Multiword discourse markers across languages: a linguistic and computational perspective

Elena Simona Apostol, Ciprian Octavian Truică, Mariana Damova, Purificação Silvano, Giedre Valunaite Oleškeviciene, Chaya Liebeskind, Dimitar Trajanov, Anna Baczkowska, Emma Angela Montecchiari, Christian Chiarcos

Angaben zur Veröffentlichung / Publication details:

Apostol, Elena Simona, Ciprian Octavian Truică, Mariana Damova, Purificação Silvano, Giedre Valunaite Oleškeviciene, Chaya Liebeskind, Dimitar Trajanov, Anna Baczkowska, Emma Angela Montecchiari, and Christian Chiarcos. 2025. "Multiword discourse markers across languages: a linguistic and computational perspective." *International Journal of Applied Linguistics* 35 (4): 2078–90. https://doi.org/10.1111/ijal.12755.

International Journal of Applied Linguistics

WILEY

ORIGINAL ARTICLE OPEN ACCESS

Multiword Discourse Markers Across Languages: A Linguistic and Computational Perspective

Elena-Simona Apostol[1,2,3] | Ciprian-Octavian Truică[1,2,3] | Mariana Damova[4] | Purificação Silvano[5] | Giedre Valunaite Oleškeviciene[6] | Chaya Liebeskind[7] | Dimitar Trajanov[8] | Anna Baczkowska[9] | Emma Angela Montecchiari[4] | Christian Chiarcos[10]

[1]Faculty of Automatic Control and Computers, National University of Science and Technology POLITEHNICA Bucharest, Bucharest, Romania | [2]RoNLP: CLARIN K-Centre for Romanian Natural Language Processing, PRECIS Research Institute, National University of Science and Technology POLITEHNICA Bucharest, Bucharest, Romania | [3]Academy of Romanian Scientists, Ilfov 3, Bucharest, Romania | [4]Mozaika, Ltd., Sofia, Bulgaria | [5]Faculty of Arts and Humanities, University of Porto, Porto, Portugal | [6]Mykolas Romeris University, Vilnius, Lithuania | [7]Jerusalem College of Technology, Jerusalem, Israel | [8]Ss. Cyril and Methodius University, Skopje, Republic of North Macedonia | [9]University of Gdansk, Gdańsk, Poland | [10]University of Augsburg, Augsburg, Germany

Correspondence: Ciprian-Octavian Truică (ciprian.truica@upb.ro)

Received: 4 December 2024 | **Revised:** 3 April 2025 | **Accepted:** 6 April 2025

Keywords: discourse markers | ISO-based annotation scheme | linguistic linked data | machine learning prediction | multiword expressions | marcatorii discursivi | expresii formate din mai multe cuvinte (locuțiuni) | schemă de adnotare bazată pe ISO | date lingvistice interconectate | predicție folosind învățarea automată

ABSTRACT

Discourse markers (DMs) are linguistic expressions that convey different semantic and pragmatic values, managing and organizing the structure of spoken and written discourses. They can be either single-word or multiword expressions (MWE), made up of conjunctions, adverbs, and prepositional phrases. Although DMs are the focus of many studies, some questions regarding the interoperability of taxonomies and automatic identification and classification require further research. We aim to tackle these issues by offering a critical analysis and discussing the constitution of a multilingual corpus in 10 languages, i.e., English, Lithuanian, Bulgarian, German, Macedonian, Romanian, Hebrew, Polish, European Portuguese, and Italian. The novel two-level annotation approach is based on (i) signaling the existence or non-existence of DMs in a given text, and (ii) applying the ISO-24617 standard to annotate the DMs' discourse relation and communicative function in the corpora. Additionally, we introduce prediction models for detecting the presence of DMs within a text.

ABSTRACT

Marcatorii discursivi (DM-uri) sunt expresii lingvistice care transmit diverse valori semantice și pragmatice, având rolul de a gestiona și organiza structura discursurilor vorbite și scrise. Aceștia pot fi fie expresii formate dintr-un singur cuvânt, fie locuțiuni, expresii formate din mai multe cuvinte (MWE), alcătuite din conjuncții, adverbe și grupuri prepoziționale. Deși marcatorii discursivi reprezintă obiectul multor studii, unele întrebări legate de interoperabilitatea taxonomiilor și de identificarea și clasificarea automată a acestora necesită cercetări suplimentare. Ne propunem să abordăm aceste aspecte printr-o analiză critică și prin discutarea constituirii unui corpus multilingv în 10 limbi, și anume: engleză, lituaniană, bulgară, germană, macedoneană, română, ebraică, poloneză, portugheză europeană și italiană. Noua abordare de adnotare pe două niveluri se bazează pe (i) semnalarea existenței sau inexistenței marcatorilor discursivi într-un text dat și (ii) aplicarea standardului ISO-24617 pentru a

https://doi.org/10.1111/ijal.12755

adnota relația discursivă și funcția comunicativă a marcatorilor în corpusuri. În plus, în acest articol, introducem modele de predicție pentru detectarea prezenței marcatorilor discursivi într-un text.

1 | Introduction

Multiword expressions can convey various types of semantic and pragmatic information, and their study is paramount to language generation and processing. Among those studies, there are some that target multiword expressions that function as discourse markers (DM) (Heeren 2022). The analysis of discourse markers plays a vital role in understanding discourse structure, making it relevant to various fields, including linguistics and computational studies. Research in this area has resulted in multiple approaches for identifying, extracting, and classifying discourse markers across monolingual and multilingual datasets. These approaches fall into two main categories: (i) corpus-based frameworks and functional taxonomies (e.g., Cuenca 2013; Gromann et al. 2024) and (ii) computational methods (Gessler et al. 2021; Khan et al. 2022). Furthermore, recently, efforts have shifted towards creating cross-lingual tools, such as queryable discourse marker inventories, to facilitate multilingual analysis (Chiarcos and Ionov, 2021).

However, the lack of interoperable taxonomies and effective automatic identification methods for multiword discourse markers, especially in multilingual corpora, presents a significant research gap. This article seeks to explore and analyze the development of a multilingual corpus containing multiword expressions that function as discourse markers in ten languages: English, Lithuanian, Bulgarian, German, Macedonian, Romanian, Hebrew, Polish, European Portuguese, and Italian. The language selection prioritized low-resource languages, with the exception of the Germanic languages (German and English), which were included to facilitate comparison and contrast with the other language families represented: Slavic (Bulgarian, Macedonian, Polish), Romance (Italian, European Portuguese, Romanian), Baltic (Lithuanian), and Canaanite (Hebrew). We address the cross-lingual aspects of the interpretation and occurrence of multiword expressions as discourse markers, show different discrepancies in the multilingual context, and come up with a multi-lingual vocabulary of multiword expressions describing discourse markers. Further, we tackle the issues of selecting a proper annotation scheme and annotating the parallel corpus with it, by showing a novel two-level annotation approach based on first signaling the existence or non-existence of a discourse marker in a given text and secondly applying ISO 24617 - language resource management—Semantic annotation framework (SemAF, part 8 - semantic relations in discourse, core annotation schema (DR-core) (Bunt and Prasad 2016) with a plug-in to part 2 dialogue acts (Bunt et al. 2020) to annotate the discourse relation and the communicative function of the discourse marker in this text (Silvano et al. 2022; Chiarcos et al. 2022). This is, to our knowledge, the first cross-lingual annotation of discourse markers using ISO 24617 parts 8 and 2. Finally, we present machine learning approaches to predict discourse markers' presence or absence in text chunks based on Transformer models and analyze the results from the perspective of MWE in a cross-lingual context.

The current study aims to answer the following research questions:

RQ1: How does the two-level annotation approach improve the accuracy and richness of discourse marker identification compared to single-level annotation schemes?

RQ2: What are the challenges and best practices encountered when constructing a multilingual corpus of discourse markers, particularly concerning MWEs, across multiple typologically diverse languages?

RQ3: How do cross-linguistic variations in the realization of discourse markers compare to instances of stable, literally translatable discourse markers across languages?

The manuscript is structured as follows. In Section 2, we discuss multiword discourse markers and current methods for their detection within textual corpora. In Section 3, we present the methodology used in this study. In Section 4, we present our findings on multiword discourse markers in 10 distinct languages. Finally, we conclude our work and hint at future research directions.

