Towards a learner need-oriented second language collocation writing assistant

Margarita Alonso Ramos[1], Roberto Carlini[2], Joan Codina-Filbà[3], Ana Orol[4], Orsolya Vincze[5], and Leo Wanner[6]

Abstract. The importance of collocations, i.e. idiosyncratic binary word co-occurrences in the context of second language learning has been repeatedly emphasized by scholars working in the field. Some went even so far as to argue that "vocabulary learning is collocation learning" (Hausmann, 1984, p. 395). Empirical studies confirm this argumentation. They show that the "collocation density" in learner corpora is nearly the same as in native corpora, i.e. that the use of collocations by learners is as common as it is by native speakers. At the same time, they also find that the collocation error rate in learner corpora is about 32% (compared to about 3% by native speakers). A CALL-based collocation writing aid could help learners to better master collocations. However, surprisingly little work has been done so far on collocation learning assistants. We propose a collocation writing assistant for American English learners of Spanish which may be used as checker of both isolated collocations and collocations in texts. In addition, it offers the possibility to actively explore and administer collocation resources.

Keywords: second language learning, writing assistant, collocations, miscollocation correction.

[1] Universidade da Coruña, Spain; lxalonso@udc.es

[2] Universitat Pompeu Fabra, Spain; roberto.carlini@upf.edu

[3] Universitat Pompeu Fabra, Spain; joan.codina@upf.edu

[4] Universidade da Coruña, Spain; ana.orol.gonzalez@udc.es

[5] Universidade da Coruña, Spain; ovincze@udc.es

[6] ICREA and Universitat Pompeu Fabra, Spain; leo.wanner@upf.edu

How to cite this article: Alonso Ramos, M., Carlini, R., Codina-Filbà, J., Orol, A., Vincze, O., & Wanner, L. (2015). Towards a learner need-oriented second language collocation writing assistant. In F. Helm, L. Bradley, M. Guarda, & S. Thouësny (Eds), *Critical CALL – Proceedings of the 2015 EUROCALL Conference, Padova, Italy* (pp. 16-23). Dublin: Research-publishing.net. http://dx.doi.org/10.14705/rpnet.2015.000304

1. Introduction

The importance of collocations, i.e. idiosyncratic binary word co-occurrences of the type "*hold* [a] *lecture*", "*give* [*a*] *hint*", "*pass exam*", "*blue skies*", "*overwhelming success*", etc. in the context of second language learning is well known (Granger, 1998; Lewis, 2000; Nesselhauf, 2005). Hausmann (1984) went so far as to argue that "vocabulary learning *IS* collocation learning" (p. 395, our emphasis). Empirical studies confirm this argumentation. According to a study by Orol and Alonso Ramos (2013), the "collocation density" in learner corpora is nearly the same as in native corpora, i.e. the use of collocations by learners is as common as it is by native speakers.

At the same time, the study finds that the collocation error rate in learner corpora is 32% (compared to 3% in native corpora). A CALL-based collocation writing assistant could help learners to better master collocations. However, surprisingly few works address CALL-oriented collocation learning assistants. Most of the existing assistants are limited to the assessment of the correctness of isolated collocations, lists of collocations extracted from a corpus with one of the elements of an assumed miscollocation, and examples of the use of a specific collocation. None of them target the identification and correction of miscollocations in the writings of language learners (as, e.g. spell and grammar checkers do), and none of them follow the active learning paradigm that assigns the learner an active role during learning. We aim to advance the state of the art in this area.

In what follows, we present the HARenES[7] collocation writing assistant, designed to support American English learners of Spanish. The assistant can be used as checker of both isolated collocations and collocations in texts. In addition, it also offers the possibility to actively administer personal collocation resources (e.g. collocation dictionaries, lists of collocations grouped in accordance with specific user criteria, etc).

In Section 2, we first clarify the notion of collocation that underlies the design of the HARenES assistant and then discuss the needs of a learner. Section 3 describes HARenES' functionality at its current state of development. In Section 4, some conclusions are drawn and possible extensions and ameliorations of HARenES are outlined.

