrases, which are then looked up in a large parallel corpus to obtain the most likely L2 co-occurrences. For this strategy, they report a precision of 38%.

Futagi et al. (2008) target the detection of miscollocations in learner writings, without considering the correction. Unlike the above proposals, they are not restricted to V+N co-occurrences. But similarly to Chang et al. (2008), they extract the co-occurrences from a learner text, variegate them and subsequently look up the original co-occurrence and its variants in a reference list to decide on its status. To obtain the variants, they apply spell checking, vary articles and inflections and use WordNet to retrieve synonyms of the collocate.

Wu et al. (2010) use a classifier to provide a number of collocate corrections. The classifier takes the learner sentence as lexical context. The probability predicted by the classifier for each suggestion is used to rank the suggestions. According to the evaluation included in Wu et al. (2010), an MRR of 0.518 for the first five correction suggestions has been achieved.

Liu et al. (2009) retrieve miscollocation correction suggestions from a reference corpus using three metrics: (i) mutual information (Church and Hanks, 1989), (ii) semantic similarity of an incorrect collocate to other potential collocates based on their distance in WordNet, and (iii) the membership of the incorrect collocate with a potential correct collocate in the same "collocation cluster".[8] A combination of (ii)+(iii) leads to the best precision achieved for the suggestion of a correction: 55.95%. A combination of (i)+(ii)+(iii) leads to the best precision, 85.71%, when a list of five possible corrections is returned.

Ferraro et al. (2011) suggest a two-stage strategy for correction of miscollocations in Spanish. The first stage is rather similar to the one proposed by Futagi et al. (2008): it retrieves the synonyms of the collocate in the miscollocation in question from a number of auxiliary resources (including thesauri, bilingual L1-L2 dictionaries, etc.) and combines them with the base of the miscollocation to candidate corrections. The

[8] Roughly speaking, members of the same "collocation cluster" are values of the same LF.

candidate corrections that are valid collocations of Spanish are returned as correction suggestions. In the case that none are, the second stage applies a metric to retrieve correction suggestions. Three metrics have been investigated: affinity metric, lexical context metric and context feature metric. The context feature metric, which uses the contextual features of the miscollocation (tokens, PoS tags, punctuation, grammatical functions, etc.), performed best in that it achieved an MRR of the top five suggestions of 0.72.

Again, we can observe that all proposed miscollocation correction strategies are assumed to be equally valid for any type of miscollocation. This can be considered a valid assumption if we dispose of (i) a universal technique to identify the meaning intended by the learner when using the miscollocation (or, in other words, to automatically classify miscollocations in terms of a (semantically-motivated) collocation error typology as proposed by, e.g., Alonso Ramos et al., 2010b); and (ii) a universal technique to identify collocations of a specific type (LF) in a reference corpus. Since, to the best of our knowledge, no collocation error classification techniques are as yet available and, as we have seen in Section 2.2, state-of-the-art techniques cannot be used to retrieve collocations of a given type (at least not with an equal accuracy), collocation type-specific miscollocation correction techniques seem more promising. In the light of the characteristics of SVCs (see above), it is especially promising to single out SVC error correction.

3. Towards SVC error correction

In this section, we present an experimental set up of SVC error detection and correction. The setup involves the following stages:

1. Detection of binary word co-occurrences that are potential SVCs.
2. Assessment of their correctness.
3. In case of being judged incorrect, suggestion of a ranked list of corrections.

Each of these stages shall now be discussed in turn.

3.1 Detection of SVC candidates

Since SVCs are verb + object co-occurrences, the most reliable way to obtain candidate SVCs is dependency parsing. However, off-the-shelf parsers tend not to perform well on non-native texts; see, e.g. Heift and Schulze (2007); Krivanek and Meurers (2011). Therefore, many authors use simpler and more reliable (although more approximate) approaches. For instance, Wanner et al. (2005) use a chunker, while Yin et al. (2008), Chang et al. (2008), Ferraro et al. (2011) and others extract N+V co-occurrences identified within a sequence of words of a specific length, i.e., PoS tags. In our preliminary experiments, we also use only PoS. Obviously, this

low-tech practice can (and should) be improved to obtain optimal candidates as it collects both collocations and free word combinations. Without any analysis, the subject/object relation between noun and verb also remains unclear. However, this is a fast and quite robust approach, the quality of which is sufficient for our first round of experiments.