2 | Discourse markers

2.1 | The Concept and Approaches

Discourse markers are a set of linguistic expressions that are an inseparable part of discourse and serve crucial purposes in the understanding of spoken and written discourse. Discourse markers may be single-words or MWEs made up of conjunctions, adverbs, and prepositional phrases (Fraser, 2009). They indicate a link between discourse units, i.e., utterances, longer stretches of text, and even the text and the extralinguistic background. Discourse markers fulfill multiple functions in both monologues and interactive communication, such as conversations and dialogues. Their roles include, but are not limited to, establishing coherence between clauses and sentences, indicating hesitation, facilitating turn-taking, signaling topic shifts, marking turn boundaries, expressing hedging, conveying attitude, managing interactions with interlocutors, seeking approval (Jucker and Ziv, 1998), and indicating transitions (Heeman and Allen 1999). Thus, in the context of this study, we define the term *discourse marker* as a linguistic element that functions primarily to structure discourse, signal relationships between utterances, and guide interpretation rather than contributing to propositional meaning. These markers help manage coherence, cohesion, and interaction in both spoken and written communication.

The concept of discourse markers has been widely debated (Zwicky 1985; Schiffrin 1987), and the term is often used interchangeably with others, such as discourse particle

(Schourup 2018), discourse connective (Blakemore 2004), pragmatic marker (Fraser 1996), and pragmatic particle (Ostman 1981). Some of these terms are typically associated with a specific theoretical approach; for example, Maschler and Schiffrin (2015)'s research adheres primarily to the early, coherence-oriented integrative approach, Blakemore (1987, 2004) represents the relevance-theoretic framework, and Aijmer (2013) discusses pragmatic markers in the variational pragmatic approach she proposes. In this paper, we shall use the most general and popular term discourse marker to mean what is also spanned by all the other terms proposed by various authors.

The study of discourse markers has been approached from various perspectives, ranging from structural and discourse organization analyses to more specific examinations of their role in establishing local dependencies (Prasad et al. 2008) and their attitudinal or affective functions (Sanders et al. 1992). While the early research highlighted discourse markers primarily as elements that help organize discourse in units or saw them as topic transition devices that bridge parts of discourse, in later discussions, such as the epistemic stance proposed by Schourup (2018), discourse markers were assigned the function of signaling mental activity (e.g., consideration). Ochs (1996), in turn, investigates the affective stance of discourse markers, i.e., attitude, mood, feeling, etc., that they can convey, while Aijmer (2013) focuses on pragmatic markers as expressions of politeness or uncertainty. Finally, Fischer (2006) treats pragmatic markers mainly as indicators of interactants' involvement. In this study, all these functions are seen as relevant in a discussion of discourse markers.

Concerning the meaning of discourse markers, there has been a recent trend of leveraging translation data to get insight into the exact meaning of the studied linguistic components. Such a cross-linguistic strategy may aid in the establishment of semantic-pragmatic domains and give insights into the multifunctionality of discourse markers and their relationship to semantic and pragmatic polysemy.

A particular discourse marker can simultaneously carry different functions that correspond to different shades of meaning (Bazzanella et al. 2007). Therefore, by selecting the equivalent, the translator emphasizes a particular meaning over other meanings, thereby aiding in the process of making explicit the various meaning components involved in the use of a particular discourse marker. In our study, we will adopt this approach, by selecting a parallel corpus with English as the pivot language. Thus, with our approach, we try to alleviate any concerns regarding the construction of text corpora (Biber 1990) and provide new insights into corpus-based analysis.

2.2 | Discourse Markers Detection

Currently, the great majority of discourse marker corpora are manually annotated, largely by qualified linguists and less so by non-specialists, while only a tiny number are automatically or semi-automatically annotated (with human supervision). In fact, discourse marker detection has seen very little development in the current literature due to ineffective extraction methods and data sparseness (Sileo et al. 2019, Damova et al. 2023). Furthermore, the focus has been on predicting discourse markers in English, other languages mostly being ignored, especially low-resource languages.

Sileo et al. (2019) use FastText and shallow features to predict discourse markers, and with this approach on a large web-collected dataset, the authors manage to identify and create a curated list of 174 English discourse markers. Ji and Huang (2021) introduce discourse-aware discrete variational transformer (DISCODVT) as a model for generating long texts. DISCODVT first learns sequences that summarize the textual data and then decodes them into a discrete latent representation that incorporates discourse-aware sentences. To embed textual data and capture the overall structure of the text, the model employs a bidirectional encoder, which generates contextualized token representations. During decoding, the latent embeddings are rescaled and added to the embedding layer of the decoder. To obtain the latent encoding of the discourse markers, the model uses the penn discourse treebank 2.0 (PDTB) (Prasad et al. 2008) to extract adjacent elementary discourse units. Sileo et al. (2020) propose DiscSense to create associations between discourse markers and various tasks' labels with a model that predicts the discourse markers between sentences using a fine-tuned bidirectional encoder representations from transformers (BERT) model (Devlin et al. 2019). Regarding the automatized process of discourse marker identification and classification, Zufferey and Degand (2017) also describe a three-step process: identifying the existence of discourse markers, assigning inferential semantic functions to discourse markers, and determining the scope of unique functions. Kurfali (2020) applies a BERT-based model to perform shallow discourse parsing (SDP), a method designed to identify explicit local discourse relations without the need for complex tree or graph structures. DisCoDisCo (Gessler et al. 2021) is a system that uses bidirectional long short-term memory (Bi-LSTM) (Hochreiter and Schmidhuber 1997) and BERT for discourse segmentation, classification, and connective detection. This model outperformed state-of-the-art solutions. However, the connective detection task was found to be notably challenging. This observation has led to the conclusion highlighting the need for further research in this area (Braud et al. 2023).

3 | Our Study

3.1 | Multilingual Corpus

The parallel corpus we developed includes data from 10 languages, utilizing publicly available TED Talk transcripts as an extension of the TED-EHL parallel corpus, which is hosted in the LINDAT/CLARIN-LT repository[1]. This multilingual corpus consists of language alignments with English serving as the pivot language, comprising 1.3 million sentences. The selection is based on the presence of multiword expressions (MWEs) that function as discourse markers, guided by theoretical insights from Maschler and Schiffrin (2015) and the classification framework provided by Fraser (2009). However, the MWEs of our selection can be ambiguous, e.g., in some contexts, they are interpreted as discourse markers, while in other contexts, they are interpreted as content words, as in Examples 1 and 2 below. In Example 1, the multiword expression "you know" functions as a discourse marker (annotated as 1), serving to introduce a new discourse

TABLE 1 | Compiled multilingual datasets.

Language	Aligned sentences with MWE
English	43 600
Macedonian-English	2 846
German-English	15 852
Lithuanian-English	4 112
Bulgarian-English	19 209
European Portuguese-English	4 398
Polish-English	17 408
Romanian-English	18 946
Hebrew-English	23 566

message. In contrast, in Example 2, it acts as a content word (annotated as 0), fully integrated into the sentence structure.

Example 1.

That's ridiculous. You know, this is New York, this chair will be empty, nobody has time to sit in front of you.

Example 2.

You know some people who say "Well".

To annotate the corpus, we have applied a two-step approach by first identifying the discourse markers' presence and then their semantic or pragmatic value.

The multilingual corpus encompasses utterances from English, Lithuanian, Bulgarian, European Portuguese, Macedonian, Polish, Romanian, Hebrew, Italian, and German. The bilingual parallel corpora—English and one other language—counted more than 10K utterances on average (see Table 1, each utterance being uniquely identified with a combination of three types of IDs. To obtain a consistent parallel corpus of 10 languages, the bilingual corpora were automatically compared, and the intersection of all 10 corpora was singled out and split again into bilingual corpora, containing the English examples and the examples of one other language.

For step one, the corpus structure in Table 2 is used. The first three columns contain IDs, followed by four columns related to the English utterance. These include the MWE, a description of the discourse marker, a brief context in which it appears, a broader context window, and an annotation indicating whether the MWE functions as a discourse marker in the text. For target languages other than English, the subsequent four columns present the same information for the respective language. The annotators were supposed to evaluate whether the MWE plays the role of a discourse marker or not and fill in column 6 with 1 (in role as discourse marker) or 0 (in role as content word) for English and in column 9 for the target language accordingly. This way of representation avoids many of the complications that conventional discourse annotation systems have, namely that different discourse markers in the same sentence can be annotated in multiple way and these annotations all need to be condensed in a coherent and human-readable format that preserves overlapping argument spans and crossing relations.

This structure allows the application of machine learning methods to predict the presence of discourse markers in an unseen text, as discussed in Section 3.2, and has been extended with further columns to allow annotation based on the annotation scheme described in Section 3.3.

