Proceedings of the 13th Conference on Language Resources and Evaluation (LREC 2022), pages 2739–2749
Marseille, 20-25 June 2022

ISO-based Annotated Multilingual Parallel Corpus for Discourse Markers

Purificação Silvano[†], Mariana Damova[⋆], Giedrė Valūnaitė Oleškevičienė[⊙]
Chaya Liebeskind[◇], Christian Chiarcos[*,+], Dimitar Trajanov[∘]
Ciprian-Octavian Truică[§,‡] Elena-Simona Apostol[§,‡], Anna Bączkowska[×]
[†]Faculty of Arts and Humanities of the University of Porto, Centre of Linguistics of the University of Porto
[⋆]Mozaika, Ltd. [⊙]Mykolas Romeris University [◇]Jerusalem College of Technology
[*]Applied Computational Linguistics, Goethe-Universität [+] Institute for Digital Humanities, Universität zu Köln
[∘]Faculty of Computer Science and Engineering, Ss. Cyril and Methodius University
[§]Department of Information Technology, Uppsala University
[‡]Faculty of Automatic Control and Computers, University Politehnica of Bucharest
[×]Institute of English and American Studies, University of Gdansk
[†]Porto, Portugal; [⋆]Sofia, Bulgaria; [⊙]Vilnius, Lithuania; [◇]Jerusalem, Israel; [*]Frankfurt am Main, Germany;
[+]Cologne, Germany; [∘]Skopje, North Macedonia; [§]Uppsala, Sweden; [‡]Bucureşti, Romania; [×]Gdańsk, Poland
[†]msilvano@letras.up.pt, [⋆]mariana.damova@mozajka.co, [⊙]gvalunaite@mruni.eu,
[◇]liebchaya@gmail.com, [*]chiarcos@cs.uni-frankfurt.de, [∘]dimitar.trajanov@finki.ukim.mk,
[§]ciprian-octavian.truica@it.uu.se, [‡]elena.apostol@upb.ro, [×]anna.k.baczkowska@gmail.com

Abstract

Discourse markers carry information about the discourse structure and organization, and also signal local dependencies or epistemological stance of speaker. They provide instructions on how to interpret the discourse, and their study is paramount to understand the mechanism underlying discourse organization. This paper presents a new language resource, an ISO-based annotated multilingual parallel corpus for discourse markers. The corpus comprises nine languages, Bulgarian, Lithuanian, German, European Portuguese, Hebrew, Romanian, Polish, and Macedonian, with English as a pivot language. In order to represent the meaning of the discourse markers, we propose an annotation scheme of discourse relations from ISO 24617-8 with a plug-in to ISO 24617-2 for communicative functions. We describe an experiment in which we applied the annotation scheme to assess its validity. The results reveal that, although some extensions are required to cover all the multilingual data, it provides a proper representation of discourse markers value. Additionally, we report some relevant contrastive phenomena concerning discourse markers interpretation and role in discourse. This first step will allow us to develop deep learning methods to identify and extract discourse relations and communicative functions, and to represent that information as Linguistic Linked Open Data (LLOD).

Keywords: multilingual corpus, discourse markers, ISO-based annotation scheme, discourse relations, communicative functions

1. Introduction

The notion of discourse marker (Zwicky, 1985; Schiffrin, 1987; Lenk, 1998) is an elusive and fuzzy concept. The term is often used interchangeably with, inter alia, discourse particle (Schourup, 1985; Kroon, 1995), pragmatic marker (Fraser, 1996; Aijmer et al., 2006), pragmatic particle (Östman, 1981) or discourse connective (Blakemore, 2006). In earlier studies, discourse markers were studied from a structural, discourse organization perspective, while more recent investigations also focus on their role as establishing local dependencies (Prasad et al., 2008), and on the epistemic stance they may encode, thus following an attitudinal or affective perspective (Sanders et al., 1992; Bączkowska, 2016). Regardless of the adopted definition or theoretical approach, discourse markers provide instructions on how to interpret the discourse (Crible and Zufferey, 2015), and its study is paramount to understand the mechanism underlying discourse organization.

Due to their relevance, discourse markers have been largely researched by different communities of practice from linguistics and computation, which has led to several proposals regarding their identification, extraction and classification in monolingual and multilingual datasets in both areas (corpus-based frameworks and functional taxonomies include, e.g. Halliday and Hasan (1976), Redeker (1990), Sweetser (1990), Cuenca (2013), Crible (2014), Mann and Thompson (1988), Prasad et al. (2008), Asher et al. (2003); for computational approaches, see Zufferey (2004), Prasad et al. (2018), Kurfali (2020), Gessler et al. (2021)). Nonetheless, many problems persist not only concerning an interoperable taxonomy, but also efficient methods for the automatic identification and classification of discourse markers, even more when dealing with a multilingual corpus.

The present study aims to contribute with a multilingual corpus with discourse markers in nine languages, English, Lithuanian, Bulgarian, German, Macedonian, Romanian, Hebrew, Polish, and Portuguese annotated using two parts of ISO 24617 - *Language resource management – Semantic annotation framework (SemAF*, part 8 - *Semantic relations in discourse, core annotation schema (DR-core)* (ISO, 2016; Bunt and Prasad, 2016) with a plug-in to Part 2 *Dialogue acts* (Bunt et al., 2020; ISO, 2020).

The harmonization across multiple theories and frameworks is a very relevant issue even more when dealing with different languages. ISO 24617-8 puts forward an interoperable core-annotation scheme for discourse relations, which with a plug-in to ISO 24617-2 can adequately represent the meaning of discourse markers, and be used cross-linguistically. To the best of our knowledge, it is the first cross-lingual application of parts 8 and 2 of standard 24617 to the annotation of discourse markers.

In this paper, first we revise some of the related work (section 2), then the annotation scheme is presented (section 3) and finally we discuss data preparation (section 4), annotation challenges (section 5) and overall results (section 6).

2. Related work

Discourse markers are single-word or multiword expressions (MWE) from syntactic classes of conjunctions, adverbials, and prepositional phrases (Fraser, 2009a), but also according to some authors, expressions such as *you know*, *you see*, and *I mean* (Schiffrin, 2001; Maschler and Schiffrin, 2015). In addition to signaling moves in conversation and dialogue (Heeman and Allen, 1999), discourse markers signal coherence relations established between clauses and sentences (Das, 2014; Taboada, 2006). For that reason, some authors have proposed a somewhat polemic, but useful, distinction (Crible, 2016) between relational and non-relational discourse markers, depending on their semantic or interactional use. Although the line that divides these two groups can be blurry, taxonomies for both types have been proposed either separately (Mann and Thompson, 1988; Sanders et al., 1992; Prasad et al., 2008), or in conjunction (González, 2005; Crible, 2014).

