Adopting ISO 24617-8 for Discourse Relations Annotation in Polish: Challenges and Future Directions

Sebastian Żurowski
Nicolaus Copernicus University in Toruń
zurowski@umk.pl

Daniel Ziembicki
University of Warsaw
daniel.ziembicki@uw.edu.pl

Aleksandra Tomaszewska
Institute of Computer Science
Polish Academy of Sciences
a.tomaszewska@ipipan.waw.pl

Maciej Ogrodniczuk
Institute of Computer Science
Polish Academy of Sciences
m.ogrodniczuk@ipipan.waw.pl

Agata Drozd
Wrocław University of Science
and Technology
agata.drozd@pwr.edu.pl

Abstract

This paper explores a discourse relations annotation project carried out under the CLARIN-PL initiative, leveraging the ISO 24617-8 standard. The goal is to boost research interoperability and foster multilingual research. Our team of three linguist-annotators tackled the annotation of a corpus spanning several genres, including e.g., literature and press articles in the Polish language. This effort was guided by a project expert and external linguists from the CLARIN-PL language technology research infrastructure. Several significant challenges emerged during the process. Ambiguities within the ISO standard's relation categories, poorly-defined definitions for certain relation categories, and the difficulty of identifying and annotating implicit discourse relations, which lack explicit discourse connectives or signaling devices, were among the key issues. To overcome these problems, we implemented strategies such as regular team meetings, collaborative annotation forms, and preliminary revisions to the annotation scheme. This paper presents the project, the annotation process, and offers initial annotation data on the discourse relations and connectives identified within the corpus. Looking forward, we discuss potential enhancements to the process, including additional revisions to the guidelines and conclude with an overview of the project's contributions and a discussion of our future development plans.

1 Introduction

As defined in the ISO-24617-8 standard, discourse relations are the relations between situations expressed explicitly or implicitly in a discourse. They are vital for achieving a comprehensive understanding of discourse that goes beyond the meaning of individual sentences or clauses. Discourse relations occur between units known as arguments. These arguments possess distinct names corresponding to the specific relation connecting them (for instance, one argument is called BROAD and another SPECIFIC in a relation known as ELABORATION). Arguments may or may not be linked by a connective. Connectives can be single-word (for instance *and*) or multi-word (*not only... but*). In the ISO standard, discourse relations can be classified as explicit or implicit. Explicit relations are overtly signaled in discourse, for example, with connectives (such as e.g., *however* and *and*). These connectives serve as indicators of the underlying discourse relation, assisting the annotation process. Implicit discourse relations, which play a vital role in the project and underscore the significance of the human factor in our research, lack such explicit signaling devices yet maintain a connection between the arguments. Annotating implicit relations necessitates a meticulous examination and comprehension of the samples, relying on context and the annotator's knowledge of the world as well as the organization of discourse in a given language.

Discourse relations are pivotal to the evolution of natural language processing (NLP), and have been used to develop NLP tools such as summarization, sentiment analysis, and complex question answering (ISO 24617-8:2016, 2016)[1]. To sup-

[1] For a comprehensive list of other applications and the correlation between discourse relations and semantic and pragmatic relations, we recommend referring

port the development of such tools, annotated resources for discourse relations have been generated through various collaborative efforts, including international initiatives. This paper presents an ongoing annotation project conducted within the CLARIN-PL consortium[2]. In addition to a description of the project, it presents preliminary annotation statistics as well as technical challenges associated with annotating discourse relations in Polish based on practical experience of the annotators to identify possible enhancements to the process.

The project focuses on annotating discourse relations in Polish. The main objective of the annotation is to deliver the first-ever Polish discourse parser.

The project relies on a triad of components:

- the ISO 24617 guidelines (ISO 24617-5:2014, 2014; ISO 24617-8:2016, 2016) for representation of semantic relations in discourse
- knowledge gathered through the creation of the Polish subcorpus of the TED Multilingual Discourse Bank (TED-MDB) (Zeyrek et al., 2020), and
- the data and preliminary annotation of the Polish Discourse Corpus (PDC) (Heliasz, 2017)[3]; see more information in Section 3.1 below.

To systematically and accurately annotate discourse relations in Polish, the project employs Inforex, a web-based annotation platform (Marcińczuk et al., 2012, 2017; Marcińczuk and Oleksy, 2019). The system has not been prepared specifically for this work, but has been configured to meet its objectives. Annotators undertake a sequence of tasks:

1. Initial identification of discourse connectives within the samples
2. Location and labeling of relevant arguments
3. Systematic correlation of discourse connectives with their corresponding arguments
4. Naming the relations
5. Approving and marking the annotations as final

