Polish Discourse Corpus (PDC): Corpus Design, ISO-Compliant Annotation, Data Highlights, and Parser Development

Maciej Ogrodniczuk[1], Aleksandra Tomaszewska[1], Daniel Ziembicki[2],
Sebastian Żurowski[3], Ryszard Tuora[1] and Aleksandra Zwierzchowska[1]
[1]Institute of Computer Science, Polish Academy of Sciences;
[2]University of Warsaw;
[3]Nicolaus Copernicus University in Toruń
maciej.ogrodniczuk@ipipan.waw.pl; aleksandra.tomaszewska@ipipan.waw.pl;
daniel.ziembicki@uw.edu.pl; zurowski@umk.pl;
ryszardtuora@gmail.com; aazwierzchowska@gmail.com

Abstract

This paper presents the Polish Discourse Corpus, a pioneering resource of this kind for Polish and the first corpus in Poland to employ the ISO standard for discourse relation annotation. The Polish Discourse Corpus adopts ISO 24617-8, a segment of the Language Resource Management – Semantic Annotation Framework (SemAF), which outlines a set of core discourse relations adaptable for diverse languages and genres. The paper overviews the corpus architecture, annotation procedures, the challenges that the annotators have encountered, as well as key statistical data concerning discourse relations and connectives in the corpus. It further discusses the initial phases of the discourse parser tailored for the ISO 24617-8 framework. Evaluations on the efficacy and potential refinement areas of the corpus annotation and parsing strategies are also presented. The final part of the paper touches upon anticipated research plans to improve discourse analysis techniques in the project and to conduct discourse studies involving multiple languages.

Keywords: Polish Discourse Corpus, discourse analysis, ISO 24617-8, discourse parsing, natural language processing, corpus linguistics, discourse annotation

1. Introduction

Understanding the nuances of human communication is pivotal in discourse analysis. Discourse relations dictate how sentences and utterances connect, forming a cohesive text. Over the years, experts in Natural Language Processing (NLP) and corpus linguists have developed a variety of annotated corpora to study discourse relations across different genres and languages. However, the methodologies and standards utilized in these corpora varied, leading to inconsistencies in how annotation processes were organized and conducted. Recently, the development of a standard framework, ISO 24617-8 (ISO, 2016), has provided a foundation for annotating and analyzing these discourse relations across various genres and languages. This indicates that cross-linguistic research in discourse analysis holds the potential to gain momentum, offering fresh perspectives for this line of study.

This paper focuses on adopting ISO 24617-8 to an existing discourse-marker annotated corpus of Polish developed in 2018, offers its key statistics, introduces an initial version of the discourse parser trained on the corpus and sets the trajectory for future endeavors.

The main contributions of the paper include:

1. the Polish Discourse Corpus (PDC), encompassing over 17,881 identified discourse relations
2. a baseline automatic parsing tool using the sequence-tagging approach, to estimate the difficulty of the task
3. pilot studies and future research avenues, merging theoretical linguistics and natural language processing, with a focus on multi-genre and cross-linguistic discourse studies.

2. Key Terms and Related Work

Various frameworks have been developed for discourse annotation over the years for different languages and genres, each with more or less different categories and methods. Examples include Hobbs' Theory of Discourse Coherence (HTDC; Hobbs, 1985), Rhetorical Structure Theory (RST; Mann and Thompson, 1988; Taboada and Mann, 2006) and (Carlson et al., 2002) the Cognitive Approach to Coherence Relations (CCR; Sanders et al., 1992), Segmented Discourse Representation Theory (SDRT; Lascarides and Asher, 2007)

LREC-COLING 2024, pages 12829–12835

and the Penn Discourse Treebank (PDTB; Prasad et al., 2008).

