Electronic lexicography in the 21st century: **post-editing lexicography**

Proceedings of the eLex 2021 conference

edited by

Iztok Kosem
Michal Cukr
Miloš Jakubíček
Jelena Kallas
Simon Krek
Carole Tiberius

virtual, 5–7 July 2021

elex.link/elex2021

Electronic lexicography in the 21st century (eLex 2021):
Post-editing lexicography
Proceedings of the eLex 2021 conference.

edited by
Iztok Kosem
Michal Cukr
Miloš Jakubíček
Jelena Kallas
Simon Krek
Carole Tiberius

published by
Lexical Computing CZ s.r.o., Brno, Czech Republic

proofreading by
Paul Steed
Dean DeVos
Jan Nagel

Brno, July 2021

ISSN 2533-5626

eLex 2021

ORGANIZERS

LEXICAL COMPUTING

cjvt

Univerza v Ljubljani

/instituut voor de Nederlandse taal/

Institute of the Estonian Language

elexis
european lexicographic infrastructure

elex.link/elex2021

CONFERENCE COMMITTEES

Organising Committee

Iztok Kosem
Michal Cukr
Miloš Jakubíček
Jelena Kallas
Simon Krek
Carole Tiberius

Scientific Committee

Andrea Abel
Špela Arhar Holdt
Kristian Blensenius
Gerhard Budin
Nicoletta Calzolari
Lut Colman
Paul Cook
Margarita Correia
Gilles-Maurice de Schryver
María José Dominguez Vazquez
Patrick Drouin
Edward Finegan
Thierry Fontenelle
Polona Gantar
Yongwei Gao
Radovan Garabik
Zoe Gavriilidou
Alexander Geyken
Kris Heylen
Aleš Horák
Miloš Jakubíček
Maarten Janssen
Jelena Kallas
Ilan Kernerman
Maria Khokhlova
Annette Klosa-Kückelhaus
Svetla Koeva
Kristina Koppel
Iztok Kosem
Vojtěch Kovář
Simon Krek
Michal Kren
Tanara Zingano Kuhn
Margit Langemets
Lothar Lemnitzer
Robert Lew
Pilar León Araúz
Henrik Lorentzen
Stella Markantonatou
John P. McCrae
Amalia Mendes
Michal Boleslav Měchura
Julie Miller
Monica Monachini
Orion Montoya
Sara Može
Christine Möhrs
Chris Mulhall
Carolin Müller-Spitzer
Lionel Nicolas
Sussi Olsen
Vincent Ooi
Jordi Porta
Adam Rambousek
Laurent Romary
Hindrik Sijens
Emma Sköldberg
Nicolai Hartvig Sørensen
Egon Stemle
Vít Suchomel
Kristina Štrkalj Despot
Arvi Tavast
Carole Tiberius
Yukio Tono
Lars Trap Jensen
Agnes Tutin
Tamas Varadi

TABLE OF CONTENTS

Automatic induction of a multilingual taxonomy of discourse markers

Rogelio Nazar

Instituto de Literatura y Ciencias del Lenguaje
Pontificia Universidad Católica de Valparaíso
rogelio.nazar@pucv.cl

Abstract

This paper describes a proposed method for the identification and classification of discourse markers (e.g., *however, therefore, by the way*) by applying statistical analysis to large parallel corpora. The objective is to build a lexical resource consisting of a multilingual taxonomy, so far in English, Spanish, German and French. A method is proposed that first separates discourse markers from the rest of the lexical units in the corpus using a measure of entropy, and then classifies them in groups by function using a clustering procedure especially designed for massive data processing. From that point onwards, the system is used to recursively identify and classify more units. Experimental evaluation shows that, in terms of precision, the automated method is able to perform as well as a team of human annotators (undergraduate students of linguistics), and it outperforms them in terms of recall.

Keywords: automatic creation of dictionary content; connectives; discourse markers; taxonomy induction; natural language processing

1. Introduction

This paper presents the first results of a lexicographic research project aimed at cataloging discourse markers (DMs) by means of statistical analysis of large parallel corpora. It describes a newly developed algorithm for the automatic induction of a multilingual taxonomy of DMs, which is then used to recursively identify and classify more units. The objective of the research is to obtain an exhaustive inventory of DMs of different languages. Some preliminary results are described, including a classifier of DMs and a first version of the multilingual taxonomy, so far in English, Spanish, German and French.

The method is solely based on the exploitation of parallel corpora by statistical algorithms. There is no human intervention in the process chain, and no external resources are used, such as POS-taggers or dictionaries. The reason for disregarding external resources, even when such resources are available for the languages considered in the present research, is in part for scientific parsimony but also to facilitate replication of experiments in other, possibly less resourced, languages. One has to take into consideration, too, that one of the outcomes of a purely corpus-based approach is that it may lead to the detection of new units, those that are currently in use in the texts but have not yet been added to dictionaries.

The method uses only co-occurrence association measures and an entropy model to identify DMs according to their distribution in the corpus. As DMs are independent of the content of the texts in which they appear, their occurrence in texts cannot be used to predict the occurrence of other units. Once they are separated from the set of vocabulary units, they are then grouped together using a clustering method which uses their shared equivalence in other languages as a similarity measure. The algorithm will classify new candidates by language, will then decide if they are effectively DMs and, if that is the case, it will assign them to a category.

