Twenty Years of Learner Corpus Research

Looking Back, Moving Ahead

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

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

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

Editorial Board
Kate Beeching (University of the West of England, Bristol)
Douglas Biber (Northern Arizona University)
Mireille Bilger (Université de Perpignan)
Benjamin Fagard (Université Paris 3)
Gaëtanelle Gilquin (Université catholique de Louvain)
Stefan Th. Gries (University of California, Santa Barbara)
Hilde Hasselgård (University of Oslo)
Philippe Hillairemann (Université catholique de Louvain)
Diana Lewis (Université de Provence Aix-Marseille 1)
Christian Mair (Universität Freiburg)
Fanny Mounier (Université catholique de Louvain)
Rosamund Moon (University of Birmingham)
Maj-Britt Mosegaard Hansen (University of Manchester)
Joanne Nef van Aesel (Universidad Complutense de Madrid)
Marie-Paule Péry-Woodley (Université Toulouse-Le Mirail)
Paul Rayson (Lancaster University)
Ted Sanders (Utrecht University)
Anne Catherine Simon (Université catholique de Louvain)

Series editors
Liesbeth Degand
Sylviane Granger

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

Published volume

Forthcoming volumes
<table>
<thead>
<tr>
<th>Title</th>
<th>Page</th>
</tr>
</thead>
<tbody>
<tr>
<td>A comparison of learner and native speaker writing in online self-presentations: Pedagogical applications</td>
<td>1</td>
</tr>
<tr>
<td>Katherine ACKERLEY</td>
<td></td>
</tr>
<tr>
<td>Classifying intermediate learner English: A data-driven approach to learner corpora</td>
<td>11</td>
</tr>
<tr>
<td>Theodora ALEXOPOULOU, Helen YANNAKOUADAKIS &amp; Angeliki SALAMOURA</td>
<td></td>
</tr>
<tr>
<td>L1 transfer versus fixed chunks: A learner corpus-based study of L2 German</td>
<td>25</td>
</tr>
<tr>
<td>Margit BRECKLE &amp; Heike ZINSMEISTER</td>
<td></td>
</tr>
<tr>
<td>Native language detection with ‘cheap’ learner corpora</td>
<td>37</td>
</tr>
<tr>
<td>Julian BROOKE &amp; Graeme HIRST</td>
<td></td>
</tr>
<tr>
<td>The Corpus of Academic Learner English (CALE) – A new resource for the study and assessment of advanced language proficiency</td>
<td>49</td>
</tr>
<tr>
<td>Marcus CALLIES &amp; Ekaterina ZAYTSEVA</td>
<td></td>
</tr>
<tr>
<td>Integrating learner corpus data into the assessment of spoken interaction in English in an Italian university context</td>
<td>61</td>
</tr>
<tr>
<td>Erik CASTELLO</td>
<td></td>
</tr>
<tr>
<td>Intonational phrasing as a potential indicator for establishing prosodic learner profiles</td>
<td>75</td>
</tr>
<tr>
<td>Evelyne CAUVIN</td>
<td></td>
</tr>
<tr>
<td>Phrasal verbs in a longitudinal learner corpus: Quantitative findings</td>
<td>89</td>
</tr>
<tr>
<td>Meilin CHEN</td>
<td></td>
</tr>
<tr>
<td>The search for sophisticated language in advanced EFL writing: A longitudinal study</td>
<td>103</td>
</tr>
<tr>
<td>Pieter DE HAAN &amp; Monique VAN DER HAAGEN</td>
<td></td>
</tr>
<tr>
<td>Referential expressions in English learner argumentative writing</td>
<td>117</td>
</tr>
<tr>
<td>Deise P. DUTRA &amp; Tony Berber SARDINHA</td>
<td></td>
</tr>
<tr>
<td>Investigating lexical difficulties of learners in the error-annotated UPF learner translation corpus</td>
<td>129</td>
</tr>
<tr>
<td>Anna ESPUNYA</td>
<td></td>
</tr>
<tr>
<td>Producing an annotated corpus with automatic spelling correction</td>
<td>139</td>
</tr>
<tr>
<td>Michael FLOR &amp; Yoko FUTAGI</td>
<td></td>
</tr>
</tbody>
</table>
If-conditionals in ICLE and the BNC: A success story for teaching or learning? 155
Costas GABRIELATOS

This and that in native and learner English: From typology of use to tagset characterisation 167
Thomas GAILLAT

