Nicoletta Calzolari, Khalid Choukri, Bente Maegaard, Joseph Mariani, Jan Odijk, Stelios Piperidis, Mike Rosner and Daniel Tapias (eds.) Proceedings of the Seventh conference on International Language Resources and Evaluation (LREC'10), 19-21 May 2010, Valletta, Malta. Language Resources Evaluation.

Towards a Motivated Annotation Schema of Collocation Errors in Learner Corpora

Margarita Alonso Ramos[1], Leo Wanner[2,3], Orsolya Vincze[1], Gerard Casamayor del Bosque[3], Nancy Vázquez Veiga[1], Estela Mosqueira Suárez[1], Sabela Prieto González[1],
[1]University of La Coruña, [2]Institució Catalana de Recerca i Estudis Avançats, [3]Pompeu Fabra University
[1]Campus da Zapateira s/n, 15071 CORUÑA, SPAIN
[3]C/ Roc Boronat, 138, 08018 BARCELONA, SPAIN
E-mail: lxalonso@udc.es, leo.wanner@upf.edu, ovincze@udc.es

Abstract

Collocations play a significant role in second language acquisition. In order to be able to offer efficient support to learners, an NLP-based CALL environment for learning collocations should be based on a representative collocation error annotated learner corpus. However, so far, no theoretically-motivated collocation error tag set is available. Existing learner corpora tag collocation errors simply as "lexical errors" – which is clearly insufficient given the wide range of different collocation errors that the learners make. In this paper, we present a fine-grained three-dimensional typology of collocation errors that has been derived in an empirical study from the learner corpus CEDEL2 compiled by a team at the Autonomous University of Madrid. The first dimension captures whether the error concerns the collocation as a whole or one of its elements; the second dimension captures the language-oriented error analysis, while the third exemplifies the interpretative error analysis. To facilitate a smooth annotation along this typology, we adapted Knowtator, a flexible off-the-shelf annotation tool implemented as a Protégé plugin.

1. Introduction: The Problem

The relevance of annotated learner corpora in second language acquisition is generally acknowledged to date; cf., among others, (Dagneaux et al., 1998; Granger, 1998, 2007; Tono, 2003). As a rule, the annotation marks grammatical, stylistic, and wording errors – including wrong idiosyncratic word co-occurrences such as, e.g., (*I*) *have a curiosity* or (*I*) *have 20 years* in English learner corpora and Sp. *salvar dinero* 'save money', Sp. *recibir un llamo* 'to receive a call', etc. in Spanish learner corpora. We focus on the latter – erroneous idiosyncratic word co-occurrences, or *collocations*. Following the common lexicographic tradition (Hausmann, 1989; Mel'čuk, 1998), we assume that a collocation is a restricted binary co-occurrence of lexical units (LUs) between which a syntactic relation holds, and that one of the LUs (the *base*) is chosen according to its meaning as an isolated LU, while the other (the *collocate*) is chosen depending on the base and the intended meaning of the co-occurrence as a whole, rather than on its meaning as an isolated LU.[1]

Currently, available learner error annotations tend to group collocation errors into a single subclass of lexical errors (Aldabe et al., 2005; Nesselhauf, 2005; Martelli, 2006; Granger, 2007; Díaz & García, 2007). A closer look at a learner corpus, in our case, the *Corpus Escrito del Español L2* (CEDEL2) from the Autonomous University of Madrid, [2] immediately reveals, however, that a considerably more detailed collocation error classification is needed in order to offer the learners more targeted (and thus more effective) learning exercises, and to facilitate the development of techniques for automatic correction of collocation errors in writings of the learners. In the scope of the COLOCATE project, we derived a detailed collocation error typology. To facilitate a smooth annotation along this typology, we adapted *Knowtator* (Knublauch et al., 2004) a flexible off-the-shelf annotation tool realized as a Protégé plugin[3]. In what follows, we present the collocation error typology, the basics of the annotation procedure with Knowtator and some preliminary findings derived from our annotation. All collocation (error) examples stem from CEDEL2.