[7] "HARenES" stands for the title of the corresponding research project *Herramienta de ayuda a la redacción en español: procesamiento de colocaciones* 'Support tool for writing in Spanish: Processing of collocations'.

2. What can a learner expect from a collocation-oriented writing assistant?

Given that "collocation" is an ambiguous term in lexicography and computational linguistics, it seems appropriate to offer an exact definition of its use in the given context.

2.1. On the notion of collocation

The term "collocation" as introduced by Firth (1957), and cast into a definition by Halliday (1961), encompasses the statistical distribution of lexical items in context: lexical items that form high probability associations are considered collocations. However, in contemporary lexicography and lexicology, an interpretation that stresses the idiosyncratic nature of collocations prevails. According to Cowie (1994), Hausmann (1984), Mel'čuk (1995) and others, a collocation is a binary idiosyncratic co-occurrence of lexical items between which a direct syntactic dependency holds and where the occurrence of one of the items (the 'base') is subject of free choice by the speaker, while the occurrence of the other item (the 'collocate') is restricted by the base. Thus, in the case of *take* [*a*] *walk*, *walk* is the base and *take* is the collocate, in the case of *high speed*, *speed* is the base and *high* the collocate. It is this notion of "collocation" that we find reflected in general collocation dictionaries and that we follow in the design of the HARenES collocation writing assistant.

2.2. Collocations in the context of language learning

As already pointed out above, numerous studies in the context of second language (L2) acquisition demonstrate that collocations in L2 pose a great challenge to any language learner: the fact that they are idiosyncratic implies that, in general[8], they cannot be learned by analogy (as grammatical constructions can be). Native speakers can use their intuition, but learners must learn an overwhelming share of them by heart. Most often, a miscollocation by a learner is a calque from L1 (as, e.g. Sp. **tomar* [*un*] *paseo*, lit. 'take [a] walk' instead of *dar* [*un*] *paseo* 'give a walk'), an incorrect construction by analogy (as, e.g. **dar* [*un*] *camino* (as *dar* [*un*] *paseo*), lit. 'give [a] path' instead of *tomar* [*un*] *camino* 'take [a] path'), or an

[8] This does not mean that there are absolutely no regularities between collocations (Mel'cuk & Wanner, 1996); for instance, we "give" a "presentation", as we do in the case of "lecture", "talk", etc.: *give* [*a*] *presentation*, *give* [*a*] *lecture*, *give* [*a*] *talk*, *give* [*an*] *outline*, … However, these regularities are, as a rule, subtle and to a large extent unpredictable.

incorrect construction used to avoid a mistrusted literal translation from L1 (as, e.g. **hacer* [*una*] *charla*, lit. 'make a talk', instead of *dar* [*una*] *charla*, lit. 'give a talk').

To address these problems, a CALL application should allow the learner to at least:

- verify whether a specific word co-occurrence is a correct collocation in L2 and if it is not, provide alternative suggestions;
- solicit examples of the use of a given collocation in context, i.e. sample sentences in which the collocation occurs;
- explore the collocation space of L2, e.g. which other collocates a given base occurs with and how prominent these collocates are; which other bases share the same collocate(s) as a given base; etc.;
- request the verification / correction of collocations in a given writing;
- administer personal collocation repositories such as collocation dictionaries, created and maintained for better memorization or other purposes.

The HARenES collocation writing assistant attempts to account for these needs.