In a subsequent processing step, the original text files had to be converted to a more sustainable and self-contained representation that provides the full textual content along with the annotations. For this purpose, we employed the CoNLL-RDF (a tool for converting between formats of annotated linguistic corpora and annotations, as well as linking and enriching these with external ontologies) format (Chiarcos and Fäth 2017) to represent the original text in tokenized form along with (optional) morphosyntactic and/or syntactic annotations and enriched the CoNLL-RDF graphs with POWLA (a generic formalism to represent linguistic annotations in an interoperable way by means of OWL/DL) (Chiarcos 2012) nodes (for arguments and discourse markers) and relations (for the relations between arguments and discourse markers). While this representation can be easily and effectively queried with SPARQL (an RDF—resource description framework—query language), it is, of course, much less human-readable than the native tabular annotation format. A key benefit of this way of representation is that the query can access both annotated data and the schema definition. As we also provide a formalization of the annotation scheme in RDF/OWL (resource description framework/web ontology language), the taxonomy of discourse relations and their features can be consulted at query time (Chiarcos et al. 2023).

3.2 | Model for Discourse Markers Prediction

We have used a manually annotated and validated segment of the parallel corpus in English, Lithuanian, Bulgarian (Valūnaitė Oleskevicienė et al., 2021), and subsequently in Italian (Montechiari et al. 2022) and trained two cross-language machine learning models based on FastText (Bojanowski et al. 2017) and XLM-RoBERTa-Large (Conneau et al. 2020) to predict the existence of discourse markers in unseen text. A learning rate of 0.00001 was used to train the model for 3 epochs. The k-train library (Maiya 2022), built on top of the HuggingFace (Wolf et al. 2020) transformer library, was used to fine-tune the model. An 80%-20% train-test split was used for training and testing the models. The fine-tuning was done for 10 iterations. The results of these experiments, provided in Table 3, show very good performance for Lithuanian with the two models, and different scores for the two models, run on Bulgarian data.

The observed performance differences could be attributed to the architectural variations between FastText and XLM-RoBERTa. FastText, being a more traditional word embedding model, might struggle with capturing contextual information and handling complex discourse marker constructions. XLM-RoBERTa, on the other hand, is a transformer-based model that excels at capturing long-range dependencies and understanding context. Also, the

TABLE 2 | Structure of the parallel corpus files.

Column	Description
id	Unique identifier
vid	Video unique identifier
lid	Line unique identifier
DM EN	Discourse marker in English
Sentence chunk EN	The sentence in English where the DM appears
Larger textual context EN	The full paragraph in English
DM presence EN	The presence of a DM in English, i.e., 1 present, 0 otherwise
Sentence chunk TL	The sentence in the TL where the DM appears
Larger textual context TL	The full paragraph in the TL
DM presence TL	The presence of a DM in the TL, i.e., 1 present, 0 otherwise DM
Target language	Discourse marker in the TL

TABLE 3 | Results achieved on datasets in different languages using cross-lingual methods.

Model	Accuracy	Precision	Recall	Specificity	F1-Score	MCC
FastText (EN)	0.4558	0.6515	0.1928	0.8467	0.2976	0.0507
FastText (BG)	0.5764	0.6457	0.6457	0.4733	0.6457	0.1191
FastText (LT)	0.9321	0.9369	0.9942	0.0548	0.9647	0.1285
FastText (IT)	0.5700	0.7400	0.5100	0.6800	0.6000	
XLM-RoBERTa (EN)	0.9180	0.8900	0.7860	0.9130	0.9030	0.8080
XLM-RoBERTa (BG)	0.8260	0.8260	0.8300	0.8220	0.8290	0.6520
XLM-RoBERTa (LT)	0.8289	0.9899	0.8242	0.8904	0.8995	0.4393
XLM-RoBERTa (IT)	0.6900	0.8000	0.6900	0.6900	0.7400	0.3700

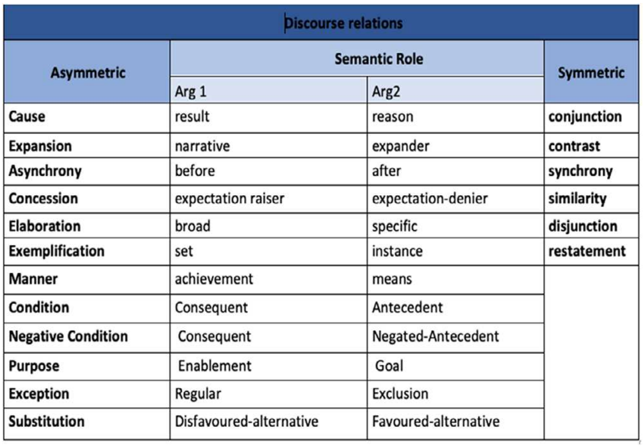
complexity of the linguistic features of each language must be considered. For example, Bulgarian, with its rich and complex morphology, might benefit more from XLM-RoBERTa's ability to capture morphological information, while Italian, with its relatively simpler morphology, shows less of a difference between the two models.

Besides using FastText and XLM-RoBERTa-Large to predict discourse markers on the Italian annotations from the annotated dataset, the language-agnostic BERT sentence embedding (LaBSE) (Feng et al. 2022) model was used. Although deep learning models have traditionally been developed by training individual languages separately, this BERT-based multilingual transformer model generates fixed-length vector representations for sentences and proves to be effective for low-resource languages. LaBSE was used to produce sentence representations to detect the predicted discourse markers within sentences using a binary classification task, i.e., 1 if the discourse marker is present and 0 otherwise (see Section 3.1). Furthermore, in this set of experiments, the model is trained on the English datasets, and, by using transfer learning, the evaluation was performed on the Italian dataset.

Further, the model, trained on English was run not only on the Italian dataset, but also on the 9 other languages from the parallel corpus, and a validation by native-speaker linguists was carried out on randomly selected 100 text contexts, different for each language for a total of 1000 validated predictions. The results are presented in Table 4, showing an average of 12 incorrect predictions and a precision rate of 88%.

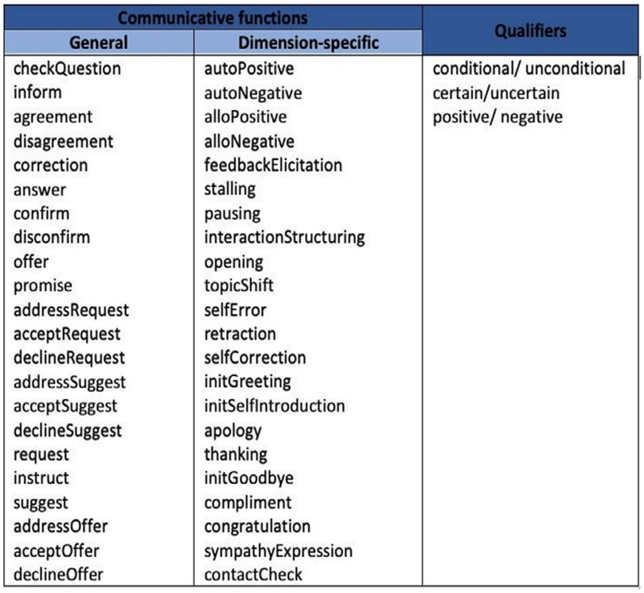
TABLE 4 | Human validation results.

Language	Number of Wrong Predictions	Total Number of Examples	Precision ratio
BG	10	100	0.90
MK	19	100	0.81
EN	16	100	0.84
HE	5	100	0.95
PT	20	100	0.80
DE	17	100	0.83
PL	10	100	0.90
LT	12	100	0.88
RO	1	100	0.99



Discourse relations			
Asymmetric	Semantic Role		Symmetric
	Arg 1	Arg2	
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 1 | The set of discourse relations outlined in ISO 24617–8 (Bunt and Prasad 2016). [Color figure can be viewed at wileyonlinelibrary.com]



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	

FIGURE 2 | The set of communicative functions and qualifiers outlined in ISO 24617-2 (Bunt et al. 2020). [Color figure can be viewed at wileyonlinelibrary.com]

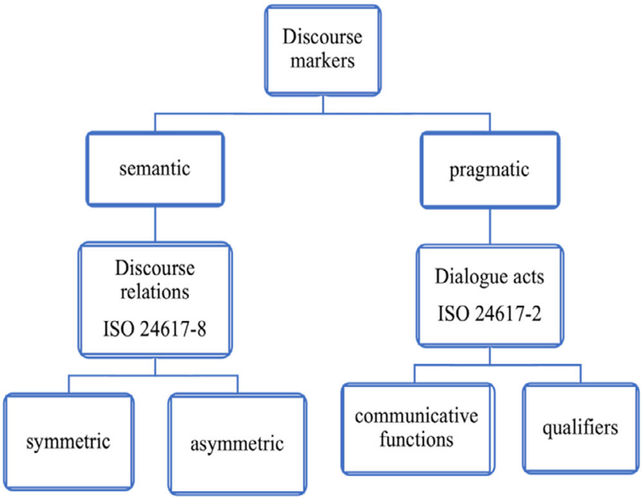


FIGURE 3 | The two-level annotation scheme (Silvano et al. (2022)). [Color figure can be viewed at wileyonlinelibrary.com]

The causes of these discrepancies in the correct prediction rate have yet to be analyzed. We anticipate that they may be linked to factors such as the nature of the texts, the expert judgment of the human analysts, and the differences in language structure compared to English.

