A Lexical Tool for Academic Writing in Spanish based on Expert and Novice Corpora

Marcos García Salido, Marcos Garcia, Milka Villayandre, Margarita Alonso-Ramos
Universidade da Coruña, Universidad de León
Facultade de Filoloxía, Rúa Lisboa 7, Campus da Zapateira, 15008 A Coruña, Spain
Facultad de Filosofía y Letras, Campus de Vegazana s/n, 24071 León, Spain
{marcos.garcias, marcos.garcia.gonzalez, margarita.alonso}@udc.gal
milka.villayandre@unileon.es

Abstract

The object of this article is to describe the extraction of data from a corpus of academic texts in Spanish and the use of those data for developing a lexical tool oriented to the production of academic texts. The corpus provides the lexical combinations that will be included in the afore-mentioned tool, namely collocations, idioms and formulas. They have been retrieved from the corpus controlling for their keyness (i.e., their specificity with regard to academic texts) and their even distribution across the corpus. For the extraction of collocations containing academic vocabulary other methods have been used, taking advantage of the morphological and syntactic information with which the corpus has been enriched. In the case of collocations and other multiword units, several association measures are being tested in order to restrict the list of candidates the lexicographers will have to deal with manually.

Keywords: academic writing, corpus, writing aid, multi-word units, collocations

1. Introduction

One of the challenges faced by university students is producing texts written following the conventions of academic discourse. Writers of English, Dutch or French – to cite a few European languages – have at their disposal several production-oriented tools (Kübler and Pecman, 2012; D'Hertefelt et al., 2014; Granger and Paquot, 2015). Writers of Spanish can find several guidelines dealing mostly with structural features of academic texts, but, to the best of our knowledge, there is no resource (on paper or in electronic format) where they can search for lexical combinations in order to compose academic texts.

This paper presents a tool oriented to the production of academic texts in Spanish – the *Herramienta de Ayuda a la Redacción de textos Académicos*, henceforth, HARTA – based on a corpus of academic texts. This tool will mainly provide indications as to how to use vocabulary typical of the academic genre in order to build complete texts. More particularly, the tool will focus on *academic lexical combinations* (ALCs). By ALCs we mean recurrent sequences of words that may or may not be semantically compositional and fulfill rhetorical functions such as giving examples, introducing conclusions, expressing certainty or probability, etc.

Three types of ALCs will form the dictionary module: *collocations, idioms* and *formulas*. The ACLs occurring in the corpus will be classified into these three categories and treated accordingly in the tool, following the criteria proposed by Mel'čuk (2015) within the framework of Explanatory and Combinatorial Lexicology. *Idioms* are non compositional multiword combinations, such as *punto de vista* 'perspective', or *sin embargo* 'however', which will have their own entries along with monolexemic lexical units. *Collocations* are compositional combinations of two lexical units, one of which – the collocate – is chosen depending on the other – the base, such as *formular (una) hipótesis* 'to formulate a hypothesis' or *adoptar un punto de vista* 'to adopt a perspective'. Collocations will be provided in the entry of their bases (e.g., *formular* under the headword *hipótesis* and *adoptar* under the headword *punto de vista*). Finally, *formulas* are compositional combinations for which neither their meanings nor their encoding are freely selected. For instance, if speakers want to express the idea «Now I will rephrase what has been said before» in Spanish, they are not free to choose the meaning 'to put', in contrast to Eng. *to put it differently*, nor can they encode the meaning of 'differently' other than *de otra manera* or *de otro modo* (cf. *dicho de otra manera/*diferentemente/??de manera diferente*). Formulas will be given entries specifying their discourse function: e.g. *dicho de otra manera*: 'used to rephrase a previously introduce idea, argument, etc.' (for more details the reader is referred to Alonso-Ramos et al. (2017)).

Along with the expert subcorpus from which the ALCs which will be included in HARTA have been extracted, a novice writers subcorpus is being compiled containing texts of students in bachelor and master degrees. In what follows, we offer a description of both subcorpora, present some of the results obtained from them, and give an account of their incorporation in HARTA.

2. Corpus description

The HARTA Corpus consists of two subcorpora. The first subcorpus is devoted to the written production of experts and is made up of research articles published in scientific journals and originally written in Spanish. The core of this subcorpus (234 research articles) stems from the Spanish part of the SERAC 2.0 corpus (InterLAE Research Group, 2008). This core has been supplemented with 180 further articles in order to obtain four balanced subsections in terms of their size (see below).

