Collocations: A Challenge in Computer Assisted Language Learning

Gabriela Ferraro (1), Rogelio Nazar (2), Leo Wanner (1, 3)
(1) Department of Information and Communication Technologies
Pompeu Fabra University, C/ Roc Boronat, 138, 08018 Barcelona
(2) Institute for Applied Linguistics, Pompeu Fabra University
(3) Catalan Institution for Research and Advanced Studies (ICREA)
<firstname>.<familyname>@upf.edu

Abstract

The correct use of collocations is one of the most difficult tasks that the student faces when learning a second language, such that one of the goals of Computer Assisted Language Learning (CALL) is to develop programs that aim to identify collocation errors in learners' writings and propose corrections. However, while statistical models currently used by most of these programs still manage to predict, with a reasonable probability, whether a given word combination is a valid collocation in the language in question or not, they fail to suggest corrections. At most, they offer a list of supposedly valid collocations of the base of the erroneous collocation, from which then the learner shall pick one. This is clearly unsatisfactory. We present ongoing work in which we aim to develop algorithms that do better in that they use the sentential context of the erroneous collocation to suggest a correction and in which we assess how crucial the use of Lexical Functions in the sense of the Explanatory Combinatorial Lexicology is in the context of CALL. All our work is tested on a corpus of American English learners of Spanish

Keywords

second language learning, CALL, collocations, lexical functions, Spanish

1 Introduction

Long time, the research in second language learning in general and in Computer Assisted Language Learning (CALL) in particular focused on difficulties of learners with grammatical constructions. The consequence of this was that while for typical grammatical errors more or less detailed analyses have been performed, all types of errors related to the lexicon have been generally classified as "lexical errors", without any further distinction (Granger, 2007). This is certainly a gross oversimplification. One of the larger classes of lexical errors is constituted by errors in the use of collocations (Granger, 1998; Nation, 2001). Since the early 2000ies, a considerable amount of work has been carried out in CALL on the development of programs

(although focused mainly on English as L2[1]) that judge a combination to be a valid or invalid collocation and, in the latter case, attempt to provide a list of correction suggestions. But, again, to consider all collocation errors to be of the same unique class is an oversimplification which does not do justice to the complexity of the problem and thus to the needs of learners. Alonso Ramos et al. (2010) presented a fine-grained collocation error typology which is based on an empirical study of a corpus of American English learners of Spanish (Lozano, 2009).[2] This typology reveals that learners often literally translate collocation elements from their native tongue, use non-existing words as collocation elements, get a wrong subcategorization for one of the elements, etc. Each of these errors requires a potentially distinct focus of the learning aid offered to the learner. Furthermore, in order to be able to correct an error in a targeted way, the meaning that the learner intended to express by the erroneous collocation must be known. In other words, we need to know the Lexical Function (LF) that the learner intended to use. In order to facilitate both learning aids that react to each type of collocation error distinctly and programs that are able to detect and correct collocation errors, the work in the COLOCATE project focuses on the following two tasks: (i) annotation of a learner corpus with collocation error types as defined in (Alonso Ramos et al., 2010) and with the corrections of the errors (tagged additionally with LF labels); (ii) development of algorithms for automatic recognition of collocation errors and their correction – in a long term, at the level of LFs. The first task is addressed by Alonso et al. (Alonso et al., 2011; Vincze et al., 2011). In what follows, we focus on the second task, presenting the state of our current effort towards this long term goal and assessing the next steps to be taken. In the next section, we discuss the related work in the area of collocation checking and correcting. In Section 3, our approach is outlined and its advances compared to the state of the art are discussed. Section 4 finally, presents the lines of our future work in this area.

2 Related Work

The research in the area of collocation checking focused so far mainly on one of the tasks related to collocation error correction: assessment whether a given word combination is a valid collocation in L2. The task of correction has been accounted for, as a rule, cursorily in that a list of collocations of the base in question to choose from has been offered.