3.3 | Two-Level Semantic Annotation Schema

The annotation scheme that we propose (Silvano et al. 2022; Silvano and Damova 2023) comprises two interlinked levels, thus enabling the representation of the semantic and pragmatic values of multiword discourse markers. After careful consideration of the different proposals and interoperability being a relevant issue for the purpose of our project, we deemed it best to utilize ISO 24617 semantic annotation framework (SemAF), more specifically, Part 8–semantic relations in discourse, core annotation schema (DR-core)–ISO 24617–8 (Bunt et al. 2020), and Part 2-dialogue acts. Accordingly, we propose that, whenever a discourse marker carries a semantic value, the annotator should resort to the set of discourse relations put forward by ISO 24617–8. Since the discourse relations can be symmetric or asymmetric, depending on the arguments having or not having the same semantic role, the annotator has to identify the role of each argument in case the discourse relation is symmetric.

Figure 1 presents the core set of discourse relations proposed by ISO 24617–8 and included in our annotation scheme. Because multiword discourse markers can convey a pragmatic value, a one-level annotation scheme does not suffice. That is why the plug-in into part 2 of ISO 24617 is crucial. ISO 24617-2 introduces a model for the annotation of dialogue acts postulating several dimensions, communicative functions, and qualifiers. For simplicity, and because it meets the annotation needs of our corpus, the annotation scheme we developed includes only the set of communicative functions and qualifiers defined in Figure 2.

Figure 3 outlines the two-level scheme that we developed for the annotation of semantic and pragmatic values of multiword discourse markers. The design of the annotation scheme was followed by the preparation of an instruction manual on how to annotate the data, which included not only the definitions of the discourse relations, communicative functions, and qualifiers but also some illustrative examples. Concurrently, the English dataset was manually annotated by an expert, and this acted as the gold standard for the other datasets. These were manually annotated by native speakers in a spreadsheet, and, whenever doubts arose, group discussions were organized to reach a consensus on how to proceed with respect to the annotation. During these group discussions, the annotation schema was presented in detail with examples in English to better grasp the concept for each discourse relation, communicative function, and qualifier. For languages that belong to the same family (e.g., Romance, Slavic, etc.), words with the same etymology were used to cross-reference the concepts from the schema.

The annotation task includes the consideration of different parameters. The annotator begins by observing the pivot language (English), in what concerns the sentence part where the discourse marker occurs, the larger textual context, and the MWE that was extracted from the corpus. Next, the annotator determines if that MWE is in fact a discourse marker or not, writing 1 or 0, respectively. The next step is analyzing the target language, looking at the context where the MWE occurs, and ascertaining whether or not, it is indeed a discourse marker. Subsequently, the sections of the text outlining the first and second arguments of the assigned discourse relation are recorded. It follows the identification of the discourse relation and the roles of arguments one and two, whenever the discourse marker carries a semantic value, and/or of the communicative function and qualifier (if necessary) when the discourse marker takes a pragmatic value.

4 | Analysis of the Dataset

In this section, we aim to provide an in-depth exploratory data analysis of our proposed multilingual corpus. For this analysis, we selected 55 examples extracted from the English dataset and the respective parallel data from the other languages. These examples represent unique matching contexts across all languages, out of a total of 44,192 distinct textual contexts for English. As a result, there are 550 examples in the set of 10 languages, which were annotated using the ISO-based annotation scheme from Figure 3. For each language, we had between 1 and 2 annotators. The annotators were recruited from the members of the NexusLinguarum COST Action.

4.1 | English

English being the pivot language for all language pairs of our corpus, we conducted a baseline annotation following the annotation framework described in the previous section. First, we determined whether the MWE signaled in the text was indeed a discourse marker. As a matter of fact, out of the 55, examples 15 were not instances of discourse markers and were excluded from the annotation. Second, we proceeded to assign an ISO discourse relation, or ISO communicative function/qualifier, or both whenever necessary, to the discourse marker of the example to represent its semantic or pragmatic value.

In the English dataset, the multiword discourse markers acted as clues to a relatively considerable variety of discourse relations, proving not only the richness of our corpus but also the range of ISO 24617–8. Although, in this sample, we came across a small set of discourse markers with a pragmatic value, nonetheless, the communicative functions and qualifiers included in our annotation scheme were relevant to properly capture the interactional meaning of those discourse markers.

In this sample, the most frequent discourse marker is “I think”, with 11 occurrences conveying the meaning of attribution, which is in accordance with the nature of the text from which the multiword discourse markers were extracted. Since TED talks are monologues mainly of an argumentative character, it is expected to have a high frequency of discourse markers with this value. The meaning of expansion is also expressed quite often (9 occurrences) by two multiword discourse markers, “in fact” and “of course”. In two instances, these two discourse markers have also a pragmatic value of certain for the latter and of confirm for the former. Exemplification realized by either “for example” or “for instance” is the third most common value in the English dataset, followed by restatement with the discourse markers “in other words” and “I mean”. As mentioned before, discourse markers with communicative functions and qualifiers occur less frequently, “of course” and “in fact” being the only two examples.

4.2 | German

One characteristic of German is the abundant use of modal adverbs, many of which can be interpreted as discourse markers. In consequence, this means that German seems to use discourse markers more abundantly than English, at least. However, our annotation is based on discourse marker candidates predicted from the English source, so that only cases have been annotated in which a multiword discourse marker was identified for the English source. Thus, discourse marker candidates without an obvious correspondent in the English text have not been considered.

German also has a reputation for being rich in morphological compounding, so that one may expect that a number of English MWE would indeed correspond to single, but morphologically complex words in German. To some extent, this is what we observed: out of a sample of 40 candidate discourse markers in German predicted from the English source text, we found that 8 (14%) cases used single-word expressions. The majority of these, however, originate from phrasal expressions: “demzufolge” (“because of that”, lit. “from that to follow”), “bislang” (‘so far’, lit. “until long”), “andererseits” (“on the other hand”, lit. “of the other side”). One is a derivation of a nominal compound: “tatsächlich” (“indeed”, lit. “fact-like”), from “Tatsache” (“fact”, lit. “deed-thing”). Only two single-word discourse markers: “natürlich” (“of course”, lit. “nature-like”), “nämlich” (“in fact”, lit. “name-like”) are morphologically derived from simple nouns.

Another characteristic of German is relatively free word order, which can also be used to express certain discourse phenomena, in particular continuity and contrast, so that there are grammatical devices that may make an overt discourse marker superfluous at times. Yet, in only four cases (of 40 in the sample), we found that German did not provide an explicit discourse marker in a case in which English did have a multiword discourse marker.

Overall, we found that in the majority of cases, an English multiword discourse marker in English corresponds to a multiword discourse marker or a phrasal expression in German, an observation that we attribute to the close linguistic relationship between both languages. There are a number of literal correspondences, including “ich denke” (“I think”), “in anderen Worten” (“in other words”), “das ist” (“that is”), and “an diesem Punkt” (“at this point”). However, these may be overrepresented in the data because of a translation bias from the English source text (as evident, for example, from the usage of “in anderen Worten” instead of the more common “mit anderen Worten”, which is clearly influenced by the English translation source ‘in other words’).

4.3 | Lithuanian

Lithuanian researchers identify conjunctions, sentential relatives, particles, adverbs, and pronouns, different verbal and nominal forms, and their constructions as the main grammatical classes for lexical units that function as discourse markers. The use of particles as discourse markers in Lithuanian is one of the characteristic features, especially in spoken discourse (Šinkūnienė et al. 2020). The analysis of the annotated data sample reveals that Lithuanian discourse markers undergo some cases of omission (out of 40 annotated discourse markers in English, 36 discourse markers can be annotated due to their presence in the Lithuanian text, the rest 4 are omitted). Another observation is related to the lexico-grammatical nature of the Lithuanian language that English multiword discourse markers become one-word discourse markers (43% of Lithuanian discourse markers are translated into one-word discourse markers). The analyzed sample contains the omission cases of discourse markers being transformed or integrated into different grammatical structures that do not include any presence of a discourse marker. What concerns the semantics of the annotated discourse relations in both languages English and Lithuanian, is that they demonstrate semantic stability, being annotated by the same semantics of discourse relations. The most common discourse relations in the analyzed sample are attribution with 11 instances (“I think” turned into “manau”) and exemplification with six instances (“for example” turned into “pavyzdžiui”). Concerning the communicative functions, there are quite a few in the annotated sample expressing just confirm and inform cases.