The ISO 24617-8 standard was developed to address the inconsistencies in linguistic annotation methodologies. Its primary advantage is its adaptability across languages and its relevance and applicability in diverse language genres/contexts. It utilizes an extensible set of relation labels called *DR-Core*. The DR-core provides a broad framework for annotation while allowing for capturing language-specific nuances (for the annotations to be clear and machine-readable). While the DR-core presents a promising framework for linguistic annotation, its adoption across the NLP/linguistic community is still in its early stages. DR-core lays the groundwork for representing and annotating local, "low-level" discourse relations. In an effort to ensure consistency and to bridge the gap between differing frameworks, the standard provides a mapping across them. Adopting the terminology provided in the standard, *discourse* is understood in this paper as a sequence of clauses or sentences in text or of utterances in speech; *situation* as eventualities, facts, propositions, conditions, beliefs, or dialogue acts that can be expressed linguistically. Following ISO, *discourse relations* are seen as relations between situations in a discourse. Moreover, despite the creation of the ISO standard for discourse relations annotation, its adoption across the resources is, as far as it is known, very sparse. To our best knowledge, one of the few exceptions is the DRIPPS corpus (Silvano et al., 2023). Similarly, there are currently no discourse parsers compatible with the ISO framework. The problem may be in the accessibility of the ISO standards in general. While their aim is increasing quality and efficiency, as well as striving for global consistency, obtaining the full standard necessitates a purchase from the official website. However, a detailed understanding can be achieved through open-access academic publications (Bunt and Palmer, 2013; Bunt and Prasad, 2016).

3. Dataset

The Polish Discourse Corpus reuses the dataset from 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 1,745 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.

3.1. Annotation Procedure and Initial Results

In the domain of NLP, advancements in text parsing methods have heightened the importance of discourse analysis (Atwell et al., 2021). The task of annotating discourse relations is intricate, demanding specific linguistic expertise. Sometimes, procedures might not achieve optimal execution.

For this endeavor, a team comprising a PhD holder in linguistics, a doctoral candidate, and a student with a bachelor's degree in applied linguistics was assembled. All annotators brought a background in linguistics and prior annotation experience. Notably, the primary annotator had engaged with test annotations previously, facilitating an early evaluation of discourse relation annotations (Heliasz and Ogrodniczuk, 2019).

To address challenges during the annotation phase, the team sought guidance from another (more senior) scholar in linguistics. Regular meetings were held, allowing for continuous discussions about annotation challenges and refining the annotation guidelines beyond initial instructions. Upon completion of the annotation process, a verification step was initiated. Subsequently, an external review on a random 20% sample of the annotations has been conducted. This review stage was feedback-oriented, enabling annotators to reconsider their annotations without immediate changes being made.

The annotation procedure was implemented using the Inforex platform. Inforex[1] is a web-based tool designed for building text corpora and an important component of the CLARIN-PL infrastructure (Marcińczuk et al., 2012; Marcińczuk et al., 2017; Marcińczuk and Oleksy, 2019). Inforex supports simultaneous online access and facilitates resource collaboration among its users. It offers features for semantic annotations, including the marking of text references and word senses, and allows for defining custom tag and relation sets to meet specific needs, as in the case of ISO annotation. For the project, a discourse relation and argument set was established in Inforex, aligned with the standard. A notable feature of Inforex is its language-independent design, simplifying the process of adopting the methodology and principles from the annotation for creating equivalent resources in other languages.

The initial annotation process (discussed in more detail by Żurowski et al., 2023) reveals the distribution of discourse relations within the corpus, with preliminary data presented in Table 1. Further analysis reveals that certain relations, such as NEGATIVE CONDITION (appearing only 9 times) and FEEDBACK DEPENDENCE (noted in just 6 instances), are no-

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

ISO 24617-8 Relation	Count
Conjunction	8247
Cause	1745
Contrast	1490
Asynchrony	1041
Disjunction	810

Table 1: Most frequent relations in the corpus.