2 Annotation Schemes and Standardization Efforts

Numerous annotation frameworks (presented in Table 1) have emerged over time, each possessing unique underpinnings and methodological approaches to annotate discourse relations. Hobbs' Theory of Discourse Coherence (Hobbs, 1985) introduces a catalog of 'coherence relations' and a methodology for constructing high-level tree structures. Rhetorical Structure Theory (RST) (Mann and Thompson, 1988; Carlson et al., 2002; Taboada and Mann, 2006) views texts as hierarchical, recursive tree structures, identifying 25 distinct types of relations. The Cognitive Approach to Coherence Relations (CCR) (Sanders et al., 1992) introduces an analytical framework that segments discourse relations into four key categories. Segmented Discourse Representation Theory (SDRT) (Lascarides and Asher, 2008) connects elementary discourse units in an acyclic directed graph, accommodating nonadjacent unit linkages. Lastly, the Penn Discourse Treebank (PDTB) (Miltsakaki et al., 2004; Prasad et al., 2008) stands out for its differentiation between explicit and implicit discourse markers.

Each of the frameworks offers unique insights and methodological approaches to discourse relation annotation. The primary divergences lie in their structural foundations, e.g., tree-based versus graph-based; focal points, e.g., rhetorical intent versus explicit and implicit markers; and flexibility[4]. Given this heterogeneity of existing frameworks, the ISO 24617-8:2016 standard was introduced to address discrepancies and facilitate interoperability and, through its flexible and extensible core relations, homogenize the annotation of relations in discourse to ensure compatibility across diverse annotation frameworks (ISO 24617-8:2016, 2016). Although ISO standards are a unified endeavor for global standardization, their accessibility paradoxically falls short of being fully universal as they are not freely available. To gain access to the complete norm, it is necessary to directly purchase the standard from

to the complete ISO-24617-8 norm, available upon payment at https://www.iso.org/obp/ui/#iso:std:iso:24617:-8:ed-1:v1:en.

[2] https://clarin-pl.eu/index.php/en/

[3] http://zil.ipipan.waw.pl/PolishDiscourseCorpus

[4] For a deeper exploration of the differences and nuances among these theories and inventories, see e.g., (Benamara and Taboada, 2015; Hoek et al., 2021)

Table 1: Overview of Discourse Relation Annotation Schemes

No.	Short Name	Full Name
1	Hobbs' Theory	Hobbs' Theory of Discourse Coherence
2	RST	Rhetorical Structure Theory
3	CCR	Cognitive Approach to Coherence Relations
4	SDRT	Segmented Discourse Representation Theory
5	PDTB	Penn Discourse Treebank
6	ISO 24617-8:2016	Semantic annotation framework Part 8: Semantic relations in discourse, core annotation schema

the website. However, for a comprehensive understanding of this norm, one can also refer to open-access publications (Bunt and Palmer, 2013; Bunt and Prasad, 2016).

The ISO 24617-8:2016 standard, titled "Language resource management – Semantic annotation framework (SemAF) – Part 8: Semantic relations in discourse", presents an extensive framework for annotating discourse relations within linguistic corpora (ISO 24617-8:2016, 2016). It delineates a set of universally applicable discourse relations that span multiple languages. The annotation scheme put forth by the ISO standard encompasses various types of relations that can emerge in discourse, including cause-effect relations, (e.g., CAUSE), temporal (e.g., SYNCHRONY, ASYNCHRONY), CONTRAST, ELABORATION, EXEMPLIFICATION, and more (ISO 24617-8:2016, 2016).

3 Annotation

3.1 The Dataset: Polish Discourse Corpus

The dataset used in our experiments is Polish Discourse Corpus (PDC), created in a previous, preliminary phase of the project in which discourse connectives were annotated (Heliasz and Ogrodniczuk, 2019) to investigate how they are used in different types of relations.

The PDC consists of 1745 texts retrieved from the Polish Coreference Corpus (Ogrodniczuk et al., 2015), each comprising 250–350 words, extracted from documents randomly selected from the National Corpus of Polish (Przepiórkowski et al., 2012) and following the original distribution of text genres in this corpus. The size of the resource contains approximately 496,000 tokens.

3.2 Annotation Procedure

Discourse analysis has recently played a crucial role in the field of NLP, particularly in the context of experimental approaches to text parsing, which has experienced a rapid growth (Atwell et al., 2021). However, the annotation procedure is not always carried out in an appropriate manner. Indeed, the process of annotating discourse relations is a very complex task, requiring specialized linguistic knowledge and careful work from annotators.

For the purposes of our project, a team of specialists in linguistics with annotation experience was formed, comprising three individuals: a PhD in linguistics, a doctoral candidate in linguistics, and a person with a bachelor's degree in applied linguistics. The first annotator had also worked on previous test annotations, which allowed for a preliminary assessment of the quality of discourse relation marking (Heliasz and Ogrodniczuk, 2019). Additionally, the team included an experienced PhD in linguistics who provided assistance in resolving substantive problems that arose during the annotation process. The level of education of the team corresponded sufficiently to the specificity of the task. Team meetings were held once a week, allowing for regular discussion of annotation problems and the establishment of annotation rules that went beyond the instructions provided to the annotators. Before starting the annotation process, the team received detailed instructions on how to mark discourse relations. After completing the process, the obtained results were verified by checking the accuracy of a random 20% sample of annotations. This verification was carried out by people from outside the team (professional linguists associated with the CLARIN-PL infrastructure) and did not

Table 2: The summary of ISO 24617-8 relations.