The other subcorpus covers the production of novice writers, and is similar in conception to BAWE (Gardner and

Nesi, 2013) or CALE (Callies and Zaytseva, 2013), among others. It is a collection of bachelor and master degree theses publicly available from institutional repositories. Currently, 125 texts have been incorporated to this subcorpus and xml-marked, 77 bachelor's degree and 48 master's degree theses, amounting to a total of ca. 1.5 million running words.

After completing the incorporation of novice texts, each subcorpus will contain a total of ca. two million words. Consequently, the completed corpus will be similar in size to other corpora exploited for the creation of lexical resources in languages such as English (Coxhead, 2000; Paquot, 2010) or French (Tutin and Kraif, 2016).

2.1. Scientific domains

Both subcorpora are equally divided into four thematic sections: (i) Arts and Humanities, (ii) Biological and Health Sciences, (iii) Physical Science and Engineering, and (iv) Social Sciences and Education. Although there are standards with regard to the classification of scientific fields (UNESCO, 1978), these standards do not seem to have gained acceptance when it comes to compile corpora of academic discourse. Thus, the French SCIENTEXT corpus is divided in ten scientific domains (Tutin and Kraif, 2016), Coxhead (2000) distinguishes four big scientific areas (Arts, Commerce, Law, and Science) in turn subdivided into 27 subareas, Paquot (2010) draws a big divide between hard and soft sciences and further breaks down the former into four sections and the latter into six, etc.

The four thematic areas in both HARTA subcorpora are balanced for size as measured in number of words, like in the case of Coxhead (2000) or Paquot (2010). In this respect, the Spanish part of SERAC 2.0 had to be modified, since it was balanced with respect to the number of texts, rather than the number of words per domain. In its final versions, each subcorpus will include ca. 500,000 words for scientific domain.

2.2. Markup

HARTA corpus is xml marked. This markup makes explicit (a) editorial metadata of the texts included (author, year, article title, journal title, and, in the case of novice writers, university and degree – bachelor vs master) and (b) descriptive metadata relative to the scientific domains and textual sections – introduction, body, etc. – (see Figure 1).

As for text structure markup, all texts must include an obligatory *body* section and optional peripheral sections or subdivisions such as *introduction, methods, conclusion*, and *footnotes*. Even though research articles' structure is quite standard (cf. Swales and Feak (2004), among others), we have decided to keep most textual sections as optional, given the relatively loose structural conventions of some domains (e.g. articles on literature). The information provided by this kind of markup will enable sophisticated searches both to users of HARTA and to researchers, so that they can restrict their queries to specific domains or to particular sections (e.g., abstracts, conclusion sections, etc.).

```
<HARTA>
    <header>
        <article title="" author=""></article>
        <deposit year="" university=""/>
        <class type="" domain="" subdomain=""/>
        <size words=""/>
    </header>
    <content>
        <abstract><p>...</p></abstract>
        <introduction><p>...</p></introduction>
        <body><p>...<i>...</i>...
                <b>...</b>...</p></body>
        <conclusion><p>...</p></conclusion>
        <footnotes><p>...</p></footnotes>
    </content>
</HARTA>
```

Figure 1: HARTA's xml-markup

2.3. Morphological annotation and parsing

The expert corpus has been tokenized and lemmatized with LinguaKit (Garcia and Gamallo, 2016) and PoS-tagged with FreeLing (Padró and Stanilovsky, 2012). Subsequently, we have used UDPipe (Straka et al., 2016) to perform dependency parsing using universal dependencies (Nivre et al., 2016).

3. Preliminary findings

The corpus described in the previous section will feed HARTA with lexical units and combinations thereof retrieved from it. Such data will constitute the raw materials of which HARTA will be made after a manual revision on the part of lexicographers. The information extracted from the corpus is being or will be pre-processed by means of different techniques before handing it to the lexicographers. In this section we give an account of such pre-processing techniques.