4.4 | Romanian

The English-Romanian parallel corpus consists of 40 unique records, and we found no omissions. The most common discourse relations in this corpus are attribution, expansion, and exemplification. To exemplify attribution discourse relation, in Romanian, we can use “cred că” (in English, depending on the context, it can mean either “I think” or “I believe”). Expansion is marked by “de fapt” (“in fact”) or “de sigur” (“of course”). These markers are also used for confirmation, i.e., the communicative function Confirm. Finally, the Exemplification discourse marker by “de exemplu” (“for example”) and “astfel” (“for instance”)

The Restatement discourse relation has a rethink function, and it is used to express again what has been said but in a different form. In Romanian, the most common Restatement discourse markers are “adică” (“that is”) and “cu alte cuvinte” (“in other words”). This can also be seen in our English-Romanian parallel corpus, where “cu alte cuvinte” (“in other words”) is the most used discourse marker in restatement discourse relations.

Different combinations of discourse markers were observed in the corpus: expansion, i.e., “de fapt” (“as a matter of fact”, or “in fapt”), “mai ales” (“especially”), or exemplification, i.e., “de exemplu” (“for example”) with attribution, i.e, “cred că” (“I believe”), to give specific examples for broad situations that inform or confirm the views of the discourse participants using certain, e.g., “de fapt, eu cred” (’in fact, I think’), or uncertain, “de exemplu, cred că” (“for example, I believe”), discourse qualifiers. In the current literature, the discourse markers “de fapt” (“in fact”) together with “de altfel” (“by the way”) and “de altminteri” (“otherwise”) are used to indicate the discourse coherence relation of specification in the rhetorical domain (Crible and Degand 2019; Ştefănescu et al. 2020).

Expansion markers, i.e., “chiar” (“even”), “până’ şi” (“even”), “adică” (“namely”), “deci” (“thus”), “bineînţeles” (“of course”), “de fapt” (“in fact”), are used mainly to expand the narrative of the speaker to create a better understanding and experience of the listener. In our corpus, we observe the presence of bineînţeles (“of course”), de fapt (“in fact”). These markers belong in the consecutive connectors class, as “deci” (“so”), “astfel” (“thus”), “de asemenea” (“likewise”), etc., which do not appear in our corpus but are worth mentioning as they are widely used in Romanian.

4.5 | Bulgarian

The Bulgarian examples of the corpus show several discrepancies with respect to the manifestation of discourse markers in the other languages. Apart from the literal lexicalizations like “разбира се” (“of course”), “например” (“for example”), and the like, in many cases the communicative nuance conveyed by the discourse marker in English appears in Bulgarian as modified word order, different tense form, interrogative phrase, imperative form or plain omission, e.g. “you see“—”разберете”, “разбирате ли”, “разбирате сами”, “можете дори да видите”, “нали разбирате”, etc. This makes the application of the described approach harder as it is difficult to single out MWE to label the discourse relation or communicative function conveyed in the text.

With respect to the most common semantic values of discourse markers, the tendency is the same as the other languages, e.g., Attribution, Expansion, and Exemplification, although the cases of Attribution are fewer than in the other datasets. On the other hand, the communicative function Confirm comes across more times in the set of Bulgarian discourse markers when compared to the other sets. Further, the analysis of the discourse relations and the communicative functions conveyed by certain discourse markers identified a tendency of interdependence between the two scales leading to classifying the discourse marker as introducing both a discourse relation and a communicative function, like Expansion and Confirmation, Certainty, like in the case of “разбира се” (“of course”).

4.6 | European Portuguese

In the European Portuguese subcorpus, out of the 40 examples of the English dataset where the MWE is utilized as a discourse marker, only in four of them is the discourse marker omitted. This omission causes the loss of a pragmatic value of certainty conveyed by the discourse marker in the English data. The other two translations transpire the same semantic value using other linguistic mechanisms. It is the case of the use of the emphasis particle ´e que instead of the discourse marker that is.

With respect to the morphosyntactic nature of the discourse marker, about 39% of the occurrences in the European Portuguese data are single-word discourse markers, contrary to the English dataset. This discrepancy is mainly a result of the fact that European Portuguese is a null-subject language, and, since there are many examples with the discourse marker "I think", the great majority of them is translated only by the verbal forms "penso" or "acho", leaving out the subject, which can be recovered by the verb morphology. The second case is the equivalent discourse marker of "of course", which in Portuguese is only one word "claro".

Similar to other datasets, we found lexical variability in discourse markers within the European Portuguese data which showed diverse word choices when translating discourse markers from English. Specifically, "de facto" and "na verdade" were both used to translate "in fact".

Being the most frequent discourse marker "penso", "eu penso" or "acho", the most frequent discourse relation is attribution with 11 occurrences, followed by exemplification marked by "por exemplo" with seven instances. Expansion with five cases is, in the European Portuguese dataset, the discourse relation signaled by the wider array of discourse marker "ou seja", "claro", "na verdade", and "de facto". Restatement is communicated mainly by the multiword expression "por outras palavras". In what concerns communicative functions, despite the vast list included in our annotation in this sample only two came out of the European Portuguese sample, that is, confirm and inform. Solely the qualifier certain was necessary to annotate the pragmatic value of the discourse markers in our corpus in cases such as "claro" or "de facto".

4.7 | Hebrew

Only four of the 40 instances in the English dataset where the MWE is used as a discourse marker are omitted in the Hebrew corpus. One omission causes the pragmatic value of assurance communicated by the of course discourse marker in the English data to be lost. The other three translations reveal the same semantic content using alternative linguistic mechanisms, personal pronouns, such "והן" ("and they" (female)) and "והוא" ("and he") replace the discourse marker that is.

Contrary to the English dataset, approximately 35% of the occurrences in the Hebrew data are single-word discourse markers. This disparity is frequently caused by the replacement of discourse markers' prepositions with prefixes. For instance, the "for example" discourse marker is translated to "ל+דוגמה". There are some examples of the discourse marker "I think". We discovered lexical variations of this discourse marker in the Hebrew examples: "חושב אני" ("I think"), "סבור אני" ("I believe"), "לי נראה" ("it seems to me"), "לדעתי" ("in my opinion").

The discourse relations and communicative functions identified in the English examples were not consistently mirrored in their Hebrew translations. For instance, the English discourse relation "expansion", marked by "ובעצם" ("in fact"), was replaced by the Hebrew the "לדוגמה" ("for example"), indicating an "exemplification" relation. Sometimes a discourse marker is replaced by another semantically related discourse marker, such as in particular ("בפרט"), which was translated to "בעיקר" ("mainly").

4.8 | Macedonian

While there are some differences between Macedonian and English discourse markers, they share many similarities regarding their function, role in discourse, and use of intonation. The observed similarities underscore the global function of discourse markers in guiding the organization and flow of communication.

In the Macedonian corpus, only one example is omitted out of the 40 examples of the English dataset where the MWE is utilized as a discourse marker. Concerning the morphosyntactic nature of the discourse marker, about 1/3 of the occurrences in the Macedonian data are single-word discourse markers, contrary to the English dataset.

In the analyzed corpus, the most common discourse marker is "I think" which is translated into "Мислам дека" in Macedonian. "Мислам дека" in the Macedonian language is used to express the speaker's opinion, belief, or assumption about a situation or event. The English discourse marker "I think" conveys a similar function of expressing the speaker's perspective. However, the use of "мислам дека" in Macedonian may have a stronger emphasis on the speaker's personal viewpoint or subjectivity compared to the more neutral or objective use of "I think" in English. The second most common discourse marker in the dataset is "На пример", which corresponds to the English "for example" and 'for instance. "На пример" serves a discourse-organizing function and can help to clarify a point, make the discussion more concrete, or make the information more accessible to the listener.

In conclusion, while Macedonian and English discourse markers may have similarities in terms of their function and use, they also have differences that reflect the linguistic, cultural, and communicative context in which they are used.

4.9 | Polish

Within the forty contexts, only 24 (60%) contained the Polish equivalent of the English discourse marker, which means the omission of the original English discourse markers at the level of 40%. Compared to other languages analyzed in this study, this number is high. The omitted discourse markers comprise: "on the one hand", "you see", "so far" (twice), "in fact" (twice), "of course", "I think" (twice), "as a result", "for example", "in other words", "that is", "I mean". Attribution is the most common discourse marker value, which adds up to ca. 33% (8 occurrences), Exemplification encodes 25% of the total contexts, ca. 13% manifests Expansion, ca. 3% takes up Contrast, the same percentage goes for Restatement, and Elaboration is represented by only one context (ca. 4%). The communicative function Confirm, with the qualifier Certain appeared three times (ca. 13%), with reference to "of course" (Pol. "oczywiście").