The task of the validation of a word combination as a collocation is closely related to the task of collocation identification. Outside CALL, the identification of collocations in corpora has been actively worked on since the late eighties. The majority of the works explore purely statistical models (Choueka, 1988; Church & Hanks, 1990; Evert, 2007; Pecina, 2008). These ("first generation") models can be more or less complex, but all of them measure in one way or the other the distribution of words in combination and in isolation. Some of the works

[1] Following the terminology in language learning, we refer to the native tongue of the learner as L1 and the language being learned as L2.

[2] The corpus is question was CEDEL2 (http://www.uam.es/proyectoinv/woslac/cedel2.htm), which has been compiled by the group directed by Amaya Mendikoetxea from the Universidad Autónoma de Madrid. It contains about 400.000 words of essays in Spanish on a predefined range of topics by native speakers of English.

combine the statistical model with the use of some syntactic features – e.g., submitting to the statistical model only words in collocation-valid syntactic structures (Smadja, 1993; Kilgarriff, 2006; Evert and Kermes, 2003). Most recent statistical proposals take the context of the words that tend to occur into account, which allows for an indirect consideration of the semantics of these words (Bouma, 2010). Another strand uses the co-occurrence range of a given word, i.e., relative frequencies of tokens that co-occur with this word most often (Wible and Tsao, 2010). Opposed to the frequency-based models above is our previous work (Wanner, 2004; Wanner et al., 2005), which uses explicit semantic information from EuroWordnet (Vossen, 1989) to identify and classify collocations with respect to the typology of LFs.

In CALL, the vast majority of the approaches uses statistical models of the first generation (see Chang et al., 2008; Chen, 2010; Park, 2008 and others) or do not use Natural Language Processing techniques at all. Since the pioneering work by Shei and Pain (2000), quite a few proposals have been made on how to improve the collocation competence of the learner of English. First of all, V+N collocations have been considered; see, for instance, (Park et al., 2008; Chang and Chang, 2004; Chang et al., 2008; Chen, 2010, Wu et al., 2003; Wu et al., 2010; Wu, 2010). Futagi et al. (2008) are among the few who treat other syntactic constructions and also consider grammatical errors related to collocations. As far as the resources used in these proposals are concerned, the tendency is to use, in addition to the learner corpus, synonym dictionaries, bilingual dictionaries (which shall facilitate the detection of calques from L1) and reference corpora for L2. S. Wu et al. (2010) and Park et al., (2008) use lists of *n*-grams as reference corpus (in the case of S. Wu et al., provided by Google as mirror of the web).

In general we can state that the current proposals on collocation error recognition and correction still suffer from three shortcomings. First, they are not able to distinguish between "true" collocations and frequent free co-occurrences. Consider, in this context, the reaction of the MUST collocation checker (Wu et al., 2010) to the erroneous collocation *make question* in Figure 1,[3] next page. Second, they are not able to offer any kind of error classification, although such a classification would be helpful to find the most adequate correction of the error. Compare, for instance:

(1) a. *Yo tengo el deseo personal de ser bilingüe*, lit. 'I have the personal wish to be bilingual'
b. *gastar todo el año estudiando español*, lit. 'spend all the year studying Spanish'
c. *hablar un lenguaje*, lit. 'speak (a formal) language'
d. *derechos mujeriles*, lit. 'women rights'

where in (1a), we encounter a register error (the collocation *tener* [*un*] *deseo* 'have [a] wish' exists, but it is not appropriate in this context), in (1b) a collocate error (*pasar* [*un*] *año* instead of **gastar* [*un*] *año*), in (1c) a base error (*hablar* [*una*] *lengua* instead of **hablar* [*un*] *lenguaje*), and in (1d) the use of a non-existent word as collocate (*derechos de las mujeres* instead of *derechos *mujeriles*).

[3] Consulted at http://miscollocation.appspot.com on July, 20th 2011.

