Annotation of lexical bundles with discourse functions in a Spanish academic corpus

Eleonora Guzzi[1], **Margarita Alonso Ramos**[1], **Marcos Garcia**[2], **Marcos Garcia Salido**[1]
[1]CITIC, Universidade da Coruña
[2] CiTIUS Research Center, Universidade de Santiago de Compostela
{eleonora.guzzi,marcos.garcias,margarita.alonso}@udc.es
marcos.garcia.gonzalez@usc.gal

Abstract

This paper describes the process of annotation of 996 lexical bundles (LB) assigned to 39 different discourse functions in a Spanish academic corpus. The purpose of the annotation is to obtain a new Spanish gold-standard corpus of 1,800,000 words useful for training and evaluating computational models that are capable of identifying automatically LBs for each context in new corpora, as well as for linguistic analysis about the role of LBs in academic discourse. The annotation process revealed that correspondence between LBs and discourse functions is not biunivocal and that the degree of ambiguity is high, so linguists' contribution has been essential for improving the automatic assignation of tags.

1 Introduction

Lexical bundles (LB) in academic English have been the object of many studies (Hyland, 2008, Douglas et al., 2004, Simpson-Vlach and Ellis, 2010). Although LBs are strictly defined as recurrent lexical sequences with high frequency and dispersion, their linguistic value comes from the discourse function that they fulfil. It is well known that the mastery of these LBs, such as *it should be noted* ('to emphasize'), *as can be seen* ('to resend'), or *it is clear that* ('to show certainty'), is crucial in academic writing. In English, lexical resources have been proposed (e.g. Granger and Paquot, 2015) in order to offer aid especially to novice writers. However, for academic Spanish few resources are available.

In light of this, the aim of this paper is to discuss the annotation of a Spanish academic corpus with the subset of LBs that have a discursive function, referred here as the umbrella term of *formula*. To the best of our knowledge, it is the first Spanish corpus with this type of annotation. Even though there is an extensive research on Spanish discourse markers, focused on a lexicographic description (Briz et al., 2008) or on its automatic identification and classification (Nazar, 2021), we do not know any Spanish corpus with annotations of academic formulae. Our research is related to Connective-lex (Stede et al., 2019), although it is based on the tagset of Penn Discourse Treebank 3.0 (Webber et al., 2019). Likewise, we must mention da Cunha et al. (2011), the Spanish corpus annotated with the discourse relations used in Rhetorical Structure Theory (Mann and Thompson, 1998).

The purpose of the annotation described here is to obtain a gold-standard corpus to train and evaluate computational models on the automatic identification and classification of academic formulae in new corpora. If generally multiword units have been especially difficult in NLP, formulae have the extra difficulty that they deal with discourse functions that seem more slippery for language models. Although many formulae are compositional, they must be also considered as phraseological units because they work as a whole and cannot be replaced by synonymous expressions that are unnatural; for instance, in English we cannot replace *to put it differently* with *to use some different expressions* or *to say it in a different way*. In our approach (Mel'čuk, 2015) *multiword expressions* (or *phrasemes*) include compositional and non-compositional phrases. Likewise, in the studies developed for academic English such as Simpson-Vlach and Ellis (2010), formulae include compositional and non-compositional expressions but all of them are considered *formulaic sequences*.

In what follows, we describe the process of annotation and human validation, where the main challenge has been to select the proper discourse function to ambiguous formulae.

Proceedings of the 19th Workshop on Multiword Expressions (MWE 2023), pages 99–105

2 Dataset

This section describes the corpus and the formulae list of academic Spanish used for the present study.