3. HARenES collocation writing assistant

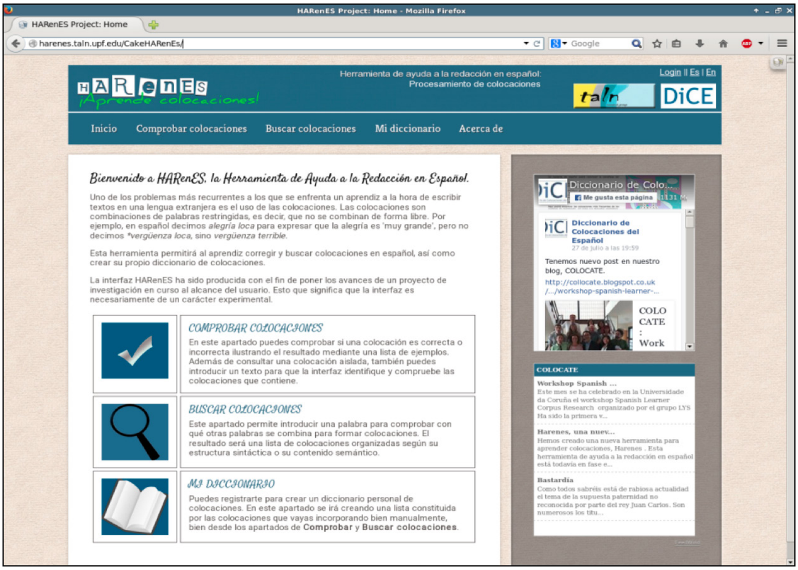
The on-line HARenES collocation writing assistant has been designed to support second language learners in their needs of the kind sketched above. From the linguistic perspective, these needs can be grouped around isolated collocations, collocations in texts and personalized collocation repositories. In the outline of the webpage of the assistant, the offered functions have been clustered into "collocation checker", "collocation search engine", and "personal collocation dictionary". Figure 1 shows the main page of the HARenES assistant.

In what follows, the three functional modules are presented in separate subsections.

3.1. Treating isolated collocations in HARenES

The basic functionality of the HARenES assistant with respect to isolated collocations concerns the validation of a given word co-occurrence; cf. the corresponding part of the main page in Figure 1. If the co-occurrence is a valid

Figure 1. Main page of the web-based HARenES collocation writing assistant



Bienvenido a HARenES, la Herramienta de Ayuda a la Redacción en Español.

Uno de los problemas más recurrentes a los que se enfrenta un aprendiz a la hora de escribir textos en una lengua extranjera es el uso de las colocaciones. Las colocaciones son combinaciones de palabras restringidas, es decir, que no se combinan de forma libre. Por ejemplo, en español decimos *alegría loca* para expresar que la alegría es 'muy grande', pero no decimos **vergüenza loca*, sino *vergüenza terrible*.

Esta herramienta permitirá al aprendiz corregir y buscar colocaciones en español, así como crear su propio diccionario de colocaciones.

La interfaz HARenES ha sido producida con el fin de poner los avances de un proyecto de investigación en curso al alcance del usuario. Esto que significa que la interfaz es necesariamente de un carácter experimental.

COMPROBAR COLOCACIONES

En este apartado puedes comprobar si una colocación es correcta o incorrecta ilustrando el resultado mediante una lista de ejemplos. Además de consultar una colocación aislada, también puedes introducir un texto para que la interfaz identifique y compruebe las colocaciones que contiene.

BUSCAR COLOCACIONES

Este apartado permite introducir una palabra para comprobar con qué otras palabras se combina para formar colocaciones. El resultado será una lista de colocaciones organizadas según su estructura sintáctica o su contenido semántico.

MI DICCIONARIO

Puedes registrarte para crear un diccionario personal de colocaciones. En este apartado se irá creando una lista constituida por las colocaciones que vayas incorporando bien manualmente, bien desde los apartados de **Comprobar** y **Buscar colocaciones**.

Diccionario de Colo...
Me gusta esta página

Diccionario de Colocaciones del Español
27 de julio a las 19:59

Tenemos nuevo post en nuestro blog, COLOCATE.
http://collocate.blogspot.co.uk/.../workshop-spanish-learner-...

