https://doi.org/10.14618/ids-pub-11183

Irene Renau/Rogelio Nazar

TOWARDS A MULTILINGUAL DICTIONARY OF DISCOURSE MARKERS

Automatic extraction of units from parallel corpus

Abstract This paper presents a multilingual dictionary project of discourse markers. During its first stage, consisting of collecting the list of headwords, we used a parallel corpus to automatically extract units from texts written in Spanish, Catalan, English, French and German. We also applied a method to create a taxonomy structure for automatically organising the markers in clusters. As a result, we obtain an extensive, corpus-driven list of headwords. We present a prototype of the microstructure of the dictionary in the form of a standard XML database and describe the procedure to automatically fill in most of its fields (e. g., the type of DM, the equivalents in other languages, etc.), before human intervention.

Keywords Computational lexicography; corpus-driven lexicography; discourse markers; multilingual lexicography

1. Introduction

In this paper we present *Dismark*, an ongoing multilingual dictionary project on discourse markers (DMs), especially oriented towards those that are used on written texts. We focus on the first stage of the project: the automatic extraction of the list of headwords of the dictionary (also called macrostructure, Hartmann 2001, p. 64). We also deal with the first tasks concerning the microstructure, that is, the organisation of the information in the entries and the way the different elements are connected to each other (ibid., pp. 64 f.).

We use a parallel corpus to detect DMs with similar functions in different languages (so far, in Spanish, Catalan, English, French and German), to obtain an extensive corpus-driven lemma list. This is a very different approach from traditional DMs' dictionaries, which are manually crafted based on previous dictionaries or classifications. For the manual creation of a prototype we used Lexonomy (Měchura 2017), an online dictionary software that provides functions to create, import and export database contents in the XML standard.

This project is motivated by the fact that online DM dictionaries are scarce, they tend to be outdated, incomplete, and often lack multilingual support. There are also general dictionaries that contain DMs among their entries, but they receive the same lexicographic treatment as regular lexical units. This is far from ideal as DMs, due to their functional nature, require specific solutions. Dealing with DMs means to consider practical aspects of written production and comprehension, such as punctuation, discursive order, register, multifunctionality, etc. In the first stages of our project, *Dismark* will contain general information about DMs covering the needs of a standard user in a literate society in which written documents are fundamental (Smith/Schryer 2013). In further stages, however, it will be possible to narrow down the type of dictionary to accommodate it to the more specific needs of particular groups of professionals or students.

In the following sections, we first provide a theoretical framework about the concept and categorisation of DMs (section 2), we then explain the method we used to extract the DMs

This paper is part of the publication: Klosa-Kückelhaus, Annette/Engelberg, Stefan/Möhrs, Christine/Storjohann, Petra (eds.) (2022): Dictionaries and Society. Proceedings of the XX EURALEX International Congress. Mannheim: IDS-Verlag.
https://doi.org/10.14618/phpy-6r66

from corpus (section 3), we later provide a preliminary description of the microstructure of the dictionary (section 4) and, finally, we arrive at some conclusions and propose a program for future work (section 5).

2. Theoretical framework

In recent years, DMs have attracted considerable attention in linguistic research (e. g. Casado Velarde 1993; Fraser 1999; Martín Zorraquino/Portolés 1999; Pons 2001; Fischer 2006; Borreguero/López 2010). Early interest on the subject began to appear in the context of text grammar and discourse analysis (van Dijk 1973, 1978; Halliday/Hasan 1976; Halliday 1985). In these preliminary studies, DMs were described as particles used to facilitate the coherent interpretation of texts. In other words, instructions to connect the different propositions in a text and to organise the argumentation. They are, for this reason, considered functional rather than lexical units, as they provide procedural instead of conceptual information. Notwithstanding this characterization, DMs do play an important role in written and oral communication. They not only connect and organise parts of discourse, but can also indicate subjectivity or attitudes, or may even be used to regulate the interaction between participants in communication (Fox Tree 2015). They are, thus, fundamental textual pieces which lay on an intermediate space between grammar and lexicon.

DMs are difficult to recognise and categorise (Cartoni/Zufferey/Meyer 2013). They can be single or multiword expressions, and they can pertain to different categories, such as conjunctions, adverbs, and prepositional phrases, among others. The most applied approach for the organisation is their functional similarity. Among the most frequently found categories, one can find for instance additive connectives (*also*, *furthermore*); contrastive connectives (*however*, *nevertheless*); causal connectives (*consequently*, *for this reason*), and a large number of other categories and examples.