2.1 Corpus

We rely on the HARTA academic corpus (HARTA-Exp) (García-Salido et al., 2019) for the annotation. It contains 2,025,092 word tokens extracted from 413 research articles published in scientific journals in Spanish from different areas. The core of this corpus derives from the Spanish part of SERAC corpus (Pérez-Llantada, 2008). Texts are classified in 4 main areas: (i) Arts and Humanities, (ii) Biology and Health Science, (iii) Physical Science and Engineering, and (iv) Social Sciences and Education. This corpus has been tokenized and lemmatized with LinguaKit (Garcia and Gamallo, 2016) and PoS-tagged with FreeLing (Padró and Stanilovsky, 2012). Lastly, UDPipe (Straka et al., 2016) was used for dependency parsing using universal dependencies (Nivre et al., 2016).

2.2 Academic formulae

The formulae selected for this study are recurrent sequences of words that are relevant for Spanish academic writing. They fulfil a discourse function, namely, they can help writers to reformulate what is said, i.e. *dicho de otro modo* ('in other words'), to indicate opposition, i.e. *no obstante* ('however'), to express certainty, i.e. *es sabido que* ('it is well known that'), and so on.

Initially, the list included 985 formulae that were identified using a semi-automatic method (García-Salido et al., 2018), although it was extended after manual revision, as we show in Section 4. We first automatically extracted from the corpus around 5,772 LBs corresponding to strings from two to six n-grams. A frequency and distribution threshold was set to 10 occurrences per million words and to ≥ 1 occurrence in each of the four areas. Secondly, LBs were exhaustively revised by lexicographers to identify relevant academic formulae. This task consisted of discarding irrelevant structures, such as LBs made up of grammatical elements or LBs that hardly fulfilled textual or interpersonal functions, and to select the candidates that they judged were relevant for academic writing. Once the list was obtained, each formula was assigned to the a discourse function based on García-Salido et al. (2019) classification.

The classification is the result of combining top-down and bottom-up approaches. It consists of 3 main groups which contain 39 discourse functions[1]: (i) bundles related to the research process, such as to 'present the conclusions', e.g. *podemos concluir que* ('we may conclude that'); (ii) text-oriented bundles, e.g. for 'ordering', such as *en primer lugar* ('first'); and (iii) interpersonal bundles, that is, expressions conveying epistemic, deontic and evaluative meanings, such as to 'mitigate', e.g. *tal vez* ('perhaps'). In case of ambiguous formulae with two possible functions, they were assigned to the most frequent function. As a result, we may find formulae such as *de acuerdo con* ('according to'), which can be assigned to two discourse functions depending on the context, or *es más* (lit. 'is more'), which sometimes behaves as a formula that fulfills a function and sometimes does not. The list of academic formulae with their discourse function tags makes the point of departure of the annotation task.

3 Annotation procedure

The procedure followed for annotating academic formulae is summarized in Fig. 1.

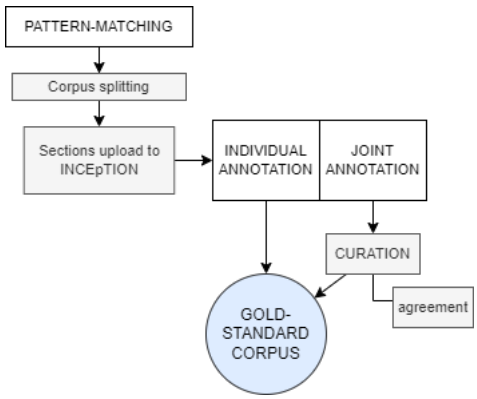

Figure 1: Annotation procedure of LBs.

The first step involved using the academic formulae list with their discourse functions to identify their occurrences in the corpus through a pattern-matching technique. As for the second step, the annotated corpus was split in 15 blocks of ca. 120,000 word tokens each, with the aim of mixing texts from different authors and disciplines. The 15 blocks were uploaded to INCEpTION (Klie et al., 2018), a tool that has been used for the manual

[1]The entire classification is shown in Appendix A, along with the most frequent formula of each discourse function.

evaluation of the automatically annotated corpus to validate the results.

As illustrated by Fig.2, once the corpus is uploaded, the main page for the annotator shows the text, the formula underlined and the discourse function's tag.