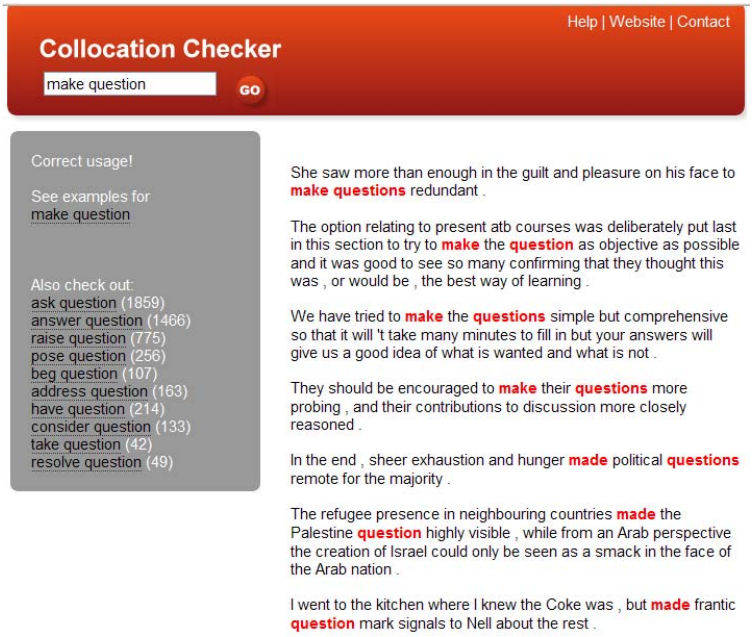

Figure 1: Output of the MUST collocation checker

Third, they are not able to correct an error and must thus stick to offering a list of possible options the learner has to choose from, without any meaningful preferences. Compare, for instance, the following list provided by MUST for the correction of the erroneous collocation *have* [*an*] *obstacle* (in the cited order):

> *overcome obstacle*, *present obstacle*, *clear obstacle*, *jump obstacle*, *prove obstacle*

As further options, the following combinations are given under the heading "Also check out":

> *remove obstacle*, *place obstacle*, *remain obstacle*, *surmount obstacle*, *face obstacle*, *avoid obstacle*, *encounter obstacle*, *eliminate obstacle*, *negotiate obstacle*, *erect obstacle*.

Apparently, MUST attempts to separate correction candidates that are closer to the sought collocation (according to a specific metric) from less likely candidates. But, at least in this example, it does not recognize the intended semantics of the erroneous collocation. The right correction, *face obstacle*, is listed as fifth in the secondary "Also check out"-list. If the level of the learner is not advanced enough, he will not be able to make the right choice.

3 A step forward in collocation error recognition and correction

In our experiments, we focused so far on the third of the three shortcomings of the state-of-the-art proposals listed above, and, since the motivations for the third and second shortcomings are at least related, partially also on the second. Our technique for the recognition of collocation errors is therefore still largely comparable with the state-of-the-art techniques in the field in that it is based on frequency oriented metrics to decide whether a given combination is a correct collocation or not. Similarly to Park et al. (2008) and S. Wu et al. (2010), we use a list of *n*-grams (with $n \leq 4$) as a reference corpus. In our case, this list has been derived from a large Spanish newspaper corpus. Furthermore, we use a number of

auxiliary resources: the Open Office thesaurus, an automatically compiled bilingual Spanish-English dictionary, the Spanish EuroWordNet, and the Web as an additional reference corpus.

In the next subsection, we present first the procedure for the assessment of the correctness of a collocation in Spanish and for the selection of the best correction candidate in case the collocation is judged incorrect, and illustrate then how the procedure performs in action.

