Towards a collocation writing assistant for learners of Spanish

Margarita Alonso Ramos
Universidade da Coruña
Faculty of Philology
Campus da Zapateira s/n
15071 A Coruña (Spain)
lxalonso@udc.es

Marcos García Salido
Universidade da Coruña
Faculty of Philology
Campus da Zapateira s/n
15071 A Coruña (Spain)
marcos.garcias@udc.es

Orsolya Vincze
Universidade da Coruña
Faculty of Philology
Campus da Zapateira s/n
15071 A Coruña (Spain)
ovincze@udc.es

Abstract

This paper describes the process followed in creating a tool aimed at helping learners produce collocations in Spanish. First we present the *Diccionario de colocaciones del español* (DiCE), an online collocation dictionary, which represents the first stage of this process. The following section focuses on the potential user of a collocation learning tool: we examine the usability problems DiCE presents in this respect, and explore the actual learner needs through a learner corpus study of collocation errors. Next, we review how collocation production problems of English language learners can be solved using a variety of electronic tools devised for that language. Finally, taking all the above into account, we present a new tool aimed at assisting learners of Spanish in writing texts, with particular attention being paid to the use of collocations in this language.

1 Introduction

This paper[1] presents the process followed in developing a tool that helps learners of Spanish as L2 to produce collocations. 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, B*) is chosen freely and the other (the *collocate, C*) is chosen idiosyncratically depending on B; cf., e.g., *take a walk, dar un paseo, faire une promenade*[2]. It has often been claimed that collocations are challenging for second language learners. In fact, the difference in collocational knowledge has been found to constitute an important factor that contributes to the difference between native and non-native language use (e.g. Howarth, 1998; Granger, 1998; Higueras García, 2006).

When producing a text, a language learner may face different types of problems relating to how words are combined in a native-like way. For instance, German learners of Spanish may wonder how to translate the collocation *einen Spaziergang machen* from their native language to Spanish, for which they need to know that in the case of this combination the verb *machen* translates to Spanish *dar* (lit. 'give'), and not *hacer* ('make'). This example shows a production problem. In other cases, learners may need information concerning the meaning of a collocation, for example, *sacar buenas notas* 'to get good grades'. Furthermore, the complexity of collocations is not limited to knowing which lexical item to combine with another, but it also concerns grammar. For instance, in order to avoid errors such as those found in the following learner sentence: *Los gays deben tener los derechos para casarse* (lit. 'Gays must have the rights in order to marry'), a learner of Spanish has to know not only that *derecho*

[2]Note that this definition does not use frequency of the combination as a determinative criterion, rather it emphasizes the lexical restriction imposed by one element on the selection of the other, in contrast with the approach promoted by corpus linguistics (Sinclair, 1991).

('right') goes with the verb *tener* ('to have'), but also that it is used in the singular form, without a determiner, and that it governs the preposition *a* (not *para*).

Given all these needs, we may raise the question of what the ideal resource designed to help learners overcome difficulties posed by collocations should be like. A straightforward answer would be the dictionary; however, we must be aware that in recent years the traditional dictionary format has been facing a serious crisis due to the challenges posed not only by online lexical and translation tools, but also by language corpora containing vast amounts of lexical information. Corpus-driven lexicography has given rise to what can be called "lexically-driven corpora", i.e. resources which do not provide lexical information in the form of a dictionary, but in the form of a concordance program exploiting language corpora. Through an appropriate user interface lexical items become pointers to the texts that reveal their meaning, blurring the boundaries between dictionaries and corpora (see Alonso Ramos, 2009). Some authors even claim that corpora can completely substitute dictionaries (e.g. Sinclair, 1987).

It is clear that the concept of the dictionary is changing towards a more flexible and dynamic tool, which aims to better address user needs, to the extent that certain authors propose alternative terms -e.g. leximat (Tarp, 2008) or lexical site (Jousse et. al. 2008)– to refer to this newly emerging concept. Jousse et al. (2008), in particular, argue that the word *dictionary* carries connotations of a linear structure, failing to describe the concept of a constantly evolving network, embodied by modern online lexical tools and constituting a better model of lexical knowledge. Independently of the term we use to refer to these new lexical resources, the fact is that dictionaries have ceased to be stand-alone products, which means that they are increasingly integrated with other resources such as corpora, other dictionaries, and glossaries. They also serve to complement and are in turn well complemented by CALL applications.

What we have described so far matches the course of the evolution taken by our research interests detailed in this paper: from an online collocation dictionary of Spanish (DiCE), the development of which began ten years ago, towards an online collocation writing assistant, integrated with the DiCE. In the next section, we briefly present the DiCE and explain the motivations behind the development of a further tool that would complement it. Section 3 focuses on the potential user of a collocation learning tool, examining the usability problems posed by the DiCE and exploring language learners' needs through a learner corpus study of collocation errors. As we will show, both of these aspects should be taken into account when designing a collocation writing assistant. Section 4 provides an overview of freely available online lexical tools for English that can potentially resolve collocation production problems. Section 5 describes in detail the architecture of a new tool aimed at assisting Spanish as L2 learners' collocation production. Finally, Section 6 draws some conclusions from the work presented here and outlines the direction of future research in the area of automatic collocation error detection and correction.