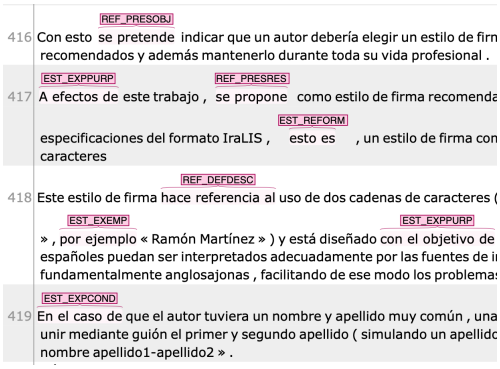

Figure 2: INCEpTION's interface for annotators.

Besides the tagged text, the annotator is provided with a panel with access to the 39 discourse functions. Here, the annotator can change the discourse function, delete it, as well as associate a new discourse function to a formula that needs to be added.

Thus, the main task for annotators has been to validate whether discourse functions were correctly tagged by pattern-matching and to revise whether annotated LBs were proper formulae in all contexts, because different situations could have emerged. A more detailed description of each situation is given in Section 4.

The 15 blocks of texts were distributed among three annotators, in such a way that each annotator had 5 individual annotation blocks, a joint annotation (two annotators who worked on the same block but independently) and, finally, a consensus annotation. The consensus annotation is obtained from applying a curation process to joint annotations. More precisely, the annotator starts a process of "neutralization" of mismatching annotations by changing the discourse function of a formula that was wrongly assigned, by adding a tag in a formula that was not identified, or by removing the formula because it does not behave as such in given contexts. Instead of errors, different annotations might be seen as plausible variations among annotators due to different reasons, as pointed out by Plank (2022).

Once this exhaustive task has been completed, an annotated corpus of ca. 1,800,000 words was obtained (88% of HARTA-Exp), including 360,000 words of consensus annotations. The product obtained from the curation process is a set of peer-reviewed texts that have been used to calculate inter-annotator agreement.

4 Results and Discussion

Manual examination of the automatically annotated corpus has been time consuming and a demanding task for annotators. It lasted around 180 hours only for the individual annotations, at least 12 hours for each block of 120,000 words. In addition to the validation in INCEpTION, we must take into account the previous long and exhaustive linguists' task of identifying formulae and assigning the proper discourse functions. Consequently, we can say that linguists' contribution has been essential to identify academic formulae and their functions in corpora as a first step, as well as to improve a part of the automatic annotation (11%)[2], which ensured the high quality of the data in the gold-standard corpus.

The time invested led to an average of 414 changes per ca. 3,858 tagged formulae in each block that underwent manual examination. Because we wanted to ensure there was coherence among decisions made by annotators, we calculated the agreement for the 3 joint annotations. Results have shown high values for the raw agreement (number of agreed items/nº of total items) of the consensus texts ranging between 89% and 92%, so it provided a positive general overview about the annotation process. Krippendorff α (Krippendorf, 2011) was also performed in order to calculate the amount of agreement that was attained above the level expected by chance or arbitrary coding. Similarly, values for joint annotations revealed a high level of agreement: α=0.885 for block 1; α=0.898 for block 2, and α=0.925 for block 3. Therefore, this agreement was considered as an acceptable reference for annotating the rest of blocks individually.

The main findings provided by the annotation process suggest that annotators dealt with four different types of changes: (i) formulae that annotators judged they do not behave as such in given contexts; (ii) ambiguous formulae associated to two discourse functions; (iii) occurrences of nested formulae where only the longest string was identified;

[2]The 11% is calculated considering that, following annotators judgments, the 89% is correctly annotated by the automatic technique, and the remainder corresponds to manual changes of the automatic annotation.

and (iv) occurrences of new formulae as different morpho-syntactic forms of existing ones.