Form	Count
i (*and*)	6829
ale (but)	939
a (*while*, *whereas*)	827
bo (*because*)	610
oraz (*and*)	542

Table 2: Most frequent connectives in the corpus.

ticeably underrepresented. These disparities can be traced back to challenges faced by annotators when aligning ISO standard definitions with the text samples. The decision was made to temporarily not to include these ambiguous relations during the project's initial phase. As the next phase of the project begins, the objective is to revisit these DR-core relation definitions for clearer identification. This also includes a focus on relations like Expansion and Evaluation. Not all relation types in Table 1 have been matched with their typical connectives. The continued analysis will address the task of associating specific connectives with their respective relations.

The annotation procedure reveals several challenges involved in annotating discourse relations. Firstly, there is the inherent ambiguity, or underdefinition of the formalism and guidelines associated with it. Secondly a large portion of discourse relations is left implicit, without using lexical markers. This entails, that the annotators have to infer (sometimes using world knowledge) the relations between discourse units, which can be prone to omissions. To tackle these issues, the procedure may be refined by coordinating individual annotators, e.g., by cooperative annotation or employing a superannotator, or by utilizing an iterative approach.

The ISO document on annotating discourse relations was created to standardize this complex type of annotation process, which is a positive development in the field. However, there are several points of the standard that, as we experienced during our work with PDC, require annotator's attention. Our observations are in line with the issues that have also been highlighted in the ISO document (section 2.16). Among them are, above all, the need for more details on the extent and adjacency of argument spans; the clarification on identification criteria of some of the discourse relations, and the assessment of the standard's applicability across multiple languages. In addition, despite discussions in literature (Hoek et al., 2018), the absence of clear signaling devices in the ISO document hinders labeling. Annotators must use linguistic expertise and world knowledge to accurately label these implicit relations.

Considering the challenges mentioned above, using the ISO standard for discourse annotation demands strategic approaches. Emphasizing clear communication within the team can mitigate discrepancies and enhance quality. Double annotation, complemented by an extra verification, could offer increased accuracy and consistency. Furthermore, the adjustment of guidelines based on collective feedback could improve the effectiveness of the annotation process.

3.2. Corpus Statistics

The annotated corpus comprises 1,794 paragraphs, amounting to a total of 537,158 tokens. 52,276 discourse nodes have been annotated. Among these, 16,955 are connectives and 35,321 are relation arguments. These nodes are interconnected by 35,915 arcs. In 926 instances of relations, the connective was implicit. Discourse annotations span 35.52% of the paragraph tokens on average.

The statistics in Table 1 rank the relations based on their prevalence. Conjunction is the most dominant, followed closely by Cause. Contrast claims the third spot, while Asynchrony and Disjunction round off the top five. The five connectives with the highest frequency are given for each relation.

The statistics provided in Table 2 categorize the connectives by their frequency and detail the discourse relations in which they appear. The total count of all connectives is also given.

4. Discourse Parser

Discourse-annotated corpora are mainly constructed with the aim of building automatic tools for discourse parsing. The task is challenging because of the richness of structures used by different formalisms (e.g., graphs with non-terminal nodes in the case of SDRT). Full discourse analysis might involve segmentation into discourse units, attachment of relations between these, and classifying the relations.

The most important recent development in discourse parsing is the series of DisRPT shared tasks (Zeldes et al., 2019, 2021; Braud et al., 2023), which addresses the problem of discourse parsing for a range of languages, in a cross-framework fashion. These tasks addressed only the most el-

TRAINING TASK	EVALUATED TASK				
	TRAINED TASK	REDUCED TASK			
		Arg	Dir_Arg		Connective
			Arg1	Arg2	
EDU	52.04	52.04	–	–	–
DIR_EDU	46.98	50.46	43.55	50.03	–
CONN	80.17	–	–	–	**80.17**
DIR_EDU+CONN	59.19	55.29	47.12	52.53	78.62
FULL	54.02	54.31	46.10	51.37	78.65
DIR_EDU+CONN → FULL	55.50	**56.02**	**48.07**	**53.95**	79.07

Table 3: Parsing evaluation results on different tasks (F1 scores).

ementary aspects of discourse parsing, i.e., discourse segmentation and relation classification.