The identification and subsequent classification of DMs is an extremely difficult task due to various factors. Even for humans (and, indeed, for specialists) it is not always clear

where the distinction between DMs and the rest of the lexical units lies, and the definition of the concept varies according to authors and theories. This is due to several reasons. Among them, there is the polyfunctionality of DMs (Pons Bordería & Fischer, 2021), i.e. the fact that the same unit can have a DM function in some contexts but not in others, and even that the same unit can have different DM functions depending on the context. Other factors that further complicate any attempt to determine a clear-cut distinction is that, while some of them operate at the discourse level (one of their characteristic features), others instead seem to be more integrated into the syntactic structure. In part, this is one of the reasons why it is important to conduct empirical research on the subject, especially when the field is dominated by theoretical approaches that rely heavily on introspection or with corpus-based research but with hand-picked examples.

The method's performance varies by language. It is fairly successful in English, Spanish and French, but less so in German, where it has been only moderately successful. On the whole, however, the results of the approach are promising, especially when a preliminary evaluation with Spanish results shows that the method outperforms a group of human annotators. This is a remarkable achievement considering that it is an extremely minimalist approach, one which is computationally inexpensive and has no dependency on linguistic resources other than a parallel corpus. In its current form, the method could be of interest to lexicographers working on DMs, for researchers applying algorithms to automate some levels of discourse analysis, and also for final users, such as translators or people writing in a first or a second language.

2. Related work

In recent years, linguistic theorists have turned their attention to DMs, with an increasing number of publications being devoted to the subject (Fraser, 1999; Pons Bordería, 2001; Schiffrin, 2001). The topic, however, is by no means new in linguistics, and appears in some early grammars, especially of the Spanish tradition. For instance, grammarians such as Antonio de Nebrija, Gregorio Garcés o Andrés Bello in the 15th, 18th and 19th centuries, respectively (Casado Velarde, 1993; Pons Bordería, 2001) all make reference to DMs in their works; more recently there is Gili Gaya (1943), who discusses DMs, albeit using different terminology.

Greater interest in the subject began to appear much later, with the advent of discourse analysis, and more specifically in the field of text grammars. Early work by van Dijk (1973), for instance, presents the main functions of what he then called connectives, which mark the logical relations between propositions, such as conjunction, disjunction, causality, condition, concession, contrast, purpose and so on. A few years later, Halliday & Hasan (1976) presented a developed categorisation of what they call conjunctive relations, with additive, adversative, causal and temporal markers, as well as other continuative or conversational units. A final important historical precedent in the study of DMs is the analysis of connectives in the field of argumentation theory by Anscombre & Ducrot (1976). They notably pointed out that the absurdity of an example such as (1) is a consequence of the use of the expression *même* ('even'):

(1) # *Une mule vaut mieux qu'un âne, même mauvais.*
(A mule is better than a donkey, even a bad donkey).

DMs are perceived to be a driving force behind the proliferation of text grammars, as they were a subject for which earlier linguistic theories proved inadequate. As Stubbs (1983: 77) puts it, DMs "provide problems for sentence based grammars, but are of great interest in a study of discourse sequences, since their functions are largely to do with the organization of connected discourse, and with the interpretation of functional categories of speech acts".

The following years saw a profusion of publications dealing with DM's defining properties and attempting to delineate their boundaries and categorisations. DMs are, probably, a universal feature of language, but they are not easily defined as a single class. They have been defined as particles that facilitate the interpretation of coherence relations in texts (Fraser, 1999; Pons Bordería, 2001). That is to say, they are instructions on how to connect propositions and organise argumentation. It must be noticed, however, that coherence relations between propositions can be inferred even in the absence of DMs, and therefore they are considered optional. However, their presence facilitates comprehension and reduces the chances of ambiguity. They also have an important function in facilitating the interaction between participants, so they have an interpersonal value beyond their textual one, by signalling changes of subject or turn taking (Mosegaard Hansen, 1998). In this sense, one must consider DMs in the context of other pragmatic particles with an interpersonal function, such as interjections, modal particles, focus particles, conjunctions, etc.

In terms of their morphology, they are formally mostly invariable. They have no inflection, do not admit modifiers and cannot be negated or coordinated (Martín Zorraquino & Portolés, 1999). They can pertain to different categories, such as conjunctions, adverbs, prepositional phrases, idioms, and so on.

Regarding their syntactic nature, Schiffrin (2001) describes them as utterance-initial and syntactically independent, although this is perhaps a too restrictive characterisation that would leave out many valid DMs. But it is true that they often are parenthetical and seem to be outside of the syntactic structure of the sentence. More critically, they do not participate directly in the sentence's propositional content, but rather affect the whole sentence or the relation between the sentence and other chunks of text. Their scope varies across different levels of discourse (Pons Bordería, 2001; Brinton, 2010).

In terms of their semantics, they have procedural rather than semantic content, i.e., no referential, propositional or truth value. Historically, though, they derive from lexical units that did have these properties (Traugott & Dasher, 2002), but lost them due to a process of grammaticalisation. It is therefore said that their propositional content has been gradually 'bleached' (Wichmann & Chanet, 2009).

DMs can be organised according to function. One of the most common classifications is counter-argumentation, with expressions such as *however* or *nevertheless*, among others. These are intended to alert the reader/listener that the following propositions will not be what might be expected based on what came before it. Other common functions are to make a cause-consequence relation explicit, such as *consequently* or *therefore*. In their well-known taxonomy, Martín Zorraquino & Portolés (1999) describe a series of broad categories that then divide into branches. Among the main classes we find the structuring type (e.g. *on the one hand, on the other, finally*), connectives (e.g. *moreover, furthermore, in the same way*), reformulatives (e.g. *in other words, better said*), and others. This

categorisation has been extremely influential not only in the Spanish tradition, but in other languages as well, e.g. in German (Blühdorn et al., 2017).