In the Polish corpus, the most common discourse marker is the equivalent of two English Exemplification discourse markers, which are "for example" and "for instance"; they are all rendered

TABLE 5 | The frequency of the discourse markers values.

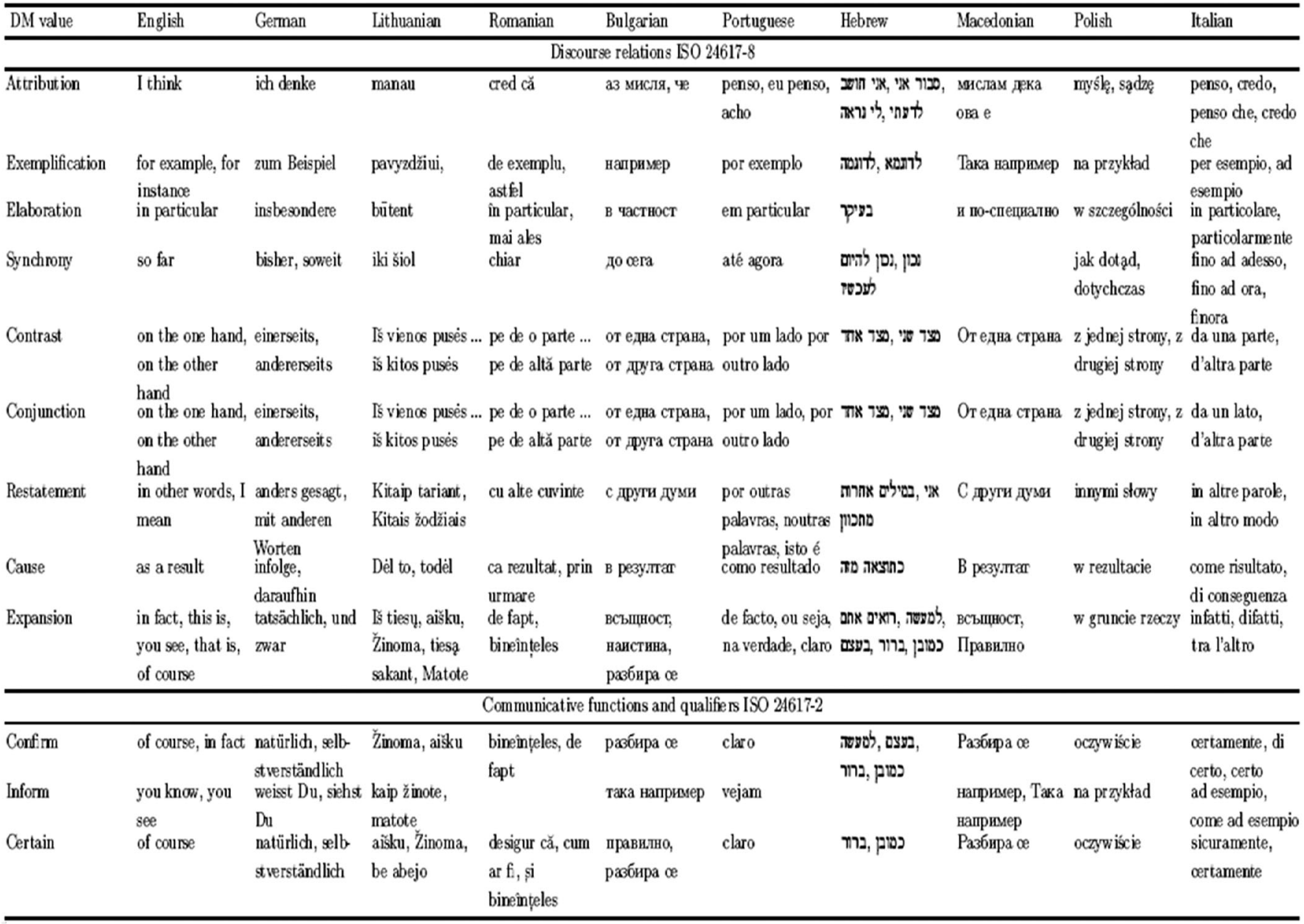
DM value	EN	GE	LT	RO	BG	EP	HE	MK	PL	IT
Discourse relations ISO 24617–8										
Attribution	11	8	11	11	6	11	11	11	8	10
Expansion	9	4	6	8	6	5	7	6	3	8
Exemplification	7	3	6	8	6	7	7	7	6	6
Restatement	4	2	4	4	3	4	4	5	2	3
Elaboration	1	1	1	2	2	1	1	3	1	2
Synchrony	1	0	1	0	0	1	1	1	0	2
Contrast	1	0	0	1	0	1	0	0	2	1
Conjunction	1	1	1	1	0	1	1	0	0	1
Cause	1	1	2	2	2	1	1	1	0	1
Concession	0	0	0	1	0	0	0	0	0	0
Manner	0	0	0	1	0	0	0	0	0	0
Communicative functions and qualifiers ISO 24617-2										
Confirm	2	0	2	6	5	2	0	1	3	2
Inform	0	0	2	0	0	1	0	0	0	0
Suggest	0	0	0	0	1	0	0	0	0	0
Certain	0	0	0	0	0	0	0	0	3	2

into Polish as “na przykład”, as there are no separate equivalents in Polish to distinguish between the two original versions. The second most popular discourse marker relates to the English “I think”, which was translated into Polish in three variants, all of which are semantically synonymous and signal “attribution”: “myślę” (5 occurrences), “uważam” (2 occurrences), and “sądzę” (1 occurrence). There is some stylistic difference in the three equivalents, however, with “myślę” being the most common and colloquial one, and “uważam” the most sophisticated option. The English “of course” occurred three times in the Polish dataset, and it was translated literally as “oczywiście” (3 occurrences). As Polish is a pro-drop language, wherein the pronouns used in the function of a subject are typically omitted, the equivalents of “I think” are transferred as one-word discourse markers (in total 8 instantiations). Consequent upon this grammatical rule, there are 11 examples (44%) of single-word discourse markers in Polish, 12 instantiations of two-word markers (48%), and 2 cases of MWE consisting of three words (8%). Of these, four discourse markers are prepositional phrases: “na przykład” (“for example”), “w szczególności” (“in particular”), “z jednej strony” (“on the one hand”), and “z drugiej strony” (“on the other hand”).

4.10 | Italian

The analysis of the Italian annotated corpus focuses on the peculiar characteristics of Italian discourse markers. The syntactic position is surely one of the main points regarding this. Unlike in English, in Italian, the interaction function in the discourse is covered by the position within the sentence and the punctuation preceding or following it. For this reason, spaces and punctuation have been included in the quotation of expressions, and the presence of capital letters, or otherwise lower case, whether at the beginning of the sentence or not. Moreover, the corpus, being made of transcripts of spoken language (oral speech), punctuation has a strong significance in describing the intonation used by the speaker. And consequently, as intonation in the Italian language is essential to signal the change of function, from an expression exclusively related to content to one linked to discursive interaction, it was taken as an important clue for the analysis. Therefore, the larger context has often been used, except in cases where punctuation undoubtedly leads to a mark-making function.

A difficulty in recognition has been experienced with verbal expressions. Verbs in several phrasal examples have an essential significance for the overall sentence meaning and at the same time a relational and underlining significance for the discursive instance. One of the most frequent examples of this type of category is that of the verb “essere” (“to be”), to which various forms of expressions are associated, covering different syntactic categories. Most of them refer to subordinate propositions that the verb to be associated with the connective “che” (“that”) connects to the main ones. When, on the other hand, the verb is associated with an adversative connective, they are usually placed at the introduction of a subordinate. In these two cases, the expression is used to support the connection between the two expressions, which in this way are well amalgamated within the discourse. Another category of expressions used as markers in Italian are those that use the point of view to emphasize the presence of the speaker within the discourse. For this category, we frequently have three verbs: “intendere” (“to intend”), “pensare”