4.1. Tasks

In the preliminary solution, a range of tasks similar to DisRPT was devised:

1. CONN: Detecting connective spans in the text.
2. EDU: Detecting EDU spans in the text (elementary discourse units, alternatively referred to as situation in the ISO standard).
3. DIR_EDU: Just like in EDU, but additionally the EDUs are labeled as Arg1 or Arg2 (reflecting the directionality of relations).
4. DIR_EDU+CONN: Both EDUs (labeled with respect to direction), and connectives are detected.
5. FULL: Connectives and EDUs are detected, and the latter are labeled with respect to what role they play in relations using the full tagset.

Notably this does not include attaching EDUs to form graphs, or linking discontinuous spans of one unit together. Also, because of low number of implicit connectives, the popular task of Implicit Discourse Relation Classification (see e.g. Xiang and Wang, 2023) was skipped.

4.2. Results

To the best of our knowledge, there are no other parsers trained within the ISO standard, and so direct comparison was impossible. On the other hand, in our opinion, comparison against different schemes, e.g. in the PDTB scheme would be misleading. A baseline sequence-tagging model, initialized from herbert-large-cased (Mroczkowski et al., 2021) was trained on the corpus data converted into the tasks as specified above. The results are listed in Table 3. Each model was evaluated against the task it was trained on. Additionally, the predictions were also mapped onto simpler variations of the tasks (in gray) e.g., the distinction between Arg1 and Arg2 could be collapsed to evaluate the DIR_EDU model on the EDU task. Connective identification seems to be the easiest task of the suite, which is expected, as connectives are usually lexicalized as such. The FULL task including relation classification is theoretically the most demanding one, but classification itself is not particularly difficult. This is clear based on the fact that a ruleset defined lexically in terms of connectives yields an accuracy of 90.26% when predicting relation types, indicating that simple lexical features give virtually all the information needed.

It is notable that, in many cases, models trained on more demanding tasks scored higher on the simpler versions than the dedicated models. This indicates that additional information coming from richer supervision has a top-down facilitating effect on the simpler tasks. This is further corroborated by considering a curriculum learning approach, in which a model is trained on the DIR_EDU+CONN and FULL tasks in succession. This improves the F1 score on the more demanding task by 1.48 p.p.

5. Conclusions and Future Work

The Polish Discourse Corpus has been successfully developed, marking it as the inaugural corpus specifically tailored for the Polish language and the first multi-genre resource annotated in alignment with ISO 24617-8. Concurrently, efforts are underway to refine the parser. Key achievements to date include the finished first annotation iteration and the establishment of the corpus, capturing over 17,881 distinct discourse relations as well as the development of an initial automatic parsing tool utilizing the sequence-tagging approach, providing preliminary assessments of the task's complexity. Looking forward, there are plans to further this project with the intent to enhance the resource and continue the exploration in this domain. As the research contin-

ues, the following strategic directions are laid out, informed by insights from preliminary efforts.

5.1. Reannotation

By leaving room for interpretation, the standard now encourages adaptability but also necessitates that basic rules be negotiated among annotators before or during the annotation process to minimize discrepancies in identifying argument spans and other key elements, ensuring higher inter-annotator agreement and consistency in the application of the standard across various texts and languages. This will be addressed in subsequent iterations of our work. Drawing lessons from ambiguities encountered in relation to the ISO standard's categories, we will seek to increase the precision of annotations and enhance the consistency and accuracy of the annotated data. To assess and monitor the quality of annotations, we plan to implement appropriate Inter-Annotator Agreement (IAA) measures. In our approach, we will consider not only the classical Cohen's Kappa measure but also other measures that are relevant to the specificity of our task, such as BLEU. By incorporating multiple measures we will seek to ensure a thorough assessment of the annotated data.