The Lexicon of Spoken Italian by Foreigners: A study on the acquisition of vocabulary by L2 Italian learners between measures of lexical richness and lexical fields 179
Francesca GALLINA

Learners of English and conversational proficiency 197
Pascale GOUTERAUX

Recurrent formulaic sequences in the speech and writing of the Lithuanian learners of English 211
Jonė GRIGALIŪNENĖ & Rita JUKNEVIČIENĖ

Underuse of syntactic categories in Falco: A case study on modification 223
Hagen HIRSCHMANN, Anke LÜDELING, Ines REHBEIN, Marc REZNICEK & Amir ZELDES

Morphology, lexical priming and second language acquisition: A corpus-study on learner Finnish 235
Jarmo Harri JANTUNEN & Sisko BRUNNI

A beginner French learner corpus 247
Régis KAVECKI

Concessive adverbial clauses in L2 academic writing 263
Elma KERZ

A comparison of spoken and written learner corpora: Analyzing developmental patterns of vocabulary used by Japanese EFL learners 277
Yuichiro KOBAYASHI

Corpus-based error analysis of Korean particles 289
Sun-Hee LEE, Markus DICKINSON & Ross ISRAEL

An LSP learner corpus to help with English radiotelephony teaching 301
Stéphanie LOPEZ, Anne CONDAMINES & Amélie JOsselin-leRAY

Corpus and experimental data: Subjects in second language research 313
Cristóbal LOZANO & Amaya MENDIKOETXEA

Cross-linguistic influence on the accuracy order of L2 English grammatical morphemes 325
Akira MURAKAMI

The TREACLE project: Profiling learner proficiency using error and syntactic analysis 335
Susana MURCIA-BIELSA & Penny MACDONALD

Communication strategies used by Norwegian students of English 345
Susan NACEY & Anne-Line GRAEDLER

The use of articles in Japanese EFL learners’ essays 357
Masumi NARITA

Analysing the use of vague language in spoken interlanguage: A corpus-based study of a group of Brazilian university students learning English as a second language 367
Barbara Malveira ORFANÔ

Writer/reader visibility in learner writing across genres: A comparison of the French and Norwegian components of the ICLE and VESPA learner corpora 377
Magali PAQUOT, Hilde HASSELGÅRD & Signe OKSEFJELL EBELING

On the use of spatial prepositions: Differences in L1 and L2 English 389
Nina RESHÖFT & Linn GRALLA

Connectors in the argumentative writing of Norwegian novice writers 401
Sylvi RØRVIK & Thomas EGAN

Shell noun patterns in student writing in English for specific academic purposes 411
Christine S. SING

The (under)use of partitive objects in Estonian, German and Dutch learners of Finnish 423
Marianne SPOELMAN

A learner corpus of Czech: Current state and future directions 435
Barbora ŠTINDLOVÁ, Svatava ŠKODOVÁ, Alexandra ROSEN & Jirka HANÁ

Inappropriate uses of psychological verbs by Japanese learners of English 447
Misuzu TAKAMI & Naoko AKAHORI

Using a learner corpus to improve distractors in multiple choice grammar questions 455
Hiroko USAMI
Acknowledgements

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

Academic partners

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

Main sponsors

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

Supporting sponsors

- Edinburgh University Press
- Oxford University Press
- Rodopi
- Routledge - Francis & Taylor Group
Annotation of collocations in a learner corpus for building a learning environment

Leo Wanner, Margarita Alonso Ramos, Orsolya Vincze, Rogelio Nazar, Gabriela Ferraro, Estela Mosqueira, Sabela Prieto

Abstract

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

Within the first strand, we carry out a detailed collocation-oriented annotation of a fragment of the corpus of learners of Spanish (CEDEL2). Within the second strand, we experiment with a number of strategies for choosing the most likely correction of a collocation error.

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

1. Introduction

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

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

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

2. Collocation error analysis and corpus annotation

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

2.1. A closer look at collocation errors in learner corpora

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

2.1.1. Analysis of the linguistic nature of collocation errors

In contrast to the impression given by previous studies, errors in the use of collocations do not always have the same scope. In particular automatic collocation error detection and correction proposals seem to implicitly assume that (a) the base is always correct and (b) the collocation with the intended meaning exists in L2, such that the learner can only err in the choice of the collocate. The material in CEDEL2 shows that the learner can err on both elements of the collocation (cf. (1) for wrong collocate uses and (2) for wrong base uses) or on the collocation as a whole (cf. (3)).

