ISO 24617-8 Applied: Insights from Multilingual Discourse Relations Annotation in English, Polish, and Portuguese

Aleksandra Tomaszewska[1], Purificação Silvano[2], António Leal[2], Evelin Amorim[3]
[1]Institute of Computer Science, Polish Academy of Sciences
[2]University of Porto/ Centre for Linguistics of the University of Porto
[3]Institute for Systems and Computer Engineering, Technology and Science

aleksandra.tomaszewska@ipipan.waw.pl;
msilvano@letras.up.pt; jleal@letras.up.pt;
evelin.f.amorim@inesctec.pt

Abstract

The main objective of this study is to contribute to multilingual discourse research by employing ISO-24617 Part 8 (Semantic Relations in Discourse, Core Annotation Schema – DR-core) for annotating discourse relations. Centering around a parallel discourse relations corpus that includes English, Polish, and European Portuguese, we initiate one of the few ISO-based comparative analyses through a multilingual corpus that aligns discourse relations across these languages. In this paper, we discuss the project's contributions, including the annotated corpus, research findings, and statistics related to the use of discourse relations. The paper further discusses the challenges encountered in complying with the ISO standard, such as defining the scope of arguments and annotating specific relation types like Expansion. Our findings highlight the necessity for clearer definitions of certain discourse relations and more precise guidelines for argument spans, especially concerning the inclusion of connectives. Additionally, the study underscores the importance of ongoing collaborative efforts to broaden the inclusion of languages and more comprehensive datasets, with the objective of widening the reach of ISO-guided multilingual discourse research.

Keywords: ISO 24617-8, discourse relations, parallel corpora

1. Introduction

Discourse relations are connections linking the meaning conveyed by two or more situations in discourse, articulated either explicitly or implicitly. The ISO-24617-8 standard provides a structured approach for annotating these relations in texts across various languages and genres. It is designed for use in natural language corpora and serves as a reference model for automated techniques in basic discourse parsing, summarization, and other related applications (ISO, 2020).

Importantly, ISO-24617-8 has the potential to advance multilingual discourse studies by offering a universal analytical framework. Despite its utility, projects utilizing this standard, especially in multilingual contexts, are rare. Our research addresses this by applying ISO-24617-8 to a corpus comprising Polish, English, and European Portuguese. The aim is to examine the distribution of discourse relations in these languages, along with the challenges of applying the standard to such data.

This study was carried out within the Multilingual Discourse Annotation Initiative (MDAI), an emerging collaboration in multilingual discourse analysis between Polish and Portuguese scholars. The initiative adopts the ISO 24617-8 standard for its versatility across different languages and genres.

In this paper, we present the inaugural study conducted by our team. Our work encompasses the development of research materials, pilot annotations on select samples, and a trilingual annotation approach, offering early insights into the nature of discourse relations. Moreover, we examine the challenges we faced, especially in complying with the ISO standard, which paves the way for further refinement of the standard and its possible extension to other languages. The subsequent sections present our accomplishments, annotation methodologies, and initial findings.

Our main contributions are as follows:

- Testing ISO-24617-8 in a trilingual corpus to enhance comparative analyses and support multilingual discourse annotation.
- Providing statistics on the use of discourse relations across the three languages.
- Identifying challenges in adhering to the ISO-24617-8 standard for discourse annotation.

The paper is organized into six sections. The first section introduces the subject and outlines the research rationale. The second section reviews related work in the field, setting the context for the research. In the third section, we present ISO 24617-8, discussing its relevance and application to our study. The fourth section describes the research methodology, including the data collection process and the methods employed. The fifth section discusses the results of the study. The paper concludes with the sixth section, where we provide final remarks and propose future work in this area.

ISA-20 Proceedings @LREC-COLING-2024, pages 99–110

2. Related Work

Discourse relations are meaning relations between discourse units essential to understanding discourse structure and explaining different linguistic problems. They integrate semantic and pragmatic theories such as Theory of Discourse Coherence (Hobbs, 1985), Rhetorical Structure Theory (RST) (Mann and Thompson, 1988), Taxonomy of Coherence Relations (Sanders et al., 1992), and Segmented Discourse Representation Theory (SDRT) (Asher and Lascarides, 2003). These theories differ along several aspects, namely discourse relations':

- designations – coherence (Hobbs, 1985; Sanders et al., 1992) or rhetorical relations (Mann and Thompson, 1988; Asher and Lascarides, 2003);
- definitions – based on semantic (Hobbs, 1985; Asher and Lascarides, 2003), pragmatic criteria (Grosz and Sidner, 1986) or a combination of the two (Mann and Thompson, 1988);
- nature – descriptive and operational constructs (Asher and Lascarides, 2003) or cognitive entities (Sanders et al., 1992; Mann and Thompson, 1988);
- number – for most proposals, an open list;
- arguments – type: clauses, single sentences, nominalizations; events, states or entities; simple or composite; adjacency: adjacent (RST) or also non-adjacent (SDRT);
- relevance – nucleus/satellite (RST) or subordinating/coordinating relations (SDRT).

Deriving discourse relations has been the subject of extensive research. One of the most comprehensive and well-founded frameworks for this purpose is SDRT, which combines a detailed formalization of the elements involved in discourse interpretation with semantic and pragmatic constraints to infer discourse relations. According to SDRT, there are two types of information sources responsible for computing a given discourse relation: linguistic sources, such as the lexicon and compositional semantics, and non-linguistic sources, such as world knowledge and the cognitive state of the participants.

Discourse relations can either be implicit, not signaled linguistically, or explicit (Taboada and Das, 2013). Explicit discourse relations are identified through the presence of a linguistic marker, which could be a word (e.g., 'because' for EXPLANATION), a lexical expression (e.g., 'with the purpose of' for RESULT), tense/mood/aspect (e.g., sequence of Simple Pasts for NARRATION), or syntactic structure (e.g., relative clause for ELABORATION). These linguistic markers are known as 'discourse relational devices', 'connectives' (van Dijk, 1979), 'discourse markers' (Schiffrin, 1987), 'cue-phrases' (Asher and Lascarides, 2003), or 'relational signs' (Das and Taboada, 2019). Discourse relational devices play a significant role in triggering discourse relations and have been extensively studied (Iruskieta et al., 2014; Das and Taboada, 2019). Different taxonomies and findings have been reported in the literature to annotate datasets.