ISO 24617-8 relation and corresponding connectives	Example with relation role names
CAUSE 3566 occurrences (bo, więc, jak... to...)	**REASON**: Las jest także olbrzymią fabryką tlenu. / *The forest is also a huge oxygen factory* **CONNECTIVE**: więc / *so* **RESULT**: zapewnia komfort oddychania / *it provides respiratory comfort.*
CONDITION 1617 occurrences (jeśli, jeżeli, gdyby)	**CONNECTIVE**: Jeśli / *If* **ANTECEDENT**: pieniądze te dostaną, / *if they get this money* **CONSEQUENT**: atmosfera w placówkach szpitalnych ulegnie poprawie. / *the atmophere in the hospital facilities will improve.*
NEGATIVE CONDITION 9 occurrences (albo... albo..., chyba że, gdyby nie)	**CONSEQUENT**: Mamy prawo odmówić dalszych napraw i zażądać zwrotu pieniędzy, / *We have the right to refuse further repairs and demand a refund* **CONNECTIVE**: chyba że / *unless* **NEGATED ANTECEDENT**: wada nie była istotna. / *the defect was not significant*
PURPOSE 1028 occurrences (żeby, aby, by)	**CONNECTIVE**: Aby / *In order to* **GOAL**: skorygować błędy w sposobie myślenia, / *correct errors in the way of thinking* **ENABLEMENT**: zacznij prowadzić wykaz codziennych zajęć. / *start keeping a record of daily activities.*
MANNER 206 occurrences (poprzez, tym samym, w taki sposób, że...)	**ACHIEVEMENT**: Szuka się więc sposobów, jak je poprawić, / *So, ways are sought to improve them* **CONNECTIVE**: między innymi poprzez / *among other things by* **MEANS**: kojarzenie leczenia chirurgicznego z pooperacyjną chemioterapią. / *associating surgical treatment with postoperative chemotherapy.*
CONCESSION 1376 occurrences (jednak, choć/chociaż, natomiast)	**EXPECTATION-RAISER**: Widzimy nieraz filmy nakręcane według wybitnego utworu, / *We often see movies based on an outstanding work* **CONNECTIVE**: a mimo to / *and yet* **EXPECTATION-DENIER**: zupełnie niepodobne, przeważnie złe. / *completely dissimilar, usually bad.*
CONTRAST 3114 occurrences (a, ale, tylko, lecz)	**ARGUMENT 1**: Nie stoją w pierwszym szeregu, / *They are not at front* **CONNECTIVE**: ale / *but* **ARGUMENT 2**: wykonują nieraz ciężkie i niewdzięczne zadania. / *they often perform hard and thankless tasks.*
EXCEPTION 68 occurrences (inaczej, w takim razie, przeciwnie)	**REGULAR**: Akcje spółki są dopuszczone do obrotu na rynku regulowanym / *The company's shares are admitted to trading on a regulated market.* **CONNECTIVE**: za wyjątkiem / *except for* **EXCLUSION**: art. 8 ust. 3. / *Article 8(3).*

Table 2: The summary of ISO 24617-8 relations (continued).