1. wrong collocate uses: *empezar una familia*, lit. ‘begin a family’ (instead of formar una familia), lit. ‘form a family’; *tomar una siesta*, lit. ‘take a siesta’ (instead of echar una siesta, lit. ‘throw a siesta’); *hacer errores*, lit. ‘make errors’ (instead of cometer errores, lit. ‘commit errors’).

2. wrong base uses: *tener *limitades*, lit. ‘have *limitades’ (instead of tener limits, lit. ‘have limits’); *conducso breve*, lit. ‘brief *conducso’ (instead of trayecto corto, lit. ‘short distance’); *tener un relacionamento*, lit. ‘have *relacionamento’ (instead of tener una relación, lit. ‘have a relationship’); *policia exterior*, lit. ‘external police’ (instead of politica exterior, lit. ‘external politics’); *hablar francia*, lit. ‘speak France’ (instead of hablar francés, lit. ‘speak French’).

3. wrong collocate uses: *fbrica de carne*, lit. ‘meat factory’ (instead of matadero ‘slaughterhouse’); *sito de acomapar*, lit. ‘location to camp’ (instead of camping); *escaparatear*, lit. ‘go window- shopping’; *hacer pinturas*, lit. ‘make pictures’ (instead of pintar ‘to paint’); *dar respecto ‘command respect’ (instead of tener respeto, lit. ‘have respect’).

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

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

4. determiner: *viene a mente*, lit. ‘comes to mind’ (instead of viene a la mente), *dar el amor*, lit. ‘give the love’ (instead of dar amor)

5. number: *tienen prejuicio*, lit. ‘[they] have prejudice’ (instead of tienen prejucicios)

6. government: *jugamos con tarjetas*, lit. ‘[we] play with cards...’ (instead of jugamos a las cartas)

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

---

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

2 When the erroneous use contains a non-existing word in Spanish (limitades, conducso, and relacionamento), we reproduce it as it is in the literal translation. The last two examples in (3) are examples for the wrong use of pre

3 correct collocations.

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

In general, we can distinguish between ‘interlingual’ (or ‘L1-L2 transfer’) and ‘intralingual L2’ errors. For grammatical errors, this distinction suffices: the error can control on, instead of han tomado control de, lit. ‘they have taken control of’ or not of vacaciones’.

Linguistic errors call for a more detailed identification of their source. In the case of the learner creates ‘receive a call’ or ir de hiking, lit. ‘go of hiking’) and cases where the learner extends without explanation of an L1 in L2. In many cases of such an extension, the L1 in one. Consider, e.g., juego de fútbol ‘game of football’ instead of partido de fútbol, extension is also often produced because an L2 is used due to its phonetic similarity with the equivalent form in L1; cf. maternal in lengua materna ‘mother tongue’ instead of materna, or when the use of an L2 is avoided precisely because hypercorrection (cf. atender el teléfono ‘attend the phone’, which is discarded by the learner in favour of acudir, lit. ‘come’: acudir el teléfono because it appears too similar to the English to attend).

The fact that buena nota ‘good grade’ is a collocation, given that the semantic characteristics of a noun like nota ‘grade’ call for a qualification adjective. This is not so in the case of the combination buena comida ‘good food’, where the meaning of the noun comida ‘food’ does not necessarily require qualification. Consider, however, the combination comida rica ‘delicious food’, where the adjective rica ‘delicious’ has a rather restricted use; it is the adjective prototypically chosen to speak about good food. From our point of view, combinations such as comida rica should be considered collocations, and, consequently, other less idiometic combinations, containing less restricted adjectives appearing with the same noun, such as buena comida ‘good food’ or even comida fantástica ‘fantastic food’ will be considered collocations as well.