Various annotated datasets, comprising different genres and languages (individual or parallel), have been created for discourse relation identification. Some examples of these datasets are the RST-DT English corpus (Carlson et al., 2003); Penn Discourse Treebank (PDTB) (Prasad et al., 2008); RST Spanish Treebank (RST-ST) (da Cunha et al., 2011); SDRT Annodis French corpus (Afantenos et al., 2012); TED multilingual discourse bank (TED-MDB) (English, German, Polish, Portuguese, Russian, Turkish) (Zeyrek et al., 2018). Most of these datasets identify discourse relations through the presence of a discourse marker, while only a few rely on other sources of information (Benamara and Taboada, 2015).

Annotation is mainly done manually, either by trained linguists or non-experts, with a small number of instances of assisted automatic/semi-automatic annotation. (e.g., Gecco (Lapshinova-Koltunski and Anna Kunz, 2014); French Discourse Treebank (FDTB1) (Abeillé et al., 2000)).

The abundance of different frameworks makes it difficult to compare annotated corpora within the same language or across languages. Proposals such as ISO (ISO, 2016) (Bunt, 2015; Prasad and Bunt, 2015; Bunt and Prasad, 2016) aim to create interoperable, language-agnostic annotation schemes to address this issue. These annotated datasets with discourse relations are vital to Natural Language Processing (NLP) applications such as automatic summarization and translation, information retrieval, sentiment analysis, and opinion mining (Webber et al., 2012).

3. ISO 24617-8

ISO 24617 - Language Resource Management – Semantic Annotation Framework (SemAF) is made up of various components that tackle distinct facets of semantic annotation, including referential, temporal, and semantic role labeling. SemAF offers comprehensive coverage of linguistic phenomena. Part 8 – Semantic Relations in Discourse, Core Annotation Schema (DR-core) – ISO 24617-8 (ISO, 2016) deals with the annotation of locally established discourse relations.

The primary aim of ISO 24617-8 is to provide an interoperable approach to local discourse relations annotation, facilitating mapping between existing frameworks (e.g., RST (Mann and Thompson, 1988), SDRT (Asher and Lascarides, 2003),

PDTB (Prasad et al., 2008)) while adhering to the principles of the Linguistic Annotation Framework (ISO, 2012). It is also designed to be applicable to any natural language.

The "low-level" discourse relations proposed by ISO 24617-8 link two arguments, which are defined based on semantic criteria rather than syntactic. Thus, an argument of a discourse relation is any situation (state, event, fact, proposition or dialogue act), regardless of whether it is expressed syntactically by, for example, a nominalization, a clause, a sentence, or a discourse segment. Regarding the extent and adjacency of argument spans, ISO 24617-8 is neutral.

Each argument of a discourse relation is assigned an interpretation or role. Some discourse relations present pairs of arguments with the same role, so they are symmetric. Other discourse relations, called "asymmetric", assign different roles to each argument. Similarly to other frameworks, in ISO 24617-8, discourse relations are established between the two arguments regardless of the existence or absence of discourse markers.

Figures 9 and 9 in the Appendix present the set of asymmetric and symmetric discourse relations put forward by ISO 24617-8.

According to ISO 24617-8, discourse relations are not a closed set, and many questions still need further research. For example, more precise distinctions are necessary for certain discourse relations, and there is a need for coverage of language-specific features, especially typologically distinct ones.

To the best of our knowledge, ISO-24617-8 has not been widely used to annotate discourse relations. Silvano et al. (2022) and Silvano and Damova (2023) propose a taxonomy grounded on ISO 24617-8 with a plug-in to Part 2 about Dialogue acts (ISO, 2020). This taxonomy represents the semantic and pragmatic meaning of discourse markers across nine different languages in a parallel corpus. Silvano et al. (2023) present an annotated corpus (DRIPPS) with discourse relations. It contains 993 sentences with adverbial perfect participial clauses in four varieties of Portuguese (European, Brazilian, Mozambican, and Angolan) and British English. The sentences were extracted from online newspapers and annotated with discourse relations following the ISO 24671-8 framework. The authors also annotated several discourse relational devices, such as connectors, the tense of the verb of the main clause, and the aspectual types of both clauses, to determine which ones contribute to the discourse relations inference.

Another resource is the Polish Discourse Corpus (PDC). Originating from a previous project that annotated discourse connectives to study their roles in various relations (Heliasz and Ogrodniczuk, 2019), it is the first corpus designed for Polish that conforms to ISO 24617-8, and is unique in its multi-genre content. Comprising 1,745 texts from the Polish Coreference Corpus (Ogrodniczuk et al., 2015), the PDC reflects the genre distribution of the National Corpus of Polish (Przepiórkowski et al., 2012). An evaluation of the ISO 24617-8 standard's application to Polish data revealed some challenges, especially with the subjective interpretation and vague definitions of discourse relations, indicating a need for clearer guidelines (Żurowski et al., 2023). The project has achieved several milestones, including the identification of over 17,881 discourse relations. Additionally, an early version of an automatic parsing tool has been developed, adopting a sequence-tagging approach to provide an initial assessment of the complexity involved in parsing discourse relations in Polish texts (Ogrodniczuk et al., forthcoming).

4. Method and Materials

The subsequent sections detail the objectives and methodology of this study, including the development of research materials, test annotations on selected samples, and the trilingual annotation of a complete text.

4.1. Objectives

The research focuses on testing the ISO-24617-8 standard's application in the context of multilingual discourse analysis, targeting a corpus of Polish, English, and European Portuguese. The standard is recognized for its potential as a comprehensive framework for analyzing discourse relations across various languages and genres. However, its practical deployment has been limited, especially in multilingual settings. The study seeks to address this gap by examining how the standard can be applied to a diverse linguistic dataset, aiming to uncover the distribution and utilization of discourse relations within these languages.