2 Starting from an online collocation dictionary

The *Diccionario de Colocaciones del Español* (DiCE), a web-based collocation dictionary of Spanish, has been available online since 2004, its database constantly being improved and expanded. Since the dictionary has been described in detail on various occasions (e.g. Alonso Ramos, 2005; 2006; 2008; Alonso Ramos et al. 2010a), here we only provide a brief presentation of its main features and focus on the reasons for developing a further tool that enables some of its drawbacks to be overcome.

The DiCE constitutes an online implementation of the principles of lexical description proposed by the Explanatory Combinatorial Lexicology (ECL, Mel'čuk et al.,1995). In addition to providing a theoretically well-founded description of collocations, it aims to be a useful tool not only for specialized researchers but also for the general public. To this end, lexical functions, the formal representation used to describe the semantic and syntactic features of collocations, are paraphrased in natural language glosses. At the same time, the web interface has been designed

to enable flexible access to the electronic lexical database, with a view to satisfying the needs of a broad range of users, from researchers through language learners to lexicographers working on DiCE.

In accordance with our framework, we conceive of collocations as restricted combinations of two lexical units, the base and the collocate. For instance, in the combination *reanudar una amistad* 'renew a friendship', the noun is the base, and it conditions the selection of the collocate verb.

The user interface of the DiCE consists of three main components: 1) the *dictionary* itself, 2) the *advanced search component*, and 3) the *learning module*. The *dictionary component* provides access to the contents in a way similar to other collocation dictionaries. Users are offered a list of lemmas, each associated with its lexical units, under which corresponding semantic and combinatorial information can be found.

In order to offer dynamic access to the information stored in the DiCE database, the advanced search component offers four options. Each of these was designed to provide the user with a more direct path of access to a specific type of information:

a) *What does it mean?*: a reception oriented module providing direct access to the entry of a specific collocation. The user is expected to introduce a base (e.g. *amistad*) and a collocate (e.g. *reanudar*) to be directed to the entry of the corresponding collocation.

b) *Writing aid*: a production oriented module, which allows the user to find collocates of a given base (e.g. *amor* 'love'), corresponding to a specific part of speech and meaning (e.g. 'felt for one another'), such as *amor mutuo* 'mutual love'.

c) *Direct search*: an option which serves to find collocations encoded by a specific Lexical Function (e.g. Sing(remordimiento) = *acceso de ˜* 'fit of remorse').

d) Inverse search: a module where the user is asked to introduce a collocate (e.g. *cumplir* 'fulfill') in order to find the bases it can be combined with (e.g. *deseo* 'wish', *esperanza* 'expectation').

Finally, the third component, the *learning module*, aims to provide the user with learning material concentrating on collocations. For the present it is limited to a few sections containing exercises related to a particular topic, one of which is an introduction to the use of the DiCE itself.

However, these learning activities do not differ consistently from those available on paper, but, just as an e-dictionary should offer more advanced features rather than being a mere electronic version of a paper dictionary, e-learning activities should be different from traditional teaching material. First of all, the collocation verification process should enable the user to access external language corpora, besides relying on the dictionary's own database. For instance, if in an exercise aimed at practising intensifier collocates, a learner provides *total* 'complete' as a collocate of *admiración* 'admiration', the current system will treat it as incorrect because this combination is not included in the DiCE database. However, a search in external corpora would enable the user to check whether the collocation is used in language and with what frequency as compared to other combinations with a similar meaning.

The use of language corpora is being promoted in language teaching since it is in line with the current trend of emphasizing autonomous learning. We also had the idea that learner autonomy could be further reinforced by the creation of a learning space in which learners can administer their personal collocation dictionaries, annotations, performance scores and problems identified in relation to specific collocations or collocation types. Ultimately, we believed that an ideal CALL environment focusing on collocations should tightly integrate a number of different components: a collocation database, a corpus interface, a collocation checker tool and other learning utilities, in order to support the users' collocation production in writing tasks.

These ideas constituted the main incentive behind the development of an interactive collocation learning environment. In order to create such tool, it was necessary to learn about its potential users, to which end we set out to gather information on users' reference skills when it comes to using a collocation database such as the DiCE, as well as on language learners' collocation proficiency. In the following section we will briefly present some findings concerning these two aspects.

3 Getting to know the user

3.1 Users' reference skills

As claimed above, the modifications of the DiCE interface were aimed at turning it into a useful tool for a wide range of users. This is the reason why a usability test was carried out to see how well different target user groups were able to perform with the dictionary. The aim of the test was to assess the different search options offered by the interface both in terms of efficiency and the adequacy of the layout, as well as to examine whether users' reference skills met those required by the DiCE.