ISO 24617-8 relation and corresponding connectives	**Example with relation role names**
SIMILARITY 278 occurrences (jeszcze, również, podobnie jak)	**ARGUMENT 1**: Koty nie lubią pływać. / *Cats don't like to swim* **ARGUMENT 2**: Mają / *They* **CONNECTIVE**: też / *also* **ARGUMENT 2**: problemy ze zmianą miejsca zamieszkania. / *have problems with changing their place of residence.*[5]
SUBSTITUTION 451 occurrences (raczej/raczej niż, wobec tego, zamiast)	**FAVOURED-ALTERNATIVE**: Powinna przecież promieniować światłem trwałym, / *After all, it should radiate with permanent light* **CONNECTIVE**: zamiast / *instead of* **DISFAVOURED-ALTERNATIVE**: urządzać jednorazowe fajerwerki. / *organizing one-time fireworks.*
CONJUNCTION 17437 occurrences (i, też/także, oraz)	**ARGUMENT 1**: Czytali gazety / *They were reading newspapers* **CONNECTIVE**: i / *and* **ARGUMENT 2**: książki. / *books.*
DISJUNCTION 1665 occurrences (czy, lub, albo)	**ARGUMENT 1**: Opuszczają pokój, w którym jest telewizor / *They leave the room with the TV* **CONNECTIVE**: lub / *or* **ARGUMENT 2**: przełączają kanał. / *switch TV channels.*
EXEMPLIFICATION 609 occurrences (na przykład, jak choćby, między innymi)	**SET**: Ksiądz ma prawo również do odpoczynku / *The priest also has the right to rest* **CONNECTIVE**: i np. / *and, for instance,* **INSTANCE**: wyjechać sobie w którąś sobotę na narty. / *go skiing on some Saturday.*
ELABORATION 509 occurrences (właśnie, w szczególności, przede wszystkim)	**BROAD**: Bergson był obiektem licznych ataków, / *Bergson was the subject of numerous attacks,* **CONNECTIVE**: w szczególności / *especially* **SPECIFIC**: po ogłoszeniu Ewolucji twórczej / *after announcing Creative Evolution.*
RESTATEMENT 210 occurrences (czyli, to jest, inaczej mówiąc)	**ARGUMENT 1**: Gdy klient nie miał już pieniędzy i przypomniał sobie o polisie, dowiadywał się w siedzibie towarzystwa o tak zwanym współczynniku wartości wykupu polisy. / *When the customer had no more money and remembered the policy, he would learn at the company's headquarters about the so-called policy surrender value coefficient.* **CONNECTIVE**: Innymi słowy, / *In other words* **ARGUMENT 2**: nie dostawał tego co wpłacił. / *he did not receive what he had paid.*
SYNCHRONY 1092 occurrences (gdy, kiedy, tymczasem)	**ARGUMENT 1**: W tym czasie siedzieli w oddzielnej sali / *At this time, they were sitting in a separate room* **CONNECTIVE**: i / *and* **ARGUMENT 2**: czytali gazetę. / *reading a newspaper.*
ASYNCHRONY 2157 occurrences (aż, wreszcie, skoro)	**BEFORE**: Córki upieką ciasta. / *The daughters will bake cakes.* **CONNECTIVE**: Potem / *Then* **AFTER**: przyjdzie czas na prezenty. / *it will be time for presents.*

[4] Split argument occurs when connective is interjected in the argument content.

Table 2: The summary of ISO 24617-8 relations (continued).

ISO 24617-8 relation and corresponding connectives	Example with relation role names
EXPANSION 56 occurrences	**NARRATIVE**: Uparła się, żebym poszedł na studia... / *She insisted that I go to college* **EXPANDER**: W czasie okupacji bardzo się narażała, żeby mnie uratować... / *During the occupation, she put herself in great danger to save me...*
EVALUATION 46 occurrences	**SITUATION**: Niewolników kazał wysłać do wiejskich ergastulów, / *He ordered the slaves to be sent to rural prisons* **JUDGEMENT**: co było karą straszniejszą niemal od śmierci. / *which was almost worse than death.*
FUNCTIONAL DEPENDENCE 86 occurrences	**ANTECEDENT-ACT**: — No jak, odpowiada wam? / *So, are you satisfied?* **DEPENDENT-ACT**: — Owszem, odpowiada. / *Yes, we are.*
FEEDBACK DEPENDENCE 6 occurrences	**FEEDBACK-SCOPE**: — A nasze dzieci są inne. / *But our children are different.* **FEEDBACK-ACT**: — Tak, one są inne. / *Yes, they are different.*

involve making changes to the annotations in the application, but consisted of providing feedback to the annotators, who were able to review the indicated samples again and possibly revise their original selection.

3.3 Inforex

The annotation process, outlined in 3.2, was executed using Inforex. Inforex[6] is an online platform for constructing text corpora, developed as an integral part of the CLARIN-PL infrastructure (Marcińczuk et al., 2012, 2017; Marcińczuk and Oleksy, 2019). It allows parallel online access and resource sharing among multiple users. The system assists semantic annotation of texts on several levels, such as marking text references and marking word senses. It also allows for the flexible definition of custom sets of tags and relations to accommodate specific requirements. In our task, we defined a new set of discourse relations in Inforex according to the ISO standard. Importantly, Inforex is language-independent, making it relatively straightforward to replicate the substantive and technical principles of our annotation and create comparable resources in different languages.

Figure 1 presents a view of the annotator's work window in Inforex. The different colors indicate the arguments of the different relations (blue is PURPOSE, green is ASYNCHRONY, orange is CONJUNCION, CONTRAST or FUNCTIONAL DEPENDENCE, etc.). Numbers denote arguments of all types of all relations identified in the text numbered sequentially from the beginning of the sample. Segments highlighted in grey are connectives, which are the central elements of each relation (while it is also possible for implicit relations to exist and be labeled where the connective is not present in the text). As can be seen, Inforex allows relations to be annotated in such a way that a relation from a connective (e.g., *żeby*) is marked to the first argument of the relation (e.g., argument 11) and from the same connective to the second argument of the relation (e.g., argument 12). This is what constitutes the annotation of a single discourse relation.