Since discourse markers are amply considered significant discourse relations' triggers, they have been researched for discourse relation detection and analysis (e.g. Sanders et al. (1992); Knott and Dale (1993); Marcu (2000); Silvano (2010) Das (2014); Bunt and Prasad (2016); Das and Taboada (2019)). As a result, a significant number of annotated corpora with discourse relations signaled by discourse markers has been developed: for example, RST-DT English corpus (Carlson et al., 2003); Penn Discourse Tree Bank (PDTB) (Prasad et al., 2008); SDRT Annodis French corpus (Afantenos et al., 2012). Presently, the large bulk of these corpora is manually annotated, mostly by trained linguists, less by non-experts (cf. Scholman et al. (2016) for the efficacy of annotation by non-experts), and only a reduced number undergoes automatic/semiautomatic annotation (with human supervision). One of the complex Natural Language Processing (NLP) tasks capturing discourse structure has been addressed by Zufferey (2004). He describes a three-step process: detecting the occurrence of discourse markers, attaching the inferential semantic functions to discourse markers, and determining the scope of the characteristic functions. The typical functions of discourse markers embrace: connecting discourse, signaling hesitation, turn-taking and theme, marking turn boundaries, hedging, disclosing attitude, and regulating relationship with the interlocutor, as well as seeking approval (Jucker and Ziv, 1998).

The research about discourse markers allows the identification of lexical items that can be incorporated into the lexicons and the definition of new characteristic features. Recently, shared tasks such as DISRPT 2019 and 2021 editions (Zeldes et al. (2019); Zeldes et al. (2021)) have also greatly improved automatic methods for discourse markers detection, as well as Discourse Relation Classification across RST, SDRT, and PDTB (see, for example, DisCoDisCo (Gessler et al., 2021) with a Transformer-based neural classifier).

In terms of language coverage, discourse markers annotated corpora vary greatly. Since language coverage is not uniform, cross-lingual methods have been explored to include less resourced languages (TED-MDB, but also already Gylling-Jørgensen and Korzen (2011)). NLP research embraces cross-lingual language models by applying the approach to multiple languages searching for the effectiveness of cross-lingual pretraining (Ding et al., 2020). The cross-lingual studies of discourse markers and linking discourse marker inventories by applying LLOD techniques allow creating interlinked, multilingual discourse marker lexicons to be used for further analysis of cross-lingual discourse structure (Chiarcos and Ionov, 2021). However, cross-lingual techniques necessarily extend to languages for which resources already exist, so it would be desirable to compare the resources that already exist for these languages, and/or to use these external data to complement the notoriously scarce training data for discourse annotations.

Focusing on the languages from our corpus, if, on the one hand, there are languages like English and German that are very well covered, on the other hand, languages such as Hebrew, Macedonian, Portuguese, Lithuanian, Romanian, Bulgarian, Polish, and Romanian are low-resourced languages.

Although with fewer language resources compared to English, German is relatively well covered. With the Potsdam Commentary Corpus (Bourgonje and Stede, 2020, PCC), the CoNaNo corpus (Stede and Heintze, 2004) and the DimLex discourse marker inventory (Scheffler and Stede, 2016), a number of resources for discourse analysis have been available for years. German is also covered in the TED-MDB corpus (Zeyrek et al., 2020). Aside from these small-scale resources, Tueba-D/Z (Gastel et al., 2011) is a large-scale newspaper corpus with discourse annotations, although it provides annotations only for a small number of selected discourse markers. There are no native SemAF discourse annotations available for German, but several of these resources adopt schemes which have mappings to SemAFso that these resources can be used for external evaluation of both manual annotation as well as future projection experiments into German. Moreover, they can be used to evaluate the mapping of external schemes to SemAF. While Bunt and Prasad (2016) provides mappings of PDTB, RST, and SDRT to SemAF, these do seem not to have been evaluated on a quantitative basis, so far.

By contrast, the other languages from our corpus, in spite of a wide variety of studies about discourse markers, lack language resources with discourse markers annotation.

In Lithuanian consistently annotated discourse related data is covered by TED-MDB, which is a comparatively small corpus (Oleskeviciene et al., 2018). There have been attempts to research spoken Lithuanian discourse markers by either focusing on certain expressions (Šinkũnienė, 2020) or carrying out a synchronic and diachronic corpus-based analysis of discourse markers (Šinkūnienė et al., 2020).

Although there are many studies on discourse markers and corresponding pragmatic functions in some Romance languages, e.g., French, Spanish, or Italian (Crible and Pascual, 2020; Lansari, 2019; Cristofaro and Badan, 2019), there are not so many on Romanian (Ionescu, 2020; Ştefănescu et al., 2020; Popescu et al., 2020). The current studies only analyze the pragmatic functions of some of these markers from a linguistic perspective. Unfortunately, large-scale computational linguistic studies for this language are lacking in the current literature. Ionescu (2020) proposes the analysis of a sub-category of discourse markers, i.e., topic shifters, from a contrastive perspective, in the Romanian language. The study also compares the strategies of topic shifting in Romanian with their French counterparts, using parallel corpora consisting of 150 discourse markers occurrences. A comprehensive analysis of two synonymous Romanian discourse markers and their main features is discussed in Ştefănescu et al. (2020). The authors use a corpus of 150 annotated sample sentences to study the main functionalities and discourse patterns of *de altfel* and *de altminteri* (*as a matter of fact*, *in fact*, *indeed*). The Romanian adverb *atunci* (*then*) has suffered discursive values changes over time, from a temporal adverb to a polyfunctional discourse marker (Popescu et al., 2020). By employing the CoRoLa corpusand CORPES, the authors compare *atunci* with its Spanish counterpart *entonces*.

In European Portuguese, there is a relatively small corpus of spoken discourse manually annotated following PDTB annotation principles (TED-PT) (Zeyrek et al., 2020). In this corpus, several implicit and explicit discourse makers are attributed a meaning by means of coherence relations. A lexicon of discourse markers was also created for European Portuguese (LDM-PT) (Mendes et al., 2018) within the project of TextLink. In Connective-Lex (Stede et al., 2019) discourse markers are annotated with discourse relations in addition to the syntactic and lexicographic information.

Bulgarian corpora with discourse markers annotations are sparse. We find collections of several examples with intonation marking of discourse markers in BulPhonC (Hateva et al., 2016), and collections of text corpora, such as the Bulgarian National Corpus (Koeva et al., 2011), and the Bulgarian-English Parallel TreeBank Simov et al. (2011), that do not contain specific information about discourse markers or their roles. The corpus of annotated examples for discourse markers presence or absence from TED talks that is reported in this paper is the first systematic work on creating such an annotated corpus with discourse markers for Bulgarian. In the course of the current effort a Macedonian and Hebrew language corpora with examples of TED talks annotated for the presence and absence of discourse markers have also been created.