In relation to user skills and preferences, the study, described in detail in Vincze and Alonso (2013), revealed that subjects were rather reluctant to explore the dictionary interface in search of different search options and that they were not familiar with certain terms applied in the dictionary. It was observed that subjects preferred to stick to familiar or more straightforwardly accessible search options, and did not show willingness to experiment with unknown or more novel functions. This could be seen in that they most frequently used the *Dictionary module* instead of more specific search options that could have provided more direct and quicker access to the items they were required to look up. The reason for this could be, on the one hand, that this access path is offered by default in the web interface, and, in addition, it allows the correct answer to be retrieved in the case of most questionnaire items; consequently when participants managed to find the required information in this way, they did not turn to the advanced search options. Furthermore, the type of access provided by this module is very similar to paper dictionaries and may therefore seem more familiar to users. Another finding pointing to the direction of users' preference for familiar search options was that the second most frequently and most successfully used query type was *What does it mean?*. It can be argued that this query type stands for the most common type of dictionary use, i.e. looking up a given lexical item in order to check its meaning or its spelling, as opposed to production oriented look-ups represented by the *Writing aid* option.

With respect to participants' reference skills, it was found that a lack of knowledge concerning the terminology applied in the dictionary caused difficulties in interpreting the dictionary content involving some of the query interfaces and the presentation of lexicographic data. Subjects were often unfamiliar with the notion of collocation and the specific terminology applied in the DiCE, leading them to confuse the elements of collocations (the base and the collocate), as well as with the more general concepts of word form and lemma, complicating the use of a number of search options.

In conclusion, the usability study of the DiCE interface showed that potential users of an online lexical learning environment 1) are more used to manipulating lexical resources in reception than in production tasks, and that 2) they might be more successful at using a tool whose functions do not differ radically from resources they are already familiar with, 3) whose search options are not highly modular, and 4) which keeps reference skill requirements to the minimum.

3.2 Language learners' collocation use

In order to design useful learning tools, it is necessary to know how learners use collocations. Previous studies (Alonso Ramos et al. 2010b, 2010c; Vincze et al., 2011; Wanner et al., 2013a), addressed the following two research questions for Spanish as L2: (1) Can errors in learners' collocation use 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 learning material targeting the type of error?

Previous work suggests that a CALL environment focusing on collocations can profit from data on learners' actual language behaviour obtained from corpus research (Shei and Pain, 2000; Chang et al., 2008). In order to gain information on the collocation knowledge and typical errors of Spanish as L2 learners, correct and erroneous collocations in a portion of the CEDEL2 corpus[3] (Lozano and Mendikoetxea, 2013) were

[3]CEDEL2 is an L1 English-L2 Spanish learner corpus containing essays written by English mother tongue Spanish L2 learners see `http://www.uam.es/`

annotated. Although currently available general learner error typologies tend to group collocation errors into a single subclass of lexical errors (Aldabe et al., 2005; Miličevič and Hamel, 2007; Granger, 2007; Díaz-Negrillo and García-Cumbreras 2007), a closer look at the learner corpus revealed that a considerably more detailed collocation error typology is needed in order to offer more targeted (and thus more effective) learning exercises, and to facilitate the development of techniques for automatic correction of collocation errors in learner writing.

Consequently, we created a detailed collocation error typology, which distinguishes three parallel dimensions (for a more detailed description see Alonso Ramos et al., 2010b and 2010c). The first of these captures the location of the error, i.e. whether it affects the base, the collocate, or the collocation as a whole. The second dimension models descriptive error analysis and distinguishes between three main types of error: lexical, grammatical and register error. Finally, the third dimension represents explanatory error analysis: it classifies errors according to their perceived source into one of the two main categories of transfer errors, namely errors reflecting L1 interference or interlanguage errors, the result of incomplete knowledge of the L2 without L1 interference.

The annotated corpus contains 46,266 words, in which a total number of 1938 collocation tokens, corresponding to 1171 collocation types were identified during the manual annotation process. Manual selection of collocations was necessary since our aim was to only examine combinations which qualify as collocations following our theoretical framework (see Section 1). Out of the total number of annotated collocation tokens, 1481 are correct and 457 are erroneous.

As for the location dimension, it was found that lexical errors most often affect the collocate, in a total of 180 collocations (62%), see (1), although a relatively large proportion, 62 collocations (21%) have erroneous bases, see (2), with cases of collocations having both an incorrect base and collocate, see (3), while 50 expressions (17%) contain a lexical error that is considered to affect the collocation as a whole. These results suggest that a genuinely effective CALL system should not be limited to recognizing errors in the collocate, as in e.g. Liu (2002) or Chang et al. (2008) (see below), but should also foresee lexical errors concerning the base or even both elements of the collocation.