COLOCATE: Work

COLOCATE

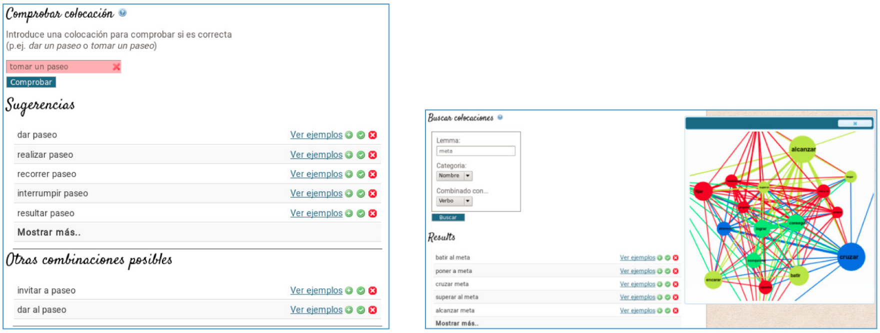
Workshop Spanish ...
Este mes se ha celebrado en la Universidade da Coruña el workshop Spanish Learner Corpus Research organizado por el grupo LYS Ha sido la primera v...

Harenes, una nuev...
Hemos creado una nueva herramienta para aprender colocaciones, Harenes . Esta herramienta de ayuda a la redacción en español está todavía en fase e...

Bastardia
Como todos sabréis está de rabiosa actualidad el tema de la supuesta paternidad no reconocida por parte del rey Juan Carlos. Son numerosos los titu...

collocation in Spanish, it is flagged in green. If it is considered incorrect, alternative suggestions are proposed; see Carlini, Codina-Filbà, & Wanner (2014) for the collocation validation metrics and the metrics used to select and order alternative suggestions retrieved from a reference corpus. The left snapshot in Figure 2 illustrates this feature. The user may also solicit the illustration of the use of a given collocation in context. To support this feature, HARenES displays some sample sentences (the user decides how many they want to see) in which this collocation occurs from the reference corpus.

Figure 2. Validation of a word co-occurrence and exploration of the collocation space



Comprobar colocación

Introduce una colocación para comprobar si es correcta
(p.ej. *dar un paseo* o *tomar un paseo*)

tomar un paseo ✗

Comprobar

Sugerencias

dar paseo Ver ejemplos
realizar paseo Ver ejemplos
recorrer paseo Ver ejemplos
interrumpir paseo Ver ejemplos
resultar paseo Ver ejemplos
Mostrar más..

Otras combinaciones posibles

invitar a paseo Ver ejemplos
dar al paseo Ver ejemplos

Buscar colocaciones

Lemma: meta
Categoria: Nombre
Combinado con... Verbo
Buscar

Results

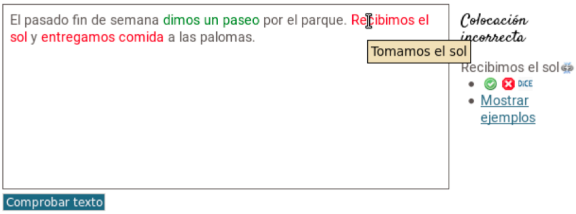
batir al meta Ver ejemplos
poner a meta Ver ejemplos
cruzar meta Ver ejemplos
superar al meta Ver ejemplos
alcanzar meta Ver ejemplos
Mostrar más..

An additional innovative feature of the HARenES writing assistant is that it uses Visual Analytics (VA) techniques to facilitate interactive exploration of the collocation space (Carlini, Codina-Filbà, & Wanner, 2015). Optimal collocation learning is active learning, and active learning is closely related to exploration. For instance, the user may want to explore all collocations of a given base and their intensity of use, contrast the collocation spaces of two different bases, or see the bases of a given collocate lexeme. The right snapshot in Figure 2 shows an example of the use of VA in HARenES.