Corpora of Polish with annotated discourse markers are scarce as well. Probably the best known is the corpus of spoken language dubbed *Gesprochene Wissenschaftssprache Kontrastiv*, available in four languages, Polish among them (also in German, English, and Bulgarian), which amounts to 760,000 words in size and is based on 92 hours of recordings. It contains annotation of discourse phenomena in texts illustrating academic interaction, and it is also PoS tagged, lemmatized, time-aligned and orthographically transcribed (available at clarin.eu).

3. Annotation scheme

The design of the annotation scheme followed a main requirement: suitability for representing the meaning of discourse markers across different languages. On the one hand, since the discourse markers extracted from the corpus served two different functions, either establishing coherence relations or conveying interactional purposes, the tag set had to properly account for their relational and non-relational uses (Crible, 2016). On the other hand, due to the multilingual nature of the corpus, the annotation scheme had to be comprehensive enough to cover different language specificities.

Despite the existence of many corpus-based frameworks, like RST (Mann and Thompson, 1988), CCR (Sanders et al., 1992), SDRT (Asher et al., 2003), PDTB (Prasad et al., 2008), and of several taxonomies, such as the ones proposed by Cuenca (2013) and Crible and Zufferey (2015), being interoperability and language-independence key factors, we deemed it best to opt for ISO 24617 - Language resource management – Semantic annotation framework (SemAF), which provides semantic annotation schemes with extensive coverage. The relevant part to codify the discourse markers meaning was Part 8 – Semantic relations in discourse, core annotation schema (DR-core)– ISO 24617-8 (ISO, 2016). Assuming that there are evident compatibilities across the semantic description of discourse relations in the different frameworks, which enables mapping between them, ISO 24617-8 defines an interoperable set of low-level semantic discourse relations according to the meaning of the relation's arguments. These discourse relations are divided into two types: symmetric, whenever the two arguments assume relation-specific semantic role, and asymmetric, whenever the arguments take the same semantic role. This part of ISO 24617 adequately represents the meaning of discourse markers that establish a relation between the content of two arguments, as illustrated by Example 1 from the English dataset.

Example 1. *For every ton of cement that's manufactured, almost a ton of CO2 is emitted into the atmosphere.* **As a result, the cement industry is the second-largest industrial emitter of CO2, responsible for almost eight percent of total global emissions.**

In this example, the two arguments are linked by the discourse relation Cause, signaled by the discourse marker *as a result*. Thus, Arg2 (in italics) bears the semantic role of *reason*, while Arg1 (in bold) plays the semantic role of *result*.

The problem arose when we tried to apply this annotation scheme to examples with discourse markers that fulfil an interactional function, as exemplified in Example 2.

Example 2. Here's a project called "Just Landed", where I'm looking at people tweeting on Twitter. "Hey! I just landed in Hawaii!" – *you know*, how people just casually try to sneakrtion that into their Twitter

In this case, the discourse marker *you know* is clearly different from *as a result* from the previous example. Instead of relating the meaning of two arguments, it intervenes in an interpersonal domain being used to manage the relation between speaker-listener (Crible, 2014). Therefore, the value of discourse markers such as this one cannot be properly described within ISO 24617-8.

Inspired in the plug-in interface designed for the second edition of Language resources management —Semantic annotation framework (SemAF) — Part 2: Dialogue acts (2020) (Bunt (2019); Bunt et al. (2020)), we decided to add a similar mechanism to our annotation scheme, but working in the inverse direction, that is, from the discourse relations set to the dialogue acts set. The second version of ISO 24617-2 (ISO, 2020) puts forward a wide-ranging metamodel for the annotation of dialogue acts that includes dimensions, communicative functions and qualifiers. As anticipated by the standard, there may be the need to customize the annotation scheme by simplifying it to the specific needs of the project, and that is what we considered to be the best path bearing in mind that our aim was to represent the meaning of discourse markers. So, whenever the annotation requires the use of the plug-in interface, only the communicative function should be identified, and, if it is relevant, the qualifier should also be registered. This means that the segmentation of the dialogue acts, the identification of the participants and of the dialogue act dimension were left out. Accordingly, in Example 2, the discourse marker *you know* would be annotated with *checkQuestion*, a general purpose function part of the information-seeking functions.

Figure 1 summarizes our proposal and Figure 2 and Table 1 specify the discourse relations, communicative functions and qualifiers that integrate our model.

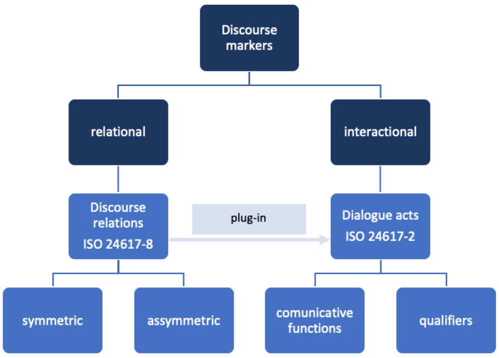

Figure 1: Annotation scheme

Discourse Relations			
Asymmetric	Semantic Role		Symmetric
	Arg 1	Arg 2	
Cause	result	reason	Conjunction
Expansion	narrative	expander	Contrast
Asynchrony	before	after	Synchrony
Concession	expectation-raiser	expectation-denier	Similarity
Elaboration	broad	specific	Disjunction
Exemplification	set	instance	Restatement
Manner	achievement	means	
Condition	consequent	antecedent	
Negative Condition	consequent	negated-antecedent	
Purpose	enablement	goal	
Exception	regular	exclusion	
Substitution	disfavoured-alternative	favoured-alternative	

Figure 2: Set of Discourse Relations (Bunt and Prasad, 2016)

To sum up, the framework that we propose includes a host annotation scheme based on ISO 24617-8 (DR-core) and an annotation plug-in to ISO 24617-2 (DA), thus combining two standardized means to represent the discourse markers' meaning.