Another objective is pinpointing the challenges encountered in adhering to the ISO-24617-8 framework for discourse annotation. Identifying them is essential for suggesting potential adjustments or enhancements to the standard, thereby improving its applicability and effectiveness in future research.

4.2. Dataset

The Multilingual Discourse Annotation Initiative (MDAI) dataset currently features 60 TED talks in English, European Portuguese, and Polish. English serves as the pivot language. The decision to use TED talks[1] was based on their accessibility, which

[1] https://www.ted.com/talks

makes publishing the annotated dataset possible. Additionally, some other TED texts were annotated with discourse relations using other frameworks (Zeyrek et al., 2018), allowing for future annotation comparisons.

Our TED talks selection process relied on three key criteria:

1. availability in all three target languages,
2. a narrative nature, and
3. a length of 600 to 800 words.

As part of our pilot study, we chose to transcribe "The History of the World According to Cats" TED talk[2].

4.3. Annotation methodology

The annotation process engaged a diverse group of participants, including two early-career researchers and two scholars with substantial academic backgrounds, all of whom shared an interest and expertise in discourse annotation. This collaborative effort laid the groundwork for examining the ISO 24617-8 standard's applicability across different languages.

The initial phase of the study involved selecting texts for annotation, designing a pilot dataset, and establishing a shared digital workspace to facilitate joint annotation and discussions. A sample text from open-source data was chosen for test annotation, aimed at aligning the annotation methods with the ISO 24617-8 standard and identifying differences in annotation strategies, particularly between the Polish and European Portuguese annotators with experience from their individual teams.

During the initial (test) annotation phase, the annotators worked on the sample text to ensure methodological consistency with the standard and to uncover any discrepancies in their approaches. Following this, a meeting was convened to discuss and resolve differences, especially concerning argument length and content, definitions and categorizations of relations, and relation hierarchies. Both teams identified these aspects as challenging.

During the pilot review, minor discrepancies were addressed, particularly in argument scope and the distinction between certain relations. After thorough discussion, consensus was reached, with Expansion maintained as a crucial component of the ISO standard, despite the omission in the Polish Discourse Corpus (PDC) annotation (Żurowski et al., 2023), for instance. Elaboration was defined to include instances where arguments pertained to the same event or situation, as stipulated by ISO, but we also incorporated insights from SDRT's interpretation of Elaboration. For this reason, cases where the second argument portrayed a subevent of an event introduced by the first argument were also annotated as Elaboration. Additionally, following Prévot et al. (2009), whenever the second argument provided more information about an entity represented in the first argument, the selected discourse relation would be Elaboration as well.

The annotation proper was conducted on an entire transcription of a TED talk titled "The History of the World According to Cats" in English, European Portuguese and Polish. We prepared a working document with the rules of the MDAI Annotation Scheme, grounded in the ISO 24617-8 standard. The process included the identification of the text span, discourse connectives, argument scopes, determination of arguments' order, identification of discourse relations, and arguments' role. The English version was annotated by three non-native annotators, fluent in English, with expertise in discourse relations and experience with ISO 24617-8. The European Portuguese and Polish translations were each annotated by two native experts.

Following the annotation proper, we have conducted the subsequent parts of the study: (i) assessing inter-annotator agreement, (ii) discussing the findings, and (iii) challenges.

5. Findings and Discussion

In this section, we describe the results of our study. We begin by presenting the results regarding the different tasks, and then we elaborate on one of the tasks, discourse relations identification, discussing some of the challenges we faced.

5.1. Results

Text spans Table 1 reveals that while there is a general consensus among annotators within each language, there are discrepancies in the number of text spans identified across languages.

	A1	A2	A3	A4
English	72	72	61	-
Polish	-	-	55	47
Portuguese	74	76	-	-

Table 1: The number of text spans identified in the three texts

Notably, the Polish subcorpus had fewer annotated text spans than the Portuguese dataset. The English and Portuguese datasets were annotated by the same individuals (A1 and A2), while A3 was among the annotators for the Polish subcorpus. It

[2]The video is available at https://www.ted.com/talks/eva_maria_geigl_the_history_of_the_world_according_to_cats/transcript.

is worth noting that all annotators had prior experience working with ISO 24617-8 in other projects. For this pilot study, they were provided with the guidelines presented by ISO 24617-8. The standard's impartiality towards text span length may account for this variance. Moreover, these results point to some indefiniteness as to what an argument should be by some annotators, who seem to have a broader notion and do not conduct a finer-grained analysis of all possible arguments and the discourse relations between them.

For the inter-annotator agreement (IAA) of the span texts, we opted to follow a pairwise BLEU-1 approach due to the difficulties associated with measuring the traditional Cohen's kappa in text span labeling (Deleger et al., 2012; Brandsen et al., 2020; Miranda, 2023). Some other scores to measure agreement are also possible. Carlson et al. (2003) mapped the hierarchical structures of the discourse into sets of units and then computed the Cohen's kappa of the categorical sets, while Zeldes (2017) employed an automatic tagger to compare with the human annotations and then obtaining the accuracy between automatic and manual labels. However, grouping in a set of units makes the agreement score not intuitive to interpret since it is necessary to detail which groups exist and their proportions. The exact accuracy of spans can also not reflect the labeling work done, because to measure the argument agreements, we allowed some minor disagreement (up to 20% in the BLEU-1 score) in the text spans. Hence, the BLEU-1 score seemed a rational choice in the context of our research.