DM value	English	German	Lithuanian	Romanian	Bulgarian	Portuguese	Hebrew	Macedonian	Polish	Italian
Discourse relations ISO 24617-8										
Attribution	I think	ich denke	manau	cred că	аз мисля, че	penso, eu penso, acho	סבור אני, אני חושב, לדעתי, לי נראה	мислам дека ова е	myślę, sądzę	penso, credo, penso che, credo che
Exemplification	for example, for instance	zum Beispiel	pavyzdžiui,	de exemplu, astfel	например	por exemplo	לדוגמה, לדוגמא	Така например	na przykład	per esempio, ad esempio
Elaboration	in particular	insbesondere	būtent	în particular, mai ales	в частност	em particular	בעיקר	и по-специално	w szczególności	in particolare, particolarmente
Synchrony	so far	bisher, soweit	iki šiol	chiar	до сега	até agora	נכון, נכון להיום לעכשו		jak dotąd, dotychczas	fino ad adesso, fino ad ora, finora
Contrast	on the one hand, on the other hand	einerseits, andererseits	Iš vienos pusės ... iš kitos pusės	pe de o parte ... pe de altă parte	от една страна, от друга страна	por um lado por outro lado	מצד אחד, מצד שני	От една страна	z jednej strony, z drugiej strony	da una parte, d'altra parte
Conjunction	on the one hand, on the other hand	einerseits, andererseits	Iš vienos pusės ... iš kitos pusės	pe de o parte ... pe de altă parte	от една страна, от друга страна	por um lado, por outro lado	מצד אחד, מצד שני	От една страна	z jednej strony, z drugiej strony	da un lato, d'altra parte
Restatement	in other words, I mean	anders gesagt, mit anderen Worten	Kitaip tariant, Kitais žodžiais	cu alte cuvinte	с други думи	por outras palavras, noutras palavras, isto é	במילים אחרות, אני מתכוון	С други думи	innymi słowy	in altre parole, in altro modo
Cause	as a result	infolge, daraufhin	Dėl to, todėl	ca rezultat, prin urmare	в резултат	como resultado	כתוצאה מה	В резултат	w rezultacie	come risultato, di conseguenza
Expansion	in fact, this is, you see, that is, of course	tatsächlich, und zwar	Iš tiesų, aišku, Žinoma, tiesą sakant, Matote	de fapt, bineînțeles	всъщност, наистина, разбира се	de facto, ou seja, na verdade, claro	למעשה, רואים אתם, כמובן, ברור, בעצם	всъщност, Правилно	w gruncie rzeczy	infatti, difatti, tra l'altro
Communicative functions and qualifiers ISO 24617-2										
Confirm	of course, in fact	natürlich, selbstverständlich	Žinoma, aišku	bineînțeles, de fapt	разбира се	claro	בעצם, למעשה, כמובן, ברור	Разбира се	oczywiście	certamente, di certo, certo
Inform	you know, you see	weisst Du, siehst Du	kaip žinote, matote		така например	vejam		например, Така например	na przykład	ad esempio, come ad esempio
Certain	of course	natürlich, selbstverständlich	aišku, Žinoma, be abejo	desigur că, cum ar fi, și bineînțeles	правилно, разбира се	claro	כמובן, ברור	Разбира се	oczywiście	sicuramente, certamente

FIGURE 4 | The annotation of discourse markers.

("to think"), and "ricordare" ("to remember"). Then there are those types that emphasize the connection with the interlocutor, particularly frequent are those that revolve around the verb "sapere" ("to know"), and "vedere" ("to see").

4.11 | An Overview of the Annotation Results

Overall, the analysis of the corpus reveals that although the nine datasets are aligned with the English dataset, many discrepancies take place with respect to not only the morphosyntactic nature of the discourse markers but also the semantic and pragmatic value they convey (Table 5 and Figure 4).

The analysis that we conducted of a small sample of our corpus is proof that the ISO-based annotation scheme that we propose is comprehensive enough to cover all the semantic and pragmatic values of all the multiword discourse markers present in our datasets.

Additionally, the observation of the two tables accounts for, on the one hand, the wide range of meanings that discourse markers can display, and, on the other hand, the polyfunctionality they exhibit, thus confirming the results of prior investigations.

Further, evidence is present of common points across languages, such as omissions of discourse markers, conveyance of the meaning of the discourse marker through other grammatical means, such as modals, interrogatives, changes in the word order, phrasal variations, and single words corresponding to MWEs describing discourse markers. Nonetheless, some multiword discourse markers are stable in their interpretation across languages, providing literal translation and introducing identical discourse relation or communicative function, like in the case “of course”, “for example”, and “in particular”.

5 | Conclusions

This paper presents a linguistic and computational approach to studying multiword discourse markers across ten languages, i.e., English, Lithuanian, Bulgarian, German, Macedonian, Romanian, Hebrew, European Portuguese, Polish, and Italian. Our approach started with creating a multilingual parallel corpus extracted from the TED Talks transcripts, with English as a pivot language. The next phase was the training of two machine learning models. Concurrently, we devised an annotation scheme based on ISO 24617–8 with a plug-in to ISO 24671–2 to describe the semantic and pragmatic values of the discourse markers in the nine datasets. The proof of concept was conducted in a sample of 55 aligned examples with multiword discourse markers in the ten languages. The proposed annotation schema can be applied to other languages to comparable corpora with the aim of developing a multilingual, interoperable lexical database centered on discourse markers—answering RQ1 and RQ2.

While discourse markers exhibit cross-linguistic variations in their realization, including omissions, grammatical substitutions, and phrasal variations, the existence of stable, literally translatable multiword discourse markers like “of course,” “for example,” and “in particular” suggests a core set of shared functions. This interplay between variation and stability highlights the potential and the challenges in developing a unified, multilingual framework for discourse marker annotation–answering RQ3.

In the future, we intend to work on the (semi)automatic extraction of the different values of discourse markers and improve the OWL-ontology for representing, linking, and querying the discourse annotations we have started building (Chiarcos et al. 2023). Ultimately, our purpose is to develop and provide researchers with a multilingual language resource of multiword discourse markers annotated with an interoperable two-dimensional taxonomy published in CLARIN.

Acknowledgments

Open access publishing facilitated by Anelis Plus (the official name of Asociatia Universitatilor, a Institutelor de Cercetare - Dezvoltare si a Bibliotecilor Centrale Universitare din Romania), as part of the Wiley - Anelis Plus agreement.

Data Availability Statement

The data that support the findings of this study are available from the corresponding author upon reasonable request.

Peer Review

The peer review history for this article is available at https://publons.com/publon/10.1111/ijal.12755

Endnotes

[1] http://hdl.handle.net/20.500.11821/34

References

Aijmer, K. 2013. *Understanding Pragmatic Markers: A Variational Pragmatic Approach*. Edinburgh University Press.

Bazzanella, B., C. Bosco, A. Garcea, B. G. Fivela, J. Miecznikowski, and F. T. Brunozzi. 2007. “Italian Allora, French Alors: Functions, Convergences and Divergences.” *Catalan Journal of Linguistics* 6, no. 1: 9.

Biber, D. 1990. “Methodological Issues Regarding Corpus-Based Analyses of Linguistic Variation.” *Literary and Linguistic Computing* 5, no. 4: 257–269. https://doi.org/10.1093/llc/5.4.257.

Blakemore, D. 1987. *Semantic Constraints on Relevance*. Blackwell.

Blakemore, D. 2004. “Discourse Markers.” In *Handbook of Pragmatics*, edited by L.R. Horn and G. Ward, 121–140. Wiley.

Bojanowski, P., E. Grave, A. Joulin, and T. Mikolov. 2017. “Enriching Word Vectors With Subword Information.” *Transactions of the ACL* 5, no. 1: 135–146.

Braud, C., Y. J. Liu, E. Metheniti, et al. 2023. “The DISRPT 2023 Shared Task on Elementary Discourse Unit Segmentation, Connective Detection, and Relation Classification.” in *The 3rd Shared Task on Discourse Relation Parsing and Treebanking*, 1–21. ACL. https://doi.org/10.18653/v1/2023.disrpt-1.1.

Bunt, B., and P. Prasad. 2016. “ISO DR-Core (ISO 24617–8): Core Concepts for the Annotation of Discourse Relations.” in *Joint ACL-ISO Workshop on Interoperable Semantic Annotation*, 45–54. ACL.

Bunt, H., V. Petukhova, E. Gilmartin, et al. 2020. “The ISO Standard for Dialogue Act Annotation, Second Edition.” *Language Resources and Evaluation Conference* 549–558. https://aclanthology.org/2020.lrec-1.69/.

Chiarcos, C. 2012. “POWLA: Modeling Linguistic Corpora in OWL/DL.” in *The 9th Extended Semantic Web Conference*, 225–239, Springer. https://doi.org/10.1007/978-3-642-30284-8_22.

Chiarcos, C., and C. Fäth. 2017. “CoNLL-RDF: Linked Corpora Done in an NLP-Friendly Way.” in *The 1st International Conference on Language, Data, and Knowledge Conference*, 74–88, Springer. https://doi.org/10.1007/978-3-319-59888-8_6 .

Chiarcos, C., and M. Ionov. 2021. “Linking Discourse Marker Inventories.” *Conference on Language, Data, and Knowledge* 40: 1–15. https://doi.org/10.4230/OASIcs.LDK.2021.40.