4. Multilingual parallel corpus

The multilingual corpus that we present in this paper contains data from nine languages English, Lithuanian, Bulgarian, German, Macedonian, Romanian, Hebrew, Portuguese and Polish, using the publicly available TED Talk transcripts. TED Talks are spoken monologues that employ a great deal of communicative instruments to render the presentation lively and interactive. This includes abundant use of discourse markers whose roles are of interest to our study. The constitution of the multilingual corpus is an ongoing expansion of TED-EHL parallel corpus LINDAT/CLARIN-

Table 1: Set of communicative functions and qualifiers (Bunt et al., 2020)

Communicative functions		**Qualifiers**
General	Dimension-specific	
checkQuestion	autoPositive	conditional/ unconditional
inform	autoNegative	certain/uncertain
agreement	alloPositive	positive/ negative
disagreement	alloNegative	
correction	feedbackElicitation	
answer	stalling	
confirm	pausing	
disconfirm	interactionStructuring	
offer	opening	
promise	topicShift	
addressRequest	selfError	
acceptRequest	retraction	
declineRequest	selfCorrection	
addressSuggest	initGreeting	
acceptSuggest	initSelfIntroduction	
declineSuggest	apology	
request	thanking	
instruct	initGoodbye	
suggest	compliment	
addressOffer	congratulation	
acceptOffer	sympathyExpression	
declineOffer	contactCheck	

LT repository.The corpus has been built as a series of bilingual datasets with English as pivot language, so as a result we have eight datasets with language pairs of English and one of the other eight languages. Each bilingual pair contains aligned examples in English and in another language, based on the occurrence of a MWE potentially in the role of discourse marker, defined according to theoretical insights by Schiffrin (2001); Fraser (2009b); Crible (2014), among others. In other words, in this work, we consider discourse markers to be not only connectives, conjunctions, but also verbs and lexicalized expressions with verbs, used to link the content of utterances or to express the stance of the speaker.

It is important to point out that many of the selected multiword discourse markers in our corpus are ambiguous, e.g. they can be discourse markers or content expressions depending on the context in which they occur. Such expressions are *you see*, *you know*, *that is*, that can be either clauses, or discourse markers, as shown in Examples 3 and 4.

Example 3. You can see *(content expression)* areas where neuronal cell bodies are being stained. And what you can see is it's very non-uniform.

Example 4. So **you see** *(discourse marker)*, we're navigating the web for the first time

To render the corpus homogeneous and uniform for experimenting the ISO-based annotation, all eight datasets of language pairs have been compared and the matching examples across all of them identified and extracted. Thus, we came up with an intersection of 55 unique matching contexts across all nine languages out of a total of 44 192 distinct textual contexts for English, representing the union of the English examples from all eight language pairs. This amounts to a total of 495 examples in the set of the nine languages that were annotated with the ISO-based annotation scheme. We deem this corpus size satisfactory as our annotation effort is the first attempt to produce ISO-based annotations of discourse markers, in addition to match them across nine languages.

The use of TED talks and their translations permits building a parallel corpus and studying contrastively the different languages. The multilingual consistently annotated corpus makes it possible to compare discourse annotated translated texts with the English discourse annotated source text in order to analyze different languages of the text, as well as to understand translation trends. Such analysis offers a better perspective of comparative studies because the same type of texts is used for annotation (eg. Zufferey (2004)). Text-type similarity is a major advantage in analyzing language discourse relations/ communicative functions and discourse markers. The other important usage scenario of multilingual discourse annotations is where cross-lingual data is being applied to languages that do not provide any native discourse annotations or only at a very small scale as is the case of Macedonian, Bulgarian, European Portuguese, Hebrew, Polish, and Romanian.

Although the pivot language is English, and the aligned examples in the other languages are the translations of the contexts where there is a multiword discourse marker in English, in fact, the former are not mere literal translations, which could result in non-naturalistic texts. As shown in section 6, the parallel corpus allows us to observe particularities with regards to the type and distribution of discourse markers in each language. For instance, discourse markers in English are in some cases omitted in the other languages, or translated by others with different characteristics.

Another relevant aspect of our corpus worth referring to is its monologic nature. As is well known, TED talks are spoken monologues, which does not mean that they are devoid of interactional discourse markers. In spite of being less frequent, they are indeed present, which is also a challenge for our annotation scheme, in particular in what concerns the plug-in to ISO 24617-2, because this part of ISO was designed primarily to account for dialogues, and not monologues.

5. Annotation process

The structure of the data in the bilingual datasets provides tables with shorter sentence chunks and larger contexts where a multiword discourse marker appears, the identified discourse marker expression aligned in English and in the other language, a classification of the occurrence of the expression as discourse marker or not, and the following breakdown of categorization of the examples, based on the adopted annotation scheme, combining ISO 24617-8 for discourse relations representation and ISO 24617-2 for communicative functions representation. Figure 3 shows the structure of the annotation table.

The analysis of each example of the dataset requires estimation whether the multiword discourse marker

<table>
<tr><td>vid</td><td>lid</td><td colspan="2">Sentence Chunk EN</td></tr>
<tr><td colspan="2">Larger Textual Context EN</td><td colspan="2">Discourse Marker EN</td></tr>
<tr><td colspan="4">Discourse Marker Presence EN</td></tr>
<tr><td colspan="4">Sentence Chunk TL (target language)</td></tr>
<tr><td colspan="4">Larger Textual Context TL</td></tr>
<tr><td colspan="2">Discourse Marker Presence TL</td><td colspan="2">Discourse Marker TL</td></tr>
<tr><td>Arg 1</td><td>Discourse marker</td><td colspan="2">Arg 2</td></tr>
<tr><td colspan="2">Discourse Relation (ISO)</td><td>Arg 1 role</td><td>Arg 2 role</td></tr>
<tr><td colspan="2">Communicative function (ISO)</td><td colspan="2">Qualifier</td></tr>
</table>

Figure 3: Annotation dataset structure

is employed as discourse marker, determination of the discourse relation it conveys, identification of the boundaries of the textual chunks, describing the arguments of the discourse relation. As a consequence, the string describing the discourse marker in the example is recorded in the field *Discourse marker*, while the parts of the text describing the first and the second arguments of the assigned discourse relation are recorded in the fields Arg1 and Arg2. The recording of the ISO discourse relation role and of the roles of its arguments is provided in the following three fields: *Discourse relation (ISO)*, *Arg 1 role*, *Arg 2 role*. Finally, the last two fields in the table are dedicated for the assignment of ISO communicative roles, e.g. dialogue acts and qualifiers, *Communicative function (ISO)* and *Qualifier*, as shown in Figure 3. Accordingly, Example 1 would be annotated as in Figure 4.