3.2 Assessment of the correctness of a candidate

As discussed in subsection 2.2, assessment of the collocation status of an extracted word combination and examination of the correctness of a collocation candidate can be done in one stage, using the same technique. For SVCs, two techniques seem most straightforward. The first is to match a given candidate co-occurrence with collocation lists compiled from existing (collocation) dictionaries (see the Introduction). Thus, for French, data from Fontenelle (1997) and *Dafles* (Selva, Verlinde and Binon, 2002) can be exploited; for our experiment, we compiled a matrix of a non-exhaustive list of 233 support verbs, combined with 673 different nominal bases. For Spanish, DICE (Alonso Ramos, 2004; Alonso Ramos et al., 2010a) currently contains 21,324 collocations, a significant part of which are SVCs. With extensive collocation lists at hand, a very high accuracy of collocation error recognition can be achieved.

The second technique is to draw on the distribution of SVCs in corpora. Thus, since we can assume that SVCs are used considerably more often than other types of collocations (see also subsection 2.1), a simple frequency-based technique is likely to suffice: a V+N co-occurrence whose context of use shows significant similarity with the average context of an SVC, but whose frequency is significantly below the average frequency of known SVCs, can be assumed to be an SVC miscollocation.

3.3 Correction of collocation errors

In order to find the most relevant suggestions for incorrect collocates, possible candidates have to be selected and ordered according to specific criteria, before they are presented to the user. Subsequently, either a list of possible corrections or the most relevant (or plausible) correction can be offered. As mentioned above, the limited accuracy of the state-of-the-art collocation correctors suggests caution and provides (ranked) correction candidate lists from which the user can choose.

Due to the observed distribution of SVCs, we can assume that a given noun is more likely to co-occur in a reference corpus with its support verbs (forming SVCs) than with any other verbs. As a consequence, we can retrieve the most likely (or most prominent) verbal co-occurrences as correction suggestions for the noun in question. In the context of our experiments on French, we explored some standard likeliness measures: frequency, an association measure (Z-score) and the product of both

metrics.[9] According to the MRR of the top five suggestions for the 673 nominal bases we analyzed, Z-score and the product of this association measure with frequency lead to the best results: 0.87 and 0.88 of MRR. Both measures are superior to simple frequency, which seems to be used, e.g., by the MUST collocation checker[10] for ranking their correction suggestions, because they give less weight to very frequent verbs (*avoir*, *être*, *faire*). They are comparable to the performance of the ranking metrics used in the *Just The Word* (jtw) collocation checker.[11]

Once the list of possible corrections of a miscollocation has been determined and ranked, we need to decide how many candidates should be proposed to the user, i.e., from which rank do we believe the uncertainty of the proposed correction to still be appropriate; and indeed an SVC is too high. The graph in Figure 1 provides some evidence on this.

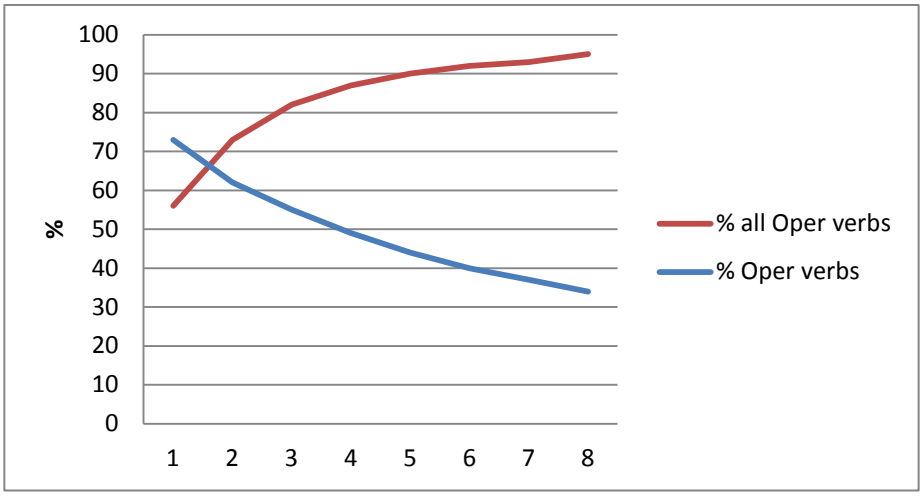

Figure 1: The quality of the SV correction suggestions, depending on the number of presented suggestions