Chiarcos, C., P. Silvano, M. Damova, et al. 2022. “An OWL Ontology for ISO-Based Discourse Marker Annotation.” in *International Scientific Interdisciplinary Conference LLOD Approaches for Language Data Research and Management (LLODREAM2022)*, 28–30, Mykolas Romeris University.

Chiarcos, C., P. Silvano, M. Damova, et al. 2023. “Building an Owl-Ontology for Representing, Linking and Querying SemAF Discourse Annotations.” *Rasprave* 49, no. 1: 117–136.

Conneau, A., K. Khandelwal, N. Goyal, et al. 2020. “Unsupervised Cross-Lingual Representation Learning at Scale.” *Annual Meeting of the Association for Computational Linguistics*, 8440–8451. ACL. https://doi.org/10.18653/v1/2020.acl-main.747.

Crible, L., and L. Degand. 2019. “Domains and Functions: A Two-Dimensional Account of Discourse Markers.” *Discours* 24, no. 1: 1–33.

Cuenca, M. J. 2013. “The Fuzzy Boundaries Between Discourse Marking and Modal Marking.” In *Discourse Markers and Modal Particles*, edited by L. Degand B. Cornillie, and P. Pietrandrea, 191–216. John Benjamins.

Damova, M., K. Mishev, G. Valūnaitė-Oleškevičienė, et al. 2023. “Validation of Language Agnostic Models for Discourse Marker Detection.” *Workshop on Discourse Studies and Linguistic Data Science (DiSLiDaS 2023) @LDK2023* 434–439. https://aclanthology.org/2023.ldk-1.46/.

Devlin, J., M. W. Chang, K. Lee, and K. Toutanova. 2019. “BERT: Pre-Training of Deep Bidirectional Transformers for Language Understanding.” in *Conference of the North American Chapter of the Association for Computational Linguistics*, 4171–4186. ACL. https://doi.org/10.18653/v1/N19-1423.

Feng, F., Y. Yang, D. Cer, N. Arivazhagan, and W. Wang. 2022. “Language-Agnostic BERT Sentence Embedding.” in *Annual Meeting of the Association for Computational Linguistics*, 878–891. ACL. https://doi.org/10.18653/v1/2022.acl-long.62.

Fischer, K. 2006. “Towards an Understanding of the Spectrum of Approaches to Discourse Particles: Introduction to the Volume.” In *Approaches to Discourse Particles*, edited by K. Fischer, 1–20. Brill.

Fraser, B. 1996. “Pragmatic Markers.” *Pragmatics* 6, no. 2: 167–190.

Fraser, B. 2009. “An Account of Discourse Markers.” *International Review of Pragmatics* 1, no. 2: 293–320. https://doi.org/10.1163/187730909X12538045489818.

Gessler, L., S. Behzad, Y. J. Liu, S. Peng, Y. Zhu, and A. Zeldes. 2021. DisCoDisCo at the DISRPT2021 Shared Task: A System for Discourse Segmentation, Classification, and Connective Detection. Shared Task on Discourse Relation Parsing and Treebanking. Association for Computational Linguistics. Accessed November, 2021. https://doi.org/10.18653/v1/2021.disrpt-1.6.

Gromann, D., E. S. Apostol, C. Chiarcos, et al. 2024. “Multilinguality and LLOD: A Survey Across Linguistic Description Levels.” *Semantic Web Journal* 15, no. 5: 1915–1958. https://doi.org/10.3233/SW-243591.

Heeman, P. A., and F. A. Allen. 1999. “Speech Repains, Intonational Phrases, and Discourse Markers: Modeling Speakers’ Utterances in Spoken Dialogue.” *Computational Linguistics* 25, no. 4: 527–572.

Heeren, J. F. 2022. *Establishing a Mechanism-Based Framework for the Corpus-Informed Analysis of Multi-Word Discourse Markers*. Springer.

Hochreiter, S., and J. Schmidhuber. 1997. “Long Short-Term Memory.” *Neural Computation* 9, no. 8: 1735–1780. https://doi.org/10.1162/neco.1997.9.8.1735.

Ji, H., and M. Huang. 2021. “DiscoDVT: Generating Long Text With Discourse-Aware Discrete Variational Transformer.” in *Conference on Empirical Methods in Natural Language Processing*, 4208–4224. ACL. https://doi.org/10.18653/v1/2021.emnlp-main.347.

Jucker, A. H., and Y. Ziv. 1998. *Discourse Markers: Descriptions and Theory, vol. 57*, 353–363, John Benjamins Publishing.

Khan, A. F., C. Chiarcos, T. Declerck, et al. 2022. “When Linguistics Meets Web Technologies. Recent Advances in Modelling Linguistic Linked Data.” *Semantic Web Journal* 13, no. 6: 987–1050. https://doi.org/10.3233/SW-222859.

Kurfali, M. 2020. “Labeling Explicit Discourse Relations Using Pre-trained Language Models.” in *International Conference on Text, Speech, and Dialogue*, 79–86, Springer Nature Switzerland AG. https://doi.org/10.1007/978-3-030-58323-1_8.

Maiya, A. S. 2022. “Ktrain: A Low-Code Library for Augmented Machine Learning.” *Journal of Machine Learning Research* 23, no. 158: 1–6.

Maschler, Y., and D. Schiffrin. 2015. “Discourse Markers Language, Meaning, and Context.” In *The Handbook of Discourse Analysis*, edited by D. Tannen H.E. Hamilton, and D. Schiffrin, 189–221. Wiley.

Montechiari, E. A., S. Stankov, K. Mishev, and M. Damova. 2022. “Machine Learning Methods for Discourse Marker Detection in Italian.” *LLOD Approaches for Language Data Research and Management* 74–79.

Ochs, E. 1996. “Linguistic Resources for Socializing Humanity.” In *Rethinking Linguistic Relativity*, edited by J. J. Gumperz and S.C. Levinson, 407–437. Cambridge University Press.

Östman, J.-O. 1981. *You Know: A Discourse-Functional Approach*. John Benjamins.

Prasad, R., N. Dinesh, A. Lee, et al. 2008. “The Penn Discourse TreeBank 2.0.” *Language Resources and Evaluation Conference* 2961–2968.

Sanders, T. J. M., W. P. M. Spooren, and L. G. M. Noordman. 1992. “Toward a Taxonomy of Coherence Relations.” *Discourse Processes* 15, no. 1: 1–35.

Schiffrin, D. 1987. Discourse Markers. Cambridge University Press.

Schourup, L. C. 2018. *Common Discourse Particles in English Conversation*. Routledge.

Sileo, D., T. Van De Cruys, C. Pradel, and P. Muller. 2019. “Mining Discourse Markers for Unsupervised Sentence Representation Learning.” in *Conference of the North American Chapter of the Association for Computational Linguistics*, 3477–3486. ACL. https://doi.org/10.18653/v1/N19-1351.

Sileo, D., T. Van De Cruys, C. Pradel, and P. Muller. 2020. “DiscSense: Automated Semantic Analysis of Discourse Markers.” in *Language Resources and Evaluation Conference*, 991–999, ELRA.

Silvano, P., and M. Damova. 2023. “ISO-DR-Core Plugs Into ISO-Dialogue Acts for a Cross-Linguistic Taxonomy of Discourse Markers.” in *Language, Data and Knowledge Conference*, 440–448, NOVA CLUNL.

Silvano, P., M. Damova, G. Valūnaitė Oleškevičienė, et al. 2022. “ISO-Based Annotated Multilingual Parallel Corpus for Discourse Markers.” in *Language Resources and Evaluation Conference*, 2739–2749, ELRA.

Šinkūnienė, J., E. Jasionytė-Mikučionienė, A. Ruskan, and A. Šolienė. 2020. “Discourse Markers in Lithuanian: Semantic Change and Functional Diversity.” *Lietuvių Kalba* 14, no. 1: 1–78.

Ştefănescu, A., S. Postolea, and V. Barbu Mititelu. 2020. “The Romanian Discourse Markers De Altfel and De Altminteri. Patterns of Use and Core Functions.” *Revue Roumaine De Linguistique* 65, no. 3: 307–322.

Wolf, T., L. Debut, V. Sanh, et al. 2020. “Transformers: State-of-the-Art Natural Language Processing.” in *Conference on Empirical Methods in Natural Language Processing*, 38–45. ACL. https://doi.org/10.18653/v1/2020.emnlp-demos.6.

Zufferey, S., and L. Degand. 2017. “Annotating the Meaning of Discourse Connectives in Multilingual Corpora.” *Corpus Linguistics and Linguistic Theory* 13, no. 2: 399–422.

Zwicky, A. M. 1985. “Clitics and Particles.” *Language* 61, no. 2: 283.