As we can see in Fig. 3 below, the most faced situations by annotators have been discarding LBs (i), that stands for the 50%, followed by changing the discourse function of ambiguous formulae (ii), that represents the 41% of the total amount of changes. Conversely, nested bundles (iii) and (iv) addition of new morpho-syntactic forms describes only the 4-5%.

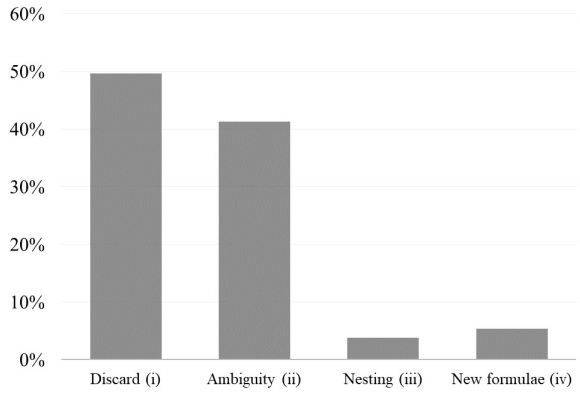

Figure 3: Frequency of each type of annotation change.

Regarding the first type of change (i), it is worth emphasizing that some of the occurrences of 12 formulae, such as *es más* ('in addition'), were discarded because in some specific contexts they were not associated to any discourse function. For instance, *es más* can be used to 'add information' (1), but in contexts such as (2) it is a LB that is not associated to any discourse function, so it must be removed:

(1) **Es más**, la misma alumna emplea este apelativo dirigiéndose a un amigo o amiga.

'**What is more**, the student uses this appellation for addressing to a friend.'

(2) [...] debido a que su fabricación **es más** sencilla.

(lit.)'[...] because its fabrication **is more** simple.'

As for the second type of change (ii), it turned out that the discourse function chosen for 27 [3] ambiguous formulae (two possible functions) was not much more frequent than the other function, so that it involved several changes in annotation. It was especially the case of strings like *en relación con* ('with regard to'), which depending on its position in the sentence is associated to different functions. Thus, *en relación con* and the like, when used sentence-initially normally serve to 'introduce the topic' of a sentence, whereas in sentence-internal distributions they usually head some 'delimiting' modifier. In this regard, the function 'introduce the topic' was substituted for 'delimiting' 499 times, way above other functions, which were modified 30 times on average during the validation process. The difference of switching times from 'delimiting' to other discourse functions in ambiguous formulae is shown in Fig. 4. In this respect, 'delimiting' frequently alternates with 'introduce the topic' (INTOPIC) as well as with 'quoting and reporting' (INDSOURCE), but hardly ever switches to 'compare' (COMP):

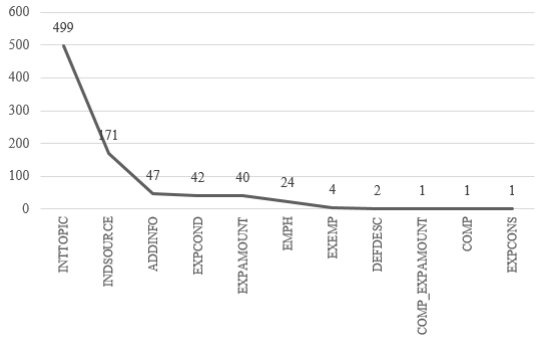

Figure 4: Frequency of changes of 'delimiting' to another discourse function.

Such type of change is reflected also in the formula *de acuerdo con* ('according to'), that is used for 'quoting and reporting' (3) or as a 'delimiting' marker (4):

(3) **De acuerdo con** Takada y Lourenço en 2004, las características generales de esta disciplina [...].

'**According to** Takada and Lourenço in 2004, general features of this discipline [...].'

(4) '[...] tiene que ver con estrategias y prioridades de actuación de cada biblioteca **de acuerdo con** su particular circunstancia local.'

'[...] it has to do with strategies and priorities of action of each library **according to** their particular local circumstance.'