The red line indicates the percentage of support verbs (SVs) that are available for a given base (noun) in the set of suggestions offered to the user, depending on the size of the set. It shows that nearly all SVs are contained in the first eight correction suggestions. The blue line indicates the percentage of SVs in the set of correction suggestions, again depending on the size of the set. It shows that in our experiment on French the first suggested collocate was indeed a support verb for 73% of the bases considered.

[9] As argued previously, our experience is that purely frequency-based measures tend to perform well for SVCs since SVCs are rather common.

[10] http://miscollocation.appspot.com/; http://candle.fl.nthu.edu.tw:9000/

[11] http://www.just-the-word.com/. Consider, for illustration, the ranking of the correction suggestions provided by *jtw* for *make a walk*: http://www.just-the-word.com/main.pl?word=make+a+walk&alternatives=alternatives&cdb=thesaurus

The more suggestions, however, the greater the number of non-relevant verbs: with eight suggestions, only 35% are SVs. At the same time, more different support verbs are displayed if the number of suggestions increases, leading to a better coverage of all uses of SVs (or Oper-LFs).

In general, each application will need to decide on where to draw the line and how many correction suggestions to show. What is important is that the decision be informed.

4. Integrated online writing assistants

The programs for collocation error detection and correction (be they collocation type-specific or generic) can be used either as collocation checker demons, integrated into an editor and switched on or off by the user as deemed appropriate, or integrated into an online writing assistant; see, e.g., StringNet,[12] MUST or Just The Word. It is the latter option that we have chosen for both French and Spanish. For French, the automatic correction of collocation errors, limited to N+V and V+N SVC combinations is due to be integrated into the *Interactive Language Toolbox* website. For Spanish, the corresponding module is integrated into the HaRenEs writing assistant environment.

In what follows, we briefly present each of these environments.

4.1 Interactive Language Toolbox

This online application offers access to the most relevant online lexicographical resources available for Dutch, English and French (*predictive writing aid*) and a spell, grammar and lexical checker for French (*corrective writing aid*)[13]; see Ziyuan (2012).

As shown in Figure 2, the application will not only display a list of alternatives for incorrect collocate selection, but will also give more information on the real use of word combinations with information on determiner use, for instance, (Figure 3) or authentic examples taken from a corpus or found on the web. Figure 3 shows that in almost 91% of corpus occurrences, the determiner *des* is used to combine *forces* (base) with *reprendre* (collocate).

[12] http://www.lexchecker.org/

[13] Similar checkers for Academic Dutch and Dutch as a foreign language are in development and we plan to conceive a similar tool for Academic English.