The BLEU score is a standard metric to evaluate the results of translation task (Papineni et al., 2002). The BLEU-1 is a variation of the BLEU score that considers the tokens in the reference and the target as one gram and computes the proportion of tokens from the target that appears in the reference. This score ranges from 0 to 1, where 0 is no match between the tokens of two texts, and 1 is the full match of the tokens of the reference and target texts. To compute the BLEU-1 of the annotated text spans, we consider one annotator as the reference, i.e., the gold standard, and the other annotator as the target. Then, we calculated the BLEU-1 score for each text span. If a text span does not present a match in the reference, then we set the BLEU-1 score of that annotation as 0. Next, we average the BLUE-1 scores of all text spans. After that, we switch the reference annotator and the target and compute the BLEU-1 score again. Finally, we average these two scores. Table 2 describes the agreement between annotators in each dataset.

Overall, the results indicate that identifying text spans, which may not necessarily be limited to minimal chunks, led to a reasonable level of agreement. However, it is worth noting that there was a different interpretation of the definition of the relevant text span for breaking down arguments by A3.

	$A_{1,2}$	$A_{1,3}$	$A_{2,3}$	$A_{3,4}$
English	.63	.65	.48	-
Polish	-	-	-	.63
Portuguese	.67	-	-	-

Table 2: The IAA of the text spans as BLEU-1 score between annotators A1 and A2 ($A_{1,2}$), A1 and A3 ($A_{1,3}$), A2 and A3 ($A_{2,3}$) and A3 and A4 ($A_{3,4}$).

Example 1 illustrates some of the divergences observed in the annotations.

Example 1

He rode to Gibraltar with the rescued crew and served as a ship cat on three more vessels – one of which also sank.

The three annotators agree that the discourse relation ASYNCHRONY should link two situations. The initial situation, which has the argument role of Before, is "he rode to Gibraltar with the rescued crew", while the second situation, which has the role of After, is "(he) served as a ship cat on three more vessels". However, the annotators showed some disagreement regarding the extent of the second argument. A2 incorporated "one of which also sank" in the second argument ("and served as a ship cat on three more vessels – one of which also sank"), whereas A1 did not include this part ("and served as a ship cat on three more vessels"). Furthermore, A3 excluded the conjunction "and".

Arguments Another task we conducted during our pilot study was identifying the arguments for the selected text spans. For the IAA, we have only considered the cases where there was agreement of at least 0.8 in the BLEU-1 score on the text span. The agreement of arguments is computed in a similar way to the text spans agreement. Table 3 describes the BLEU-1 score for the arguments identified by the annotators.

The IAA for identifying arguments is higher compared to the IAA for identifying text spans. ISO 24617-8 has established clearer criteria for identifying arguments, which is not the case for identifying text spans. However, in certain cases, the absence of specific information can lead to inconsistent annotation of arguments. ISO 24617-8 defines an argument as an event, state, fact, proposition or dialogue act, but it does not address problematic cases, like the example 2.

Example 2

A população estava a aprender a dominar a natureza

The population was learning to dominate nature

	$A_{1,2}$	$A_{1,3}$	$A_{2,3}$	$A_{3,4}$
English	.83/.83	.88/.88	.76/.80	-
Polish	-	-	-	.89/.82
Portuguese	.84/.84	-	-	-

Table 3: The IAA agreement of the arguments (arg1/arg2) as BLEU-1 score between annotators A1 and A2 ($A_{1,2}$), A1 and A3 ($A_{1,3}$), A2 and A3 ($A_{2,3}$) and A3 and A4 ($A_{3,4}$).

One of the annotators identified two sentence fragments, "The population was learning" and "mastering nature", and connected them using the discourse relation SYNCHRONY. However, the other annotator believed that "to learn" and "to master" conveyed the same idea, so they only identified one sentence fragment.

Sometimes, there is disagreement because of how the connective is included in the sentence fragment. Various annotator teams have different practices; for instance, the team of Polish-language annotators treated connectives as a separate category and did not include them within sentence fragments. On the other hand, annotations from Portuguese annotators consistently show that connectives are always part of the second sentence fragment. The following examples demonstrate the discrepancy in how different annotators interpreted the same sentence fragment.

Example 3

(Arg 1) For the next several months this cat hunted rats and raised British morale (Arg 2) until a sudden torpedo strike shattered the hull and sank the ship. [Connective marked and included in Argument 2]

(Arg 1) For the next several months this cat hunted rats and raised British morale until (Arg 2) a sudden torpedo strike shattered the hull and sank the ship. [Connective marked and not included in Argument 2]

While this difference may slightly affect the argument span, it does not clearly lead to divergent interpretations of discourse relations.

ISO 24617-8 provides a flexible and neutral (core) framework, accommodating diverse interpretations of i.a., number of events in text spans. Each annotation project necessitates the development of tailored guidelines to adapt the ISO framework to its specific requirements, including addressing unique cases. Nonetheless, for enhanced interoperability, it would be better if these were addressed directly within the ISO standard, ensuring consistency and ease of application across different projects and languages.

Discourse relations Following the identification of the arguments, the annotators identified discourse relations. To compute the agreement of the discourse relations, we employed Cohen's kappa metric, which is a traditional way to analyze the inter-rater reliability of categorical data (McHugh, 2012). Since the discourse relations comprise a set of classes, i.e. categorical data, we chose this methodology. Cohen's kappa values range from -1 to +1, where -1 represents total disagreement and +1 total agreement. Furthermore, when Cohen's kappa results in values around 0, then the amount of agreement expected is no more than what could occur by random chance. The IAA regarding the identification of discourse relations is presented in Table 4.

	$A_{1,2}$	$A_{1,3}$	$A_{2,3}$	$A_{3,4}$
English	.52	.76	.56	-
Polish	-	-	-	.73
Portuguese	.52	-	-	-

Table 4: The IAA of discourse relations as Cohen Kappa score between annotators A1 and A2 ($A_{1,2}$), A1 and A3 ($A_{1,3}$), A2 and A3 ($A_{2,3}$) and A3 and A4 ($A_{3,4}$).

Concerning the English text, the measurement of Cohen's kappa relative to A1/A2 and A2/A3 is moderate, while for A1/A3 is substantial. Within the same language, we observe different results. Cohen's kappa is moderate in Portuguese annotators, while it is substantial in the case of Polish annotators.