3.4 Annotation Results

The annotation process offers an initial glance into the frequency of distinct discourse relations within the corpus. Initial phase statistics, as gleaned from this annotation, are detailed in Table 2. Upon initial review, certain concerns may arise due to the noticeably limited representation of certain relations. For instance, NEGATIVE CONDITION shows up in just 9 instances, while FEEDBACK DEPENDENCE is observed in a mere 6 cases. This scarcity stems from the hurdles our annotators

[6] http://inforex-work.clarin-pl.eu

faced when trying to apply the ISO standard definitions to the corpora samples. Identifying some of the relations within them proved to be particularly challenging. Given these circumstances, we consciously decided to sideline these problematic relations during the first phase of our work. As we kick off the second stage, our initial task will be to reevaluate and clarify definitions of discourse relations before making another attempt to recognize them within the texts. This focus includes EXPANSION and EVALUATION, in addition to the ones previously mentioned. As a result, not all relation types highlighted in Table 2 are paired with typical connectives. The assignment of specific connectives to their corresponding relationships is a task that will be addressed in the process of our ongoing analysis.

4 Using ISO Annotation Framework to Annotate Discourse Relations: Challenges

An important challenge that arises in implementing the ISO standard for annotating discourse relations is ambiguity of relation categories and unclear definitions for some of the relations. Firstly, the standard includes several relation categories that are ambiguous, making it difficult for annotators to determine which category to apply in a given context. This issue can lead to inconsistent (potentially erroneous) annotation, hindering the reliability and validity (and replicability) of research results. Secondly, some of the relation categories are not well-defined, resulting in confusion and inconsistency in the annotation process.

Thirdly, identifying and annotating implicit discourse relations also poses a challenge, although some of these relations have already been discussed in the literature ((Zikánová et al., 2019), (Demberg et al., 2019), (Hoek et al., 2018)), their labelling in the context of the ISO standard is still hampered by the lack of clear connectives/signaling devices. Accurately labeling implicit relations requires expertise and intuition on the part of the annotators, as they must rely on their knowledge of the language (especially discourse organization) and world events to identify and label these relations accurately. The following sections 4.1 and 4.2 present challenges related to distinguishing discourse and syntagmatic relations as well as discourse and semantic relations we have also encountered during the process.

4.1 Discourse Relations vs. Syntagmatic Relations

Although the syntagmatic structure of text segments has been studied quite extensively (Lüngen et al., 2010), the differences between discourse and syntagmatic relations may turn out to be much more blurred than anticipated. Syntagmatic relations exist between the elements of syntagmas and connect elements of different grammatical functions, such as predicates, subjects, complements, adjuncts, and attributes. However, they are limited to a single (simple or complex) sentence. In contrast, discourse relations can extend beyond a single sentence, linking different situations (expressed by different clauses / syntagmas) throughout the whole text, and thus making it coherent. These relations primarily indicate logical or temporal connections between situations. The challenge lies in distinguishing between a situation connected by a discourse relation and an adjunct linked to a predicate by a syntagmatic relation. Let's look at the following example:

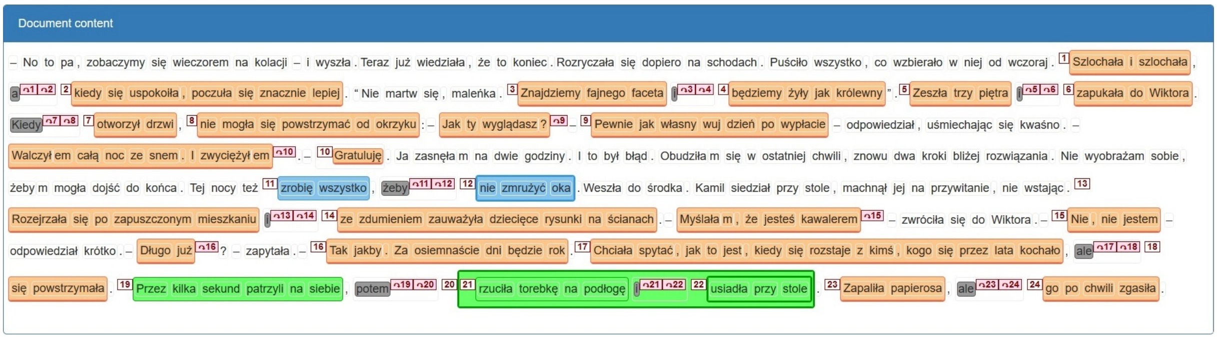
(1) PL *Jan kupił rower podczas dorocznego jarmarku.*
EN *Jan bought the bike during the annual fair.*

In cases similar to (1) annotators were not sure whether they were dealing with syntagmatic or discourse relation. This indicates that a more precise, or rather, more practical definitions of both syntagmatic and discourse relations are needed. It is possible that a lot of these relations exist alongside corresponding syntagmatic ones, but clear guidelines on how to handle them are necessary. Annotators encountered uncertainty regarding whether they should annotate discourse relations between elements such as a predicate and an adjunct within the same clause, especially when the adjunct could be interpreted as a nominalized descriptor of an independent situation.