3.1. Academic Word List

An Academic Word List (AWL) has been extracted from the HARTA expert subcorpus. The items of the list fulfill two requirements: (i) being specific or "key" to the academic corpus and (ii) having an even distribution through the whole expert subcorpus in order to discard discipline specific terms. The *keyness* or specifitiy of the list has been established by comparing the distribution of the lemmas corresponding to content words present in the HARTA expert subcorpus with their distribution in a non-academic corpus. Following Paquot (2010), we have used a corpus of fiction narrative as a strongly contrasting reference corpus: the fiction narrative part of LEXESP (Sebastián-Gallés et al., 2000). To determine whether each lemma was significantly more frequent in the academic corpus we used the Wilcoxon-Mann-Whitney (WMW) rank test (Paquot and Bestgen, 2009; Lijffijt et al., 2014) and discarded the pairs yielding a p-value equal or greater than 0.001. In order to apply this test, each lemma has been assigned a series of ranks derived from their frequency in the sections of the academic and fiction corpora. As sections of the academic corpus we used its very division in articles and we divided the fiction corpus in fragments of 5,000 words. Subsequently we obtained the frequency of each lemma in each

section, normalised it per 5,000 words and transformed it into the ranks that fed the WMW test.
Although the test results ultimately derive from information relative to counts in corpus sections, it is not clear whether it is sensitive to evenness of distribution – see Paquot and Bestgen (2009), who attribute this quality only to the t-test. For that reason, we also used Gries's Deviation of Proportions (DP) (Gries, 2008), a coefficient indicative of the evenness of distribution of the elements of a corpus. DP values near 0 correspond to the absence of differences with respect to the expected distribution, whilst values close to 1 are suggestive of highly skewed distributions. For our AWL we have kept lemmas with values under 0.5.
The resulting academic list contains 1080 lemmas of content words, i.e. nouns, adjectives, verbs and adverbs. The breakdown of the list into these four parts of speech can be seen in Table 1.

Part of Speech	No. of instances
noun	333
adjective	235
verb	384
adverb	128

Table 1: AWL breakdown by part-of-speech

This list contains the potential candidates for the collocation bases that will be included in HARTA. The final decision, however, will be made after a manual exam by expert lexicographers. Predictably, not all four categories will be equally productive in this respect. Thus, nouns are much more interesting than the other three categories as collocational bases.

3.2. Formulas and uninflected idioms list

Some ALCs tend to occur as invariable strings. This is the case of idioms with prepositional structures, such as *sin embargo* 'however', *a través de* 'by means of', *en lo que respecta a* 'as far as X is/are concerned' and formulas such as *en otras palabras* 'in other words', *como se ha visto (anteriormente/más arriba, etc.)* 'as seen (before)'. To obtain such sequences, we extracted n-grams and filter them by frequency (10 occurrences per million words, one of the thresholds conventionally used for the identification lexical bundles; see Biber et al. (1999)) and by dispersion with the same criteria indicated in Section 3.1. So far, we have extracted bi-grams, tri-grams and four-grams.
A frequency threshold alone in the case of bi-grams performs poorly and does not even distinguish combinations produced by chance from others (Bestgen, 2014). However, among bi-grams some interesting ALCs can be found, e.g.: *no obstante* 'notwithstanding', *cabe esperar* 'it should be expected', etc. For that reason, we resisted the idea of discarding bi-gram extraction and added to the frequency threshold other association measures – namely, pointwise Mutual Information, Backwards Transition Probability and Forward Transition Probability, cf. Appel and Trofimovich (2017) for the latter two. Whereas the precision in retrieving phraseological expressions benefits from these measures in the case of bi-grams, this effect is not so evident with longer n-grams. Table 2 shows the result of manually checking the top hundred items from n-grams lists sorted by the above-mentioned association measures.

	Freq.	MI	BTP	FTP
bi-gr	.09	.25	.67	.05
tri-gr	.39	.74	.59	.54
four-gr	.48	.52	.60	.38

Table 2: Precision of association measures in identifying phraseological n-grams

3.3. Collocations and inflected idioms list

Collocations and certain idioms (especially verbal ones) are not necessarily continuous invariable strings. This is especially evident in the case of collocations. Thus, for instance, *formular hipótesis* may occur as *nos lleva a formular la siguiente hipótesis* 'leads us to formulate the following hypothesis', *formulen y revisen sus hipótesis* 'formulate and revise their hypotheses', etc.
In order to extract the relevant information from such highly variable configurations we have resorted to the syntactic annotation of the expert subcorpus. We have extracted dependency triples "relation(head,dependent)" of the following relations: "amod(*estudio* 'study', *prospectivo* 'prospective')", "obj(*base* 'foundation', *sentar* 'lay')" and "nsubj(*consenso* 'consensus', *existir* 'exist')". Currently, we are running tests with several association measures (t-score, pointwise mutual information, etc.) in order to determine which one performs better in identifying collocations from the corpus. Once the novice corpus is completed, our intention is to extract the same types of ACLs in the novice writer subcorpora as well and compare their use in both types of writers in order to discover differences between them and know better the needs of the possible users of the tool.