<table>
<tr><td>66 954</td><td>14</td><td>As a result,</td></tr>
<tr><td colspan="3">One of the byproducts of that process is carbon dioxide, or CO2. For every ton of cement that's manufactured, almost a ton of CO2 is emitted into the atmosphere. As a result, the cement industry is the second-largest industrial emitter of CO2, responsible for almost eight percent of total global emissions.</td></tr>
<tr><td colspan="2">as a result</td><td>1</td></tr>
<tr><td colspan="3">Como resultado,</td></tr>
<tr><td colspan="3">Por cada tonelada de cimento produzido,é emitida para a atmosfera quaseuma tonelada de CO2. Como resultado, aindústria do cimento é a segunda maioremissora industrial de CO2, responsávelpor quase 8% do total de emissõesglobais.</td></tr>
<tr><td>1</td><td colspan="2">como resultado</td></tr>
<tr><td>Por cada tonelada de cimento produzido, é emitida para a atmosfera quase uma tonelada de CO2.</td><td colspan="2">como resultado</td></tr>
<tr><td colspan="3">Como resultado, a indústria do cimento é a segunda maior emissora industrial de CO2, responsável por quase 8% do total de emissões globais.</td></tr>
<tr><td>cause</td><td>result</td><td>reason</td></tr>
<tr><td>N/A</td><td colspan="2">N/A</td></tr>
</table>

Figure 4: Annotation dataset example

This structure allows us to capture all the relevant information related to discourse relations or communicative functions/ qualifiers conveyed by the discourse markers. So, if the discourse marker was of a relational nature, the annotator had to identify the arguments, the discourse relation that linked both arguments and, in case of asymmetric discourse relations, the arguments role. When the discourse marker had an interactional meaning, the annotator had to identify its communicative function and, if pertinent, the appropriate qualifier. After designing the annotation scheme, a manual of annotation with instructions, definitions and illustrative examples was prepared. The following step was the annotation of the English dataset by a linguist, which worked as gold standard. Next, the 55 parallel text segments with discourse markers from each language pair was annotated by one native speaker (all authors of this paper) following the annotation guidelines. Although the annotators work in different field of expertise - either Linguistics, Computer Science, or both -, all have been developing research about discourse markers. Whenever the annotators encountered some difficulties or issues not included in the annotation manual, these were addressed by the group and a solution was agreed on.

6. Discussion and Results

This section presents the results of the annotation process along with discussion of different phenomena single and cross-language and evidence issued from the linguistic analysis carried out while interpreting and applying the ISO 24617-8 for discourse relations and the plug-in to ISO 24617-2 for communicative functions determination and assignment. The aim of this section is not so much to provide a quantitative systematic and detailed analysis of the annotation results as to give an overall description of the different phenomena we encountered. The quantitative analysis will be carried out on larger sets of data. The main objective of the paper is to present a multilingual parallel corpus for discourse markers as evidence that a combination of ISO 25617-8 with ISO 25617-2 can account for the semantic and pragmatic values of discourse makers.

Baseline As English has been a pivot language for all language pairs of our corpus, a baseline annotation has been provided for the English examples. This annotation abides the principles set for the annotation scheme to either assign an ISO discourse relation, or ISO communicative function/ qualifier, or both when required, to the discourse marker of the example. Additionally, the examples have been annotated with information whether the string of the MWE identified in the text has actually the role of discourse marker or not. Out of the 55 examples, 11 have turned to contain a MWE that is actually not uttered as discourse marker in the text, as shown in Example 5 below, conversely to Example 6, where the expression *you know* is uttered as discourse marker.

Example 5. And Tiger Woods, for a long time, the perfect brand ambassador. Well, **you know** the story.

Example 6. Just stay here for a second. (Applause) **You know**, when I heard Simon's – please sit down;

The annotated English dataset identifies examples with different ISO 24617-8 discourse relations, as shown in Table 2, which testifies for the variety of the corpus, for the quite extensive coverage of the ISO 24617-8 with it, and hence for the representativeness of the selection of our dataset. The encountered communicative functions and qualifiers of interactional discourse markers in the English dataset are in smaller number.

Table 2: Annotation of discourse markers in the English dataset.

Discourse markers meaning	English DM
Discourse relations ISO 24617-8	
Exemplification	for example, for instance
Elaboration	in particular, to sum up
Synchrony	so far
Contrast	on the one hand
Concession	on the other hand
Conjunction	on the other hand
Restatement	in other words, I mean
Cause	as a result
Expansion	in fact, this is, that is, of course
Communicative functions and qualifiers ISO 24617-2	
CheckQuestion	you know
Confirm	of course, in fact
Opening	You know
AlloPositive	you see
Certain	of course

Cross-lingual annotation analysis While provided as a baseline, each of the other eight languages have been annotated according to proper single language analysis of the dataset texts, so the English dataset and the single examples in it have been used as validation for the judgement of the annotators in the other languages. As it will become clear in the following paragraphs, the work on this multilingual corpus has come up with a series of interesting phenomena with respect to the expression of discourse markers, discourse relations and communicative functions in the different languages, on the one hand, and with respect to the determination and assignment of proper ISO defined tags for each single example.

Omissions. When looking at the cross language examples contrastively, several generalizations appear. Lexicalized expression of discourse markers as MWE in English can disappear in any given target language, but the discourse relation it conveys remains present. In these cases, the discourse relation is established by grammatical means, e.g., modal adverbs in German; modal verbs and questions in Bulgarian; changes in the word order in the languages where this is possible, for instance, in German to convey most commonly contrast or continuity; and other grammatical and phrasal variations, like in Hebrew, the discourse marker *in fact* is expressed by a personal pronoun, such as **והן** (and *they* (female)) and **והוא** (and *he*). Lexical variations of the discourse markers are also abundant. For instance, the discourse marker *in fact* in European Portuguese occurs as the direct translation (*de facto*), but also as *na verdade* (*in true*). There are also instances where in English the discourse marker is a multiword expression and in the targeted languages it is used a single word discourse marker like *pavyzdžiui* in Lithuanian.

Different discourse relations/communicative functions and different discourse markers in one and the same context across languages. The first observation when comparing the baseline annotations to the single examples of the other eight languages is that the identified discourse relations and communicative functions in the English set of examples do not correspond always to their juxtaposed counterparts in the parallel datasets. For example, a discourse relation *expansion*, expressed with the MWE *in fact* in the English text (Ex. 7) occurs as conveying the communicative function *confirm* in the Bulgarian text, reflected by the utterance правилно (*right*), (Ex. 8), while in German and Lithuanian it remains *expansion* signaled by the expressions *tatsächlich*, *tiesą sakant*. In Hebrew, while in this example (i.e., Ex. 7) the same discourse relation remains (**ובעצם**) (Ex. 9), a similar text with the same MWE *in fact* (Ex. 10) is replaced by the *for example* exemplification relation (**לדוגמה**) (Ex. 11).