The vast majority of the literature on DMs has been devoted to the qualitative study of individual cases, e.g. Urgelles-Coll (2010) in the case of the English DM *anyway* or Llopis-Cardona (2014) in the case of several DMs in Spanish. Fewer are the attempts to compile extended lists of DMs. Two exceptions are Knott (1996) and Stede (2002) who took on this task in English and German, respectively. More work was carried out later in the case of Spanish, for instance dictionaries such as those by Santos Río (2003) or Briz et al. (2008). Recent years have seen an increase in activity in this area. For instance, the material provided by Roze et al. (2012) for French, Feltracco et al. (2016) for Italian, Mírovský et al. (2017) for Czech and Mendes et al. (2018) for Portuguese. Special mention must be made of the contribution by Stede et al. (2019), who are centralising a multilingual taxonomy of DMs in a single database: http://connective-lex.info/.

The computational linguistics community that deals with discourse analysis has paid comparatively less attention to the topic of DMs, Stubbs (1996) being among the exceptions. When these researchers do mention DMs, they use different terminology to refer to them, for instance "discourse cues" (Moore & Wiemer-Hastings, 2003). The field has seen a renewed interest in DMs as of late, in part motivated by recent progress in the field of discourse parsing (Xue et al., 2016), but there is still much to be done. Lopes et al. (2015: 1), for instance, note that "little has been said on their cross-language behavior and, subsequently, on building an inventory of multilingual lexica of discourse markers".

A driving force in this renewed interest seems to be the application of parallel corpora and machine translation. Versley (2010) used an English-German parallel corpus to transfer linguistic annotations from English to German. In a similar way, Lopes et al. (2015) used machine translation to obtain a list of equivalent DMs in different languages from an original list of 427 markers in English.

Also using parallel corpora, but taking a different approach, one similar to that being presented in this study, Robledo & Nazar (2018) described a method based on clustering to offer a bottom-up taxonomy of Spanish DMs. There, as in the current paper, the functional equivalence of different DMs is based on their shared translation as shown in the corpus alignment. Using that method, 587 Spanish DMs were obtained, with evaluation figures showing 0.93 precision and 0.78 recall in the task of identifying false DMs in a list with mixed genuine and false items. A limitation is that the method requires a variety of language-dependent resources, such as POS-taggers, syntax-based rules to filter out improbable candidates and a *gazetteer* used as a stoplist for the same purpose. The main drawback, however, is the hierarchic clustering method that is used. Based on a distance matrix, it entails great computational expense when dealing with large datasets.

More recently, Sileo et al. (2019) used a curated list of 174 markers for English in order to discover sentence initial, parenthetical, high-frequency DMs using contextual cues (word ngrams). After a complex and computationally expensive machine learning procedure involving sentence selection, tokenising, tagging and finally classification with the Fasttext library, they discovered 243 DM candidates, but their results are modest in terms of accuracy.

This study continues in the same vein as the aforementioned ones in that it is an empirical method, based on the statistical analysis of large corpora. The difference is that the present one is comparatively a very simple method, and with a focus on a multilingual and language-agnostic approach. With regards to earlier qualitative studies on DMs, the main difference is that the present one is an empirical method, i.e., a bottom-up rather than a top-down approach. This is important for practical reasons, as the automation saves a lot of effort, but also, and most importantly, for scientific reasons, as the quantitative method favours objectivity. Also, in contrast with the manually compiled DM lexicons existing today, which comprise only a few hundred entries, in this project thousands of them are discovered, which are offered to the public in an open database online. All these are reasons to believe that the present paper represents a substantial contribution to the state of the art on DM research methodology.

3. Methodology

As already anticipated, the methodology consists of first identifying DMs in corpora by separating them from the rest of the vocabulary and then classifying them in a bottom-up functional taxonomy. It is a minimalist approach based solely on statistical measures and without any type of external resource apart from a parallel corpus. Section 3.1 explains how DMs are identified according to their distribution in the corpus by exploiting one of their characteristics, which is to be independent of the content of the texts in which they appear. In operational terms, this means that their occurrence cannot be used to predict the occurrence of other lexical units. Section 3.2 describes the subsequent step, i.e. their classification, which is performed using an original clustering algorithm. Section 3.3 shows how the clusters are tagged and organised. Finally, section 3.4 explains how, once this core taxonomy is built, it is then used to further populate it by classifying new DMs obtained from corpora in a recursive manner.

3.1 Separating DMs from the rest of the vocabulary

The same parallel corpus was used for all steps of the procedure: the Opus Corpus (Tiedemann, 2012), a large collection of parallel corpora in different languages, freely available and organised by corpus in different TMX files, a standard format in the field of translation. The number of corpora varies according to the language pairs, but is close to 30 files per pair. Each corpus presents a different specialised technical domain and/or discourse genre. It is aligned at 'translation units', which generally correspond to sentences but sometimes larger segments, like paragraphs. The corpus does not include lemmatisation or POS-tagging annotations but that is not a problem since such data is not needed for the method presented here.

For the first step, only the target language segment is used, ignoring the alignments. An initial set of vocabulary units is obtained from the corpus by sorting ngrams, defined as sequences of one, two and three words not including punctuation marks. These are not used as a means to determine the boundaries of the ngrams because doing so would lead to the obtainment of only parenthetical DMs, which are merely a subset of all existing DMs. Moreover, DMs do not behave in this way in all languages. For instance, German DMs are not used parenthetically as frequently as in the other languages.