Another example of ambiguity is found within the formula *en torno a* ('around'), that in some contexts it is used for 'delimiting' (5), but in other contexts to 'mitigate' a quantity (6):

(5) Desde el análisis de contenido, hemos normalizado las respuestas **en torno a** cuatro categorías identificativas.

'From the analysis of content, we normalized the responses **around** four identifying categories.'

[3] It should be noted that if we treat ambiguous formulae separately in the final list, the total number of formulae would be 1,023 instead of 996, since 27 formulae have two different entries.

(6) La temperatura media de la capital se sitúa **en torno a** los 15º C.

'The capital's average temperature is **around** 15º C.'

Concerning the third type of change (iii), annotators dealt with some cases where two formulae were nested but only the longest one was automatically tagged by pattern-matching. For instance, in *como podemos observar en la tabla* ('as we can see in the table'), we find *como podemos observar en* (4-gram) and *en la tabla* (3-gram), so the preposition *en* ('in') belongs to both formulae. In those cases, annotators selected the formula they considered the most relevant for each context and assigned them its discourse function.

Finally, the fourth type of change (iv) relates to new formulae that were not identified in the automatic extraction but were of particular interest. New formulae were selected if they met the frequency criterion and were morpho-syntactic variants of already registered ones. For instance, expert writers tend to use the complete and discontinuous formula *por una parte, por otra parte* ('on the one hand, on the other hand'), but we found instances where the abbreviated and grammatically correct counterpart was used (*por otra*; lit. 'on the other') and that were not in our initial list. Thus, 11 different types of morpho-syntactic variants identified during this phase were added to the initial list of 985, that sums up a total amount of 996 formulae.

5 Conclusions

This paper described the annotation process of a new Spanish academic corpus of 1,800,000 words annotated with 996 formulae, that are assigned to 39 different discourse functions. This process is the result of a combination of an automatic annotation and a manual validation. The corpus obtained can be considered a valuable resource because besides of being manually validated, inter-annotator agreement showed high values of coincidence between decisions made by annotators.

Automatic techniques used to identify specific vocabulary from corpus are a good starting point to provide researchers with preliminary data to work with. The same applies for annotating occurrences of formulae in corpora. However, we found that identification and annotation procedures still needed a human validation in order to obtain a gold-standard corpus as a benchmark. Especially in the annotation, ambiguity has demonstrated to be present: many instances with LBs that behaved as a formula in some contexts but not in others were found, as well as different formulae that are associated to two possible discourse functions depending on the context were frequent. Further work aims to use the gold-standard corpus obtained from this study to train and evaluate computational models that are capable of identifying automatically adequate lexical bundles in new corpora, as well as for lexicographic and linguistic studies.

Limitations

This study has two main limitations that are size-related. On the one hand, it is widely accepted the larger the corpora, the better the results, but the annotated corpus used for building the gold-standard is only ca. 1,800,000 words. Therefore, it might be criticized that language models can be trained properly with sufficient amount of data, but in the near future we expect to complete the annotation of the entire corpus. Once completed, we plan to make it available for research purposes. On the other hand, because it was too time-consuming, consensus annotations covered only a part of texts, so we cannot fully ensure the reliability and validity of the entire annotation. However, consensus annotations were made in a triangular way, so that joint annotations from mixed annotators were chosen, and agreements among different annotators were analysed.

Regarding the Inter-Annotator Agreement (IAA), we must also mention some weakness of the manual evaluation since it departs from automatically pre-annotated data and the manual task is only an edition of the result. In this sense, there might be unexpected bias (e.g. the annotator may not read carefully the unannotated part for finding a missing annotation, but focuses only on the pre-annotated part) that can lead to trust and overestimate the IAA. In light of this, a complementary IAA study on a subset of data without pre-annotation is planned for further work.