3.1 Collocation error and correction procedure

Given a V+N, V+Adv, Adj+N or Adj+Adv combination *C* (extracted from the learner corpus or introduced via an on-line interface) the procedure is as follows:

1. Check whether the relative frequency f_C of C:=Co+B[4] in the *n*-gram list is higher then an empirically determined threshold *T*
2. IF $f_c > T$, *C* is considered a correct collocation of Spanish
   ELSE do
   Collect the synonyms Co_{syn} of *Co* from the auxiliary resources.[5]
   Check whether any $c_{syn} \in Co_{syn}$ forms together with *B* a valid collocation (again, in terms of relative frequency).
   IF there are several valid collocation candidates c_{syn}+B, choose the one judged best according to a number of metrics.

Three different metrics have been applied to judge which of the candidates c_{syn}+B is the best correction of the supposedly erroneous *C*. We present each of them in what follows.

A. Affinity metrics: For each c_{syn}, its affinity is calculated as the product of association strength to *B* and graphic similarity to *Co*, plus the synonymy factor with respect to *Co*. The association strength between c_{syn} and *B* is obtained using the standard *log*-likelihood measure:

$$f(c_{syn} + B) / \ (\text{sqrt}\ (\ f(c_{syn})) * \text{sqrt}\ (f(B))\ )$$

The graphic similarity between c_{syn} and Co is calculated as the Dice coefficient:

$$sim(Co, c_{syn}) = 2\ |Co| \cap |\ c_{syn}| \ / \ |Co| \cup |\ c_{syn}|$$

The synonymy factor of c_{syn} with respect to *Co* is '1' if c_{syn} is among the synonyms of *Co* in the synonym list obtained from the auxiliary resources and '0' otherwise.

B. Lexical context metrics: The lexical context-oriented metrics is grounded in the assumption of distributional semantics, namely that the semantics of a word combination can

---

[4] 'Co' stands for "collocate" and 'B' for "base". In V+N and Adj+N combinations, N is considered the base and V respectively Adj the collocate. In V+Adv, V is the base and Adv the collocate and in Adj+Adv, Adj is the base and Adv the collocate.

[5] In the case of the bilingual Spanish-English dictionary, the "synonyms" of *Co* are the Spanish translations of the English translation equivalents of *Co*. We are aware that this procedure provokes a lot of noise since it ignores the problem of polysemy. However, it has the advantage that it allows us to capture calques from L1.

be approximately deduced from the sentential context in which this word combination appears. Consider, for illustration, the following sentences (from the web) in which one of the words has been removed:

(1) a. *She * a conference on the situation of women rights …*
b. *Mr. White responded to the changing industry and * a conference of critical success.*
c. *Eventcorp** a conference *that met the Conference Committee's criteria.*

(2) a. *The mailman** apples*, bananas, and coconuts.*
b. *Oo baby, here I am, signed, sealed *, I'm yours, oh I'm yours*… [Stevie Wonder song]
c. *Fast Flowers * fresh flowers for every occasion*.

In (1a-c), we can deduce with a certain probability that the missing word is [*to*] *deliver* or any other support verb that goes with *conference*. Why? Distributional semantics suggests that it is the context that allows us to come up with [*to*] *deliver*. In contrast, in (2a-c), this is not the case: we cannot reliably guess the missing verb. This gives us a hint that in (2a-c) the missing verb does not participate in a collocation. We can thus hypothesize that context can be useful for the detection of collocations, or, in our case, for the search of the most adequate correction candidate. More precisely, we assume that given the sentential context c_1, c_2, …, c_n of *Co* in the original sentence of the learner, the candidate c_{syn} with the highest affinity to c_1, c_2, … c_n in a reference corpus is the most adequate correction of *Co* (with "affinity" meaning the highest relative co-occurrence frequency).[6] In contrast to information retrieval oriented search, we do not eliminate from the context the functional words (which are otherwise considered as "stop words" that do not contribute to the quality of the search) since they are essential for our task. For instance, in the learner sentence (3)