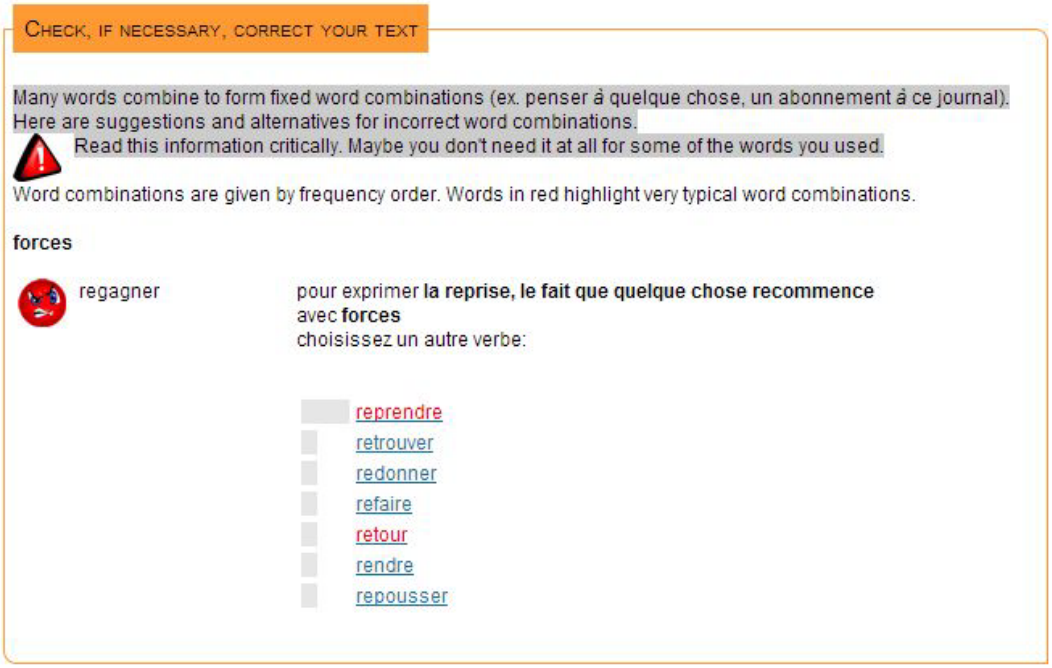

Figure 2: Interactive Language Toolbox: automatic collocation error detection and correction

		pourcentage des cas enregistrés
reprendre	des forces	90.95
reprendre	ses forces	3.41
reprendre	quelques forces	2.89
reprendre	les forces	1.03
reprendre	mes forces	0.77
reprendre	leurs forces	0.63
reprendre	vos forces	0.31

Figure 3: Interactive Language Toolbox: usage notes

In the advanced version, contextual information will aim to be even more extensive, similar to *StringNet* (Wible and Tsao, 2012).

4.2 HaRenEs Writing Assistant

The HaRenEs Writing Assistant is currently being developed in a common project by the University of La Coruña and Pompeu Fabra University. It allows the learner to verify the correctness of a specific Spanish collocation and, in the case of incorrectness, solicit correction suggestions, solicit examples of the use of a given collocation in context (in the reference corpus), and solicit the correction of collocations in a writing, etc. Figure 4 shows a snapshot of the user interface of the prototypical implementation of HaRenEs.

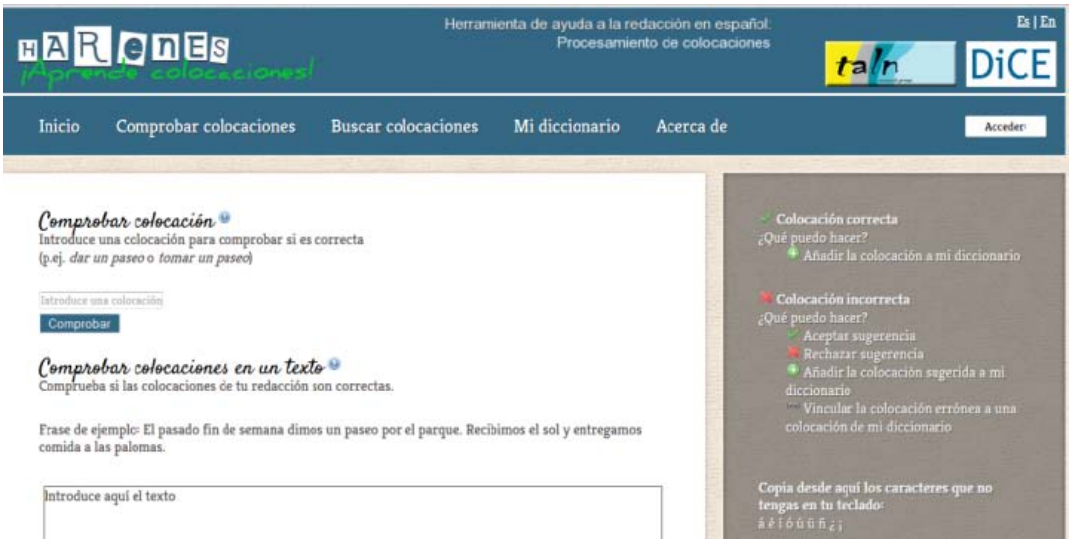

Figure 4: User interface of HaRenEs

Figure 5 shows the correction suggestions provided for the erroneous collocation *tomar* [*un*] *paseo*, lit. 'take [a] walk'.