Different ways to categorise DMs have been discussed in discourse studies (Fuentes Rodríguez 1987; Fraser 1999; Martín Zorraquino/Portolés 1999; Pons 2001), but they have not yet been described in dictionaries with sufficient detail and precision, probably due to their complexity and discursive nature. Attempts to create extensive catalogues or dictionaries of DMs are comparatively less numerous. In Spanish, some prominent examples are Santos Río (2003), Briz (2008) and Holgado Lague (2017). For other languages, there are taxonomies in English (Knott 1996), German (Stede 2002), French (Roze 2012), Portuguese (Mendes et al. 2018) and Italian (Feltracco et al. 2016), among others. In addition, an important initiative has appeared in recent years, to integrate different resources in a large, manually curated, multilingual database of DMs (Stede/Scheffer/Mendes 2019).

Efforts for the elaboration of taxonomies and catalogues of these units have been made in the past mostly by qualitative means, often by introspection, and sometimes resorting to qualitative analysis of corpora. A well-known example of this traditional approach in Spanish is the taxonomy of DMs by Martín Zorraquino/Portolés (1999), which is also valid for other languages as well. The limitations of this methodology, however, are that it can only produce a limited number of examples per category. Comparatively, less bibliography exists regarding their computational treatment, particularly using quantitative and empirical methods. This is rather surprising, considering the advantages that such methods offer. For instance, they help to overcome the subjective bias of introspection and, with efficient automation, it is possible to process massive corpora, which may lead then to the retrieval of thousands of particular DMs and also to the potential discovery of patterns of use.

This paper is part of the publication: Klosa-Kückelhaus, Annette/Engelberg, Stefan/Möhrs, Christine/Storjohann, Petra (eds.) (2022): Dictionaries and Society. Proceedings of the XX EURALEX International Congress. Mannheim: IDS-Verlag.
https://doi.org/10.14618/phpy-6r66

In contrast to our present research, which is based on a lexicographic perspective, most publications in the field of computational linguistics dealing with DMs are concentrated in the area most closely related to discourse analysis (e.g., Stubbs 1996; Marku 1998; Moore 2003, Webber et al. 2019). This means that most researchers in this trend are less interested in extracting and organising full inventories of DMs than in analysing instances of texts to find cases of coherence relations expressed by these units. Both problems are of course related, but they are not the same, as one deals with types and the other with tokens. The relation is given by the fact that, to analyse DMs in particular texts, one needs some form of dictionary, and this results in the need to create this type of resources. For instance, there have been some categorization attempts using techniques such as clustering and machine learning (Alonso/Castellón/Padró 2002; Hutchinson 2005; Debortoli et al. 2016), although limited to certain types of units and consuming considerable external resources, such as manual annotation, which has the potential for a biassed classification.

Regarding the specific use of parallel corpora for the study of DMs in computational linguistics, previous research is even more scarce. Some authors have used parallel corpora as a method to discover ambiguous DMs (Versley 2010; Zhou et al. 2012), and Robledo/Nazar (2018) used a clustering method from parallel corpora, but limited to parenthetical markers and using a variety of external resources. In contrast to these methods, our current proposal is conceptually and computationally simple, more generalizable, and less dependent on external knowledge. The method presented here is a further development of ideas suggested earlier by Nazar (2021).

3. Methodology for the compilation of DMs using a parallel corpus

We propose a method to obtain an extensive inventory of the DMs of a given language, provided that a sufficiently large parallel corpus is available for that language and some other. We describe an algorithm to fully automatise all the process, starting from the corpus and finishing with a ready-to-use database. This database contains a hierarchical organisation of categories of DMs, populated with many examples in the languages under examination. In addition, our method is designed for a dynamic process, because once a first version of the database is created, it is then used to provide examples for the automatic categorisation of new DMs, thus further populating said database. These new DMs may come from other sources, not necessarily the same initial corpus.

The core idea of the method is to first separate the DMs from the rest of the vocabulary of the corpus, and then classify them according to a novel clustering algorithm. Classifying, in this case, means also finding out which are the categories, as they are not predefined. The categories are thus a product of the process, as much as the specific DMs populating them.

To facilitate future replication in other languages, we also avoid all forms of explicit knowledge of a particular language, even POS-taggers. The proposed method is thus purely statistical using only corpus as input. The only sense in which we use predetermined knowledge is regarding the names for the categories, which we borrow from Martín Zorraquino/Portolés (1999), but we consider these names can be applied with independence of the language.