4. Incorporating CLA's into HARTA

After having been extracted from the expert subcorpus, CLA candidates will be examined by lexicographers who will decide on their inclusion in the tool. In order to facilitate the examination of candidates, two applications have been developed: one for the treatment of collocations and another for idioms and formulas. In the case of collocations, a list of candidates extracted as explained in Section 3.3 is displayed, so that the lexicographer can easily select a particular candidate for its inclusion in HARTA or discard it. Collocations are sorted by their base, and collocates of the same base are displayed in decreasing order of association strength – as determined by an association measure. For each collocation candidate the sentences in which it occurs in the corpus are also displayed so that lexicographers can select representative examples of its use (Figure 2). After being revised by the lexicographers' team, collocations will be sent to HARTA and to the *Diccionario de colocaciones del español* (Alonso-Ramos, M., 2004).
The treatment of formulas and idioms will be slightly different. First, the lexicographers will encounter lists of n-

grams and will be able to choose among the association measures mentioned in Section 3.2 to sort the candidates.

Given the apparent interaction of association measures and n-gram length (cf. Table 2 above), this feature of the application seems particularly interesting. Once the lexicographers decide that a given n-gram qualifies either as a formula or an idiom, they will proceed to manually edit them. Formulas will be assigned a rhetorical function (e.g. presenting conclusions, expressing a contrast, quoting other pieces of research, etc.; see Figure 3, field FUNCIÓN DISCURSIVA). Likewise, possible variants of one formula will have to be introduced manually. Such variants are cases where two or more n-grams show slight formal differences, but perform the same rhetorical functions (e.g. *como se ha dicho más arriba/como hemos dicho más arriba* 'as has been said before/as we have said before'). Since determining whether two n-grams perform the same rhetorical function or not will in all probability require manual analysis of concordances (Salazar, 2014), this process cannot be automatized.

Idioms will be given part-of-speech information (in the field CLASE DE PALABRA). In principle, it could be problematic to define idioms in terms of a single part-of-speech, since at some level they have internal phrase structure. In terms of their meaning, however, they behave as single lexical units and syntactically they can be assimilated to adjectives, adverbs, prepositions, etc. (Mel'čuk, 2006). Idioms' part-of-speech in HARTA will reflect their behavior as blocks, rather than their internal structure.

5. Conclusion

The present paper describes the treatment and exploitation of a corpus compiled as a data source for a dictionary cum writing-aid directed to writers of academic texts in Spanish (HARTA). The project is currently ongoing research. So far, the expert subcorpus has been marked-up, part-of-speech tagged and parsed according to the design described above. The novice writers subcorpus is on the process of being compiled and marked-up.

The expert subcorpus has provided material for the HARTA: an AWL and an idiom list of Spanish have been extracted which provide the candidates for entry headwords of the dictionary. Likewise, we extracted a set of collocations according to the methods indicated in Section 3.3 and containing lemmas of the AWL, taking advantage of the annotation and parsing of the expert subcorpus. Future research will include the manual revision of these results in order to include them in HARTA with different entry structures depending on their phraseological status. Additionally, studies comparing the use of phraseology by experts and novice writers will be carried out with a view to knowing better the needs of the latter group when writing academic texts.

We plan to make the tool accessible online once it is completed. The users of this tool will not only have access to a writing aid, but also to the expert and novice writers' subcorpora through the tool interface.