(1) **interrumpir una regla* 'interrupt a rule' instead of *romper una regla* 'break a rule'

(2) **lograr un gol* 'achieve a goal (in sport)' instead of *lograr un objetivo* 'achieve an aim

(3) **pasar un testemuño* 'pass a testimony (from Portuguese)' instead of *dar testimonio* 'give testimony'

Automatic correction of the third error type included in the location dimension may present a considerable challenge. Errors affecting the collocation as a whole include incorrect collocation-like expressions that should be correctly expressed by a single word (4) and cases of incorrect single-word forms used instead of a collocation (5)

(4) **poner apasionado* 'make passionate' instead of *apasionar* 'to fascinate'

(5) **misenterpretación* 'misinterpretation' instead of *mala interpretación*

With respect to the explanatory error type dimension, of the 292 lexical collocation errors found in the corpus (note that a collocation can contain more than one error), 60% were labeled as transfer errors, while 40% were annotated as interlanguage errors. This is in line with the findings of other authors such as Liu (2002), Nesselhauf (2005), etc. Our corpus data also corroborates the hypothesis that in most lexical collocation errors, the erroneous element can be conceived of as a synonym or a translation synonym of its correct counterpart for correction purposes, a feature that can be made use of by automatic tools (Liu, 2002; Chang et al., 2008; Futagi, 2010). Remarkably, our data shows this to be true both in the case of L1 transfer and interlanguage errors. Nevertheless a small number of error types do not fit into this picture.

Errors resulting from the phenomenon commonly known by language learners and teachers

proyectoinv/woslac/cedel2.htm.

as 'false friends (6) constitute such a case. Similarly, in the case of errors involving the use of lexical elements which constitute non-words in the target language (7), using translation equivalents or synonyms to provide correction suggestions may be problematic and/or insufficient. Here, the introduction of a strategy involving edit-distance should be considered.

(6) Hemos **licenciado en el colegio* (from college) en la vecina ciudad Lit. We earned a degree in the primary school in the neighbor town

(7) En Oaxaca se puede **ir de hiking* (instead of *hacer senderismo*) Lit. In Oaxaca one can go hiking

In addition to lexical errors, learner tools aimed at the correction of collocations should also take grammatical errors into account. From our point of view, certain grammatical errors are to be considered proper collocation errors, due to the fact that they affect the correct formulation of a lexical combination. In fact, grammatical collocation errors (see (8), (9) and (10)) were found rather frequently in the corpus, concerning 198 (45%) of the 457 erroneous collocations annotated.

(8) determination error: **tomar sol* instead of *tomar el sol* 'to sunbathe;

(9) incorrect government: **montar a bicicleta* instead of *montar en bicicleta* 'to ride a bike; *asisto la Universidad* instead of *asisto a la Universidad* 'I attend the university;

(10) incorrect number: **estamos en vacación* instead of *estamos de vacaciones* we are on holiday.

As we have shown in this section, learner errors affecting collocations can be of many kinds, and can be systematized in a specific typology. A sufficiently fine-grained distinction of error types can not only provide useful input for the design of teaching material, but can also be made use of when determining the strategies to be implemented in a tool offering automatic correction suggestions for collocation errors. Once we have a clearer idea of the difficulties learners have to face at the moment of using a collocation learning tool, as well as of the diversity of collocation errors made by learners of Spanish as L2, we can go on to examine some existing lexical tools for learners of English in order to verify whether they can solve some of the problems posed by collocations.

4 Facing the difficulties of writing texts through the use of online lexical tools

When producing a text in English, learners have at their disposal a number of online tools that help them cope with some of the problems described above. In this section, we examine a number of these tools, since, to the best of our knowledge, there are no resources of this kind for learners of Spanish. Depending on the type of information sought by learners and the output these resources produce, we have classified them into three groups, the first of which includes those tools that in some respects resemble conventional combinatorial dictionaries; in the case of the second group, the query interface is similar to that found in an electronic dictionary, but the output consists roughly of n-grams or strings of word forms; and finally, the third group consists of tools that enable users to verify whether a combination produced by them is correct or not.

Dictionary-like tools. If a learner is interested in finding out about the combinatorial properties of already known lexical units, they may use a collocation dictionary or tools such as the *Learning collocations* component of FLAX[4] (Wu et al., 2010), the automatic collocation dictionary *For better English*[5] or the *Combinations* utility of *Just the word*[6]. When using these tools, in much the same way as with a collocation dictionary, users look up the word they are interested in, and obtain its collocates sorted according to their syntactic structure (e.g. V+N, Adj+N, etc.). In one case (*Just the word*), the collocations are also grouped according to semantic proximity. Additionally, *Learning collocations* and *Just the word* provide frequency information for each collocation.

[4] http://flax.nzdl.org/greenstone3/flax?a=fp&sa=collAbout&c=collocations&if=flax

[5] http://forbetterenglish.com/

[6] http://www.just-the-word.com/

The way the user accesses a collocation dictionary like the DiCE is very similar, since, as explained above, the *Dictionary Module* provides access to collocates by looking up a lemma. Likewise, the information provided by the DiCE (syntactic structure, semantic grouping, frequency of the collocation) is as complete as that offered by the tools examined. With some of these tools, however, users' access to corpus information is more direct, since it is not filtered by the lexicographer's criterion. In addition to this, one of the tools examined (*Learning collocations*) offers the possibility of picking examples from corpora and storing them in the users' personal dictionary.