The only input is thus a parallel corpus, and in our case, we used Tiedemann's (2012) Opus Corpus, which offers large samples of aligned sentences of a wide variety of languages and

This paper is part of the publication: Klosa-Kückelhaus, Annette/Engelberg, Stefan/Möhrs, Christine/Storjohann, Petra (eds.) (2022): Dictionaries and Society. Proceedings of the XX EURALEX International Congress. Mannheim: IDS-Verlag.
https://doi.org/10.14618/phpy-6r66

genres. This material is freely available in TMX format, which specifies the alignment of translation segments (TS), a unit of measure that typically corresponds to a sentence. There are circa 30 files per language pair in the case of European languages, and each file compresses large samples of texts (circa 3,500 million tokens) of a certain genre or discipline. The corpus is representative of a great variety of written genres.

Oral speech is only indirectly represented in files containing literature and TV subtitles, which also offer large samples of general vocabulary.

The method can be synthesised as follows.

3.1 Extracting DMs from corpus

DMs are automatically separated from the rest of the vocabulary using a co-occurrence association measure that feeds an entropy model. DMs are visible because they show a particular distribution in the corpus, a characteristic pattern that is a consequence of the fact that they are independent of the content of the text in which they occur. In operational terms, this means that their occurrences show a uniform distribution, with a very wide, non-restrictive set of co-occurring words. We say they are uninformative because they cannot be used to predict the occurrence of other lexical units. In contrast, a more informative lexical unit could be *democracy*, as it shows a clear pattern of co-occurrence with a set of words such as *respect*, *freedom*, *rights*, and so on. In contrast, the word *anyway* does not have these "friends", as it only has a functional value. This difference is measured by coefficient (1) where x is a DM candidate and Rx the set of its co-occurrences.

(1) $$I(x) = \frac{\log_2 \sum_{i=1}^{n} R_{x,i}}{\log_2 |m(x)|}$$

The symbol $m(x)$ refers to the contexts candidate x, and $R(x,i)$ is the frequency of the word in position i of the ranked list of the n most frequent words that co-occur with x in the same sentences (in our experiments, n = 20). In one extreme, such coefficient will produce a very low score for function words such as articles, conjunctions, prepositions, etc. At the opposite extreme of this continuum, the most specialised vocabulary units begin to appear, because these are the ones that will typically point to a limited set of other units. An arbitrary threshold k determines if x is classified as a lexical or functional unit. For illustration, consider Figures 1 and 2, showing the co-occurrence profile of the Spanish word *electricidad* ('electricity') and *de todas maneras* ('anyway'), respectively. One can see the different shapes of both curves, the first one having a greater surface under the curve. It should be noted that the method could also be of interest for specialised lexicography because it may be implemented as a term-extractor, as suggested by Nazar/Lindemann (2022).

This paper is part of the publication: Klosa-Kückelhaus, Annette/Engelberg, Stefan/Möhrs, Christine/Storjohann, Petra (eds.) (2022): Dictionaries and Society. Proceedings of the XX EURALEX International Congress. Mannheim: IDS-Verlag.
https://doi.org/10.14618/phpy-6r66

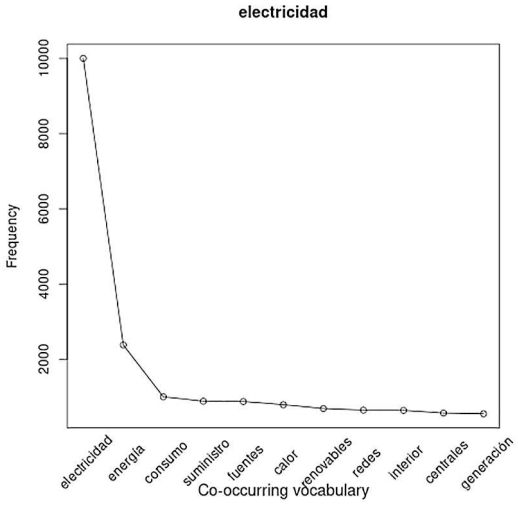

Fig. 1: Co-occurrence profile of the Spanish lexical unit *electricidad* ('electricity')