Acknowledgements

This research was funded by MCIN/AEI (PID2019-109683GB-C21 and PID2021-128811OA-I00), by Xunta de Galicia (ED431C 2020/11 and ED431F 2021/01), and by FEDER GALICIA 2014-2020 (ED431G 2019/01 and ED431G 2019/04). EG is funded by the Programa de Axudas á Etapa predoutoral da Xunta de Galicia, and MG by a Ramón y Cajal grant (RYC2019-028473-I).

References

Antonio Briz, Salvador Pons, and José Portolés. 2008. Diccionario de partículas discursivas del español.

Iria da Cunha, Juan-Manuel Torres-Moreno, and Gerardo Sierra. 2011. On the development of the RST Spanish Treebank. In *Proceedings of the 5th Linguistic Annotation Workshop*, pages 1–10.

Biber Douglas, Susan Conrad, and Viviana Cortes. 2004. If you look at....: lexical bundles in university teaching and textbooks. *Applied Linguistics*, 25(3):371–405.

Marcos Garcia and Pablo Gamallo. 2016. Yet another suite of multilingual NLP tools. In *Communications in Computer and Information Science*, pages 65–75, Cham. Springer.

Marcos García-Salido, Marcos Garcia, and Margarita Alonso-Ramos. 2019. Identifying lexical bundles for an academic writing assistant in spanish. In *Computational and Corpus-Based Phraseology: Third International Conference, Europhras 2019, Malaga, Spain, September 25–27, 2019, Proceedings 3*, pages 144–158. Springer.

Marcos García-Salido, Marcos Garcia, Milka Villayandre-Llamazares, and Margarita A. Ramos. 2018. A lexical tool for academic writing in spanish based on expert and novice corpora. In *Proceedings of the Eleventh International Conference on Language Resources and Evaluation (LREC 2018)*.

Sylviane Granger and Magali Paquot. 2015. Electronic lexicography goes local design and structures of a needs-driven online academic writing aid. *International Annual for Lexicography*, 31(1):118–141.

Ken Hyland. 2008. As can be seen: lexical bundles and disciplinary variation. *English for Specific Purposes*, 27(1):4–21.

Jan-Christoph Klie, Michael Bugert, Beto Boullosa, Richard Eckart de Castilho, and Iryna Gurevych. 2018. The INCEpTION platform: Machine-assisted and knowledge-oriented interactive annotation. In *Proceedings of the 27th International Conference on Computational Linguistics: System Demonstrations*, pages 5–9, Santa Fe, New Mexico. Association for Computational Linguistics.

Klaus Krippendorf. 2011. Computing Krippendorff's alpha-reliability. *Departmental Papers (ASC). University of Pennsylvania*.

William C. Mann and Sandra A. Thompson. 1998. Rhetorical structure theory: Towards a functional theory of text organization. *Text. Interdisciplinary Journal for the Study of Discourse*, 8(3):243–281.

Igor Mel'čuk. 2015. Clichés, an understudied subclass of phrasemes. *Yearbook of Phraseology*, 6(1):55–86.

Rogelio Nazar. 2021. Automatic induction of a multilingual taxonomy of discourse markers. *Electronic lexicography in the 21st century: postediting lexicography*, pages 440–454.

Joakim Nivre, Marie-Catherine Marneffe, Filip Ginter, Yoav Goldberg, Cristopher D. Manning, Ryan Mcdonald, Slav Petrov, Sampo Pyysalo, Natalia Silveira, Reut Tsarfaty, and Daniel Zeman. 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.

Lluís Padró and Evgeny Stanilovsky. 2012. Freeling 3.0: Towards wider multilinguality. In *Proceedings of the 8th International Conference on Language Resources and Evaluation (LREC2012)*, page 2473–2479.

Carmen Pérez-Llantada. 2008. Humans vs. machines? a multiperspective model for esp discourse analysis in intercultural rhetoric research. *ESP Across Cultures*, 5:91–104.