In an advanced version of the HaRenEs environment, users will be able to configure strategies for collocation error recognition and correction, choosing to either focus on selected types of collocations or to capture all collocations, but apply collocation type-specific error detection and correction strategies, to the extent available.

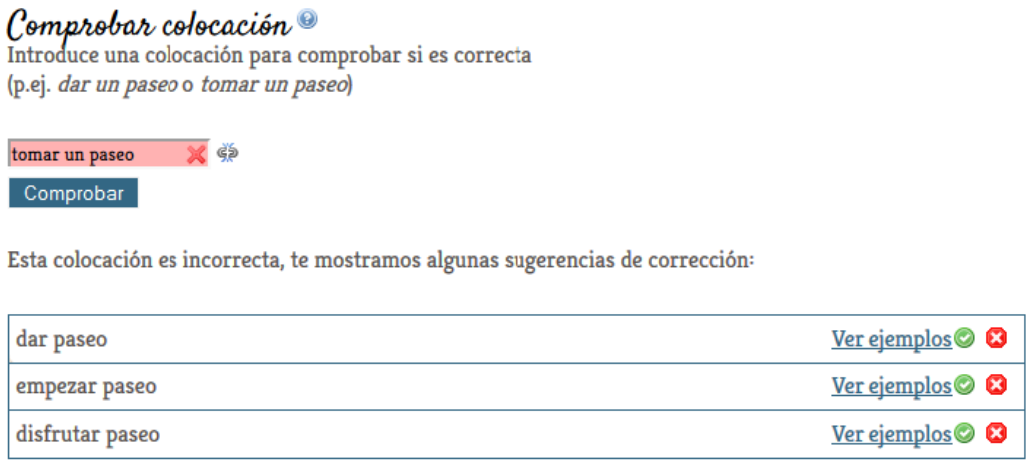

Figure 5: Correction suggestions provided by HaRenEs
for *tomar* [*un*] *paseo*, lit. 'take [a] walk'

5. Conclusions

As already demonstrated by previous works (see, e.g., Nesselhauf, 2004, Bolly, 2010) and as further supported by the studies presented in section 2 above, not all types of collocations are equally used by language learners and not all types pose the same

difficulty to learners. SVCs are the most problematic collocations for learners (at least for English learners of Spanish), such that the detection of their erroneous use and correction is of high priority.

To account for this need, we argue for collocation type specific error detection and correction strategies. For SVC verification and correction (collocation) dictionaries can play an important role. Thus, parallel corpora available on the Internet may fill existing gaps in traditional (translation) dictionaries, but non-native users will feel equally uncomfortable facing the amount of data provided by these applications. The need for a translation dictionary which offers translations in a more systematic way, for instance according to Mel'čuk's lexical functions as suggested by Kjaersgaard (2006), has been expressed for some time (see, e.g. Atkins, 1996; Danlos and Samvelian, 1992), but remains urgent.

On the other hand, corpus-based metrics that draw upon the insight that SVCs are the most common collocations are of relevance. A combination of lexicographic data, corpus analysis tools, and results and statistics combined with NLP-derived data thus provides new opportunities for (e-)lexicography.

In general, collocation error correction programs are a crucial writing aid for any L2 speaker. Such programs can be either used in a stand-alone sense (in the way the Interactive Language Toolbox and HaRenEs are currently conceived) to be consulted during writing, or be integrated into editor environments, such that an erroneous collocation is automatically highlighted and correction suggestions are offered upon request of the writer.

6. Acknowledgements

The work by M. Alonso Ramos, L. Wanner and their teams presented in this paper has been partially supported by the Spanish Ministry of Economy and Competitiveness (MINECO) and the FEDER Funds of the European Commission under the contract number FFI2011-30219-C02-01/02.

7. References

Alonso Ramos, M. (2004). *Diccionario de colocaciones del español* http:/www.dicesp.com.