3.2. Validation of collocations in the writings of learners

An important feature of any advanced learner supporting assistant is that it is able to analyze the learners' writings and correct them with respect to the targeted phenomena. For HARenES, these phenomena are collocations; see Figure 3. The collocation in green (*dar un paseo* 'take a walk') has been detected as correct. Two collocations in red have been detected as erroneous (**recibir sol*, lit. 'receive sun' and *entregar comida*, lit. 'deliver food'). When the user passes with the mouse over a miscollocation, one or several correction suggestions are displayed.

Figure 3. Correction of collocations in learner writings



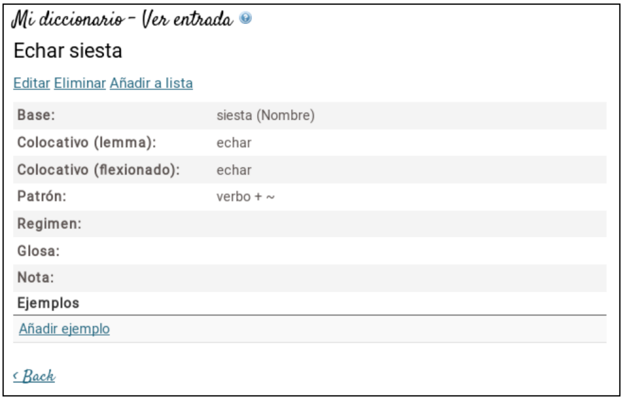
El pasado fin de semana dimos un paseo por el parque. Recibimos el sol y entregamos comida a las palomas.

3.3. Maintenance of personal collocation resources

Any learner who actively learns a language compiles lists of collocations that appear especially important, or difficult to them, writes down examples, takes notes, etc. To account for this need, HARenES offers the possibility to actively administer personal collocation resources such as collocation dictionaries (to store collocations, miscollocations and their corrections, examples of collocation use, personal notes on individual (mis)collocations, etc.), lists of collocations grouped in accordance with specific user criteria, etc. Internally, the organization of the database of these personalized collocation repositories is identical to that of the Spanish collocation dictionary DiCE (Alonso Ramos, Nishikawa,

& Vincze, 2010) – which facilitates the import of collocations and example sentences from there. See Figure 4 for an example of a fragment of a personal collocation dictionary.

Figure 4. Maintenance of personal collocation repositories of the learner



Mi diccionario - Ver entrada

Echar siesta

Base:	siesta (Nombre)
Colocativo (lemma):	echar
Colocativo (flexionado):	echar
Patrón:	verbo + ~
Regimen:	
Glosa:	
Nota:	

Ejemplos

4. Conclusions

We presented the collocation writing assistant HARenES, which supports a language learner by a verification of isolated collocations and collocations in text, provision of correction suggestions for miscollocations, illustration of the use of collocations in context, interactive exploration of collocation spaces and maintenance of personalized collocation repositories.

Some of the features of the current HARenES assistant are about to be further improved. For instance, the ordering of the correction suggestions is still not optimal and the design of the VA module still needs to undergo a revision by a professional interaction designer. Furthermore, the assistant should be further extended by some central features, such as grouping of collocations with respect to a semantic typology and classification of miscollocations encountered in the learner's writing.

5. Acknowledgements

The presented work has been partially funded by the Spanish Ministry of Economy and Competitiveness (MINECO) under the contract numbers FFI2011-30219-C02-01/02.

References

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* (pp. 369-374). Louvain-la-Neuve: Presses universitaires de Louvain.

Carlini, R., Codina-Filbà, J., & Wanner, L. (2014). Improving collocation correction by ranking suggestions using linguistic knowledge. *Proceedings of the 3rd Workshop on NLP for computer-assisted language learning, Uppsala, Sweden*.

Carlini, R., Codina-Filbà, J., & Wanner, L. (2015). Improving the use of electronic collocation resources by visual analytics techniques. *Proceedings of the eLex 2015 Conference*. Herstmonceux Castle.

Cowie, A. (1994). Phraseology. In R. Asher & J. Simpson (Eds.), *The Encyclopedia of language and linguistics, Vol. 6* (pp. 3168-3171). Oxford: Pergamon.