(3) *Afortunadamente, su profesora estuvo dispuesta a venderlas y pudo comprar dos máscaras para extender nuestra colección*
lit. 'Fortunately, his professor was willing to sell them and he could buy two masks to extend his collection'

the collocation **extender* [*una*] *colección*, lit. 'extend a collection [of art]' is not correct; this is identified in the first stage of the program. To find the right correction, the contexts of valid collocations of *colección* from our *n*-gram list are examined in the reference corpus with respect to the occurrence of *máscaras*, *para*, and *nuestra* in their neighbourhood. The strongest lexical affinities of *completar* [*una*] *colección*, lit. 'complete a collection' and *ampliar* [*una*] *collection*, lit. 'extend a collection' suggest that the program is accurate in this case.

C. Context feature metrics: As the lexical context metrics, the context feature metrics is based on the idea of distributional semantics. However, in contrast to the lexical context metrics, it allows for a more flexible implementation and the consideration of other features than concrete words. Given the sentential context c_1, c_2, c_n of *Co* in the original sentence of the learner and a list of candidates C_{syn}, the idea is to assess whether any of the contextual

[6] In our preliminary experiments, we used n ≤ 8 (with maximally 2 tokens to the left and 2 tokens to the right of each element of the combination, always within the boarders of a single sentence; duplicates are eliminated).

features $c \in Co$ speaks for the preference of one of the candidates, c_{syn}. For this purpose, we find the maximal probability of each feature c, given a collocation candidate (c_{syni}, b_i). (c_{syni}, b_i) can be calculated as:

$$argmax_{\,i=1,\ldots,n;\, c \in Co}\ (N(c_{syni},b_i) / \Sigma_{j=1,\ldots n} N(c_{synj}\ b_j)) \times (N(c, (c_{syni}, b_i)) / N(c_{syni}, b_i)),$$

where $N(c_{syni}, b_i)$ stands for the number of times the combination (c_{syni}, b_i) occurs in the corpus, and $N(c, (c_{syni}, b_i))$ for the number of times the feature c and the combination (c_{syni}, b_i) co-occur in the corpus at a distance of at most three tokens from each other. For instance, in the learner sentence (4), the collocation **sacarse* [*una*] *operación*, lit. 'take off an operation' is not correct:

(4) *Es fácil, sólo hay que sacarse una operación como Michael Jackson*
lit. 'It is easy, you have only to take off an operation as Michael Jackson'.

To find the right correction, the affinity between the candidate collocations of *operación* 'operation' and each of the contextual features is examined in the reference corpus; e.g.,

[*hay*] … *realizar una operación*, [*que*] *realizar una operación*
[*hay*] … *hacer una operación*, [*que*] *hacer una operación*

(the contextual features are here *hay* and *que*, respectively). The candidate collocation which achieves the highest score is considered to be the correct one.

3.2 Examples of the collocation error correction procedure in action

Let us illustrate the application of the procedure described above to a real world example. (5) is a sentence taken from our learner corpus:

(5) *En mi nueva posición, yo hice planes de viajar para los grupos, acudí el teléfono e hice citas para conferencias con otras compañías para Gary.*
lit. 'In my new position, I made plans to travel for groups, [I] turned to the phone and made appointments for conferences with other companies for Gary'.

One of the potential collocations detected by the program is the V+N combination *hacer citas*, lit. 'make appointments'. Due to its low frequency in the reference corpus, the combination is judged to be a collocation error. In order to find the appropriate correction, all verbal co-occurrences of the base *cita* 'appointment' are retrieved from the reference corpus and filtered; only combinations with synonyms (according to our auxiliary resources) of *hacer* 'make' are kept. The remaining combinations are assessed with respect to their collocation status and non-collocations are removed. The remaining set of combinations includes:

realizar [*una*] *cita* 'realize [an] appointment', *producir* [*una*] *cita* 'produce [an] appointment', *dar* [*una*] *cita* 'give [an] appointment', *tener* [*una*] *cita* 'have [an appointment]', *ir* [*a una*] *cita* 'go [to an] appointment', *acudir* [*a una*] *cita* 'turn [to an] appointment', *declarer* [*una*] *cita* 'declare [an] appointment', *haber* [*una*] *cita* 'receive [an] appointment', *concertar* [*una*]

cita 'arrange [an appointement', *ser* [*una*] *cita* 'be [an] appointment', *agenciar* [*una*] *cita*[7] 'mediate an appointment'.