Alonso Ramos, M., Nishikawa, A; Vinczthe, O. (2010a): "DiCE in the web: An online Spanish collocation dictionary", S. Granger, M. Paquot (eds.), *eLexicograpy in the 21st century: New Challenges, New Applications. Proceedings of eLex 2009, Cahiers du Cental 7*, Louvain-la-Neuve, Presses universitaires de Louvain, pp. 367-368.

Alonso Ramos, M. L. Wanner, O. Vincze, G. Casamayor, N. Vázquez, E. Mosqueira, S. Prieto, (2010b): "Towards a Motivated Annotation Schema of Collocation

Errors in Learner Corpora", *7th International Conference on Language Resources and Evaluation (LREC)*, La Valetta, Malta, pp. 3209-3214.

Agustin Llach, M.P. (2011). *Lexical Errors and Accuracy in Foreign Language Writing*. Bristol, Buffalo, Toronto, Multilingual Matters. (Second Language Acquisition, 58).

Atkins, B.T.S. (1996). Bilingual dictionaries. Past, present and future. In M. Gellerstam, J. Järborg, S.-G. Malmgren, K. Norén, L. Rogström, C. RöjderPapmehl (eds). *Euralex '96 Proceedings. Papers submitted to the seventh EURALEX international congress on lexicography in Göteborg, Sweden*, Part II, pp. 515-546.

Benson, M., Benson, E. and Ilson, R. (1997). *The BBI Dictionary of English Word Combinations*. Benjamins Academic Publishers, Amsterdam.

Bolly, C. (2010). *Phraséologie et collocations. Approche sur corpus en français L1 et L2*. Brussels: P.I.E. Peter Lang.

Chang, Y.C., J.S. Chang, H.J. Chen, and H.C. Liou. (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.

Choueka, Y. (1988). Looking for needles in a haystack or locating interesting collocational expressions in large textual databases. In *Proceedings of the RIAO*, pp. 34–38.

Church, K.W. & P. Hanks. (1989). Word Association Norms, Mutual Information, and Lexicography. In *Proceedings of the 27th Annual Meeting of the ACL*, pp. 76–83.

Dafles – Dictionnaire d'apprentissage du français langue étrangère ou seconde (http://ilt.kuleuven.be/inlato).

Dahlmeier, D. and H.T. Ng. (2011). Correcting semantic collocation errors with L1-induced paraphrases. In *Proceedings of the 2011 Conference on Empirical Methods in Natural Language Processing*, pp. 107-117, Edinburgh, Scotland.

Danlos, L., Samvelian, P. (1992). Translation of the predicative element of a sentence: category switching, aspect and diathesis. In *TMIMT-92, Proceedings of the Fourth International Conference on Theoretical and Methodological Issues in Machine Translation*. Montréal, CWARC, pp. 21-34.

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 Theory (Barcelona, September 8-9, 2011)*, pp. 69-79.

Fontenelle, Th. (1997). *Turning a bilingual dictionary into a lexical-semantic database*. Tübingen, Niemeyer. (Lexicographica, Series maior, 79).

Futagi, Y., P. Deane, M. Chodorow, and J. Tetreault. (2008). A computational

approach to detecting collocation errors in the writing of non-native speakers of English. *Computer Assisted Language Learning*, 21(1):353-367.

Gouws, R. (2011). Learning, unlearning and innovation in the planning of electronic dictionaries.In P.A. Fuertes-Olivera, H. Bergenholtz. *e-Lexicography. The internet, digital initiatives and lexicography*. London, New York, Continuum, pp. 17-29.

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

Granger, S. (2003). Error-tagged Learner Corpora and CALL: A Promising Synergy. *Calico Journal,* 20(3), pp. 465-480.

Hausmann, F.-J. (1984). Wortschatzlernen ist Kollokationslernen. Zum Lehren und Lernen französischer Wortwendungen. *Praxis des neusprachlichen Unterrichts*, 31(4):395-406.

Heid, U. (1996). "Using Lexical Functions for the Extraction of Collocations from Dictionaries and Corpora". In L. Wanner (ed.): *Lexical Functions in Lexicography and Natural Language Processing*, Amsterdam/ Philadelphia: Benjamins.