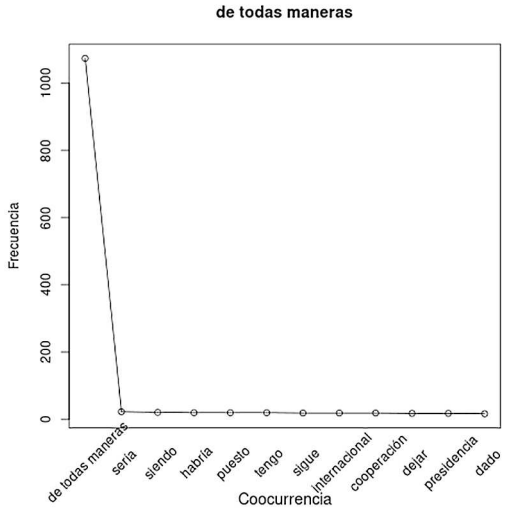

Fig. 2: Co-occurrence profile of the Spanish DM *de todas maneras* ('anyway')

3.2 Clustering DMs

We developed a clustering algorithm that uses the equivalence of the DMs in another language as a similarity measure, hence the parallel corpus. This is effectively to use the parallel corpus as a semantic mirror. For instance, *nevertheless* and *however* can be considered similar because they share the same equivalences in a second language (e. g., *sin embargo* or *no obstante*, in the case of Spanish). To find the equivalences in the parallel corpus, we used an association coefficient based on the co-occurrence of DMs in the aligned sentences (2).

(2) $$A(MD_{es,i}, MD_{ca,j}) = \frac{f(MD_{es,i}, MD_{ca,j})}{\sqrt{f(MD_{es,i})} \cdot \sqrt{f(MD_{ca,j})}}$$

Once with the list of aligned DMs at hand, the clustering algorithm proceeds as follows: it takes the pairs of aligned DMs one by one, e. g. *por esa razón* and *for that reason*. If in a subsequent pair the English DM is repeated, as in the case of *por esta razón* ~ *for that reason*, then it is assumed that *por esa razón* and *por esta razón* are equivalent, that is, they have the same function and can be used in the same context. We see no need, at this point, to exploit lexical or orthographic similarity here but in any case, that is a possibility we leave for future work. If the DMs are similar, they form a new cluster. For illustration, consider a more advanced stage in this process, in which we have a situation such as the one depicted in Figure 3, with *por esta razón* already being a member of a previously formed cluster containing units such as *por ese motivo* or *por este motivo*. In such a case, the newly arrived DM *por esa razón* is added to said cluster. The process finishes when there are no more DM pairs to process.

This paper is part of the publication: Klosa-Kückelhaus, Annette/Engelberg, Stefan/Möhrs, Christine/Storjohann, Petra (eds.) (2022): Dictionaries and Society. Proceedings of the XX EURALEX International Congress. Mannheim: IDS-Verlag.
https://doi.org/10.14618/phpy-6r66

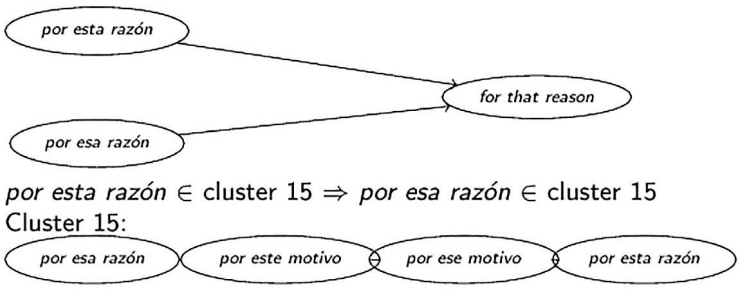

Fig. 3: A moment of the DMs clustering process

3.3 Labelling the clusters

The previous step results in several clusters of similar DMs in each language, but the system is not able at this point to produce a name for these clusters. At this point they are instead only identified with numerical codes. To give these clusters a meaningful name, we used the names of the categories in the taxonomy by Martín Zorraquino/Portolés (1999). Using the few examples they provide for their categories, we can automatically find the match with our clusters and tag them accordingly (3). Also, as all clusters are aligned by language (we keep the initial alignment obtained from the parallel corpus), the same labels are also used for the rest of the languages.