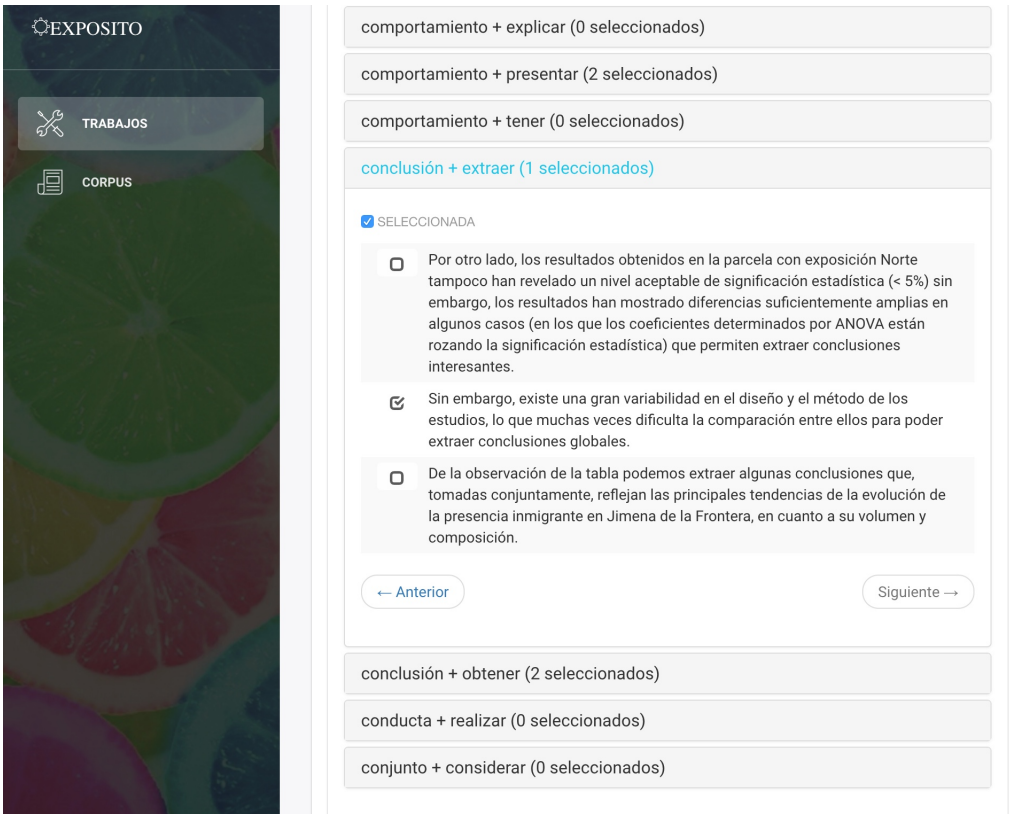

Figure 2: Editing collocations

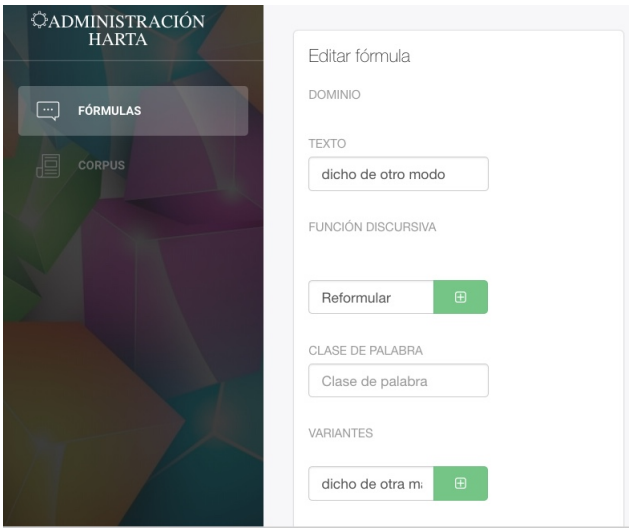

Figure 3: Editing idioms and formulas

6. Acknowledgements

The work presented in this paper has been partially supported by the Spanish Ministry of Economy and Competitiveness (MINECO) and FEDER Funds of the European Commission under the contract number FFI2016-78299-P and the post-doctoral grant Juan de la Cierva (IJCI-2016-29598), and by the Galician Government (post-doctoral grant ED481D 2017/009; RELEX network ED431D R2016/046).This publication reflects the views only of the authors, and InterLAE cannot be held responsible for any use which may be made of the information contained herein.

7. Bibliographical References

Alonso-Ramos, M., García-Salido, M., and Garcia, M. (2017). Exploiting a Corpus to Compile a Lexical Resource for Academic Writing : Spanish Lexical Combinations. In I. Kosem, et al., editors, *Proceedings of 2017 eLex Conference*, pages 571–586, Leiden. eLex.

Appel, R. and Trofimovich, P. (2017). Transitional probability predicts native and non-native use of formulaic sequences. *International Journal of Applied Linguistics*, 27:1–20.

Bestgen, Y. (2014). Extraction automatique de collocations : Peut-on étendre le test exact de Fisher à des séquences de plus de 2 mots ? In *12es Journées internationales d'Analyse statistique des Données Textuelles*, Paris.

Biber, D., Finegan, E., Johansson, S., Conrad, S., and Leech, G. (1999). *Longman Grammar of Spoken and Written English*. Longman, 1 edition.

Callies, M. and Zaytseva, E. (2013). The Corpus of Academic Learner English (CALE). *Dutch Journal of Applied Linguistics*, 2(1):126–132.

Coxhead, A. (2000). A new academic word list. *TESOL Quarterly*, 34(2):213–238.