The identification of the argument role was the subsequent task of the annotators. Tables 5 and 6 present the IAA scores for identifying Argument 1 and Argument 2 roles.

	$A_{1,2}$	$A_{1,3}$	$A_{2,3}$	$A_{3,4}$
English	.95	1.0	.94	-
Polish	-	-	-	1.0
Portuguese	.92	-	-	-

Table 5: The IAA of Arg1 as Cohen Kappa score between annotators A1 and A2 ($A_{1,2}$), A1 and A3 ($A_{1,3}$), A2 and A3 ($A_{2,3}$) and A3 and A4 ($A_{3,4}$).

The identification of the arguments' role was for the most part consistent with the IAA scores reaching perfect agreement in the three languages and between all the annotators.

	$A_{1,2}$	$A_{1,3}$	$A_{2,3}$	$A_{3,4}$
English	.91	.94	1.0	-
Polish	-	-	-	1.0
Portuguese	.92	-	-	-

Table 6: The IAA of Arg2 as Cohen Kappa score between annotators A1 and A2 ($A_{1,2}$), A1 and A3 ($A_{1,3}$), A2 and A3 ($A_{2,3}$) and A3 and A4 ($A_{3,4}$).

5.2. Discourse Relations Identification: Challenges

We can draw some conclusions by zooming in on the results of the discourse relations's annotation. The statistics in the table 7 rank the relations based on their prevalence across the three languages.

In terms of quantity, the Polish and Portuguese subcorpora showed little variation in the number of discourse relations identified (55 and 47 in Polish, and 76 and 74 in Portuguese), suggesting consistency within individual languages. In the English corpus, the counts were similar for two annotators (72 each) with similar annotation experience and lower for the third (61) from another team.

The analysis indicates that the agreement among annotators ranged from moderate to substantial, highlighting the variety in their interpretations. When reviewing the annotations across three languages, clear patterns emerged, especially in the frequency of certain discourse relations, suggesting a need for more specific discourse relations. Notably, the EXPANSION discourse relation exhibited significant variability, with counts of 22, 19, and 7 instances by different annotators within the English corpus. This variation points to different interpretations of this relation by the annotators, indicating an area for guideline improvement. In contrast, the CONCESSION and ELABORATION relations showed more consistency among annotators. For instance, in the English corpus, CONCESSION was marked 4 times by two annotators and 3 times by another, while ELABORATION was noted 5, 2, and 4 times, respectively. This suggests that the definitions for these relations might be clearer or more intuitive for the annotators. Relations such as CAUSE, ASYNCHRONY, and CONJUNCTION were annotated more frequently, possibly indicating clearer definitions or boundaries for these categories. This higher frequency could be due to the explicit nature of these relations, which often occur with connectives such as "and" in the case of CONJUNCTION or "because" in the case of CAUSE. Conversely, FUNCTIONAL DEPENDENCE, MANNER, and EXCEPTION were less commonly noted, and several discourse relations like EXEMPLIFICATION, CONDITION, NEGATIVE CONDITION, EXCLUSION, SUBSTITUTION, and FEEDBACK DEPENDENCE were not identified at all. This observation might relate to the dataset's nature or size but also may suggest a need to reassess the clarity and practicality of the definitions for these less frequently identified discourse relations. The initial analysis suggests that disagreements on annotated discourse relations often arise when an example can be interpreted according to the definitions of two distinct discourse relations. This underscores the nuanced nature of discourse relation annotation and highlights the need for more precise guidelines. Such is the case with example 4 from the Portuguese text:

Example 4

(Arg1) os gatos têm trabalhado lado a lado com os humanos há milhares de anos (Arg2) ajudando-nos, assim como nós os ajudamos

(Arg1) cats have worked side by side with humans for thousands of years (Arg2) helping us, just as we help them.

In this example, annotators agreed on the spans of both arguments and decided that Arg1 and Arg2 denoted the same situation. However, A1 identified ELABORATION, considering that Arg2 provides more detail about this situation than Arg1. A2 identified RESTATEMENT, clearly interpreting Arg2 from a different perspective.

One of the most significant differences between annotators concerns SYNCHRONY. In European Portuguese, A2 identified seven instances of SYNCHRONY, while A1 only identified three. The same annotators chose the same relations in the English subcorpus. A3 concurred with A1, identifying the feature in three instances. In the Polish subcorpus, each annotator recognized SYNCHRONY five times. The consistency observed in the Polish-language examples may stem from the explicit presence of connectives or cue phrases that indicate events occurring simultaneously, thereby easing the identification of this particular relation. The example presented in 5 illustrates this observation.

Example 5

(Arg 1) Oswojenie kota domowego miało miejsce 10 tysięcy lat temu na terenie starożytnego Bliskiego Wschodu wraz z (Arg 2) początkiem Neolitu.

(Arg 1) The domestication of the house cat took place 10 thousand years ago in the territory of the ancient Near East, together with (Arg 2) the beginning of the Neolithic period.

A different case may be observed in European Portuguese, illustrated by example 6.

Example 6

(Arg1) um gato preto e branco agarrado a uma tábua (Arg2) que flutuava

(Arg1) a black and white cat clinging to (Arg2) a floating board

Table 7: Comparative Annotation Frequencies Across Annotators for discourse relations.

	English			Portuguese		Polish	
Discourse Relation	**A1**	**A2**	**A3**	**A1**	**A2**	**A3**	**A4**
Expansion	7	22	19	23	22	3	3
Asynchrony	9	12	13	16	15	7	6
Conjunction	11	8	7	7	7	13	11
Cause	9	8	12	11	8	9	8
Elaboration	5	2	4	3	2	3	2
Concession	4	3	4	4	4	4	4
Synchrony	3	7	3	3	7	5	5
Contrast	2	4	1	2	4	2	1
Similarity	3	1	2	2	1	0	0
Restatement	2	1	3	3	1	2	1
Manner	2	0	0	0	0	1	1
Purpose	2	1	1	0	1	2	2
Exception	1	1	1	0	0	0	0
Functional Dependence	0	1	1	1	1	0	0
Disjunction	0	0	0	0	0	3	2

In this case, A2 considered that Arg2 expanded on the setting relevant for interpreting Arg1 (Expansion), while A1 annotated Synchrony. It is worth noting that temporal overlapping characterizes both Synchrony and Expansion. A similar distinction in assigning temporal and non-temporal relations can be observed for Polish. One of the annotators uses Conjunction for the discourse relation in example 7 whereas the other uses Asynchrony for a similar instance with "oraz" (and), indicating a temporal sequence rather than a simple conjunction.