4.2 Discourse Relations vs. Semantic Relations

Distinguishing between discourse and semantic relations can pose a challenge as the boundary between the two often appears vague and context-dependent. An example of a relation that was problematic in the annotation process is the causal relation. As we read in the ISO 24617-8 standard,

Figure 1: View of the annotator’s work window in Inforex.



– No to pa , zobaczymy się wieczorem na kolacji – i wyszła . Teraz już wiedziała , że to koniec . Rozryczała się dopiero na schodach . Puściło wszystko , co wzbierało w niej od wczoraj . 1 Szlochała i szlochała ,
a ↷1 ↷2 2 kiedy się uspokoiła , poczuła się znacznie lepiej . “ Nie martw się , maleńka . 3 Znajdziemy fajnego faceta i ↷3 ↷4 4 będziemy żyły jak królewny ” . 5 Zeszła trzy piętra i ↷5 ↷6 6 zapukała do Wiktora .
Kiedy ↷7 ↷8 7 otworzył drzwi , 8 nie mogła się powstrzymać od okrzyku : – Jak ty wyglądasz ? ↷9 – 9 Pewnie jak własny wuj dzień po wypłacie – odpowiedział , uśmiechając się kwaśno . –
Walczył em całą noc ze snem . I zwyciężył em ↷10 . – 10 Gratuluję . Ja zasnęła m na dwie godziny . I to był błąd . Obudziła m się w ostatniej chwili , znowu dwa kroki bliżej rozwiązania . Nie wyobrażam sobie ,
żeby m mogła dojść do końca . Tej nocy też 11 zrobię wszystko , żeby ↷11 ↷12 12 nie zmrużyć oka . Weszła do środka . Kamil siedział przy stole , machnął jej na przywitanie , nie wstając . 13
Rozejrzała się po zapuszczonym mieszkaniu i ↷13 ↷14 14 ze zdumieniem zauważyła dziecięce rysunki na ścianach . – Myślała m , że jesteś kawalerem ↷15 – zwróciła się do Wiktora . – 15 Nie , nie jestem –
odpowiedział krótko . – Długo już ↷16 ? – zapytała . – 16 Tak jakby . Za osiemnaście dni będzie rok . 17 Chciała spytać , jak to jest , kiedy się rozstaje z kimś , kogo się przez lata kochało , ale ↷17 ↷18 18
się powstrzymała . 19 Przez kilka sekund patrzyli na siebie , potem ↷19 ↷20 20 21 rzuciła torebkę na podłogę i ↷21 ↷22 22 usiadła przy stole . 23 Zapaliła papierosa , ale ↷23 ↷24 24 go po chwili zgasiła .

this relation is asymmetric, with the second argument (REASON) providing an explanation for the first argument (RESULT). Let’s examine the following example from ISO 24617-8:

(2) PL *Być może dlatego, że wygrali, napastnicy pana Borka są bardziej wyraziści niż jego obrońcy.*

EN *Perhaps because they won, Mr. Bork’s attackers come through more vividly than his defenders.*

Example 2 shows a CAUSE relation, but it could be argued that the expression *because* is a pragmatic comment that conveys the causal relation solely by its meaning. In other words, during annotation, the phenomenon that posed challenges to annotators is sometimes referred to in the literature as the ’semantic-pragmatic’ distinction (Van Dijk, 1979; Miltsakaki et al., 2008).

The current annotation process allows for a preliminary overview of the frequency of individual relations in the Corpus. Table 2 presents basic statistics resulting from the first phase of annotation.

4.3 Addressing Challenges

Several solutions can be implemented to navigate the challenges encountered in adhering to the ISO standard for discourse relation annotation. First and foremost, robust teamwork and open communication between annotators and supervisors are vital to reconcile discrepancies and refine the annotation process. This would entail regular meetings and discussions, where annotators can exchange insights and pinpoint potential issues within the annotation scheme. This cooperative approach is likely to enhance the overall quality of annotations while reducing potential errors.

Secondly, to curb the subjectivity that is innate in discourse annotation tasks, double annotation and adjudication could be applied in future. This would require multiple annotators working on the same sample, with a third person, possibly a supervisor (also referred to as an ’adjudicator’ or ’superannotator’), tasked with resolving any disagreements between annotators. This could serve to boost the reliability and overall quality of the annotations.

Lastly, an iterative refinement strategy can be employed to progressively enhance the annotation process. This would involve the incorporation of feedback from annotators, supervisors, and users of the annotated resources. This input, which would also encompass uncertainties and observations related to overlapping categories and challenging definitions, can then be utilized to improve the annotation guidelines, resulting in a more robust and reliable annotation scheme.

5 Towards Further Work

The annotation process has been divided into several phases, with the current phase forming a singular step within the comprehensive process. In this phase, each sample has been annotated once. Planned future phases will incorporate cross-annotation, designed to bolster data credibility and replicability. Presently, the results are under scrutiny for identification and correction of any errors or flaws.