Example 7. But most people don't agree. And **in fact**, because their minds don't fit (EN)

Example 8. Но повечето хора не са съгласни. Правилно, защото техните умове не се вписват, в това което обществото смята за нормално, често биват избягвани и неразбрани. (BG)

Example 9. **אבל רוב האנשים לא חושבים כך. ובעצם. בגלל שמוחותיהם לא מתאימים לה** (HE)

Example 10. So I can drill into what I've done over specific time frames. Here, in fact, is the state of all the demo that I just gave. (EN)

Example 11. **אז אני יכול להתמקד במה שעשיתי בזמן מסויים. הנה, לדוגמה, כל המצב של ההדגמה שכרגע הצגתי.** (HE)

Different discourse relations conveyed by one and the same discourse marker depending on the context. A closer look and analysis of the examples from the single languages point to a range of variations in the way discourse relations, introduced by one and the same MWE, are interpreted. For example, the expression *on the other hand* occurs to introduce *conjunction*, but also *concession* and *contrast*, as shown in the Examples 12-14 below, respectively.

Example 12. Pe de-o parte am calculat câtă energie primește o primată pe zi din mâncarea crudă, **iar pe de altă parte**, câtă energie necesită un corp de o anumită mărime și câtă energie necesită un creier cu un anumit număr de neuroni.(RO)

Example 13. que mal andam ou falam, lhe davam as bolachas se ela gostasse das bolachas, contudo davam-lhe os brócolos se ela preferisse os brócolos. **Por outro lado**, os bebés de 15 meses ficavam a olhar para ela durante muito tempo caso ela agisse como se preferisse os brócolos. (PT)

Example 14. C. P. Snow sprach von den beiden Kulturen: Naturwissenschaften **auf der einen**, Geisteswissenschaften auf der anderen Seite, niemals würden die beiden zueinander finden. Ich sage also, das Spiegelneuronensystem ist die Basis der Schnittstelle. (DE)

Identical discourse markers and discourse relations assignment across the different languages. Some MWE discourse markers show stable interpretation across languages. These are *of course* as *confirm*, *for example* as *exemplification*, *in particular* as *elaboration* and *in other words* as *restatement*. They are consistently present in the examples of the other eight languages and convey the same discourse relation. In some languages like Romanian, restatement discourse relation is referred to with other expressions, such as *adică* (*that is*).

Subjective vs. neutral speaker's stance expression in different languages. Speaking of *that is*, we touch upon another phenomenon that is worth noting. The English expression *I mean* implies clear subjectivity by introducing an element of speaker's stance or attitude towards what has been said. This expression has been conveyed in the European Portuguese equivalent with the neutral *isto é* (*that is*) with respect to the speaker's stance expression. The Romanian *de fapt* (*in fact*) and its equivalent in other languages or *of course* and its equivalents also convey a nuance of speaker's attitude.

Expression of communicative functions and discourse relations in a single utterance. The cross-lingual analysis showed that there are cases where the discourse marker conveys both a discourse relation and a communicative function. For example, the discourse marker *of course* can signal a discourse relation of *expansion*, and simultaneously, express the communicative function *confirm* with the qualifier *certain*, as it is the case of the Portuguese example (Ex. 15), the translation of the English (Ex. 16).

Example 15. Em vez disso, até agora, as medições vindas do GCH não mostram sinais de novas partículas ou fenómenos inesperados. **Claro**, o veredicto não é definitivo. (PT)

Example 16. Instead, so far, the measurements coming from the LHC show no signs of new particles or unexpected phenomena. **Of course**, the verdict is not definitive. (EN)

It is important to note that such correlations between discourse relations and communicative functions are worth studying as they provide sufficient ground for generalizations.

This non-exhaustive list of the cross-lingual and interpretation phenomena gives a glance at the complexity and the interest in carrying this experimental pioneering effort of applying in practice the ISO annotation standard guidelines to a multilingual corpus composed of languages from different language families. Although the sample is rather small, it is evidence that the scheme that we propose is able to successfully represent the relational and interactional meaning of discourse markers across languages.

7. Conclusion and Future work

Our paper presents an ISO-based annotated multilingual corpus for discourse markers annotated with discourse relations and communicative roles, which we intend to publish in CLARIN[1]. We propose an annotation scheme that combines two parts of ISO 24617, Part 8 for discourse relations with a plug-in to Part 2 for communicative functions and qualifiers. To assert the feasibility of the designed scheme, we carried out an experiment by applying it to a set of examples from a multilingual parallel corpus that comprises nine languages, English, Lithuanian, Bulgarian, German, Macedonian, Romanian, Hebrew, Portuguese and Polish, from the publicly available TED Talk transcripts.

The annotation of 55 examples for each language (495 examples in total) enabled some conclusions regarding the annotation scheme. First, we can conclude that, although ISO 24617-8 presents only the *core* discourse relations, overall, they can account for the semantic use of the discourse markers occurring in our corpus. Nevertheless, there are some instances for which new, or more specific, discourse relations are needed. For instance, the discourse relation *expansion* covers different meaning relations, and, therefore, should be divided into more distinct discourse relations. For these reasons, as future work, we will extend the list of discourse relations so that they can properly represent the discourse markers' meaning cross-linguistically.

Another problem that we encountered concerns the *Attribution* discourse relation. ISO 24617-8 does not include it in the set of discourse relations, and ISO 24617-6 suggests a separate layer for its annotation. In our corpus, several examples of *Attribution* come about signaled by what some authors consider to be also discourse markers, like *I think*. In this first experiment, we decided not to proceed with the annotation of this type of discourse markers. However, due to its frequency, and relevance, in the future, we intend to add a separate layer to codify this information.

As follow-up, we will focus on the automatic identification of discourse markers, extraction of discourse relations, and identification of the their arguments in line with what has been done within shared tasks such as DISRPT 2019 and 2021 editions (Zeldes et al. (2019); Zeldes et al. (2021)), and work on the representation of the ISO-based annotation scheme as LLOD, extending (Chiarcos and Ionov, 2021), to enable further semantic processing of discourse relations and communicative functions.

8. Acknowledgements

This research has carried out within the COST Action CA18209 - NexusLinguarum. It has been partially funded by national funds through FCT – Fundação para a Ciência e a Tecnologia, I.P., within the project UIDB/00022/2020.

[1] In this repository, as far as we know, there are no similar multilingual language resources. The existing lexicons are monolingual or bilingual, as it is the case of the *Lexicon of discourse markers for European Portuguese LDM-PT* or the dictionary for Czech and German.

9. Bibliographical References

Afantenos, S., Asher, N., Benamara, F., Bras, M., Fabre, C., Ho-Dac, L.-M., Le Draoulec, A., Muller, P., Péry-Woodley, M.-P., Prévot, L., et al. (2012). An empirical resource for discovering cognitive principles of discourse organisation: the ANNODIS corpus. In *Proceedings of the Eight International Conference on Language Resources and Evaluation (LREC'12)*. European Language Resources Association (ELRA).

Aijmer, K., Foolen, A., and Simon-Vandenbergern, A.-M. (2006). Pragmatic markers in translation: a methodological approach. *Approaches to Discourse Particles*, pages 101–114.