String-searching tools. Like the previous ones, tools of this kind can be used to obtain information about the combinations of a certain word or phrase. Their output, however, is less refined than that of a collocation-searching utility, since it lists strings of all kinds in which the target word or phrase is found. If users want to narrow down their search because they are only interested, for instance, in finding occurrences of the target word as the object of a certain verb, they can refine their query by specifying certain categorial or distributional features. Thus, the *Lexcheker* of *StringNet*[7] (Wible et al., 2011) allows its users to exploit different degrees of specification by combining word class information and word-forms (e.g. [verb] *step*), whilst in the *Web Phrases* component of FLAX users can specify the distribution and length of the strings that combine with the target word or phrase.

Besides providing information about the correctness or the frequency of a particular combination, these tools can be especially useful for raising learners' awareness about grammatical restrictions related to the combination at hand (e.g. whether a certain verb takes a to+infinitive complement or gerund; preposition selection, etc.).

Collocation checkers. By means of the resources examined so far, a learner aiming to use a certain lexical item and wanting to know which other words can be combined with it can find the correct word choices and discard incorrect ones. With a collocation checker, however, learners who have already come up with a certain combination that they believe expresses the meaning they want to convey can seek a confirmation or a rejection of their hypothesis. Tools such as the *Collocation checker*[8] (Chang et al., 2008) or *Just the word* (when searching for a phrase instead of a single word) can be employed to this end, since they provide the user with feedback concerning the correctness of the combination introduced (based on its attestation in corpora) together with frequency information and suggestions of other possible combinations.

Some limitations of this type of tools have to do with the (lack of) coverage of all possible types of learner errors. The *Collocation checker*, for instance, focuses on V+N collocations and gives feedback on whether a verb can be combined with a certain noun. Thus, if the collocation proposed by the learners is attested in corpora, they will receive a message stating its correctness and a list of related constructions. If the verb does not occur with the noun, the application will indicate either that the collocation "might not be appropriate" or that it does not recognize such an expression and will provide alternatives with other verbs. However, as shown above, collocation errors can affect different parts of a combination. Thus, if we search for a combination of a verb plus a non-existent noun (e.g. **make cite*, instead of *make an appointment*, cf. Sp. *cita* 'appointment'), the tool will not provide any useful feedback to our query. Besides, the feedback given to infelicitous searches contains linguistic or lexicographic terminology (e.g. *lemma, support verb*) that may be unfamiliar to users, as the DiCE usability test has suggested.

After having observed some tools that help learners find or check collocations, the following section presents a collocation learning assistant for learners of Spanish.

5 Getting closer to a collocation writing assistant

As already pointed out above, collocation errors can be of different types and degrees of complexity. As stated in Wanner et al. (2013b), the differing complexity of collocation errors has further consequences for the prospects of successful au-

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

[8] http://miscollocation-richtrf.rhcloud.com/

tomatic recognition and correction in case of erroneous use: some of them will be more easily and more accurately recognized and corrected by state of the art techniques than others, whilst some of them require a further step to be taken. In what follows, we first introduce the requirements for a collocation checker tool, after which we provide a brief presentation of the HaRenEs[9] interface under development, a learning tool focusing on Spanish collocations[10].

5.1 Requirements for a collocation writing assistant

On the basis of the conclusions drawn from the usability and learner corpus studies previously presented, as well as the overview of existing online lexical tools provided, it is possible to formulate a list of requirements for the learning environment we aim to create. These can be organized in the following way:

- The target of the learning tool: the proposed tool should focus on collocations as understood within our theoretical framework (see Section 1). This means that we do not wish to treat phraseological strings that are produced as non-compositional chunks, such as *de acuerdo con* 'in accordance with'. We will concentrate strictly on restricted lexical co-occurrence phenomena, as in e.g. *acuerdo tácito* 'tacit agreement'[11].

- Accuracy of correction: the learning tool must in all cases provide feedback regarding the correctness of a collocation introduced, and, in the case of incorrect combinations, it should provide accurate correction suggestions. By this we mean that the collocation checker has to determine the nature of the error, including grammatical errors (e.g. **asistir la universidad* '*assist university*').

- Integration with other resources: the learner tool should be integrated with corpora and dictionaries. All suggested collocations should be illustrated with corpus examples, and the user should be redirected to existing entries in the DiCE or other online dictionaries.

- Features supporting usability and learning: users should have at their disposal a personalized collocation dictionary in which they can include new collocations accompanied by examples, as well as collocation errors. Collocation look-up and checking should be available by introducing either a stand-alone collocation or a text. When the interface is used to verify collocations in running text, the user should be able to further edit the text once it has been verified. Dictionary look-ups should be available both through the syntactic pattern and the semantic content of a collocation. Users should be provided with a number of learning activities for practicing collocations learnt through the collocation checker (similarly to FLAX).