Given that the remaining set contains more than one option, the best correction candidate is chosen applying the metrics introduced above. The affinity metric suggests *realizar* [*una*] *cita*, while the lexical and context feature metrics suggest *concertar* [*una*] *cita*, which is, in fact, the most appropriate correction of *hacer citas*. Consider a number of further examples summarized in Table 1.

Collocation error	**Suggested correction (collocate)**		
	Affinity metrics	**Lexical metrics**	**Context feature metrics**
realizar meta 'realize goal'	**hacer* 'make'	+*alcanzar* 'reach'	+*alcanzar* 'reach'
cambiar [*al*] cristianismo 'change to Christianity'	+*convertir* 'convert'	+*convertir* 'convert'	+*convertir* 'convert'
comer café 'eat coffee'	+*tomar* 'take'	+*tomar* 'take'	**estar* 'be'
quedar [*la*] *tradición* 'remain [the] tradition'	+*seguir* 'follow'	+*seguir* 'follow'	**pasar* 'pass'
utilizar [*la*] *oportunidad* 'use [the] oportunity'	+*aprovechar* 'take advantage'	**ver* 'see'	**dar* 'give'
concluir [*un*] *problema* 'conclude [a] problem'	+*resolver* 'resolve'	+*solucionar* 'solve'	+*acabar* 'terminate'
empezar [*una*] *familia* 'begin [a] family'	**acomodar* 'accomodate'	+*formar* 'form'	+*formar* 'form'
interrumpir [*una*] *regla* 'interrupt [a] rule'	**establecer* 'establish'	**imponer* 'impose'	+*violar* 'violate'

Table 1: Examples of the correction of collocation errors by our program ('*' stands for wrong correction suggestion and '+' for correct correction suggestion)

Note that some wrong correction suggestions might be valid collocations (as, e.g., *dar* [*una*] *oportunidad*), but with a different semantics than the one required.

3.3 Evaluation

A quantitative evaluation of the procedure for the identification of collocation errors reveals that we are able to judge whether a combination is a correct or incorrect collocation in Spanish with an accuracy of 0.90. Thus, from 61 samples, the procedure fails in six cases. In five of these six cases, correct collocations have been judged to be incorrect. This is mainly due to our purely frequency-based collocation criteria. For instance, *apretar* [*los*] *dientes*, *contar cuentos*, *dar* [*la*] *bienvenida*, and *preparer* [*la*] *comida*, are correct collocations in Spanish, but their frequencies in our reference corpus are too low to consider them valid. On the other hand, for example, *pasar* [*la*] *navidad* is judged by the program to be a correct

[7] The suggestion of **agenciar* [*una*] *cita* as a possible correction candidate is due to the wrong PoS tagging of the bigram *agencia cita* 'agency cites', which is very common in a newspaper corpus as ours.

collocation due to its high frequency in our corpus, although it is questionable in European Spanish.[8]

For the second stage, i.e., the error correction stage, we performed so far two evaluations. First, in order to be able to compare our framework directly with other approaches, we evaluated the accuracy with which we are able to provide lists of valid collocations within which the right correction is encountered. This accuracy amounts to 0.73: in 73% of the trials, the right correction was encountered in the list of possible options offered by our program. Second, we evaluated the capacity of our algorithm to offer the right correction using the context feature metrics, with features being simply words in the original sentence of the learner (the metrics was thus equivalent to the lexical context metrics). The accuracy was 0.542. This is certainly still too low to be used in practical CALL. However, it is to be pointed out that the potential of the contextual features has not been fully explored as yet: the use of concrete words is too restrictive. The experience from statistical NLP (e.g., parsing and generation) teaches us that combinations of morpho-syntactic categories, grammatical functions and words are more promising. We will carry out experiments in this respect in the near future. Furthermore, it needs to be pointed out that this is the first proposal that attempts to suggest the exact correction of a collocation error (rather than to offer a list of suggestions from which the learner has then to choose).