Asher, N., Asher, N. M., and Lascarides, A. (2003). *Logics of Conversation*. Cambridge University Press.

Blakemore, D., (2006). *Discourse Markers*, chapter 10, pages 221–240. John Wiley & Sons, Ltd.

Bourgonje, P. and Stede, M. (2020). The Potsdam commentary corpus 2.2: Extending annotations for shallow discourse parsing. In *Proceedings of the 12th Language Resources and Evaluation Conference*, pages 1061–1066.

Bunt, H. and Prasad, R. (2016). ISO DR-Core (ISO 24617-8): Core concepts for the annotation of discourse relations. In *Proceedings 12th Joint ACL-ISO Workshop on Interoperable Semantic Annotation (ISA-12)*, pages 45–54.

Bunt, H., Petukhova, V., Gilmartin, E., Pelachaud, C., Fang, A., Keizer, S., and Prevot, L. (2020). The ISO standard for dialogue act annotation. In *Proceedings of the 12th Language Resources and Evaluation Conference*, pages 549–558.

Bunt, H. (2019). Plug-ins for content annotation of dialogue acts. In *Proceedings 15th Joint ISO-ACL Work- shop on Interoperable Semantic Annotation (ISA-15)*, pages 34–45.

Bączkowska, A. (2016). Well as a discourse marker in learner's inter-lingual subtitles. *Empirical Translation Studies*, pages 149–179.

Carlson, L., Marcu, D., and Okurowski, M. E. (2003). Building a discourse-tagged corpus in the framework of rhetorical structure theory. In *Current and new directions in discourse and dialogue*, pages 85–112. Springer.

Chiarcos, C. and Ionov, M. (2021). Linking Discourse Marker Inventories. In *3rd Conference on Language, Data and Knowledge (LDK 2021)*. Schloss Dagstuhl-Leibniz-Zentrum für Informatik.

Crible, L. and Pascual, E. (2020). Combinations of discourse markers with repairs and repetitions in English, French and Spanish. *Journal of Pragmatics*, 156:54–67, jan.

Crible, L. and Zufferey, S. (2015). Using a unified taxonomy to annotate discourse markers in speech and writing. 04.

Crible, L. (2014). Identifying and describing discourse markers in spoken corpora. annotation protocol v.8. Technical report.

Crible, L. (2016). Discourse Markers and Disfluencies: Integrating Functional and Formal Annotations. 05.

Cristofaro, E. D. and Badan, L. (2019). The Acquisition of Italian Discourse Markers as a Function of Studying Abroad. *Corpus Pragmatics*, 5(1):95–120, nov.

Cuenca, M. J., (2013). *The fuzzy boundaries between discourse marking and modal marking*, page 191–216. John Benjamins, Amsterdam.

Das, D. and Taboada, M. (2019). Multiple signals of coherence relations. *Discours. Revue de linguistique, psycholinguistique et informatique. A journal of linguistics, psycholinguistics and computational linguistics*, (24).

Das, D. (2014). *Signalling of coherence relations in discourse*. Ph.D. thesis, Arts & Social Sciences: Department of Linguistics.

Ding, P., Wang, L., Liang, Y., Lu, W., Li, L., Wang, C., Tang, B., and Yan, J. (2020). Cross-Lingual Transfer Learning for Medical Named Entity Recognition. In *International Conference on Database Systems for Advanced Applications*, pages 403–418. Springer.

Fraser, B. (1996). Pragmatic markers. *Pragmatics*, 6:167–190.

Fraser, B. (2009a). An account of discourse markers. *International review of Pragmatics*, 1(2):293–320.

Fraser, B. (2009b). An account of discourse markers. *International Review of Pragmatics*, 1(2):293–320.

Gastel, A., Schulze, S., Versley, Y., and Hinrichs, E. (2011). Annotation of explicit and implicit discourse relations in the TüBa-D/Z treebank. *Multilingual Resources and Multilingual Applications, Proceedings of the German Society of Computational Linguistics and Language Technology (GSCL) 2011*, pages 99–104.

Gessler, L., Behzad, S., Liu, Y. J., Peng, S., Zhu, Y., and Zeldes, A. (2021). DisCoDisCo at the DISRPT2021 shared task: A system for discourse segmentation, classification, and connective detection. In *Proceedings of the 2nd Shared Task on Discourse Relation Parsing and Treebanking (DISRPT 2021)*, pages 51–62, Punta Cana, Dominican Republic, November. Association for Computational Linguistics.

González, M. (2005). Pragmatic markers and discourse coherence relations in English and Catalan oral narrative. *Discourse Studies*, 7(1):53–86.

Gylling-Jørgensen, M. and Korzen, I. (2011). On Discourse Structure in Italian and Danish. null ; Conference date: 14-09-2011 Through 16-09-2011.

Halliday, M. A. K. and Hasan, R. (1976). *Cohesion in English*. Longman.

Hateva, N., Mitankin, P., and Mihov, S. (2016).

BulPhonC: Bulgarian Speech Corpus for the Development of ASR Technology. In Nicoletta Calzolari (Conference Chair), et al., editors, *Proceedings of the Tenth International Conference on Language Resources and Evaluation (LREC 2016)*, Paris, France, may. European Language Resources Association (ELRA).

Heeman, P. A. and Allen, J. (1999). Speech repains, intonational phrases, and discourse markers: modeling speakers' utterances in spoken dialogue. *Computational Linguistics*, 25(4):527–572.

Ionescu, A. (2020). Topic shifters in Romanian: A contrastive analysis. *Journal of Pragmatics*, 156:110–120, jan.

ISO. (2016). Language resource management-Semantic annotation framework (SemAF) - Part 8 - Semantic relations in discourse,core annotation schema (DR-core). Standard, Geneva, CH.

ISO. (2020). Language resource management-Semantic annotation framework (SemAF) - Part 2 - Dialogue acts. Standard, Geneva, CH.

Jucker, A. H. and Ziv, Y. (1998). Discourse Markers: introduction. *Discourse Markers*, pages 1–12.

Knott, A. and Dale, R. (1993). Using linguistic phenomena to motivate a set of rhetorical relations. *Human Communication Research Centre, University of Edinburgh*.

Koeva, S., Lezeva, S., Rizov, B., and Tarpomanova, E. (2011). Design and Development of the Bulgarian Sense-Annotated Corpus. In *LAS TIC: PRESENTE Y FUTURO EN EL ANÁLISIS DE CORPUS*, pages 142–150.

Kroon, C. (1995). *Discourse Particles in Latin. A Study of nam, enim, autem, vero, and at*. Giebven, Amsterdam.