Our annotation work has highlighted differing interpretations of relations among annotators, despite their shared expertise in the field. This variability can be partly ascribed to the broad scope of the ISO standard, which provides limited examples of sentences with distinct relations. Moreover, many phenomena observed in discourse remain relatively under-researched. Such factors can cause annotator uncertainty, potentially impacting the quality of annotation (Hovy and Lavid, 2010; Beck et al., 2020). Yet, we anticipate persistent discrepancies among annotators in such a complex task, even with more precise annotation guidelines. This may be attributed to the inherent ambiguity and multifunctionality of many discourse relations and connectives within the text - a recognized complexity in the field (Spooren and Degand, 2010). One interesting line of work would be to systematically gather the annotators' differing decisions and then classify these differences and possibly try to explain the reasons for the discrepancies.

The ongoing annotation phase has enabled us to identify and address potential challenges, preparing us for the subsequent round of annotation. This next phase will involve cross-annotation. Currently, we are analyzing the results to detect any errors and establish a suitable procedure for future annotation tasks.

6 Conclusions

This study represents a considerable advancement in Polish language processing, marking the successful completion of a comprehensive annotation of discourse relations. Through the course of our project, we highlighted prevalent linguistic relations which emerged as promising focal points for future investigations. The potential for optimizing annotation efficiency and quality through these findings underscores their significance.

Our exploration of the annotation process uncovered various complexities, largely attributed to the inherent subjectivity in text interpretation and the expansive remit of the ISO standard. This finding highlights the necessity of a skilled, diverse team of annotators, which is a critical factor in safeguarding data quality in linguistic research. During the project, we also navigated unique challenges related to ambiguity specific to the Polish language. One of the characteristics of the Polish language is the possible discontinuity of relational arguments. In Table 2 in the example illustrating the relation (SIMILARITY), it can be seen that argument 2 is discontinuous. Its two parts are separated by a conjunction *zaś*. There is a certain group of Polish expressions that syntactically behave in such a way that they do not need to be in front of an argument (e.g. *zaś*, *jeszcze*, *zatem*). These instances underscore the need for context-aware annotation strategies, hinting at the future development of innovative approaches tailored to address such language-specific issues.

The paper also highlighted the theoretical distinctions between discourse, syntagmatic, and semantic relations. This observation indicates that these aspects require further exploration, which will inform future work and advance practical applications of language annotation.

Thanks to the universal recognition and global accessibility of ISO standards, the utilization of one of them in the study as an alternative to less widespread and standardized criteria significantly enhances the reliability and replicability of our findings. The only drawback is that access to the standard is not provided free of charge. However, the availability of the ISO standard in multiple languages further contributes to its broader applicability. The use of the ISO standard establishes a solid foundation for fostering cross-linguistic cooperation and strengthens the potential for future multilingual research endeavors.

In sum, our project will unveil significant insights into Polish language processing, open up promising avenues for future exploration, and lay a solid groundwork for the continuation of work in this domain. We trust that our contributions will serve as a catalyst for further research advancements and fruitful collaborations in the years to come.

Acknowledgements

The work was financed 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 and the Polish Ministry of Education and Science grant 2022/WK/09.

References

Katherine Atwell, Junyi Jessy Li, and Malihe Alikhani. 2021. Where Are We in Discourse Relation Recognition? In *Proceedings of the 22nd Annual Meeting of the Special Interest Group on Discourse and Dialogue*, pages 314–325, Singapore and online. Association for Computational Linguistics.

Christin Beck, Hannah Booth, Mennatallah El-Assady, and Miriam Butt. 2020. Representation Problems in Linguistic Annotations: Ambiguity, Variation, Uncertainty, Error and Bias. In *Proceedings of the 14th Linguistic Annotation Workshop*, pages 60–73, Barcelona, Spain. Association for Computational Linguistics.

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.

Harry Bunt and Martha Palmer. 2013. Conceptual and Representational Choices in Defining an ISO standard for Semantic Role Annotation. In *Proceedings Ninth Joint ISO–ACL SIGSEM Workshop on Interoperable Semantic Annotation (ISA-9)*, pages 41–50, Potsdam.

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-11)*, pages 45–54, Portoroz, Slovenia.

Lynn Carlson, Daniel Marcu, and Mary Ellen Okurowski. 2002. RST Discourse Treebank. LDC catalog number: LDC2002T07.

Vera Demberg, Merel CJ Scholman, and Fatemeh Torabi Asr. 2019. How compatible are our discourse annotation frameworks? Insights from mapping RST-DT and PDTB annotations. *Dialogue & Discourse*, 10(1):87–135.

Celina Heliasz. 2017. Projekt schematu badań nad relacjami (meta)tekstowymi i narzędzia do ich analizy i opisu. Technical report, CLARIN-PL.

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 No. CSLI-85–37, Center for the Study of Language and Information, Stanford University.