Example 7

(Arg 1) Został ochrzczony Niezatapialnym Samem, popłynął na Gibraltar z ocalałymi członkami załogi oraz (Arg 2) pełnił służbę jako kot pokładowy na trzech innych okrętach

(Arg 1) He was named Unsinkable Sam, sailed to Gibraltar with the surviving crew members, and (Arg 2) served as a ship's cat on three other ships.

In another example, one of the annotators interprets the use of *czy* (whether/or/ and) in the phrase *nie były chętne do kontaktu z innymi kotami czy ludźmi* (were not keen on contact with other cats or people) as indicating a Disjunction, assigning the roles of "disjunction 1" and "disjunction 2". Conversely, another annotator views a similar usage of *czy* as an indicator for Conjunction, thus labeling it with the roles "conjunction 1" and "conjunction 2", illustrating the variability in understanding the connective's function in discourse.

The following example is evidence of the complexity of the annotation and of how disagreement can occur. In Portuguese, as in other languages, the same verb can occur as main or auxiliary without morphological differences. The example 8 illustrates this feature.

Example 8

um contratorpedeiro inglês veio recolher os prisioneiros

an English destroyer came to collect the prisoners

In this case, A2 interpreted the sequence as denoting two distinct situations represented by two main verbs, Arg1 being "an English destroyer came" and Arg2 "to collect the prisoners", linked by the discourse relation Purpose. A1 annotated this text span as representing one situation, assigning to the verb "came" an auxiliary role, and for that reason, the discourse relation Purpose was not identified. Once again, although the guidelines established for each project can specify how to proceed in ambiguous cases, we argue that such instructions could be given by the ISO to allow for a more standardized approach.

6. Conclusions and Future Work

This study applied the ISO 24617-8 standard to a parallel corpus in English, Polish, and European Portuguese, aiming to explore the potential and challenges of using this framework for multilingual discourse analysis. The primary contribution is the annotated corpus, which offers insights into the use of discourse relations and connectives across the three languages.

During the initiative, we have encountered the challenge of operating without specific ISO-based guidelines for individual languages, prompting us to discuss and converge on collective interpretations. The DR-core, while foundational, presents moments of neutrality and ambiguity that required careful consideration. The annotation process was inherently time-consuming. Additionally, the

scarcity of existing multilingual discourse annotations emphasized the innovative aspect of our work, though it also meant we had no direct benchmarks for comparison.

Our analysis revealed varying interpretations and applications of the ISO standard, highlighting the need for more explicit guidelines, especially in defining the scope of arguments and categorizing specific types of relations. Transitioning from the challenges encountered, the outcomes of the project have so far been promising. The findings offer initial insights into the use and nature of discourse relations in the three languages, along with an analysis of the challenges encountered in adhering to the standard.

Future efforts will focus on expanding the corpus to include a broader range of languages and genres, which could help in understanding the universality and flexibility of the ISO standard in diverse linguistic contexts. Refining the annotation guidelines based on the experiences and challenges encountered in this study will be a priority, with an aim to improve the clarity and applicability of the ISO framework for discourse analysis as well as inter-annotator agreement.

Acknowledgements

This article is based upon work from COST Action NexusLinguarum[3] — European network for Web-centered linguistic data science (CA 18209)[4], supported by COST (European Cooperation in Science and Technology)[5]. The work was supported by the European Regional Development Fund as a part of the 2014–2020 Smart Growth Operational Programme, CLARIN — Common Language Resources and Technology Infrastructure, project no. POIR.04.02.00–00C002/19[6], the Polish Ministry of Education and Science grant 2022/WK/09 and as part of the investment CLARIN ERIC — European Research Infrastructure Consortium: Common Language Resources and Technology Infrastructure (period: 2024–2026) funded by the Polish Ministry of Science and Higher Education (Programme: "Support for the participation of Polish scientific teams in international research infrastructure projects"), agreement number 2024/WK/01. Portuguese national funds also funded this paper through FCT – Fundação para a Ciência e a Tecnologia, I.P., within the project UIDB/00022/2020.

[3] https://nexuslinguarum.eu/
[4] https://www.cost.eu/actions/CA18209/
[5] https://www.cost.eu/
[6] https://clarin.biz/

7. Bibliographical References

Anne Abeillé, Lionel Clément, and Alexandra Kinyon. 2000. Building a treebank for French. In *Proceedings of the Second International Conference on Language Resources and Evaluation (LREC'00)*, Athens, Greece. European Language Resources Association (ELRA).

Stergos Afantenos, Nicholas Asher, Farah Benamara, Myriam Bras, Cécile Fabre, Mai Ho-dac, Anne Le Draoulec, Philippe Muller, Marie-Paule Péry-Woodley, Laurent Prévot, Josette Rebeyrolles, Ludovic Tanguy, Marianne Vergez-Couret, and Laure Vieu. 2012. An empirical resource for discovering cognitive principles of discourse organisation: the ANNODIS corpus. In *Proceedings of the Eighth International Conference on Language Resources and Evaluation (LREC'12)*, pages 2727–2734, Istanbul, Turkey. European Language Resources Association (ELRA).

Nicholas Asher and Alex Lascarides. 2003. *Logics of Conversation*. Cambridge University Press, United States.

Farah Benamara and Maite Taboada. 2015. Mapping different rhetorical relation annotations: A proposal. In *Proceedings of the Fourth Joint Conference on Lexical and Computational Semantics*, pages 147–152, Denver, Colorado. Association for Computational Linguistics.