2. Collocation Error Typology

Our collocation error typology distinguishes three parallel dimensions. The first dimension (the "location" dimension) captures whether the error concerns the collocation as a whole or one of its two elements (the *base* or the *collocate*). As in the multilevel-annotation of the Falko-corpus (Lüdeling et al., 2005) and in accordance with, e.g., Tono (2003), the second dimension models the analytical (or linguistic) error analysis, and the third the interpretative (or explanatory) analysis. Each dimension is captured by a typology tree the intermediate nodes of which are error classes and whose leaves stand for concrete error types used as annotation labels in the corpus (in Figures 1 to 3 below given in square brackets). Figure 1 displays the location dimension.

[1] A different definition of the notion of collocation that is not compatible with ours is based on frequency: lexical items that co-occur sufficiently often together form a collocation

[2] CEDEL2 (http://www.uam.es/proyectoinv/woslac/cedel2.htm) has been compiled by the group directed by Amaya Mendikoetxea. It contains about 400.000 words of essays in Spanish on a predefined range of topics by native speakers of English. The essays were written in a web interface; no information is available to us whether bilingual dictionaries or any other reference books were used. The essays are classified with respect to the proficiency level of the authors. The samples underlying our study stem from learners with intermediate or advanced level of knowledge of Spanish.

[3] http://knowtator.sourceforge.net/index.shtml

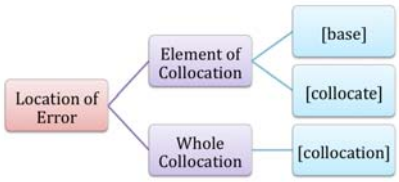

Figure 1: Location dimension of the error typology

In what follows, we first introduce the descriptive dimension and then the explanatory dimension of the error typology.

2.1 Descriptive level of the error typology

At the descriptive level, a collocation can be erroneous from the register, lexical or grammatical perspective; cf. Figure 2. Register errors capture inappropriate use of *per se* correct collocations, as, e.g., *#Yo tengo el deseo personal de ser bilingüe*, lit. 'I have a personal wish to be bilingual'. *Tener* [*un*] *deseo*, lit. 'have [a] wish' is a correct collocation in Spanish. However, it is used in a formal, emotionally charged context, which is not given in the setting of the student.

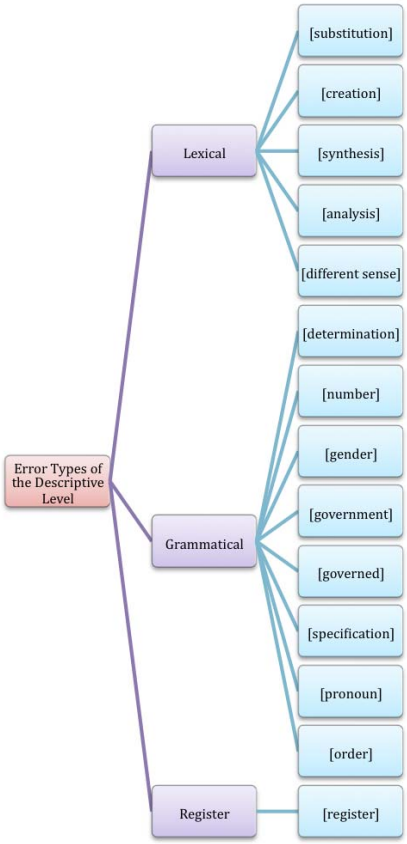

Figure 2: Descriptive dimension of the error typology

The two subclasses of lexical errors that may concern both the collocate and the base are *substitution* and *creation*. Substitution captures the incorrect replacement of a collocate or of a base by another existing word in Spanish, as, e.g., *gastar todo el año*, lit. 'to spend all the year' instead of *pasar todo el año* and *hablar un lenguaje*, lit. 'to speak a language' instead of *hablar una lengua*. *Creation* captures the use of a non-existing word as collocate or as base. Consider, for instance, *derechos mujeriles* 'women's rights' with a non-existent collocate *mujeril* (instead of *derechos de las mujeres*) and *recibí un llamo*, lit. '(I) received a call' with the non-existent base *llamo* (instead of *recibí una llamada*). The erroneous choice of the base can be considered a problem of lexical selection and not of collocation selection. Thus, in the case of *recibí un llamo*, by choosing the non-existent *llamo* instead of *llamada*, the learner shows that she does not know the LU *llamada*. However, since *llamo* makes part of a collocation, we opt to treat such cases as collocation errors. The same applies to substitution discussed above.