Jet Hoek, Jacqueline Evers-Vermeul, and Ted J. M. Sanders. 2018. Segmenting discourse: Incorporating interpretation into segmentation? *Corpus Linguistics and Linguistic Theory*, 14(2):357–386.

Jet Hoek, Merel Scholman, and Ted J. M. Sanders. 2021. Is there less agreement when the discourse is underspecified? In *Proceedings of the Integrating Perspectives on Discourse Annotation (DiscAnn) Workshop*, University of Tübingen, Germany.

Eduard Hovy and Julia Lavid. 2010. Towards a 'science' of corpus annotation: a new methodological challenge for corpus linguistics. *International Journal of Translation*, 22(1):13–36.

ISO 24617-5:2014. 2014. Language resource management – Semantic annotation framework (SemAF) – Part 5: Discourse structure (SemAF-DS). International Organization for Standardization.

ISO 24617-8:2016. 2016. Language resource management – Semantic annotation framework (SemAF) – Part 8: Semantic relations in discourse, core annotation schema (DR-core). International Organization for Standardization.

Alex Lascarides and Nicholas Asher. 2008. *Segmented Discourse Representation Theory: Dynamic Semantics With Discourse Structure*, volume 83 of *Computing Meaning. Studies in Linguistics and Philosophy*, pages 87–124. Springer Netherlands, Dordrecht.

Harald Lüngen, Maja Bärenfänger, Mirco Hilbert, Henning Lobin, and Csilla Puskás. 2010. *Discourse Relations and Document Structure*, volume 41 of *Linguistic Modeling of Information and Markup Languages: Contributions to Language Technology*, pages 97–123. Springer Netherlands, Dordrecht.

William C. Mann and Sandra A. Thompson. 1988. Rhetorical Structure Theory: Toward a functional theory of text organization. *Text — Interdisciplinary Journal for the Study of Discourse*, 8(3):243–281.

Michał Marcińczuk and Marcin Oleksy. 2019. Inforex — a collaborative system for text corpora annotation and analysis goes open. In *Proceedings of the International Conference on Recent Advances in Natural Language Processing (RANLP 2019)*, pages 711–719, Varna, Bulgaria. INCOMA Ltd.

Michał Marcińczuk, Jan Kocoń, and Bartosz Broda. 2012. Inforex – a web-based tool for text corpus management and semantic annotation. In *Proceedings of the Eight International Conference on Language Resources and Evaluation (LREC 2012)*, pages 224–230, Istanbul, Turkey. European Language Resources Association (ELRA).

Michał Marcińczuk, Marcin Oleksy, and Jan Kocoń. 2017. Inforex — a collaborative system for text corpora annotation and analysis. In *Proceedings of the International Conference Recent Advances in Natural Language Processing (RANLP 2017)*, pages 473–482, Varna, Bulgaria. INCOMA Ltd.

Eleni Miltsakaki, Rashmi Prasad, Aravind Joshi, and Bonnie Webber. 2004. The Penn Discourse Treebank. In *Proceedings of the Fourth International Conference on Language Resources and Evaluation (LREC 2004)*, pages 2237–2240, Lisbon, Portugal. European Language Resources Association (ELRA).

Eleni Miltsakaki, Livio Robaldo, Alan Lee, and Aravind Joshi. 2008. Sense Annotation in the Penn Discourse Treebank. In *Proceedings of the 9th International Conference: Computational Linguistics and Intelligent Text Processing (CICLing 2008)*, pages 275–286. Springer.

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

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 Sixth International Conference on Language Resources and Evaluation (LREC 2008)*, pages 2961–2968, Marrakech, Morocco. European Language Resources Association (ELRA).

Adam Przepiórkowski, Mirosław Bańko, Rafał L. Górski, and Barbara Lewandowska-Tomaszczyk, editors. 2012. *Narodowy Korpus Języka Polskiego*. Wydawnictwo Naukowe PWN, Warszawa.

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

Wilbert Spooren and Liesbeth Degand. 2010. Coding coherence relations: Reliability and validity. *Corpus Linguistics and Linguistic Theory*, 6(2):241–266.

Maite Taboada and William C. Mann. 2006. Applications of Rhetorical Structure Theory. *Discourse Studies*, 8(4):567–588.

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

Deniz Zeyrek, Amália Mendes, Yulia Grishina, Murathan Kurfalı, Samuel Gibbon, and Maciej Ogrodniczuk. 2020. TED Multilingual Discourse Bank (TED-MDB): A parallel corpus annotated in the PDTB style. *Language Resources and Evaluation*, 54(2):587–613.

Šárka Zikánová, Jiří Mírovskỳ, and Pavlína Synková. 2019. Explicit and Implicit Discourse Relations in the Prague Discourse Treebank. In *Proceedings of 22nd International Conference Text, Speech, and Dialogue (TSD 2019)*, volume 11697 of *Lecture Notes in Computer Science*, pages 236–248, Cham. Springer.