5.2 HaRenEs Writing Assistant

The HaRenEs Writing Assistant is currently being developed in a joint project at the University of A Coruña and Pompeu Fabra University. The current learning environment consists of three main components: 1) the collocation checker, 2) the collocation search and 3) the personal dictionary. The collocation checker allows users to verify the correctness of a specific Spanish collocation and, in the case of incorrect combinations, to request correction suggestions, as well as usage examples of a given collocation in context. Users can introduce a single collocation in the search box, not necessarily in the lemma form (e.g *dimos un paseo* 'we took a walk'); and they can also request the verification of collocations in running text. Figure 1 shows a screenshot of the HaRenEs interface in use.

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

[10] A demo version of the HaRenEs interface can be seen at: `http://harenes.taln.upf.edu/CakeHARenEs`

[11] We are aware of the fact that a sharp distinction cannot always be drawn between full idioms and collocations. However, we believe that the learning of these two types of multiword units differs considerably: among other things, full idioms are difficult to understand, but collocations are difficult to produce. The learner needs to know the collocation *acuerdo tácito* to speak about a kind of agreement, i.e. one that is implicit, not overtly expressed. On the contrary, *de acuerdo con* is learnt as a whole string since it does not contain the meaning 'acuerdo', but expresses a completely different meaning: [X] de acuerdo con Y: '[X] following the rule or the system Y or Y's wishes'.

Figure 1: The HaRenEs user interface

Unlike other proposals, our checker will offer accurate corrections of collocation errors, rather than lists of possible combinations ranked according to frequency. Furthermore, the system provides the option of linking any frequent learner error to the personal dictionary. Even though the different identification techniques used by the collocation checker are still in development (Ferraro et al., 2011; Moreno et al., 2013; Wanner et al., 2013b; Ferraro et al., 2014), the results obtained so far are promising. The system is being trained with data from CEDEL2. In Table 1 we provide examples of learner errors found in the corpus together with the corrections automatically suggested by the tool (see Ferraro et al., 2014).

Error	**Suggested Correction**
realizar meta lit. 'to realize an aim'	*alcanzar una meta* 'achieve an aim'
cambiar al cristianismo 'to change to Christianity'	*convertirse al cristianismo* 'to convert to Chistianity'
concluir un problema 'to conclude a problem'	*resolver* 'solve a problem'

Table 1: Suggested corrections of collocation error provided by HaRenEs

In order to verify the effectiveness of the collocation checker with running text, we carried out a test with full sentences taken from the learner corpus. For instance:

(11) *La hija está tratando de *capturar la atención de su madre*
lit. 'The daughter intends to capture the attention of her mother.'

In this case, the checker tool detects the incorrect collocation **capturar la atención* lit. 'capture the attention' and proposes *llamar la atención* lit. 'call the attention'. The interface allows the user to accept or reject each of the multiple suggestions, consult examples of the suggested collocation, add it as a new entry to the personal dictionary, and link the collocation error to an existing dictionary entry.

The second component, *Collocation search*, is also still under development. It is designed to be similar to the dictionary-like lexical tools using corpora introduced in Section 4. However, in contrast to these, our goal is not only to provide access to collocations via their syntactic pattern (e.g. verb+*miedo* 'fear' or *miedo*+adj), but also through a semantic typology. For instance, if a user is searching for a way to express the meaning related to the starting phase of fear, it would be desirable to find verb+object collocacions such as *coger miedo* 'take fear of sg', as well as subject+verb collocations like *entrarle miedo* 'fear enters sb', *asaltarle miedo* 'fear assaults sb', or *invadirle el miedo* 'fear invades sb'. Note that in existing lexical resources these combinations are normally not found in the same category, since they are classified according to syntactic pattern.

Concerning the third component, the personal dictionary, we believe that it is highly useful to provide the option of linking erroneous collocations with their correct counterparts. Similarly to FLAX, users can be given the option of creating and organizing collocation lists at will. In our case, however, by default each collocation included in the personal dictionary by a user will be automatically registered in an entry with a standardized structure including the following fields: base, collocate, syntactic pattern, semantic class, examples and observations.

Unlike some of the other tools presented in Section 4, we do not allow the use of wild card operators in queries, since we try to keep user interactions as simple as possible for the sake of usability. Another point of difference with other lexical tools is that HaRenEs focuses on collocations, not on government: no direct queries can be carried out to find the preposition governed by

a given verb (e.g. *depender de* 'to depend on'). However, information on government that concerns a given collocation can be found. For instance, if a user wants to know whether a collocation such as *sentir miedo* 'feel fear' governs the preposition *a* or *de*, they can find this information in the examples coming from the corpus and also in the dictionary component.

An approach similar to that of StringNet would also be possible to implement, given that our reference corpus is tagged. However, before implementing this functionality, we need to test its efficiency with users. As we have seen in the usability test of the DiCE interface, we cannot take users' knowledge of technical linguistic terms or notions, such as e.g. names of parts of speech, for granted. And, ultimately, as mentioned above, the target of the HaRenEs environment is constituted by collocations, not merely frequent lexical combinations. However, although the metrics behind our tool are based on lexical frequencies, as is the case with other lexical checkers, we have set ourselves the challenge of automatically distinguishing between phraseological combinations such as *de acuerdo con* 'in accordance with' and genuine collocations such as *un acuerdo tácito* 'tacit agreement'.