The other three subclasses of lexical errors concern collocations as a whole. They are: (a) *synthesis*, when the learner creates a new LU instead of using a collocation; cf. *escaparatear* for *ir de escaparates* 'to go window-shopping'; (b) *analysis*, when the learner creates a new expression with the structure of a collocation instead of using a single LU; cf., *hacer de cotilleos*, lit. 'to make of gossips' for *cotillear* 'to gossip'; (c) *different sense*, when the learner uses a correct Spanish collocation, only that this has a meaning different from the intended one; cf., the use of *el próximo día* 'the next day' to express the meaning of *al día siguiente*. In a narration of the past, only the second one is possible: *al día siguiente/ *el próximo día fuimos a Toledo* 'the next day we went to Toledo'.

Grammatical errors also concern the base, the collocate or the collocation as a whole. Among the base-related errors are those that affect: the determiner (erroneous absence or presence), as, e.g., *tienen el derecho de* 'they have the right of'; the wrong number, as, e.g., *tienen prejuicio* 'they have prejudice'; the wrong gender, as, e.g., *días festivas*, lit. 'celebration days' (holidays); the government, as, e.g., *tengo planes a*, lit. 'I have plans to'; and what we call *specification*. *Specification* stands for cases where an obligatory modifier is missing; cf., for instance, the incorrect *hacer un aterrizaje*, lit. 'to make a landing' vs. *hacer un aterrizaje difícil l /forzoso*, lit. 'to make a difficult / forced landing'.

As for the collocate, errors affect especially the government – such as, e.g., the use of a collocate verb which requires a preposition as a transitive verb: *asisto la universidad*, lit. 'I attend the university' instead of *asisto a la Universidad*, lit. 'I attend to the university'. The *governed* error type captures cases where a wrong preposition is chosen as governor of the collocate or a preposition is used where no preposition is admitted; for instance, *en el año pasado hacía cosas interesantes*, lit. 'in the last year I did interesting things', the correct form of the collocation *año pasado* would be without preposition: *el año pasado hacía cosas interesantes.*

The *pronoun* error type captures cases of an erroneous use / omission of the reflexive pronoun with the verbal collocate; cf., e.g., the incorrect *muero de ganas* 'I am dying for' instead of the correct pronominal form *me muero de ganas*.

The only grammatical error affecting the whole collocation that we identified concerns the order between the base and the collocate; cf., for instance, *amigos mejores*, lit. 'friends best' instead of *mejores amigos*.

2.2 Explanatory level of the typology

Contrary to other authors such as Granger (2007: 467), who limit the learner error typology to the descriptive dimension, we have decided to make the possible error sources explicit, especially in the case of lexical errors. Such an interpretative (or explanatory) dimension of the error typology is of great use in the design of didactic exercises that are supposed to target the individual errors. To allow for more flexibility, we foresee the possibility that more than one source is assigned to a single error (see, for instance, *hacer citas* discussed below).

At the explanatory level, the most generic distinction is between 'interlingual' (or 'L1-L2 transfer') errors and 'intralingual L2' errors.[4] This distinction concerns all three major types of errors introduced at the descriptive level: lexical, grammatical and register; cf. Figure 3. In order to avoid such abstract error class labels as *interlingual error* or *intralingual error*, we prefer the introduction of such class labels as *interlingual lexical error*, *intralingual lexical error*, etc. – even if this might appear to imply a certain redundancy.

As for grammatical and register errors, no further explanatory distinction is made, i.e., *interlingual gram.*, *intralingual gram.*, *interlingual reg.*, and *intralingual reg.* are used as annotation labels. For instance, a grammatical government error can be described as interlingual or as a intralingual. In the first case, we can observe the influence of English; consider, e.g., in *terminé escuela*, lit.'I finished school'. In the second case, the wrong government cannot be straightforwardly attributed to L1, cf., e.g., in *montar el autobús*, lit. 'to mount the bus'. In this case, we consider it intralingual.