The result is a very large initial vocabulary set, denoted as $InVoc$, which is then reduced in size in subsequent steps by filtering units according to their distribution in the corpus

and according to a measure of information. As DMs are procedural instead of semantic, that means that their appearance in a text is not related to the semantic content and they cannot be used to predict the co-occurrence of other vocabulary units. Thus, a subset of $InVoc$ called $FiVoc$ contains units that appear in at least seven of the 30 TMX files with a minimum frequency of 50 occurrences, all thresholds being arbitrary but empirically motivated.

This first operation results in a dramatic decrease in the size of the vocabulary lists, from an average of half a million units per language to fewer than 5,000. Yet, fewer than a third of the latter are genuine DMs, as the majority of these are words or sequences of words bearing a very general semantic content. In the case of English, these would be high frequency words such as *property* or *language* as well as names of places like cities or countries (e.g. *Paris, the Netherlands*), among others. As a consequence, a more refined procedure is then applied, which is computationally more expensive but justifiable considering that it is applied to only a few thousand units.

The second filtering operation consists of determining a measure of information of the candidates. This measure aims to tell how informative a word is in relation to its ability to predict the appearance of other words. A word with a clear semantic content, e.g. *Paris*, should exhibit a tendency to co-occur in large numbers of contexts with other units that are semantically related, e.g. *France*. The contrary would be the case of the units we are interested in, the DMs, which should score very low with this type of measure. Therefore, given a target unit x, it is possible to obtain a set $M(x)$ consisting of a sample of contexts of occurrence of x from the corpus and then sort all the vocabulary units[1] in a ranking R_x by decreasing order of frequency. One can then use the relation between this frequency and the sample size in order to obtain a distinction between semantic and procedural units. The coefficient used to calculate this is shown in (1). The parameter n is arbitrary, but experimentally fixed at 20. The decision to accept or reject x as a member of the candidate set C is based on another empirically parameter t, as shown in (2). Alternatively, one could also keep the best k candidates in C.

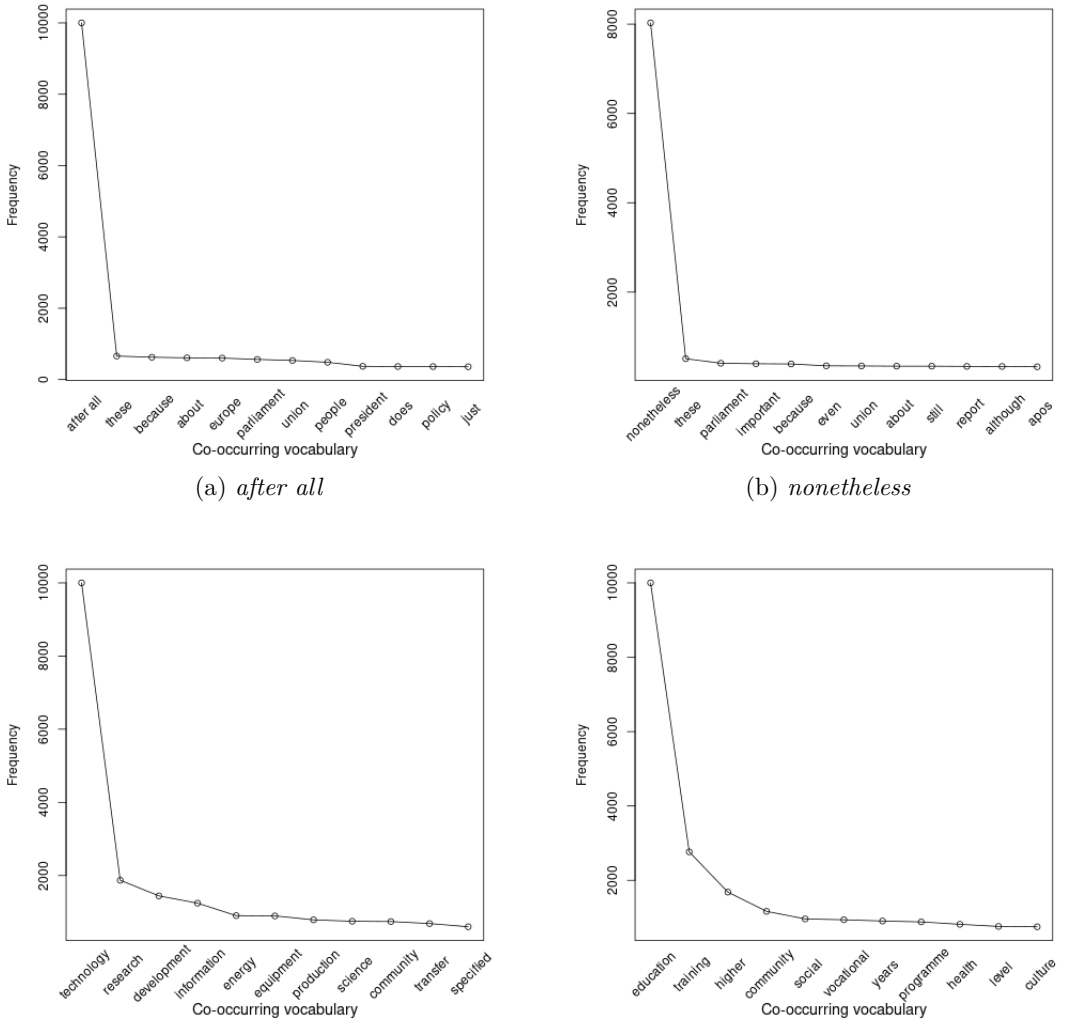
$$I(x) = \frac{\log_2 \sum_{i=1}^{n} R_{x,i}}{\log_2 |M(x)|} \tag{1}$$

$$x \in C = \begin{cases} true & I(x) < t \\ false & \text{otherwise} \end{cases} \tag{2}$$