(3) $$sim(MZP_p, CMD_q) = \frac{|M\vec{Z}P_p \cap C\vec{M}D_q|}{|M\vec{Z}P_p|}$$

3.4 Populating the taxonomy with new DMs

Once a basic taxonomy of DMs is built this way, it is then used to classify new DMs in a recursive manner. The algorithm will first classify a DM candidate by language, it will then decide if it is effectively a DM and, if this is the case, it will assign a category to it. For both tasks we used the initial parallel corpus: if a Spanish candidate is a genuine DM, its condition will be signalled by the parallel corpus, because it will be associated with English DMs of the corresponding category. For instance, given a new candidate in Spanish such as *de la misma manera*, we will find that in the Spanish-English parallel corpus this appears aligned with already known English DMs such as in *the same way*, *likewise*, *similarly*, etc. We must conclude, then, that 1) the Spanish candidate is indeed a DM, and 2) it belongs to the same category as its English counterparts. Here lies also the possibility of discovering polyfunctional cases, i. e., the possibility that this Spanish DM is also associated with a different group of English DMs but, again, we leave that challenge for future research.

3.5 Evaluation of the taxonomy of DMs

As a result of this method, we have now obtained 619 candidates for Spanish, 733 for English, 556 for French, 677 for German, and 312 for Catalan, all distributed in 70 different functional categories. The taxonomy of DMs can be consulted at http://www.tecling.com/dismark (last access: 26-05-2022). A campaign for the manual evaluation of the whole col-

This paper is part of the publication: Klosa-Kückelhaus, Annette/Engelberg, Stefan/Möhrs, Christine/Storjohann, Petra (eds.) (2022): Dictionaries and Society. Proceedings of the XX EURALEX International Congress. Mannheim: IDS-Verlag.
https://doi.org/10.14618/phpy-6r66

lection was undertaken with the collaboration of a group of linguists that are native speakers of each of the languages, with two or three linguists per language. The revision involved periodic discussions between members of the different teams, to keep a uniform criterion in all languages.

The evaluation was conducted in two phases. The first one was to determine precision, defined as the proportion of correct DMs found in the newly created DM taxonomy. The second phase, in turn, was to estimate recall, defined as the proportion of DMs that exist in a language that are included in said taxonomy.

For the evaluation of precision, we reviewed all DMs contained in the taxonomy counting the number of cases in which a) an element is not genuine DMs; b) a multi-word DM was not correctly segmented (typically missing an initial or final part) or c) the element is actually a DM but it appears in the wrong cluster or category. The revision revealed that the percentage of errors is less than 5% in all languages except in German, where we found 16% false positives, mostly with segmentation faults. In terms of precision, we believe this result is sufficiently accurate to constitute the core for a list of headwords of the dictionary.

For the evaluation of recall, the method we devised was to obtain random samples of texts and find the proportion of DMs that are in those texts and not in the DM taxonomy, divided by the sum of said number and the total count of DMs in those texts that do also appear in the taxonomy. In a sample of ten texts per language, 88% of the DMs were already documented in our database. This does not translate directly into a measure of recall, but it indicates that at least we have the majority of the most frequent exemplars.

4. Preliminary lexicographic proposal

As stated in section 1, a first stage of the *Dismark* project contemplates creating a core of DM units and microstructural information. The target users of the dictionary are, at this stage, professional communicators such as journalists, screenwriters, translators, lawyers, scientists, etc. (Schriver 2012) and college students in need of acquiring expertise in communication as part of their professional formation (Lea/Street 2006), e. g., students of Journalism, Law, Translation, etc. All these users share common needs, for example, what a specific DM is used for, how should they use punctuation, the orthography of the DM, etc. A second phase of the project contemplates the creation of sub-products, such as a specific version of the dictionary designed for lawyers or journalists.

The dictionary is unidirectional (Atkins/Rundell 2008, p. 40), with Spanish as the target language, and equivalents in English, French, German, and Catalan. The microstructure has the following types of information (see some sample entries in the online prototype: https://www.lexonomy.eu/#/dismark, last access: 05-26-2022):

- **Headword.** As the dictionary is made from scratch for the Internet, the lemmatisation does not contain any change of order, typical from the constraints of the alphabetic order in paper dictionaries. Thus, *aun así* 'still' is lemmatised *aun así* and not *así, aun.*
- **Type of DM.** The different types of DM are categorised according to Martín Zorraquino/Portolés (1999). This field will have a hyperlink to an external webpage containing extended information about the type of DM.