In the lexical error branch, the interlingual errors are divided into two subclasses: (a) *importation*: the learner creates in L2 an LU from an LU in L1 or from another language,[5] most often adapting it to the form of L2 (as in *recibir un llamo* 'receive a call'); cf. however also *hicimos wakeboarding* 'make wakeboarding, where no adaptation is made; (b) *extension*: the learner extends the meaning of an existing LU in L2. In many cases of extension, the LU in L1 is a valid translation of the L2 LU, but with a different meaning than the intended one. Consider, e.g., *gastar tiempo* 'spend time' instead of *pasar tiempo*, where *gastar* is chosen because of a possible translation of *spend* as *gastar*. An error of extension is often produced because an L2 LU is used due to its phonetic similarity with the equivalent form in L1; cf. *maternal* in *lengua maternal* 'mother tongue' instead of *materna* (in Figure 3 labelled as *phonetic similarity*) or when the use of an L2 LU is avoided precisely because it seems formally too similar to its L1 equivalent (in Figure 3 labelled as *L1-avoidance*) – what can be considered a case of hypercorrection (cf. *convertirse* in *convertirse al cristianismo* 'to convert to Christianity', which is discarded by the learner in favour of *cambiar*, lit. 'to change': *cambiar al cristianismo* because it appears too similar to the English *to convert*).

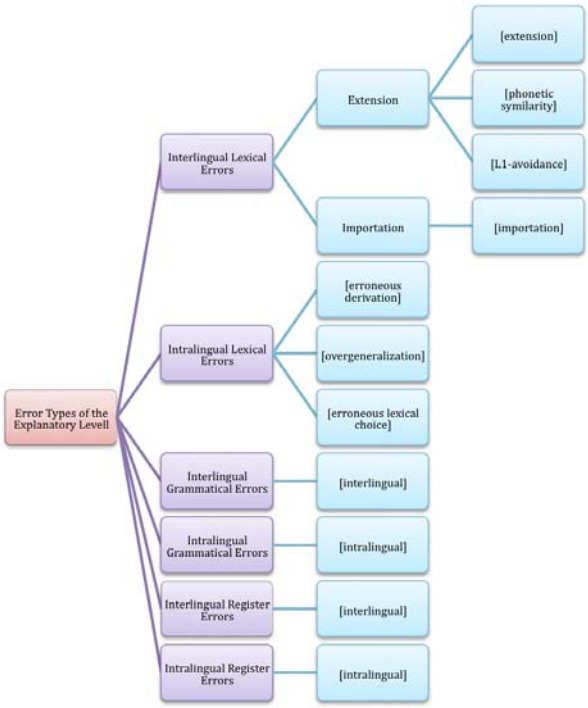

Figure 3: Explanatory dimension of the error typology

The lexical intralingual error class is divided into three subclasses: (a) *erroneous derivation*: the learner produces an inexistent form in L2 as a result of a process of erroneous derivation by analogy with another form in L2 (cf. *enseñanza segundaria*, lit. 'secondary education', instead of *secundaria*); (b) *overgeneralization*: the learner selects a vaguer or a (more) generic LU than required (cf. *hacer citas*,[6] lit. 'to make appointments' instead of *concertar citas* 'to arrange appointments'); (c) *erroneous lexical choice*: the learner selects a wrong LU without a clear reason and without intervention of L1 (cf. *escribir el examen* 'to write the exam', instead of *hacer el examen* 'to make the exam').

3. The corpus annotation tool

For the annotation of the corpus, we use the *Knowtator* annotation tool (Knublauch et al. 2004). Knowtator allows us to define an annotation schema that accommodates for the above error typology and the

[4] In accordance with the terminology in Applied Linguistics, we refer to the mother tongue of the learner as 'L1' and to the language learned as 'L2'.

[5] The meta information in CEDEL2 records all languages spoken by a learner. In some cases, the error could be traced to one of these languages rather than to L1.

[6] This is an example of an error with two possible interpretations since it is unclear whether the learner uses the verb *hacer* due to the influence of L1 (English in this case), as in *I'd like to make an appointment to see the doctor, please*, or simply, because *hacer* often functions as a light verb in Spanish.

(correct) collocation typology as given by Lexical Functions of the Explanatory Combinatorial Lexicology (Mel'čuk, 1996).

Figure 4 illustrates the definition of the annotation schema in Knowtator. The frame at the left hand side displays the general classification of the classes of phenomena we are interested in in the COLOCATE project. One of these classes is Error. Error possesses three slots (cf. the window of the Knowtator's Class Editor in Figure 4), which are the three dimensions of our error typology: 1. 'localization', 2. 'descriptive', and 3. 'explanatory'. The window that superimposes the main Class Editor window in Figure 4 suggests that the types of the errors of each dimension as exemplified in Figures 2 and 3 are defined as possible values of the slots (in Figure 4, the values of the 'explanatory' slot are displayed).