For illustration, Figure 1 presents how it is possible to obtain an almost clear-cut separation between the two classes. Functional units such as *after all* (Panel a) or *nonetheless* (Panel b) are very different from semantically-charged vocabulary units such as *technology* (Panel c) or *education* (Panel d), and the difference is revealed by their co-occurrence pattern. E.g., in the case of *technology*, one can say that if this word is found in a sentence, then there is a relatively high probability of finding other words[2],

[1] The units considered here are only single-words instead of word-ngrams. This is done this way for simplicity and to reduce computational cost, but the possibility of using larger-than-word units is worth exploring in future research.

[2] Function words (i.e., those that would appear in any random sentence such as *with, that, from, this,* etc.) are also ignored precisely because they are themselves very uninformative



(a) *after all* (b) *nonetheless*

(c) *technology* (d) *education*

Figure 1: The shape of the co-occurrence frequency curve is used to predict the semantic or procedural nature of lexical units

such as *research, development, information* and so on. This does not happen in the case of DMs. An item like *after all* shows an extremely low frequency of co-occurrence with other units. Thus, finding this phrase in a sentence does not make it possible to predict the occurrence of any other lexical item.

3.2 Induction of a functional taxonomy of DMs

The previous phase yielded a set $C(l)$ of DM candidates for each language l (en, fr, es, de). In this phase, in turn, for each l, a functional taxonomy of DMs will be created in the form of a hierarchic clustering, for which the parallel corpus is used. At this point, languages are paired together. It is irrelevant which languages are used in each pair, but for practical reasons English is used as one of the languages for each pair, as it is usually the language for which more material is available. Thus, with the English-French pair, for instance, the algorithm produces an alignment of sets C_{en} and C_{fr}. The alignment of the units in both lists can be achieved with the use of a co-occurrence measure such as $A(i, j)$, shown in (3).

$$A(C_{en,i}, C_{fr,j}) = \frac{f(C_{en,i}, C_{fr,j})}{\sqrt{f(C_{en,i})}.\sqrt{f(C_{fr,j})}} \tag{3}$$

This coefficient compares the frequency of co-occurrence of the vocabulary units in the aligned segments with their independent frequency in the whole corpus. Thus, if, for instance, $C_{en,i}$ is *nonetheless* and $C_{fr,j}$ is *néanmoins*, the algorithm contrasts the number of times they appear in translated sentences with the number of times they appear in general, that is, alone or together. For each unit in C_{en} there will be a limited number of equivalent candidates in C_{fr}. The top three candidates, as long as they have a score greater than 0.20, are kept. This parameter is again arbitrary but empirically defined.

The purpose of aligning the DM candidates in this fashion is only to allow for their organisation in a taxonomy, a result that is achieved by means of a clustering procedure. This procedure is conducted using the aligned pairs as a similarity measure, i.e., two units are considered similar for the clustering if they share the same equivalent markers in the parallel corpus. To continue with the same example, English items like *nonetheless* and *nevertheless* are considered similar because they share the same equivalence in a second language, such as *néanmoins* in the case of French.

The exact procedure of the clustering is as follows. It consists of a greedy-matching, graph-based clustering algorithm that has the property of being very efficient in comparison with regular hierarchic clustering algorithms such as those used in previous studies (Robledo & Nazar, 2018), which suffer from quadratic complexity and are not scalable to many thousands of objects. The option applied here is simpler, and is called 'the cocktail-party algorithm'. One often sees, at conference cocktail parties or coffee-breaks, that people tend to cluster together as they arrive on the basis, at least initially, of their mutual acquaintance. If the DM candidates have been aligned, one can imagine them as people coming to the cocktail in pairs. For instance, first Paul (*nonetheless*) and Eva (*néanmoins*) arrive together, followed by Robert (*of course*) and María (*évidemment*), who also arrive together. The two pairs do not know each other, so they stay apart and keep to themselves. Then, however, Eva sees that Michael (*nevertheless*) just arrived, and

since she already knows him (*néanmoins* and *nevertheless* were also found to be equivalent according to the parallel corpus), she introduces him to Paul. Now, Paul, Eva and Michael form a single cluster, as depicted in Figure 2.

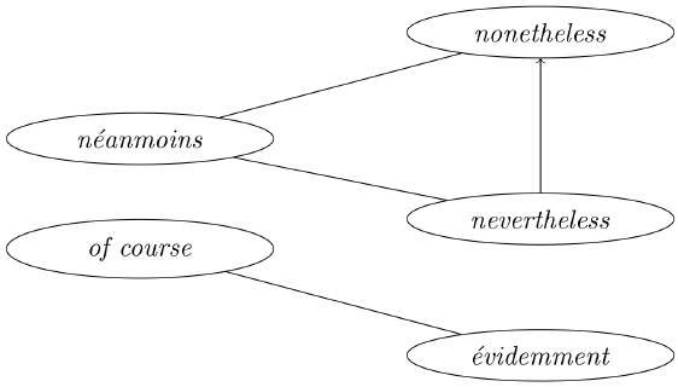


Figure 2: Illustration of a moment of the graph-based clustering process

If someone else arrives and knows at least one of these three in the cluster, she will also join the group, unless she finds another group with more acquaintances. This process goes on, and more clusters are produced during the event as more people/DMs arrive, and the result will be a bilingual taxonomy.