D'Hertefelt, M., De Wachter, L., and Verlinde, S. (2014). Writing aid dutch. supporting students' writing skills by means of a string and pattern matching based web application. In Susan Zvacek, et al., editors, *Proceedings of the 6th International Conference on Computer Supported Education: Vol. 1*, pages 486–491. SCITEPRESS.

Garcia, M. and Gamallo, P. (2016). Yet another suite of multilingual NLP tools. In J. P. Leal J. L. Sierra-Rodríguez et al., editors, *Languages, Applications and Technologies. Communications in Computer and Information Science*, pages 65–75. Springer, Cham.

Gardner, S. and Nesi, H. (2013). A classification of genre families in university student writing. *Applied Linguistics*, 34(1):25–52.

Granger, S. and Paquot, M. (2015). Electronic lexicography goes local: Design and structures of a needs-driven online academic writing aid. *Lexicographica: international annual for lexicography*, 31(1):118–141.

Gries, S. T. (2008). Dispersions and adjusted frequencies in corpora. *International Journal of Corpus Linguistics*, 13(4):403–437.

Kübler, N. and Pecman, M. (2012). The ARTES bilingual LSP dictionary. from collocation to higher order phraseology. In S. Granger et al., editors, *Electronic Lexicography*, pages 187–210. Oxford University Press, Oxford.

Lijffijt, J., Nevalainen, T., Saily, T., Papapetrou, P., Puolamaki, K., and Mannila, H. (2014). Significance testing of word frequencies in corpora. *Digital Scholarship in the Humanities*, 31:374–397.

Mel'čuk, I. (2006). Parties du discours et locutions. *Bulletin de la Société de linguistique de Paris*, 101(1):29–65.

Mel'čuk, I. (2015). Clichés, an Understudied Subclass of Phrasemes. *Yearbook of Phraseology*, 6(1):55–86.

Nivre, J., Marneffe, M.-C. D., Ginter, F., Goldberg, Y., Manning, C. D., Mcdonald, R., Petrov, S., Pyysalo, S., Silveira, N., Tsarfaty, R., and Zeman, D. (2016). Universal Dependencies v1: A Multilingual Treebank Collection. In *Proceedings of the 10th International Conference on Language Resources and Evaluation (LREC 2016)*, pages 1659–1666. European Language Resources Association (ELRA).

Padró, L. and Stanilovsky, E. (2012). Freeling 3.0: Towards wider multilinguality. In Nicoletta Calzolari, et al., editors, *Proceedings of the 8th International Conference on Language Resources and Evaluation (LREC2012)*, pages 2473–2479. European Language Resources Association (ELRA).

Paquot, M. and Bestgen, Y. (2009). Distinctive words in academic writing: a comparison of three statistical tests for keyword extraction. In Andreas H. Jucker, et al., editors, *Corpora: Pragmatics and Discourse*, pages 247–269, Amsterdam. Rodopi.

Paquot, M. (2010). *Academic vocabulary in learner writing*. Continnuum, London.

Salazar, D. (2014). *Lexical Bundles in Native and Non-native Scientific Writing: Applying a Corpus-based Study to Language Teaching*. Studies in corpus linguistics. John Benjamins Publishing Company, Amterdam/Philadelphia.

Sebastián-Gallés, N., Martí, M. A., Carreiras, M. F., and

Vega, F. C. (2000). *LEXESP: léxico informatizado del español*. Edicions de la Universitat de Barcelona.

Straka, M., Hajic, J., and Straková, J. (2016). Udpipe: Trainable pipeline for processing conll-u files performing tokenization, morphological analysis, pos tagging and parsing. In *Proceedings of the 10th International Conference on Language Resources and Evaluation (LREC 2016)*, pages 1659–1666. European Language Resources Association (ELRA).

Swales, J. and Feak, C. (2004). *Academic Writing for Graduate Students: Essential Tasks and Skills*. Academic Writing for Graduate Students. University of Michigan Press.

Tutin, A. and Kraif, O. (2016). Routines semantico-rhétoriques dans l'écrit scientifique de sciences humaine: l'apport des arbres lexico-syntaxiques récurrents. *Revue de linguistique et de didactique des langues*, 53:119–141.

UNESCO. (1978). Recommendation concerning the international standardization of statistics on science and technology. Technical report, UNESCO.

8. Language Resource References

Alonso-Ramos, M. (2004). *Diccionario de colocaciones del español*. http://www.dicesp.com/paginas.

InterLAE Research Group. (2008). *Spanish English Research Article Corpus*. http://www.interlae.com/serac.