4 Towards an advanced collocation-oriented CALL

In our experiments, we used so far only a limited amount of linguistic information, namely the morpho-syntactic categories of the elements of the combinations. While this information is necessary it is by far not sufficient. Thus, with only this information at hand, we not able to distinguish between *aprovechar* [*de una*] *oportunidad* 'take advantage [of an] opportunity' and *dar* [*una*] *oportunidad* [*a alguien*] 'give [an] opportunity [to so]' – the first being $Real_1$ and the second $CausOper_1$ in terms of LFs. We need to have access to the semantics of collocations! So far, no techniques have been developed that are able to address this challenge without depending on external lexico-semantic resources. On the other hand, the experiments in (Wanner, 2004) demonstrated that even WordNet, as the biggest resource of this kind, is by far not sufficient. This means that the only promising alternative is the use of stochastic techniques based on *distributional semantics* of the collocations in corpora. Our context feature metric is the first try in this direction, but more and additional features need to be exploited to be able to distinguish between the use of $Real_1$, $Oper_1$, $CausOper_1$, etc.

Our future work in the area of CALL will follow three different strands: first, development of techniques for automatic classification of collocation errors according to Alonso Ramos et al.'s typology (2010); second, development of techniques for automatic semantic classification of collocations identified in corpora; and third, amelioration of our techniques for the automatic correction of collocation errors. The learner corpus annotated by LFs and collocation error types by the group LYS at the University of La Coruña (Alonso Ramos et

[8] However, it is a standard collocation in Argentinean Spanish.

al., 2010) and the LF corpus of the Spanish collocation dictionary DICE (Alonso Ramos, 2009) will be essential for all three tasks.

Acknowledgements

Our experiments have been partially run on the Argo cluster of the Department of Communication and Information Technologies, UPF. Many thanks especially to Silvina Re and Iván Jiménez for their help. This work has been supported by the Spanish Ministry of Science and Innovation and the FEDER Funds of the European Commission under the contract number FFI2008-06479-C02-02 in the scope of the project COLOCATE. COLOCATE is a joint effort by the groups LYS, University of La Coruña and TALN, University Pompeu Fabra, Barcelona. We would like to thank the director of the LYS team Margarita Alonso Ramos for the very fruitful collaboration.

Bibliography

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

Alonso Ramos, M., L. Wanner, O. Vincze, G. Casamayor, N. Vázquez, E. Mosqueira & S. Prieto (2010). Towards a Motivated Annotation Schema of Collocation Errors in Learner Corpora. In *Proceedings of LREC 2010*, Malta.

Alonso Ramos, M, L. Wanner, O. Vincze, R. Nazar, G. Ferraro, E. Mosqueira & S. Prieto (2011) Annotation of Collocations in a Learner Corpus for Building a Learning Environment. In *Proceedings of the Learner Corpus Research Conference*, Louvain-la-Neuve.

Bouma, G. (2010): Collocation Extraction beyond the Independence Assumption. In *Proceedings of the ACL 2010 Conference Short Papers*, Uppsala, pp. 109–114

Chang, J. S. & Y.C. Chang (2004): Computer Assisted Language Learning Based on Corpora and Natural Language Processing: the experience of project CANDLE. En *Proceedings of IWLeL 2004: An Interactive Workshop on Language e-Learning*, pp. 15–23.