Knowtator also supports the process of annotation. A corpus can be loaded into a Knowtator window. For each detected collocation error, the annotator can choose the appropriate value for each of the three error dimension slots.

As mentioned above, in addition to the annotation of errors, the annotator can also annotate correct collocations encountered in the corpus with Lexical Function information. Figure 5 displays a fragment of the corpus in which both correct (in green) and incorrect (in red) collocations are tagged. The zoom in the figure focuses on the process of tagging the wrong collocation *llenar puestos*, lit. 'to fill positions', instead of *ocupar puestos* with the tag of 'lexical - extension' for the explanatory dimension and with the tag 'substitution' for the descriptive dimension.

4. Findings and conclusions

Although we are still in the process of the annotation of CEDEL2 following the annotation schema presented above, we can already report some interesting observations and draw some conclusions. Over the total of tagged collocations, 61% are correct and 39% are incorrect. Most of the incorrect collocations reveal lexical errors (62%), 33% reveal solely grammatical errors, and 5% contain both lexical and grammatical errors. 54% of the lexical errors concern the collocate, 20% the base, and 26% the whole collocation. The evaluation of the explanatory dimension reveals that 52% of the lexical errors are due to extension, 22% represent erroneous choice, 14% extension due to phonetic similarity, 4% extension due to L1 avoidance, 2% importation, 2% importation/erroneous derivation, 2% extension/ erroneous choice, 2% extension/ overgeneralization. Regarding the grammatical errors, the most significant information is that most of the grammatical errors (41%) concern the government of one of the elements of the collocation.