Alex Brandsen, Suzan Verberne, Milco Wansleeben, and Karsten Lambers. 2020. Creating a dataset for named entity recognition in the archaeology domain. In *Proceedings of the Twelfth Language Resources and Evaluation Conference*, pages 4573–4577, Marseille, France. European Language Resources Association.

Harry Bunt. 2015. On the principles of semantic annotation. In *Proceedings of the 11th Joint ACL-ISO Workshop on Interoperable Semantic Annotation (ISA-11)*, London, UK. Association for Computational Linguistics.

Harry Bunt and Rashmi Prasad. 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.

Lynn Carlson, Daniel Marcu, and Mary Ellen Okurowski. 2003. *Building a Discourse-Tagged Corpus in the Framework of Rhetorical Structure Theory*, pages 85–112. Springer Netherlands, Dordrecht.

Iria da Cunha, Juan-Manuel Torres-Moreno, Gerardo Sierra, Luis-Adrián Cabrera-Diego, Brenda-Gabriela Castro-Rolón, and Juan-Miguel Rolland Bartilotti. 2011. The RST Spanish treebank on-line interface. In *Proceedings of the International Conference Recent Advances in Natural Language Processing 2011*, pages 698–703, Hissar, Bulgaria. Association for Computational Linguistics.

Debopam Das and Maite Taboada. 2019. Multiple signals of coherence relations. *Discours [En ligne]*, 24:1–38.

Louise Deleger, Qi Li, Todd Lingren, Megan Kaiser, Katalin Molnar, Laura Stoutenborough, Michal Kouril, Keith Marsolo, and Imre Solti. 2012. Building gold standard corpora for medical natural language processing tasks. *AMIA ... Annual Symposium proceedings. AMIA Symposium*, 2012:144–153.

Barbara J. Grosz and Candace L. Sidner. 1986. Attention, intentions, and the structure of discourse. *Computational Linguistics*, 12(3):175–204.

Celina Heliasz and Maciej Ogrodniczuk. 2019. Eksplicytność a implicytność w świetle analizy korpusowej (meta)tekstu. *Linguistica Copernicana*, 16:75–100.

Jerry R. Hobbs. 1985. On the coherence and structure of discourse. Technical report, CSLI-85-37, Center for the Study of Language and Information.

Mikel Iruskieta, Arantza Díaz de Ilarraza, and Mikel Lersundi. 2014. The annotation of the central unit in rhetorical structure trees: A key step in annotating rhetorical relations. In *Proceedings of COLING 2014, the 25th International Conference on Computational Linguistics: Technical Papers*, pages 466–475, Dublin, Ireland. Dublin City University and Association for Computational Linguistics.

ISO. 2012. ISO 24612. 2012. Language resource management, Linguistic annotation framework. Standard, International Organization for Standardization, Geneva, CH.

ISO. 2016. ISO 24617-8. 2016. Language resource management, part 8: Semantic relations in discourse (DR-Core). Standard, International Organization for Standardization, Geneva, CH.

ISO. 2020. ISO 24617-2. 2020. Language resource management-Semantic annotation framework (SemAF) - part 2 - Dialogue acts. Standard, International Organization for Standardization, Geneva, CH.

Ekaterina Lapshinova-Koltunski and Kerstin Anna Kunz. 2014. Annotating cohesion for multilingual analysis. In *Proceedings of the 10th Joint ISO - ACL SIGSEM Workshop on Interoperable Semantic Annotation*, pages 57–64, Reykjavik, Iceland. Association for Computational Linguistics.

William Mann and Sandra Thompson. 1988. Rethorical structure theory: Toward a functional theory of text organization. *Text*, 8:243–281.

William C. Mann and Sandra A. Thompson. 1987. Rhetorical structure theory: A theory of text organization. Technical Report ISI/RS-87-190, Marina del Rey, CA: Information Sciences Institute.

Mary L McHugh. 2012. Interrater reliability: the kappa statistic. *Biochemia medica*, 22(3):276–282.

Lester James V. Miranda. 2023. Developing a named entity recognition dataset for tagalog.

Maciej Ogrodniczuk, Katarzyna Głowińska, Mateusz Kopeć, Agata Savary, and Magdalena Zawisławska. 2015. *Coreference in Polish: Annotation, Resolution and Evaluation*. Walter De Gruyter.

Maciej Ogrodniczuk, Aleksandra Tomaszewska, Daniel Ziembicki, Sebastian Żurowski, Ryszard Tuora, and Aleksandra Zwierzchowska. forthcoming. Polish Discourse Corpus (PDC): Corpus Design, ISO-Compliant Annotation, Data Highlights, and Parser Development. In *Proceedings of The 2024 Joint International Conference on Computational Linguistics, Language Resources and Evaluation (LREC COLING 2024)*.

Kishore Papineni, Salim Roukos, Todd Ward, and Wei-Jing Zhu. 2002. Bleu: a method for automatic evaluation of machine translation. In *Proceedings of the 40th annual meeting of the Association for Computational Linguistics*, pages 311–318.

Rashmi Prasad and Harry Bunt. 2015. Semantic relations in discourse: The current state of ISO 24617-8. In *Proceedings of the 11th Joint ACL-ISO Workshop on Interoperable Semantic Annotation (ISA-11)*, London, UK. Association for Computational Linguistics.

Rashmi Prasad, Nikhil Dinesh, Alan Lee, Eleni Miltsakaki, Livio Robaldo, Aravind K. Joshi, and Bonnie L. Webber. 2008. The Penn Discourse Treebank 2.0. In *Proceedings of the International Conference on Language Resources and Evaluation, LREC 2008, 26 May - 1 June 2008, Marrakech, Morocco*. European Language Resources Association.

Laurent Prévot, Laure Vieu, and Nicholas Asher. 2009. Une formalisation plus précise pour une annotation moins confuse: la relation d'élaboration d'entité. *J. Fr. Lang. Stud.*, 19(2):207–228.