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

2.2. Challenges in collocation error analysis and annotation

Collocation is known to be a notoriously difficult language phenomenon. Thus, problematic. The acceptance of a co-occurrence as a correct collocation by speakers of a language that have no collocation equivalent in Spanish. For instance, the erroneous expression Humo de segunda mano corresponds to the English collocation second hand smoke, which can only be translated into Spanish by a complex phrase expressing the same meaning without constituting a phraseological expression: humo del tabaco de otras personas ‘smoke from other people’s tobacco’. In contrast, some expressions used by the learners do not constitute collocations, while the correct form to be should be a collocation in Spanish. An example for this is tengo curiosidad lit. ‘I have curiosity’: *estoy curiosa conocerlo, lit. ‘I am curious to get to know it’, where the expression using the copulative verb and the adjective curioso ‘curious’ should be corrected as a collocation. Thirdly, two coexisting category labels

---

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

2 According to Mel’Eu (2012), idioms represent a major subclass of phraseological multiword expressions: non-compositional multiword expressions.
had to be allowed in the cases where the source of the error could not be determined unambiguously. For instance, in the case of the incorrect collocation ‘hice citas, lit. ‘I made appointments’, the annotators found it feasible to treat the error both as a direct translation from English and as a generalization error, whereby the generic verb ‘hacer’ ‘make/do’ is used instead of the correct and more restricted ‘concertar’ ‘arrange’.

3. Automatic collocation error correction

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

3.1. Insights from the collocation error analysis for CALL

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

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

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

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

3.2. Experiments on the automatic collocation error correction

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

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

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

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

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

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

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

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

8 For the sake of simplification, we use henceforth ‘X+N’ for both ‘X+N’ and ‘N+X’.

9 See Ferraro et al. (2011) for formal details of the metrics.
The affinity metric is a local metric that takes into account the co-occurrence (or association) strength of $C$ with $N$, the graphic similarity of $C$ with $X$, and synonymy of $C$ with $X$. As association strength measure, we use log-likelihood. The graphic similarity (which we calculate as Dice coefficient) captures mistyped collocates or erroneously chosen collocates due to their graphic similarity to the intended one (as, e.g., raise instead of rain in English). The lexical context metric takes into account the context in which $X+N$ occurs (in our case, the corresponding sentence in CEDEL2). It is thus grounded in the assumption that the semantics of a collocation can be approximately deduced from the sentential context in which this collocation appears. More precisely, we assume that given the sentential context $c_1, c_2, \ldots, c_n$ of $X$ in the original sentence of the learner, the candidate $C$ with the highest affinity to $c_1, c_2, \ldots, c_n$ is the most adequate correction of $X$ with “affinity” meaning here the highest co-occurrence frequency. The context feature metric is similar to the lexical context metric in that it draws upon the context of $X+N$ in the original sentence of the student. However, there are also two significant differences: Firstly, it may take into account not only lexical tokens (although this is what we tested it with so far; see below), but also kind of contextual features (POS tags, grammatical functions, punctuation, etc.). Secondly, its interpretation of these features is very different: Given the sentential context $c_1, c_2, \ldots, c_n$ of $X$ in the original sentence of the learner and the candidate collocate $C$, the idea is to assess whether any of the contextual features $c$ of $X$ speaks for the preference of $C$. For this purpose, we find the maximal probability of each feature $c$, given a candidate $C+N$. The suggestions of the three metrics tend to diverge, although common suggestions are also observed. Thus, for the learner’s erroneous concluir ‘conclude’ in *concluir un problema, lit. ‘conclude a problem’ the affinity metric suggests resolver ‘resolve’, the lexical context metric solucionar ‘solve’, and the context feature metric acabar ‘terminate’ (all three suggestions are correct); for realizar in *realizar una meta, lit. ‘realize a goal’, the affinity metric suggests an erroneous *hacer ‘make’, while the other two propose the correct alcanzar ‘reach’; and for cambiar in *cambiar al cristianismo, lit. ‘change to Christianity’, all three metrics suggest the correct convertir ‘convert’.

A preliminary evaluation demonstrated that our procedure is able to judge whether a combination is a correct or an incorrect collocation in Spanish with an accuracy of 0.90. When the procedure failed, it tended to judge a correct collocation as incorrect. This is due to our purely frequency-based collocation criteria, which need to be improved. For the error correction stage, we evaluated first the accuracy with which we are able to provide lists of collocations containing the right correction. This accuracy amounts to 0.75 for the context feature metric; the other two metrics have lower accuracies. The mean reciprocal rank (MRR) of the top five suggestions with the contextual feature metric is 0.72 (to compare: Wu et al. (2010) achieved on their experimental data an MRR of 0.518). Then, we evaluated the capacity of our procedure to offer the right correction using the context feature metrics, with features being simply words in the original sentence of the learner. The accuracy was 0.542.

4. Conclusion

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

Acknowledgements

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

References