3.3 Tagging the clusters

One limitation of the taxonomy created so far is that clusters have no meaningful names. They are identified by numeric codes that bear no relation to their content. Also, there is the problem that some of these clusters should be grouped in order to form larger categories. Since it would be too laborious to manually tag each cluster with a name, it was decided to resort to an automatic tagging procedure based on the taxonomy originally proposed by Martín Zorraquino & Portolés (1999) because, as already mentioned, it has been extensively used, even in languages other than Spanish.

Using the examples provided by these authors, a matching algorithm was developed to tag a given cluster from the induced taxonomy with the names of the categories they provide. For example, if there is a cluster that consists of contrastive connectors, it will probably include some of the examples mentioned by those authors, such as *sin embargo, no obstante*, etc. Thanks to these shared examples, the tagging algorithm can recognise the relationship between the cluster and said category and confidently assign a meaningful name to each cluster.

As the examples are in Spanish, the Spanish side of the taxonomy has to be used to do the tagging. But, since all the taxonomy is multilingually aligned, a tag assigned to a cluster in one side of the taxonomy is inherited by the other sides as well. The tagging also has the effect of aggregating similar clusters in larger categories.

In any case, the content of the clusters is kept separate, although hierarchically organised. For example, there is one broad category in the terminology of Martín Zorraquino & Portolés (1999) called *Estructuradores de la información*, referring to DMs used

for information structuring, and within this category there is a subcategory called *Comentadores*, referring to DMs used to introduce commentary. It happens that this algorithm finds new divisions within this category, and there are different clusters under the tag of *Comentadores*. For example, one of these clusters contains DMs like *arguably, certainly, presumably, probably*, among other units, while another contains DMs such as *at this point, at this stage, at this time*, etc. Keeping them separate allows one to obtain a layered categorisation, which in turn can be used as the basis for the further categorisation of new DMs.

3.4 Further population of the taxonomy

Once a basic or core multilingual taxonomy of DMs has been obtained (hereinafter *Dismark*), it is then possible to use such material as the basis for the categorisation of new DMs, done recursively. For this final part of the procedure, an input candidate x is needed ($x \notin Dismark$) for the algorithm to perform the following three subtasks:

1. Classify x by language
2. Decide if x is effectively a DM
3. If 2 is true, assign x to a category in *Dismark*

For subtask 1, one is of course limited to the available languages. The algorithm will retrieve contexts of occurrence of x in the corpora of the different languages and select the one with the highest number of hits. For subtask 2 it will use the parallel corpora. If x appears in the aligned sentences with other DMs already registered in the taxonomy, then this is taken as indication that x is a true DM. Once this has been decided, the algorithm has to find the best matching category for x, and this is done in a way reminiscent of the method explained in Section 3.3. That is, using the equivalences for x in a different language that were just obtained from the parallel corpus, the best category is selected on the basis of their matching. For instance, if x is *in that sense* and is not already in *Dismark*, its analysis in the parallel corpus will reveal that valid French equivalents are, among others, units like *à cet égard* and *dans ce sens*, which are already in the taxonomy. If this is the case, then the algorithm can safely place x on the English side of this cluster.

This taxonomy operates automatically and without supervision. Moreover, the larger the taxonomy becomes, the better the result of its predictions because it has a better knowledge base. Thus we can see how, from nothing more than a parallel corpus and a set of category names for the clusters, it is possible to obtain a taxonomy of DMs thanks to a system that is characterised by a virtuous cycle and that can incrementally improve in precision and thoroughness.

4. Evaluation

At the time of writing, the database contains a total of 2,463 different DMs classified in 20 different categories and 71 subcategories. Tables 1 and 2 show examples of two clusters belonging to different categories. These are meant to be read as groups of DMs that are functionally equivalent, and no correspondence is implied in their horizontal alignment. They share the same cluster simply because they can be used with the same function.

English	**Spanish**	**French**	**German**
• in a manner similar • in a similar manner • in the same manner • in the same way • likewise • similarly	• de forma similar • de la misma forma • de la misma manera • de manera similar • de modo similar • del mismo modo • forma similar • manera similar	• de la même façon • de la même manière • de même • même façon • même manière	• auf dieselbe Weise • desgleichen • dieselbe Weise • ebenso • gleiche Weise • gleichen Weise • in ähnlicher Weise • ähnlicher Weise

Table 1: An example of a subcategory (cluster) of the category 'additive connectives'

English	**Spanish**	**French**	**German**
• after all • at last • at some point • at some time • at the end • but after all • eventually • in a few words • in a word • in brief • in short • in sum • in summary • in the end • on balance • sooner or later • to sum up • to summarise • ultimately • upon the whole	• a fin de cuentas • a la larga • al final • así pues • de forma resumida • después de todo • en algún momento • en definitiva • en fin • en pocas palabras • en resolución • en resumen • en resumidas cuentas • en suma • en una palabra • en última instancia • en último término • eventualmente • tarde o temprano	• après tout • au bout du compte • au final • en bref • en définitive • en fin de compte • en résumé • en somme • enfin • finalement • forme résumée • forme résumée ou agrégée	• am Ende • erweitert • irgendwann • kurz gefasst • kurz gesagt • kurzum • letzten Endes • letztendlich • letztlich • schließlich

Table 2: Another example of subcategory (cluster) of the category 'recapitulation connectives'

The first look on the results reveals that there is a considerable mismatch in quality between languages. While the results on English, Spanish and French seem very impressive (on average 95% of the DMs are correct), in German, instead, one can claim only that there has been moderate success, with 84% of the DMs being correct. A worse performance in German was in part to be expected, as this language presents more challenges for automatic processing. This is due to the fact that the syntactic behaviour of DMs in German is different from the other languages regarding position, punctuation and the use of cases (e.g. nominative, accusative, dative). Many of the problems were also related to segmentation faults (e.g., the system retrieves *solchen Fällen* instead of the correct form *in solchen Fällen*).