This paper is part of the publication: Klosa-Kückelhaus, Annette/Engelberg, Stefan/Möhrs, Christine/Storjohann, Petra (eds.) (2022): Dictionaries and Society. Proceedings of the XX EURALEX International Congress. Mannheim: IDS-Verlag.
https://doi.org/10.14618/phpy-6r66

- **Register.** We added this field to separate standard from formal DMs. As the dictionary is focused on functional writing, there are not many cases of DMs used in colloquial language.
- **Function.** In this section, we synthetically describe the function of the DM. An extended explanation of the function of the DM is already offered as hyperlink in the *Type of DM*. In this field, we want to cover the need of the user to obtain a quick and clear explanation.
- **Examples.** We provide 1–2 examples of usage, containing at least two sentences, in order to provide enough discursive context. We also provide the source of the examples, which can be different corpora or obtained from documentation or the Internet.
- **Punctuation and position.** We provide the patterns of punctuation and position that the user can find when using or reading the DM. Patterns are expressed by the punctuation sign before and/or after the DM. For example, for *sin embargo* 'however', two common patterns are:

  . *Sin embargo*,
  ; *sin embargo*,

  This allows to solve other orthographic doubts, such as capitals or blank spaces.

  Each pattern is complemented by one or more *Examples*.
- **Spanish equivalents.** A list of all DMs of the same type of the headword are offered here. These groups have been automatically extracted, as explained in the previous section, but are later manually revised. Each DM in this list contains a link to the correspondent entry.
- **Translations** to Catalan, English, French and German. The group of equivalent DMs in these languages are offered. They will also be linked to the multilingual part of the dictionary.

All types of information detailed in this list required expert human supervision. However, most of it can be automatically filled in, e. g., the list of headwords, the types of DMs, the equivalents and the patterns of punctuation and position. As for the examples, a random sample of corpus concordances of each type of pattern is added to the field, so that the lexicographer can easily select convenient examples. All this information can be automatically added, as Lexonomy allows us to work with independent XML files that can be uploaded to the database.

Figure 4 shows one of the entries of the sample prototype, *sin embargo* 'however'.

This paper is part of the publication: Klosa-Kückelhaus, Annette/Engelberg, Stefan/Möhrs, Christine/Storjohann, Petra (eds.) (2022): Dictionaries and Society. Proceedings of the XX EURALEX International Congress. Mannheim: IDS-Verlag.
https://doi.org/10.14618/phpy-6r66

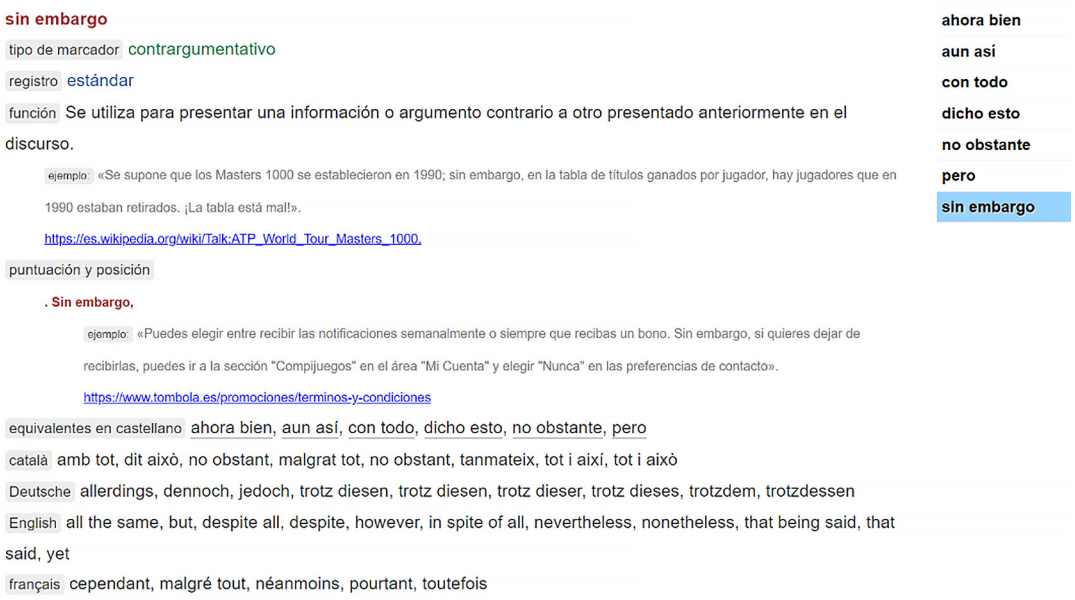

Fig. 4: Example of a *Dismark* entry in the sample prototype

5. Conclusions and further steps

In this paper, we presented our first steps towards a corpus-driven online dictionary of DMs with inter-linked entries in five languages. The method of extraction of DMs from a parallel corpus has enough precision and recall to obtain a large list of headwords for the dictionary that we are planning.