6 Conclusions

Genuine lexical writing assistants that attempt to detect collocation errors have much less tradition in CALL than spelling and grammar checkers. In general they are not as mature as the latter: many of them are not successful enough in recognizing and correcting errors. However, this is not only due to the immaturity of the technologies. As we have shown, collocation errors are very heterogeneous and thus rather difficult to deal with.

Furthermore, the challenge not only lies in developing techniques capable of identifying and correcting collocation errors in a sufficiently accurate and efficient way, but also in designing an interface which any L2 learner can manipulate with ease. As pointed out above, there is a general tendency to blur the boundaries between dictionary and corpus and, going even further, to make the lexical tool itself almost invisible to the user, hoping that the user will be able to find any desired answer with a single click of the mouse. This design strategy is already operational but only in the case of language comprehension, not for production purposes. We would like to draw attention to this important difference and to make an appeal for a concerted effort to be made to build an efficient writing assistant.

Acknowledgments

The work presented in this paper has been supported by the Spanish Ministry of Economy and Competitiveness (MINECO) and the EFRD Funds of the European Commission under the contract number FFI2011-30219-C02-01, as well as the Spanish Ministry of Education under the FPU grant AP2010-4334 and the Galician Government under the post-doctoral grant POS-A/2013/191.

References

Aldabe, I., B. Arrieta, A. Díaz De Ilarraza, M. Maritxalar, M. Oronoz and L. Uria. 2005. Propuesta de una clasificación general y dinámica para la definición de errores. *Revista de Psicodidáctica*, 10/2: 47-60.

Alonso Ramos, M. 2005. Semantic Description of Collocations in a Lexical Database. In F. Kiefer et al. (eds.), *Papers in Computational Lexicography COMPLEX 2005*. Budapest: Linguistics Institute and Hungarian Academy of Sciences, 17-27.

Alonso Ramos, M. 2006. Towards a Dynamic Way to Learn Collocations in a Second Language. In Corino, E., C. Marello and C. Onesti (eds.), *Proceedings of the Twelfth EURALEX International Congress*. Accademia della Crusca, Università di Torino, Edizioni dellOrso Alessandria, Torino: 909-923.

Alonso Ramos, M. 2008. Papel de los diccionarios de colocaciones en la enseanza de español como L2. In Bernal, E. and J. De Cesaris (eds.), *Proceedings of the XIII EURALEX International Congress*. IULA, Documenta Universitaria, Barcelona: 1215-1230.

Alonso Ramos, M. 2009. Hacia un nuevo recurso léxico ¿fusión entre corpus y diccionario? In Cantos Gómez, P. and A. Sánchez Pérez (eds.), *A Survey of Corpus-based Research. Panorama de investigaciones basadas en corpus*. AELINCO, Murcia: 1191-1207.

Alonso Ramos, M., A. Nishikawa and O. Vincze. 2010a. DiCE in the web: An online Spanish collocation dictionary In S. Granger, M. Paquot

(eds.), *eLexicograpy in the 21st century: New Challenges, New Applications. Proceedings of eLex 2009*. Cahiers du Cental 7, Presses universitaires de Louvain, Louvain-la-Neuve: 367-368.

Alonso Ramos, M., L. Wanner, N. Vázquez, O. Vincze, E. Mosqueira and S. Prieto. 2010b. Tagging collocations for learners. In S. Granger, M. Paquot (eds.), *eLexicograpy in the 21st century: New Challenges, New Applications. Proceedings of eLex 2009*. Cahiers du Cental 7, Presses universitaires de Louvain, Louvain-la-Neuve: 369-374.

Alonso Ramos, M., L. Wanner, O. Vincze, G. Casamayor, N. Vázquez, E. Mosqueira and S. Prieto 2010c. Towards a Motivated Annotation Schema of Collocation Errors in Learner Corpora. *7th International Conference on Language Resources and Evaluation (LREC)* La Valetta, Malta: 3209-3214.

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.

Díaz-Negrillo, A. and M. A. García-Cumbreras. 2007. A tagging tool for error analysis on learner corpora. *ICAME Journal*, 31/1: 197-203.

Ferraro, G., R. Nazar and L. Wanner. 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)*: 69-79.

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

Futagi, Y. 2010. The effects of learner errors on the development of a collocation detection tool. In *AND'10 Proceedings of the fourth workshop on Analytics for noisy unstructured text data*. ACM, New York: 27-34.

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

Granger, S. 2007. Corpus d'apprenants, annotation d'erreurs et ALAO: une synergie prometteuse. *Cahiers de lexicologie*, 91/2: 465-480.