In order to offer a more precise evaluation, we conducted a small experiment to compare the performance of the algorithm with a group of human annotators in the task of identifying DMs. After a university semester course on Text Grammar which deals extensively on the subject of DMs, seven of the best performing students were selected to participate in the task. Their training consisted of both theoretical lessons on the subject and practical exercises in which they had to identify and classify DMs using the taxonomy by Martín Zorraquino & Portolés (1999).

For the task, the annotators received a list of 709 expressions, roughly two thirds of which were mixed DMs and one third of which were lexical units of other types, in alphabetical order. The students, unaware of the composition of the list, were asked to place a number one beside every unit that they considered not to be a DM. They were asked to perform the task alone, without asking their classmates, and to refrain from using corpora, dictionaries or any other type of lexicographic resource. It was emphasised to them that they should follow their intuition. Table 3 shows the results.

Annotator	**Precision**	**Recall**	**F1**
Dismark	97	94	95
Student 1	96	51	66
Student 2	95	61	74
Student 3	95	41	57
Student 4	94	59	72
Student 5	93	66	77
Student 6	92	32	47
Student 7	91	75	82

Table 3: Comparing the performance of algorithm vs. humans in the task of identifying DMs

In general, they all performed fairly well in terms of precision, and as the table shows, when they selected something as a DM, they were almost always correct. They tended, however, to be more conservative. A series of follow-up interviews with the students revealed that they were unwilling to select something as a DM unless they were very sure it was one. That is, the students marked DMs that were prototypical, meaning highly grammaticalised and showing no sign of morphological variation. They tended to reject genuine cases such as *en estas circunstancias* ('in these circumstances') or *en términos más generales* ('in broader terms').

Another reason for them to reject genuine DMs was the fact that they found them too polysemous or polyfunctional, in the sense that they were elements that could function as DMs but only in certain contexts. In this regard, the lack of lines of context certainly put humans at a disadvantage. An interesting direction for future research would be to present the participants with the task of detecting DMs in a particular text. This, however, would be a different type of research, because it would not be about classifying DMs in abstract. Instead, its focus would be the classification of particular instances of DMs. That would require totally different sets of measures, such as contextual cues, to determine in which contexts something is used as a DM and in which not. Such an endeavour would be out of the scope of a lexicography project and closer to the area of discourse analysis.

At any rate, what is to be learned from this experiment is that distinguishing between a DM and a non-DM element is not an easy task and that, perhaps, the way forward would be to follow the same criterion as Rysová & Rysová (2018) with the Prague Discourse Bank. This would be to establish a distinction between primary DMs, with those more prototypical or grammaticalised units, and other categories with secondary and free DMs, to accommodate those units that fulfil the same function but are less prototypical.

5. Conclusions

This paper presented a newly developed method for the automatic induction of a multilingual taxonomy of DMs, including a description of its first results. The method is simple and effective. It is also computationally inexpensive and easy to replicate in different languages. The method is, in fact, robust to language varieties, as it could provide useful results even in German, which is, morphologically speaking, a language very different from the others.

Also, in comparison with manually curated classifications of DMs, which in most cases offer a few hundred items, the multilingual taxonomy already offers thousands of them, including items of medium to low frequency in the corpus. The results of the project, including the full database of DMs and a demo for the DM classifier, are offered at the project's website[3]. Even though this is still work in progress, the results currently available can be useful for lexicographers interested in DM projects as well as NLP professionals working on text understanding or text generation. Final users, such as writers or translators, can benefit from this collection in order to improve vocabulary richness.

With respect to future research, the priorities would be the following: 1) to continue evaluating and exploring variations in the method; 2) to continue populating the taxonomy with new, maybe less frequent items and 3) to incorporate new languages, first from Europe and later from other language typologies, taking advantage of the fact that no external resources are needed.

6. Acknowledgments

This paper has been made possible thanks to funding from a research grant by the Government of Chile: *Proyecto Fondecyt Regular 1191481. Inducción automática de taxonomías de marcadores discursivos a partir de corpus multilingües* (Automatic induction of taxonomies of discourse markers from multilingual corpora). I would like

[3] http://www.tecling.com/dismark

to thank the anonymous reviewers for pointing out different ways to improve the paper. I would also like to express my gratitude to Maureen Noble for proofreading and for her valuable comments.

7. References

Anscombre, J.C. & Ducrot, O. (1976). L'argumentation dans la langue. *Langages*, 42, pp. 5–27.

Blühdorn, H., Foolen, A. & Loureda, O. (2017). Diskursmarker: Begriffsgeschichte – Theorie – Beschreibung. Ein bibliographischer Überblick. In H. Blühdorn, A. Deppermann, H. Helmer & T. Spranz-Fogasy (eds.) *Diskursmarker im Deutschen. Reflexionen und Analysen*. Göttingen: Verlag für Gesprächsforschung.