There are different tasks to be addressed in the immediate future of this project. We have to test the prototype with users and, after validation, we have to prepare a final version. There is also work to do in describing each type of DMs present in the dictionary, which will not be part of the dictionary itself, but will be connected to it by hyperlinks. In relation to this, another important aspect to address is the design of the mediostructure (Hartmann 2001, pp. 65 f.), that is, the system of cross-references connecting different entries, parts of the dictionary with external resources, etc. We also must address the problem that some DMs can have multiple functions, as Cartoni/Zufferey/Meyer (2013) show. Finally, and as already mentioned, a long-term project will be to create sub-types of the same dictionary to address the specific needs of different types of users.

References

Alonso, L./Castellón, I./Padró, L. (2002): Lexicón computacional de marcadores del discurso. In: Procesamiento del Lenguaje Natural 29, pp. 239–246.

Atkins, S./Rundell, M. (2008): The Oxford guide to practical lexicography. Oxford.

Borreguero, M./López, A. (2010): Marcadores del discurso: De la descripción a la definición. Madrid/Frankfurt a. M.

Briz, A./Pons, S./Portolés, J. (coords.) (2008): Diccionario de partículas discursivas del español. http://www.dpde.es.

This paper is part of the publication: Klosa-Kückelhaus, Annette/Engelberg, Stefan/Möhrs, Christine/Storjohann, Petra (eds.) (2022): Dictionaries and Society. Proceedings of the XX EURALEX International Congress. Mannheim: IDS-Verlag.
https://doi.org/10.14618/phpy-6r66

Cartoni, B./Zufferey, S./Meyer, T. (2013): Annotating the meaning of discourse connectives by looking at their translation: the translation spotting technique. In: Dialogue and Discourse 4 (2), pp. 65–86.

Casado Velarde, M. (1993): Introducción a la gramática del texto del español. Madrid.

Debortoli, S./Müller, O./Junglas, I. A./vom Brocke, J. (2016): Text mining for information systems researchers: an annotated topic modeling tutorial. In: Communications of the Association for Information Systems 39 (1), pp. 1–30.

Feltracco, A./Jezek, E./Magnini, B./Stede, M. (2016): LICO: A lexicon of Italian connectives. In: Proceedings of Third Italian Conference on Computational Linguistics (CLiC-it 2016) & Fifth Evaluation Campaign of Natural Language Processing and Speech Tools for Italian. Final Workshop (EVALITA 2016), Napoli, Italy, December 5–7, 2016, volume 1749 of CEUR Workshop Proceedings. CEUR-WS.org.

Fischer, K. (ed.) (2006): Approaches to discourse particles. Amsterdam.

Fox Tree, J. E. (2015): Discourse markers in writing. In: Discourse Studies 17 (1), pp. 64–82.

Fraser, B. (1999): What are discourse markers? In: Journal of Pragmatics 31, pp. 931–952.

Fuentes Rodríguez, C. (1987): Enlaces extraoracionales. Sevilla.

Halliday, M. A./Hasan, R. (1976): Cohesion in English. London.

Halliday, M. A. K. (1985): An introduction to functional grammar. London.

Hartmann, R. R. K. (2001): Teaching and researching lexicography. Harlow.

Holgado Lage, A. (2017): Diccionario de marcadores discursivos para estudiantes de español como segunda lengua. New York.

Hutchinson, B. (2005): The automatic acquisition of knowledge about discourse connectives. PhD thesis. Edinburgh.

Knott, A. (1996): A data-driven methodology for motivating a set of coherence relations. PhD thesis. Edinburgh.

Lea, M. R./Street, B. V. (2006): The "academic literacies" model: theory and applications. In: Theory Into Practice 45 (4), pp. 368–377.

Martín Zorraquino, M. A./Portolés, J. (1999): Los marcadores del discurso. In: Bosque, I./Demonte, V. (eds.): Gramática descriptiva de la lengua española, Vol. 3. Madrid, pp. 4051–4213.

Měchura, M. B. (2017): Introducing lexonomy: an open-source dictionary writing and publishing system. In: Electronic Lexicography in the 21st Century: Lexicography from Scratch. Proceedings of the eLex 2017 Conference, 19–21 September 2017, Leiden, The Netherlands.

Mendes, A./del Rio, I./Stede, M./Dombek, F. (2018): A lexicon of discourse markers for Portuguese – LDM-PT. In: Proceedings of the Eleventh International Conference on Language Resources and Evaluation (LREC 2018). Miyazaki, Japan.