Kurfali, M. (2020). Labeling Explicit Discourse Relations using Pre-trained Language Models. *ArXiv*, abs/2006.11852.

Lansari, L. (2019). *A Contrastive View of Discourse Markers: Discourse Markers of Saying in English and French*. Springer.

Lenk, U. (1998). Discourse markers and global coherence in conversation. *Journal of Pragmatics*, 30:245–257.

Mann, W. and Thompson, S. (1988). Rethorical Structure Theory: Toward a functional theory of text organization. *Text*, 8:243–281, 01.

Marcu, D. (2000). *The theory and practice of discourse parsing and summarization*. MIT press.

Maschler, Y. and Schiffrin, D. (2015). Discourse markers: Language, meaning, and context. *The handbook of discourse analysis*, 1:189–221.

Mendes, A., del Rio, I., Stede, M., and 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, May. European Language Resources Association (ELRA).

Oleskeviciene, G. V., Zeyrek, D., Mazeikiene, V., and Kurfalı, M. (2018). Observations on the annotation of discourse relational devices in TED talk transcripts in Lithuanian. In *Proceedings of the workshop on annotation in digital humanities co-located with ESSLLI*, volume 2155, pages 53–58.

Popescu, C. M., , and Duță, O. A. (2020). Rom. Atunci and Sp. Entonces: from Adverbs to Discourse Markers. Some Convergences and Divergences. *Studia Universitatis Babeș-Bolyai Philologia*, 65(2):47–62, may.

Prasad, R., Dinesh, N., Lee, A., Miltsakaki, E., Robaldo, L., Joshi, A., and Webber, B. (2008). The Penn Discourse TreeBank 2.0. In *Proceedings of the Sixth International Conference on Language Resources and Evaluation (LREC'08)*, Marrakech, Morocco, May. European Language Resources Association (ELRA).

Prasad, R., Webber, B., and Lee, A. (2018). Discourse Annotation in the PDTB: The Next Generation. In *Proceedings 14th Joint ACL - ISO Workshop on Interoperable Semantic Annotation*, pages 87–97, Santa Fe, New Mexico, USA, August. Association for Computational Linguistics.

Redeker, G. (1990). Ideational and pragmatic markers of discourse structure. *Journal of Pragmatics*, 14(3):367–381. Special Issue: 'Selected papers from The International Pragmatics Conference, Antwerp, 17-22 August, 1987'.

Sanders, T., Spooren, W., and Noordman, L. (1992). Toward a Taxonomy of Coherence Relations. *Discourse Processes - DISCOURSE PROCESS*, 15:1–35, 01.

Scheffler, T. and Stede, M. (2016). Adding semantic relations to a large-coverage connective lexicon of german. In *Proceedings of the Tenth International Conference on Language Resources and Evaluation (LREC'16)*, pages 1008–1013.

Schiffrin, D. (1987). *Discourse Markers*. Cambridge University Press, Cambridge.

Schiffrin, D. (2001). Discourse markers: Language, meaning, and context. *The handbook of discourse analysis*, 1:54–75.

Scholman, M. C., Evers-Vermeul, J., and Sanders, T. J. (2016). A step-wise approach to discourse annotation: Towards a reliable categorization of coherence relations. *Dialogue & Discourse*, 7(2):1–28.

Schourup, L. (1985). *Common Discourse Particles in English Conversation*. Garland, New York.

Silvano, M. d. P. M. (2010). *Temporal and rhetorical relations: the semantics of sentences with adverbial subordination in European Portuguese*. Ph.D. thesis.

Simov, K., Osenova, P., Laskova, L., Savkov, A., and Kancheva, S. (2011). Bulgarian-English Parallel Treebank: Word and Semantic Level Alignment. In

Proceedings of The Second Workshop on Annotation and Exploitation of Parallel Corpora, pages 29–38, Hissar, Bulgaria, September. Association for Computational Linguistics.

Šinkūnienė, J., Jasionytė-Mikučionienė, E., Ruskan, A., and Šolienė, A. (2020). Discourse markers in Lithuanian: semantic change and functional diversity. *Lietuvių kalba*, (14).

Stede, M. and Heintze, S. (2004). Machine-assisted rhetorical structure annotation. In *COLING 2004: Proceedings of the 20th International Conference on Computational Linguistics*, pages 425–431.

Stede, M., Scheffler, T., and Mendes, A. (2019). Connective-Lex: A Web-Based Multilingual Lexical Resource for Connectives. *Discours, 24*.

Sweetser, E. (1990). *From Etymology to Pragmatics: Metaphorical and Cultural Aspects of Semantic Structure*. Cambridge University Press.

Taboada, M. (2006). Discourse markers as signals (or not) of rhetorical relations. *Journal of Pragmatics*, 38(4):567–592. Focus-on Issue: The Pragmatics of Discourse Management.

Zeldes, A., Das, D., Maziero, E. G., Antonio, J., and Iruskieta, M. (2019). Introduction to discourse relation parsing and treebanking (DISRPT): 7th workshop on Rhetorical Structure Theory and related formalisms. In *Proceedings of the Workshop on Discourse Relation Parsing and Treebanking 2019*, pages 1–6, Minneapolis, MN, June. Association for Computational Linguistics.

Zeldes, A., Liu, Y. J., Iruskieta, M., Muller, P., Braud, C., and Badene, S. (2021). The DISRPT 2021 shared task on elementary discourse unit segmentation, connective detection, and relation classification. In *Proceedings of the 2nd Shared Task on Discourse Relation Parsing and Treebanking (DISRPT 2021)*, pages 1–12, Punta Cana, Dominican Republic, November. Association for Computational Linguistics.

Zeyrek, D., Mendes, A., Grishina, Y., Kurfalı, M., Gibbon, S., and Ogrodniczuk, M. (2020). TED Multilingual Discourse Bank (TED-MDB): a parallel corpus annotated in the PDTB style. *Language Resources and Evaluation*, 54(2):587–613.

Zufferey, S. (2004). Une analyse des connecteurs pragmatiques fondée sur la théorie de la pertinence et son application au TALN. *Nouveaux Cahiers de Linguistique Française*, 25:257–272.

Zwicky, A. (1985). Clitics and particles. *Language*, 61:283–305.

Östman, J.-O. (1981). *You Know: a discourse-functional study*. Giebven, Amsterdam.

Šinkűnienė, J. (2020). Konstrukcijos su nederinamuoju neveikiamosios rűšies dalyviu “galima": diskurso žymiklio link. *Deeds and Days*, 74:77–96.

Ştefănescu, A., Postolea, S., and Mititelu, V. B. (2020). The romanian discourse markers de altfel and de altminteri. patterns of use and core functions. *Revue Roumaine de Linguistique*, 65(3):307–322.