Firth, J. (1957). Modes of meaning. In J. Firth (Ed.), *Papers in linguistics, 1934–1951* (pp. 190-215). Oxford: Oxford University Press.

Granger, S. (1998). Prefabricated patterns in advanced EFL writing: collocations and formulae. In A. Cowie (Ed.), *Phraseology: theory, analysis and applications* (pp. 145-160). Oxford: Oxford University Press.

Halliday, M. A. K. (1961). Categories of the theory of grammar. *Word, 17*, 241-292.

Hausmann, F.-J. (1984). Wortschatzlernen ist Kollokationslernen. Zum Lehren und Lernen französischer Wortwendungen. *Praxis des neusprachlichen Unterrichts, 31*(1), 395-406.

Lewis, M. (2000). *Teaching collocation. Further developments in the lexical approach*. London: LTP.

Mel'čuk, I. (1995). Phrasemes in language and phraseology in linguistics. In M. Everaert, E.-J. van der Linden, A. Schenk, & R. Schreuder (Eds.), *Idioms: structural and psychological perspectives* (pp. 167-232). Hillsdale: Lawrence Erlbaum Associates.

Mel'cuk, I. A., & Wanner, L. (1996). Lexical functions and lexical inheritance for emotion lexemes in German. In L. Wanner (Ed.), *Lexical functions in lexicography and natural language processing* (pp. 209-278). Amsterdam: Benjamins Academic Publishers.

Nesselhauf, N. (2005). *Collocations in a learner corpus*. Amsterdam: Benjamins Academic Publishers. doi:10.1075/scl.14

Orol, A., & Alonso Ramos, M. (2013). A comparative study of collocations in a native corpus and a learner corpus of Spanish. *Procedia–Social and Behavioural Sciences, 96*, 563-570.

Research-publishing.net

Published by Research-publishing.net, not-for-profit association
Dublin, Ireland; info@research-publishing.net

Critical CALL – Proceedings of the 2015 EUROCALL Conference, Padova, Italy
Edited by Francesca Helm, Linda Bradley, Marta Guarda, and Sylvie Thouësny

Disclaimer: Research-publishing.net does not take any responsibility for the content of the pages written by the authors of this book. The authors have recognised that the work described was not published before, or that it is not under consideration for publication elsewhere. While the information in this book are believed to be true and accurate on the date of its going to press, neither the editorial team, nor the publisher can accept any legal responsibility for any errors or omissions that may be made. The publisher makes no warranty, expressed or implied, with respect to the material contained herein. While Research-publishing.net is committed to publishing works of integrity, the words are the authors' alone.

Trademark notice: product or corporate names may be trademarks or registered trademarks, and are used only for identification and explanation without intent to infringe.

Copyrighted material: every effort has been made by the editorial team to trace copyright holders and to obtain their permission for the use of copyrighted material in this book. In the event of errors or omissions, please notify the publisher of any corrections that will need to be incorporated in future editions of this book.

Typeset by Research-publishing.net
Fonts used are licensed under a SIL Open Font License

ISBN13: 978-1-908416-28-5 (Paperback - Print on demand, black and white)
Print on demand technology is a high-quality, innovative and ecological printing method; with which the book is never 'out of stock' or 'out of print'.

ISBN13: 978-1-908416-29-2 (Ebook, PDF, colour)
ISBN13: 978-1-908416-30-8 (Ebook, EPUB, colour)

Legal deposit, Ireland: The National Library of Ireland, The Library of Trinity College, The Library of the University of Limerick, The Library of Dublin City University, The Library of NUI Cork, The Library of NUI Maynooth, The Library of University College Dublin, The Library of NUI Galway.

Legal deposit, United Kingdom: The British Library.
British Library Cataloguing-in-Publication Data.
A cataloguing record for this book is available from the British Library.

Legal deposit, France: Bibliothèque Nationale de France - Dépôt légal: décembre 2015.