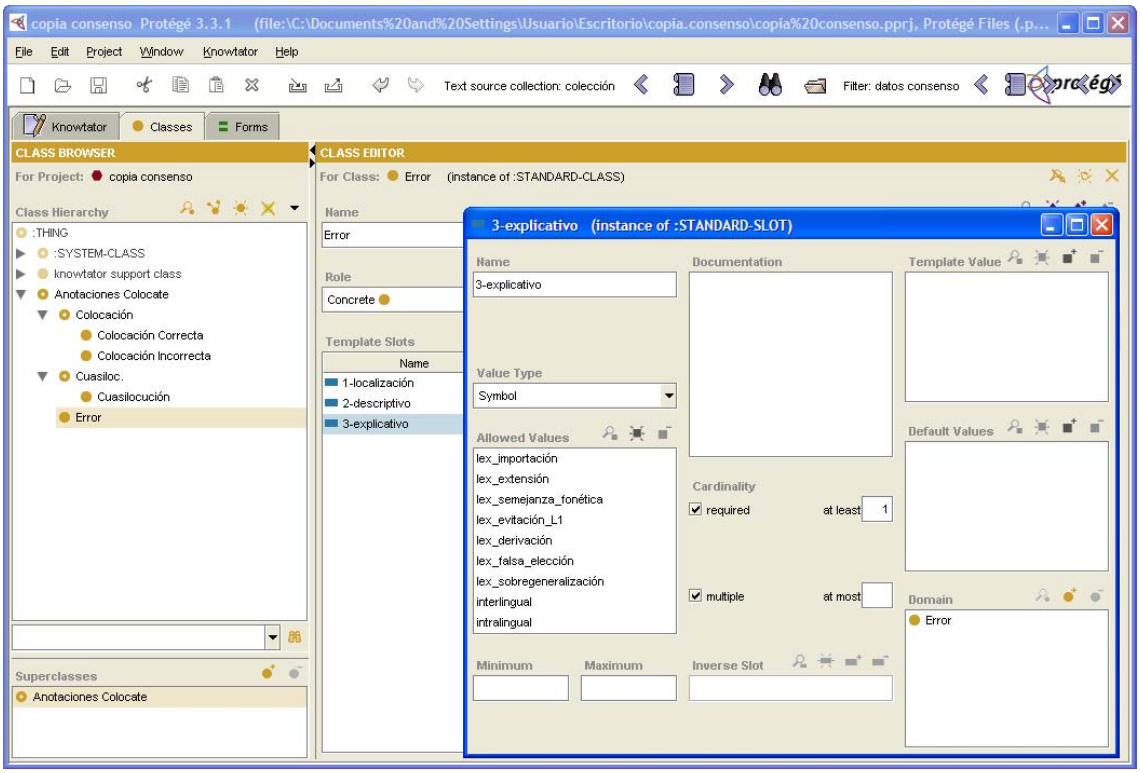

Figure 4: The definition of the annotation schema in Knowtator

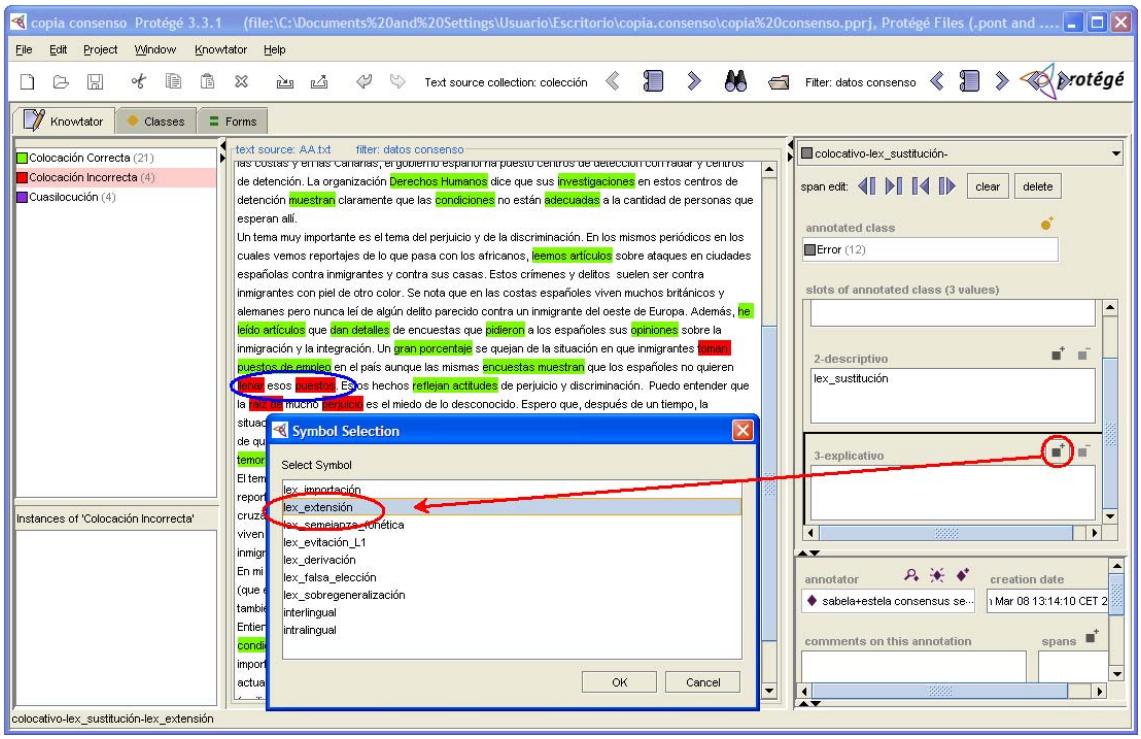

Figure 5: The process of annotation with Knowtator

The distribution of the errors across the different types shows the dominance of lexical errors and, within the lexical error class, of collocate errors. However, the distribution also shows that a considerable share of the errors belongs to the other types within our typology. In other words: the learners do make different kinds of collocation errors. These errors are too different to be simply tagged as "lexical errors". Rather, a sufficiently fine-grained distinction of collocation errors of the kind as offered in this paper is needed in order to provide efficient didactic means for learners. Given that our typology is language-independent, we expect it to be of use for the CALL community in general.

It is worth to be mentioned that the learners could have avoided the majority of the types of collocation errors with an appropriate collocation dictionary at hand. Such dictionaries are already in development for various languages and, in particular, for Spanish; cf. the *Diccionario de Colocaciones del Español* (DiCE), a web-based collocation dictionary of Spanish (http://www.dicesp.com). This dictionary makes use of the typology of Lexical Functions, together with natural language glosses to describe the semantic content of collocates and to provide syntactic information. It consists of two main components: the dictionary itself and the advanced search component that allows the user to make specific queries and to surf the corpus contained in the dictionary (Alonso Ramos, 2006; Alonso Ramos et al., 2010).