Barbara Plank. 2022. The "problem" of human label variation: On ground truth in data, modeling and evaluation. In *Proceedings of the 2022 Conference on Empirical Methods in Natural Language Processing*, pages 10671–10682, Abu Dhabi, United Arab Emirates. Association for Computational Linguistics.

Rita Simpson-Vlach and Nick C. Ellis. 2010. An academic formulas list: New methods in phraseology research. *Applied linguistics*, 31(4):487–512.

Manfred Stede, Tatjana Scheffler, and Amália Mendes. 2019. Connective-lex: A web-based multilingual lexical resource for connectives. *Discours. Revue de linguistique, psycholinguistique et informatique*, 24.

Milan Straka, Jan Hajic, and Jaja Straková. 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 4290–4297.

Bonnie Webber, Rashmi Prasad, Alan Lee, and Aravind Joshi. 2019. The Penn Discourse Treebank 3.0 annotation manual. *Philadelphia, University of Pennsylvania*, 35:108.

A Classification of discourse functions

Discourse Function	NF	Example
Añadir información 'Add information'	49	así como 'as well as'
Comparar 'Compare'	34	igual que 'as'
Delimitar 'Delimiting'	75	respecto a 'regarding to'
Ejemplificar 'Give examples'	25	por ejemplo 'for instance'
Expresar causa 'Express cause'	32	ya que 'because'
Expresar condición 'Express condition'	20	en función de 'depending on'
Expresar consecuencia 'Express consequence'	60	por lo que 'therefore'
Expresar finalidad 'Express purpose'	18	para que 'in order to'
Expresar oposición 'Express opposition'	31	sin embargo 'however'
Expresar concesión 'Express concession'	14	a pesar de 'in spite of'
Hacer referencia al propio trabajo 'Reference to the own work'	16	en este trabajo 'in this work'
Introducir un tema 'Introduce the topic'	9	respecto a 'with respect to'
Introducir una alternativa 'Introduce an alternative'	3	o bien 'or'
Introducir una excepción 'Introduce an exception'	7	a excepción de 'except for'
Ordenar 'Organize'	19	por otro lado 'on the other hand'
Reenviar 'Resend'	30	en la tabla 'in the table'
Reformular 'Reformulate'	19	es decir 'that is'
Resumir 'Summarize'	10	en la práctica 'in practice'
Definir y describir 'Defining and describing'	37	se trata de 'it is about'
Denominar 'Naming'	7	conocido como 'known as'
Establecer grupos 'Listing items'	11	de este tipo 'of this type'
Expresar cantidad 'Express amount'	112	el número de 'the number of'
Expresar frecuencia 'Express frequency'	10	a veces 'sometimes'
Expresar progresión 'Express progression'	3	a medida que 'as'
Expresar correlación 'Express correlation'	1	cuanto más 'the more'
Expresar tiempo 'Express time'	50	después de 'after'
Presentar datos 'Present data'	36	se observa 'it is observed'
Presentar el objeto de estudio 'Present the object of study'	5	se centra en 'focused on'
Presentar la hipótesis 'Present hypothesis'	4	se estima que 'it is estimated that'
Presentar la metodología 'Introduce methodology'	83	a través de 'through'
Presentar las conclusiones 'Introduce conclusions'	24	se encontró 'it was found'
Presentar los objetivos 'Introduce goals'	7	se pretende 'it is intended'
Atenuar 'Mitigate'	21	la mayoría de 'most of'
Expresar necesidad 'Express need'	8	debe ser 'it must be'
Expresar una evaluación 'Evaluate'	4	es importante 'it is important'
Hacer hincapié 'Emphasize'	30	sobre todo 'especially'
Indicar certeza 'Express certainty'	30	de hecho 'in fact'
Indicar la fuente 'Quoting and reporting'	37	de acuerdo con 'according to'
Indicar posibilidad 'Express possibility'	5	puede ser 'it may be'

Table 1: Classification of 39 *Discourse Functions*, number of formulae at type level in each discourse function (*NF*), and the most frequent formulae of each one (*Example*).