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

Chen, H. H-J. (2010): Developing an English Collocation Retrieval Web Site for ESL Learners, pp. 25–34

Choueka, Y. (1988) Looking for Needles in a Haystack. In *Proceedings of RIAO '88*, pp. 609–623.

Church, K. W., Y P. Hanks (1990): Word Association Norms, Mutual Information, and Lexicography. *Computational Linguistics*, 16(1):22–29.

Evert, S. (2007) Corpora and Collocations. Extended Manuscript of Chapter 58 of A. Lüdeling and M. Kytö, 2008, *Corpus Linguistics. An International Handbook*, Berlin: Mouton de Gruyter.

Evert, S. & H. Kermes (2003) Experiments on Candidate Data for Collocation Extraction. In *Companion Volume to the Proceedings of the 10th Conference of the EACL*, pp. 83–86.

Futagi, Y., P. Deane, M. Chodorow & J. Tetreault (2008) A Computational Approach to Detecting Collocation Errors in the Writing of Non-Native Speakers of English. *Computer Assisted Language Learning*, 21(4):353–367.

Granger, S. (2007) Corpus d'apprenants, annotations d'erreurs et ALAO : une synergie prometteuse. Cahiers de lexicologie, 91(2):465–480.

Granger, S. (1998) Prefabricated Patterns in Advanced EFL Writing: Collocations and Formulae. In *Phraseology: Theory, Analysis, and Applications*, A. P. Cowie (ed.), Oxford : Oxford University Press, pp. 145–160.

Kilgarriff, A. (2006): Collocationality (and how to measure it). In *Proceedings of the 12th EURALEX International Congress*, Torino, Italy.

Lozano, C. CEDEL2: Corpus Escrito del Español L2. In: Bretones Callejas, C. M. et al. (eds) *Applied Linguistics Now: Understanding Language and Mind / La Lingüística Aplicada Hoy: Comprendiendo el Lenguaje y la Mente*. Almería: Universidad de Almería. Almería, pp. 197–212.

Nation, I.S.P. (2001): *Learning Vocabulary in Another Language*, Cambridge: CUP.

Park, 2008 Park, T., E. Lank, P. Poupart & M. Terry (2008): "Is the sky pure today?" AwkChecker: An assistive tool for detecting and correcting errors. In *UIST '08: Proceedings of the 21st annual ACM symposium on User interface software and technology*, 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)*, Marrakech, pp. 54–57.

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

Smadja, F. (1993): Retrieving collocations from text: Xtract. *Comput. Linguistics*, 19(1):143–177.

Vincze, O., M. Alonso Ramos, E. Mosqueira & S. Prieto (2011) Exploiting a Learner Corpus for the Development of a CALL Environment for Learning Spanish Collocations. In *Proceedings of the eLEX 2011*, Bled, Slovenia.

Vossen, P. (1998) *EuroWordNet: A Multilingual Database with Lexical Semantic Networks*. Kluwer Academic, Dordrecht.

Wanner, L. (2004) Towards Automatic Fine-Grained Semantic Classification of Verb-Noun Collocations. *Natural Language Engineering Journal*, 10(2): 92–143

Wanner, L., B. Bohnet, M. Giereth. 2006. Making sense of collocations. *Computer Speech & Language*, 20(4): 609–624

Wible, D. & N.L Tsao (2010) Stringnet as a Computational Resource for Discovering and Investigating Linguistic Constructions. In *Proceedings of the NAACL-HLT Workshop on Extracting and Using Constructions in Computational Lingistics*, Los Angeles.

Wu, S. (2010) *Supporting collocations learning*. PhD Thesis, University of Waikato, Hamilton, NZ.

Wu et al., 2003 Wu, J. C., K.C. Yeh, T.C. Chuang, W.C. Shei & J. S. Chang (2003) TotalRecall: A bilingual concordance for computer assisted translation and language learning. In *Proceedings of the 41st ACL Conference*, Sapporo.

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