As far as the error annotation procedure is concerned, our experience coincides with that of Lüdeling et al. (2005) that the annotation agreement may largely vary, depending on training and background. With the aim to achieve a maximum consensus among the annotators, we have established a methodology to be followed in the process of annotation. The methodology foresees annotators in charge of continuous annotation, consensus annotators and expert annotators. Researchers in charge of the first task annotate the same texts. A consensus annotator verifies the agreements or the disagreements between them. If the disagreement is whether a given expression is a collocation or not, the consensus annotator sends the sample to an expert annotator for a final verdict. If the disagreement concerns the correction of a presumably erroneous collocation, the consensus annotator verifies the expression in question in a reference corpus (in our case: *Corpus de referencia del español actual* http://www.rae.es). In case the information provided by the corpus does not dispel the doubts, the consensus annotator asks three native speakers who have been trained in collocations for a final verdict.

Currently, our annotation team consists of two researchers in charge of continuous annotation, one consensus annotator and one expert annotator.

An aspect we still did not consider so far in our annotation exercise is the degree of acceptability of an erroneous collocation: not all errors are equally unacceptable. While some of them are even hard to understand for a native speaker (as, e.g., *futura cerca*, lit 'future fence' instead of *futuro cercano* 'near future'), others can be "tolerated". Consider, for instance, *cambiar al cristianismo* already cited above. While *convertirse al cristianismo* is better, *cambiar al cristianismo* is transparent enough to be understood. A marker of the degree of acceptability would provide valuable feedback to the learner and help her to focus first on the critical errors. However, an in-depth study is needed to determine the optimal scale of error

acceptability grades. Thus, it is to be investigated whether, for instance, a three grade scale 'bad' – 'improvable' – 'good' is enough or whether a more fine-grained scale is required.

Acknowledgements

The work described in this paper has been carried out in the framework of the project COLOCATE, partially funded under the contract number FFI2008-06479-C02-01 by the Spanish Ministry of Science and Innovation (*MICINN*) and FEDER, European Union.

References

Alonso Ramos, M. (2006). Towards a dynamic way to learn collocations in a second language. In E. Corino, C. Marello, C. Onesti. (Eds.) *Proceedings XII EURALEX International Congress*, Torino, Italia, September 6th–9th 2006. Alessandria: Edizioni Dell'Orso, pp. 909-923.

Alonso Ramos, M., Nishikawa, A. & Vincze, O. (2010). 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, Louvain-la-neuve, Presses universitaires de Louvain.

Aldabe, I., Arrieta, B., Díaz de Ilarraza, A., Maritxalar, M., Oronoz, M. & Uria, L. (2005). Propuesta de una clasificación general y dinámica para la definición de errores. *Revista de Psicodidáctica*, 10/2, pp. 47–60.

Dagneaux, E., Denness, S. & Granger, S. (1998). Computer-aided error analysis. *System*, 26, pp. 163–174.

Díaz-Negrillo, A. & García-Cumbreras, M.A. (2007). A tagging tool for error analysis on learner corpora. *ICAME Journal*, 31/1, pp. 197–203.

Granger, S. (ed.) (1998). *Learner English on computer*. Oxford: Oxford University Press.

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

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

Knublauch, H., R. Fergerson, N. F. Noy & M. A. Musen (2004). The Protégé OWL Plugin: An Open Development Environment for Semantic Web Applications. In *Proceedings of the Third International Semantic Web Conference* Hiroshima, Japan.

Lüdeling, A. et al. (2005), Multi-level error annotation in learner corpora, *Proceedings of the Corpus Linguistics Conference*. Birmingham.

Martelli, A. (2006). A corpus-based description of English lexical collocations used by Italian advanced learners. In E. Corino, C. Marello & C. Onesti (Eds.) *Proceedings XII EURALEX International Congress*, Torino, Italia, September 6th–9th 2006. Alessandria: Edizioni Dell'Orso, pp. 1005–1012.

Mel'čuk, I. (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*. Amsterdam and Philadelphia: John Benjamins, pp. 37–102.

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

Nesselhauf, N. (2005). *Collocations in a learner corpus.* Amsterdam and Philadelphia: John Benjamins.

Tono, Y. (2003). Learner corpora: Design, development and applications. In D. Archer et al. (Eds.) *Proceedings of the Corpus Linguistics 2003*, Lancaster, United Kingdom, March 28th–31th 2003, pp.323–343. Lancaster: Lancaster University, University Centre for Computer Corpus Research on Language.