Heift, T., Schulze, M. (2007). *Errors and intelligence in computer-assisted language learning. Parsers and pedagogues*. New York, Routledge.

Kjaersgaard, P. J. (2006). Esquisse d'un dictionnaire bilingue idéalisé. In Th. Szende (ed). *Le français dans les dictionnaires bilingues*. Paris, Champion, pp. 269-282.

Krivanek, J. and D. Meurers (2011). "Comparing Rule-Based and Data-Driven Dependency Parsing of Learner Language." In *Proceedings of the Int. Conference on Dependency Linguistics (Depling 2011)*, Barcelona.

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

Lewis, M. 2000. *Teaching Collocation. Further Developments in the Lexical Approach*. LTP, London.

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

Lozano, C. 2009. CEDEL2: Corpus escrito del español L2. In C.M. Bretones Callejas, editor, *Applied Linguistics Now: Understanding Language and Mind*. Universidad de Almería, Almería, pp. 197-212.

Mel'čuk, I., Clas, A., Polguère, A. (1995). *Introduction à la lexicologie explicative et combinatoire*. Louvain-la-Neuve, AUPELF-UREF/Duculot.

Moreno, P., G. Ferraro and L. Wanner. (2013). "Can we determine the semantics of

collocations without semantics?". In *Proceedings of eLex 2013*, Tallinn.

Nesselhauf, N. (2004). How learner corpus analysis can contribute to language teaching: A study of support verb constructions. In G. Aston, S. Bernardini, D. Stewart (eds). *Corpora and language learners*. Amsterdam, Philadelphia, Benjamins, pp. 109-124.

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

Park, T., E. Lank, P. Poupart, and M. Terry. (2008). Is the sky pure today? AwkChecker: An assistive tool for detecting and correcting errors. In *Proceedings of the 21st Annual ACM Symposium on User Interface Software and Technology (UIST '08)*, New York.

Pecina, P. (2008). A machine learning approach to multiword expression extraction. In *Proceedings of the LREC 2008 Workshop Towards a Shared Task for Multiword Expressions (MWE 2008)*, pp. 54–57. Marrakech.

Selva, Th., Verlinde, S., Binon, J. (2002). Le Dafles, un nouveau dictionnaire électronique pour apprenants du français. In A. Braasch, C. Povlsen (eds). *Proceedings of the Tenth EURALEX International Congress, EURALEX 2002, Copenhagen, August 13-17, 2002*. Copenhagen, CST, vol. I, pp. 199-208.

Smadja, F. (1993). Retrieving Collocations from Text: X-Tract. *Computational Linguistics*.19.1:143–177.

Verlinde, S., Peeters, G. (2012). Data access revisited : the Interactive Language Toolbox. In S. Granger, M. Paquot (eds). *Electronic lexicography*. Oxford, Oxford University Press, pp. 147-162.

Wanner, L. (2004). Towards Automatic Fine- Grained Semantic Classification of Verb-Noun Collocations. *Natural Language Engineering Journal*. 10.2:95–143.

Wanner, L., Alonso Ramos, M., Vincze, O., Nazar, R., Ferraro, G., Mosqueira, E., Prieto, S. (2011). *Annotation of Collocations in a Learner Corpus for Building a Learning Environment*. Paper presented at Learner Corpus Research conference, 2011.

Wanner, L., B. Bohnet, M. Giereth and V. Vidal. (2005). 'The first steps towards the automatic compilation of specialized collocation dictionaries'. *Terminology*, 11(1):137-174, 2005.

Wible, D., Nai-Lung, T. (2012). Towards a new generation of corpus-derived lexical resources for language learning.In F. Meunier, S. De Cock, G. Gilquin, M. Paquot (eds). *A taste for corpora. In honour of Sylviane Granger*. Amsterdam, Philadelphia, Benjamins , pp. 237-254. (Studies in corpus linguistics, 45).

Ziyuan, Y. (2012). *Breaking the language barrier: a game-changing approach*. (https://sites.google.com/site/yaoziyuan/publications/books/breaking-the-language-barrier-a-game-changing-approach)