Moore, J. D./Wiemer-Hastings, P. (2003): Discourse in computational linguistics and artificial intelligence. In: Graesser, A. C./Gernsbacher, M. A./Goldman, S. R. (eds.): Handbook of discourse processes. London.

Nazar, R. (2021): Automatic induction of a multilingual taxonomy of discourse markers. In: Kosem, I. et al. (eds.): Electronic lexicography in the 21st century: post-editing lexicography. Brno, pp. 440–454.

Nazar, R./Lindemann, D. (2022): Terminology extraction using co-occurrence patterns as predictors of semantic relevance. In: Proceedings of the Workshop on Terminology in the 21st century (Term21), LREC 2022, Marseille, France.

This paper is part of the publication: Klosa-Kückelhaus, Annette/Engelberg, Stefan/Möhrs, Christine/Storjohann, Petra (eds.) (2022): Dictionaries and Society. Proceedings of the XX EURALEX International Congress. Mannheim: IDS-Verlag.
https://doi.org/10.14618/phpy-6r66

Pons, S. (2001): Connectives/discourse markers. An overview. In: Quaderns de Filologia. Estudis Literaris 6, pp. 219–243.

Robledo, H./Nazar, R. (2018): Clasificación automatizada de marcadores discursivos. In: Procesamiento del Lenguaje Natural (61), pp. 109–116.

Roze, C./Danlos, L./Muller, P. (2012): LEXCONN: a French lexicon of discourse connectives. In: Discours – Revue de linguistique, psycholinguistique et informatique. Laboratoire LATTICE, 2012. Multidisciplinary perspectives on signalling text organisation, pp. 1–15. https://hal.inria.fr/hal-00702542.

Santos Río, L. (2003): Diccionario de partículas. Salamanca.

Schriver, K. (2012): What we know about expertise in professional communication. In: Berninger, V. W. (ed): Past, present, and future contributions of cognitive writing research to cognitive psychology. New York, pp. 275–312.

Smith, D. E./de Schryer, C. F. (2013): On documentary society. In: Bazerman, Ch. (ed.): Handbook of research on writing. Amsterdam/New York, pp. 113–117.

Stede, M. (2002): DiMLex: a lexical approach to discourse markers. In: Lenci, A./Tomaso, V. D. (eds.): Exploring the lexicon – theory and computation. Alessandria.

Stede, M./Scheffer, T./Mendes, A. (2019): Connective-Lex: a web-based multilingual lexical resource for connectives. In: Discours – A Journal of Linguistics, Psycholinguistics and Computational Linguistics 24. https://journals.openedition.org/discours/10098.

Stubbs, M. (1996): Text and corpus analysis. Oxford.

Tiedemann, J. (2012): Parallel data, tools and interfaces in OPUS. In: Proceedings of the Eighth International Conference on Language Resources and Evaluation (LREC'12). Istanbul, pp 2214–2218.

van Dijk, T. (1973): Text grammar and text logic. En Studies in Text Grammar. Dordrecht, pp. 17–78.

van Dijk, T. (1983): La ciencia del texto: un enfoque interdisciplinario. Barcelona.

Versley, Y. (2010): Discovery of ambiguous and unambiguous discourse connectives via annotation projection. In: Proceedings of the Workshop on Annotation and Exploitation of Parallel Corpora (AEPC). Tartu, pp. 83–92.

Webber, B./Prasad, R./Lee, A./Joshi, A. (2019): The Penn Discourse Treebank 3.0 annotation Manual. tech report. University of Pennsylvania.

Zhou, L./Gao, W./Li, B./Wei, Z./Wong, K.-F. (2012): Cross-lingual identification of ambiguous discourse connectives for resource-poor language. In: Proceedings of the 24th International Conference on Computational Linguistics (COLING 2012). Mumbai, pp. 1409–1418.

Contact information

Irene Renau
Pontificia Universidad Católica de Valparaíso, Chile
mailto:irene.renau@gmail.com

Rogelio Nazar
Pontificia Universidad Católica de Valparaíso, Chile
mailto:rogelio.nazar@pucv.cl

This paper is part of the publication: Klosa-Kückelhaus, Annette/Engelberg, Stefan/Möhrs, Christine/Storjohann, Petra (eds.) (2022): Dictionaries and Society. Proceedings of the XX EURALEX International Congress. Mannheim: IDS-Verlag.
https://doi.org/10.14618/phpy-6r66