Hausmann, F. J. 1989. Le dictionnaire de collocations. In F. J. Hausmann et al. (eds.) *Wörterbücher-Dictionaries-Dictionnaires*, vol. 1. Gruyter, Berlin: 1010-1019.

Higueras García, M. 2006. *Las colocaciones y su ensenñanza en la clase de ELE*. Arco Libros, Madrid.

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

Jousse, A. L., A. Polguere and O. Tremblay. 2008. Du dictionnaire au site lexical pour l'enseignement/apprentissage du vocabulaire. In Grossmann, F. and S. Plane (eds), *Les apprentissages lexicaux. Lexique et production verbale*. Presses universitaires du Septentrion, Villeneuve dAscq : 141157.

Knublauch, H., R. W. Fergerson, N. F. Noy and M. A. Musen. 2004. *The Protégé OWL Plugin: An Open Development Environment for Semantic Web Applications. Third International Semantic Web Conference, Hiroshima, Japan*.

Liu, L. E. 2002. *A corpus-based lexical semantic investigation of verb-noun miscollocations in Taiwan learners' English*. Masters thesis, Tamkang University, Taipei.

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*. Boulder, CO: 47-50.

Lozano, C., and A. Mendikoetxea. 2013. Learner corpora and Second Language Acquisition: The design and collection of CEDEL2. In A. Díaz-Negrillo, N. Ballier and P. Thompson (eds.), *Automatic Treatment and Analysis of Learner Corpus Data*. John Benjamins, Amsterdam: 65-100.

Mel'čuk, I. A. 1996. Lexical Functions: A tool for the description of lexical relations in the lexicon. In L. Wanner (ed.), *Lexical Functions in Lexicography and Natural Language Processing*. John Benjamins, Amsterdam: 37-102.

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

Melčuk, I., A. Clas and A, Polguère. 1995. *Introduction à la lexicologie explicative et combinatoire*. AUPELF-UREF/Duculot, Louvain-la-Neuve.

Miličevič, J., and M.-J. Hamel. 2007. Un dictionnaire de reformulation pour apprenants avancés du franais langue seconde. *Revue de lUniversité Moncton*, numéro hors série: 145167.

Moreno, P., G. Ferraro and L. Wanner. 2013. Can we determine the semantics of collocations without semantics?. In Kosem, I., Kallas, J., Gantar, P., Krek, S., Langemets, M., Tuulik, M. (eds.), *Electronic lexicography in the 21st century: thinking outside the paper*. Proceedings of the eLex 2013 conference, 17-19 October 2013, Tallinn, Estonia.

Trojina, Institute for Applied Slovene Studies/Eesti Keele Instituut, Ljubljana/Tallinn: 106-121.

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

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.

Shei, C.C. and H. Pain. 2000. An ESL writer's collocational aid. *Computer Assisted Language Learning*, 13(2):167182.

Sinclair, J. M. 1987. The Dictionary of the Future. Collins English Dictionary Annual Lecture. University of Strathclyde, 6 May 1987.

Sinclair, J. M. 1991. *Corpus, concordance, collocation* . Oxford University Press.

Tarp, S. 2008. *Lexicography in the Borderland between Knowledge and Non-Knowledge*. Niemeyer, Tübingen.

Vincze, O., M. Alonso Ramos, E. Mosqueira Suárez and S. Prieto González. 2011. Exploiting a learner corpus for the development of a CALL environment for learning Spanish collocations. In Kosem, I. and K. Kosem (eds.), Electronic lexicography in the 21st century: New applications for new users. *Proceedings of eLex 2011*. Institute for Applied Slovene Studies.

Vincze, O. and M. Alonso Ramos. 2013. Testing an electronic collocation dictionary interface: Diccionario de Colocaciones del Español. In Kosem, I., Kallas, J., Gantar, P., Krek, S., Langemets, M., Tuulik, M. (eds.), *Electronic lexicography in the 21st century: thinking outside the paper. Proceedings of the eLex 2013 conference, 17-19 October 2013, Tallinn, Estonia*. Institute for Applied Slovene Studies, Trojina/Eesti Keele Instituut, Ljubljana/Tallinn: 328-337.

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

Wanner, L., S. Verlinde and M. Alonso Ramos 2013b. Writing assistants and automatic lexical error correction: word combinatorics. In Kosem, I., Kallas, J., Gantar, P., Krek, S., Langemets, M., Tuulik, M. (eds.), *Electronic lexicography in the 21st century: thinking outside the paper. Proceedings of the eLex 2013 conference, 17-19 October 2013, Tallinn, Estonia*. Institute for Applied Slovene Studies/Eesti Keele Instituut, Ljubljana/Tallinn: 472-487.

Wible, D., Nai-Lung Tsao. 2011. 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 (Studies in corpus linguistics, 45)*. Benjamins, Amsterdam, Philadelphia: 237-254.

Wu, J.-C., Y.C. Chang, T. Mitamura and J. S. Chang 2010. Automatic Collocation Suggestion in Academic Writing. In *Proceedings of the ACL Conference*.