Brinton, L. (2010). Discourse Markers. In A. Jucker & I. Taavitsainen (eds.) *Historical Pragmatics*. Berlin: Gruyter Mouton.

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

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.

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

Halliday, M. & Hasan, R. (1976). *Cohesion in English*. London: Longman.

Knott, A. (1996). *A data-driven methodology for motivating a set of coherence relations*. Ph.D. thesis, University of Edinburgh, UK. British Library, EThOS.

Llopis-Cardona, A. (2014). *Aproximación funcional a los marcadores discursivos. Análisis y aplicación lexicográfica*. Frankfurt am Main: Peter Lang.

Lopes, A., de Matos, D.M., Cabarrão, V., Ribeiro, R., Moniz, H., Trancoso, I. & Mata, A.I. (2015). Towards Using Machine Translation Techniques to Induce Multilingual Lexica of Discourse Markers. arXiv 1503.09144.

Martín Zorraquino, M.A. & Portolés, J. (1999). Los marcadores del discurso. In I. Bosque & V. Demonte (eds.) *Gramática Descriptiva de la Lengua Española*. Madrid: Espasa, pp. 4051–4214.

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: European Language Resources Association (ELRA).

Moore, J.D. & Wiemer-Hastings, P. (2003). Discourse in Computational Linguistics and Artificial Intelligence. In A.C. Graesser, M.A. Gernsbacher & S.R. Goldman (eds.) *Handbook of Discourse Processes*. Routledge.

Mosegaard Hansen, M.B. (1998). *The Function of Discourse Particles : A study with special reference to spoken standard French*. Amsterdam/Philadelphia: John Benjamins.

Mírovský, J., Synková, P., Rysová, M. & Poláková, L. (2017). CzeDLex – A Lexicon of Czech Discourse Connectives. *The Prague Bulletin of Mathematical Linguistics*, (109), pp. 61–91.

Pons Bordería, S. (2001). Connectives/Discourse markers. An Overview. *Quaderns de Filologia. Estudis Literaris*, (6), pp. 219–243.

Pons Bordería, S. & Fischer, K. (2021). Using discourse segmentation to account for the polyfunctionality of discourse markers: The case of well. *Journal of Pragmatics*, 173, pp. 101–118.

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

Roze, C., Danlos, L. & Muller, P. (2012). LEXCONN: a French lexicon of discourse connectives. *Discours - Revue de linguistique, psycholinguistique et informatique*. URL https://hal.inria.fr/hal-00702542.

Rysová, M. & Rysová, K. (2018). Primary and secondary discourse connectives: Constraints and preferences. *Journal of Pragmatics*, 130, pp. 16–32.

Santos Río, L. (2003). *Diccionario de partículas*. Salamanca: Luso-española de ediciones.

Schiffrin, D. (2001). Discourse Markers: Language, Meaning, and Context. In D. Schiffrin, D. Tannen & H. Hamilton (eds.) *The Handbook of Discourse Analysis*. Oxford: Blackwell, pp. 54–75.

Sileo, D., van de Cruys, T., Pradel, C. & Muller, P. (2019). Mining Discourse Markers for Unsupervised Sentence Representation Learning. In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers)*. Minneapolis, Minnesota: Association for Computational Linguistics, pp. 3477–3486.

Stede, M. (2002). DiMLex: A Lexical Approach to Discourse Markers. In A. Lenci & V.D. Tomaso (eds.) *Exploring the Lexicon - Theory and Computation*. Alessandria: Edizioni dell'Orso.

Stede, M., Scheffler, T. & Mendes, A. (2019). Connective-Lex: A Web-Based Multilingual Lexical Resource for Connectives. *Discours*. URL https://journals.openedition.org/discours/10098.

Stubbs, M. (1983). *Discourse Analysis. The Sociolinguistic Analysis of Natural Language*. Chicago: University of Chicago Press.

Stubbs, M. (1996). *Text and Corpus Analysis*. Oxford: Blackwell.

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, Turkey: European Language Resources Association (ELRA), pp. 2214–2218.

Traugott, E. & Dasher, R. (2002). *Regularity in semantic change*. New York: Cambridge University Press.

Urgelles-Coll, M. (2010). *The Syntax and Semantics of Discourse Markers*. London: Continuum.

van Dijk, T. (1973). Text Grammar and Text Logic. In J. Petöfi & H. Rieser (eds.) *Studies in Text Grammar*. Dordrecht: Reidel, pp. 17–78.

Versley, Y. (2010). Discovery of Ambiguous and Unambiguous Discourse Connectives via Annotation Projection. In L. Ahrenberg, J. Tiedemann & M. Volk (eds.) *Proceedings of the Workshop on Annotation and Exploitation of Parallel Corpora (AEPC)*. Tartu: Northern European Association for Language Technology, pp. 83–92.

Wichmann, A. & Chanet, C. (2009). Discourse markers: A challenge for linguists and teachers. *Nouveaux cahiers de linguistique française*, 29(4), pp. 23–40.

Xue, N., Ng, H.T., Pradhan, S., Rutherford, A., Webber, B., Wang, C. & Wang, H. (2016). CoNLL 2016 Shared Task on Multilingual Shallow Discourse Parsing. In *Proceedings of the CoNLL-16 shared task*. Berlin, Germany: Association for Computational Linguistics, pp. 1–19.