Ted Sanders, Wilbert Spooren, and Leo Noordman. 1992. Toward a taxonomy of coherence relations. *Discourse Processes*, 15(1):1–35.

Deborah Schiffrin. 1987. *Discourse markers*. 5. Cambridge University Press.

Purificação Silvano, João Cordeiro, António Leal, and Sebastião Pais. 2023. DRIPPS: a corpus with discourse relations in perfect participial sentences. In *Proceedings of the 4th Conference on Language, Data and Knowledge*, pages 470–481, Vienna, Austria. NOVA CLUNL, Portugal.

Purificação Silvano and Mariana Damova. 2023. ISO-DR-core plugs into ISO-dialogue acts for a cross-linguistic taxonomy of discourse markers. In *Proceedings of the 4th Conference on Language, Data and Knowledge*, pages 440–448, Vienna, Austria. NOVA CLUNL, Portugal.

Purificação Silvano, Mariana Damova, Giedre Valunaite Oleskeviciene, Chaya Liebeskind, Christian Chiarcos, Dimitar Trajanov, Ciprian-Octavian Truica, Elena Simona Apostol, and Anna Bączkowska. 2022. Iso-based annotated multilingual parallel corpus for discourse markers. In *Proceedings of the Thirteenth Language Resources and Evaluation Conference, LREC 2022, Marseille, France, 20-25 June 2022*, pages 2739–2749. European Language Resources Association.

Maite Taboada and Debopam Das. 2013. Annotation upon annotation: Adding signalling information to a corpus of discourse relations. *Dialogue Discourse*, 4:249–281.

Teun A. van Dijk. 1979. Pragmatic connectives. *Journal of Pragmatics*, 3(5):447–456.

Bonnie Webber, Markus Egg, and Valia Kordoni. 2012. Discourse structure and language technology. *Natural Language Engineering*, 18(4):437–490.

Amir Zeldes. 2017. The gum corpus: Creating multilayer resources in the classroom. *Language Resources and Evaluation*, 51(3):581–612.

Deniz Zeyrek, Amália Mendes, and Murathan Kurfalı. 2018. Multilingual extension of PDTB-style annotation: The case of TED multilingual discourse bank. In *Proceedings of the Eleventh International Conference on Language Resources and Evaluation (LREC 2018)*, Miyazaki, Japan. European Language Resources Association (ELRA).

Sebastian Żurowski, Daniel Ziembicki, Aleksandra Tomaszewska, Maciej Ogrodniczuk, and Agata Drozd. 2023. Adopting ISO 24617-8 for Discourse Relations Annotation in Polish: Challenges and Future Directions. In *Proceedings of the 4th Conference on Language, Data and Knowledge*, pages 482–492, Vienna, Austria. NOVA CLUNL, Portugal.

8. Language Resource References

Przepiórkowski, Adam and Bańko, Mirosław and Górski, Rafał L. and Lewandowska-Tomaszczyk, Barbara. 2012. *Narodowy Korpus Języka Polskiego [En. National Corpus of Polish]*. Wydawnictwo Naukowe PWN. PID http://nkjp.pl/.

9. Appendix

	DR-core relations	Definition	Semantic Role	
			Arg 1	Arg2
asymmetric	**Cause**	Arg2 is an explanation for Arg1.	result	reason
	Expansion	Arg2 is a situation involving some entity/entities in Arg1, expanding the narrative of which Arg1 is a part, or expanding on the setting relevant for interpreting Arg1. The Arg1 and Arg2 situations are distinct.	narrative	expander
	Asynchrony	Arg1 temporally precedes Arg2.	before	after
	Concession	An expected causal relation between Arg1 and ¬Arg2 is cancelled or denied by Arg2.	expectation raiser	expectation-denier
	Elaboration	Arg1 and Arg2 are the same situation, but Arg2 provides more detail.	broad	specific
	Exemplification	Arg1 is a set of situations; Arg2 is an element of that set.	set	instance
	Manner	Arg2 specifies how Arg1 comes about or occurs.	achievement	means
	Condition	Arg2 is an unrealized situation which, when realized, would lead to Arg1.	Consequent	Antecedent
	Negative Condition	Arg2 is an unrealized situation which, when "not" realized, would lead to Arg1.	Consequent	Negated-Antecedent
	Purpose	Arg2 is the goal or purpose of the situation described by Arg1.	Enablement	Goal
	Exception	Arg2 indicates one or more circumstances in which the situation(s) described by Arg1 does not hold.	Regular	Exclusion
	Substitution	Arg1 and Arg2 are alternatives, with Arg2 being the favored or chosen alternative.	Disfavored-alternative	Favored-alternative
	Functional dependence	Arg2 is a dialogue act with a responsive communicative function; Arg1 is the dialogue act(s) that Arg2 responds to.	Antecedent-act	Dependent-act
	Feedback dependence	Arg2 is a feedback act that provides or elicits information about the understanding or evaluation by one of the dialogue participants of Arg1.	Feedback-scope	Feedback-act

Asymmetric discourse relations (ISO, 2016; Bunt and Prasad, 2016).

	DR-core relations	Definition
symmetric	**Conjunction**	Arg1 and Arg2 bear the same relation to some situation evoked in the discourse, explicitly or implicitly. Their conjunction indicates that they both hold with respect to that situation.
	Contrast	One or more differences between Arg1 and Arg2 are highlighted with respect to what each predicates as a whole or to some entities they mention.
	Synchrony	Some degree of temporal overlap exists between Arg1 and Arg2. All forms of overlap are included.
	Similarity	One or more similarities between Arg1 and Arg2 are highlighted with respect to what each predicates as a whole or to some entities they mention.
	Disjunction	Arg1 and Arg2 bear the same relation to some other situation evoked in the discourse, explicitly or implicitly. Their disjunction indicates that they are alternatives with respect to that situation, with the disjunction being non-exclusive so that both Arg1 and Arg2 may hold.
	Restatement	Arg1 and Arg2 describe the same situation, but from different perspectives.

Symmetric discourse relations (ISO, 2016; Bunt and Prasad, 2016).