5.2. Development of a Multilingual Ontology

One of the main goals of the future work is the formulation of an universal (multilingual) ontology based on the ISO 24617-8 standard. This ontology will synthesize information on discourse markers, relation arguments, and relation types across languages. With collaboration from linguists versed in twelve European languages, the ambition is to create an ontology bridging the gap between disparate discourse representation theories, making it a practical reference for researchers and increasing chances for international projects, research replicability, and data comparability.

While the research initially concentrates on the Polish language, the ontology and datasets developed will soon be used in fostering multilingual initiatives, thereby broadening the horizons of discourse relations research.

5.3. Prototyping a Multilingual Discourse Parser

Given that ISO 24617-8 is an extensible language-agnostic formalism, solutions developed for automatic parsing can be generalized to other environments. The present version of the parser tackles only the basic subtasks involved in discourse parsing, but the preliminary results show that combining training on different tasks (in the form of multi-task learning or curriculum learning) can yield improvements. For this reason, extension to the task of attachment, and handling discontinuous entities appears promising. Moreover, as shown in (Shi and Demberg, 2019), further improvements can be expected by incorporating pre-training objectives, which are better aligned with high-level relations in text.

Acknowledgements

This article is based upon work from COST Action NexusLinguarum[2] — European network for Web-centered linguistic data science (CA 18209)[3], supported by COST (European Cooperation in Science and Technology)[4].

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[5], 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.

6. Bibliographical 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.

Chloé Braud, Yang Janet Liu, Eleni Metheniti, Philippe Muller, Laura Rivière, Attapol Rutherford, and Amir Zeldes. 2023. The DISRPT 2023 Shared Task on Elementary Discourse Unit Segmentation, Connective Detection, and Relation Classification. In *Proceedings of the 3rd Shared Task on Discourse Relation Parsing and Treebanking (DISRPT 2023)*, pages 1–21, Toronto, Canada. Association for Computational Linguistics.

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

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.

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.

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

Alex Lascarides and Nicholas Asher. 2007. Segmented Discourse Representation Theory: Dynamic Semantics with Discourse Structure. In H. Bunt and R. Muskens, editors, *Computing Meaning: Volume 3*, pages 87–124. Kluwer Academic Publishers.

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

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, 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. INCOMA Ltd.

Robert Mroczkowski, Piotr Rybak, Alina Wróblewska, and Ireneusz Gawlik. 2021. HerBERT: Efficiently Pretrained Transformer-based Language Model for Polish. In *Proceedings of the 8th Workshop on Balto-Slavic Natural Language Processing*, pages 1–10, Kiyv, Ukraine. Association for Computational Linguistics.

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

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

Wei Shi and Vera Demberg. 2019. Next sentence prediction helps implicit discourse relation classification within and across domains. In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP)*, pages 5790–5796, Hong Kong, China. Association for Computational Linguistics.

Purificação Silvano, João Cordeiro, António Leal, and Sebastião Pais. 2023. DRIPPS: a Corpus with Discourse Relations in Perfect Participial Sentences. In *Language, Data and Knowledge 2023 (LDK 2023): Proceedings of the 4th Conference on Language, Data and Knowledge*, pages 470–480.

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

Wei Xiang and Bang Wang. 2023. A Survey of Implicit Discourse Relation Recognition. *ACM Computing Surveys*, 55:1–34.

Amir Zeldes, Debopam Das, Erick Galani Maziero, Juliano Antonio, and Mikel Iruskieta. 2019. The DISRPT 2019 shared task on elementary discourse unit segmentation and connective detection. In *Proceedings of the Workshop on Discourse Relation Parsing and Treebanking 2019*, pages 97–104, Minneapolis, MN. Association for Computational Linguistics.

Amir Zeldes, Yang Janet Liu, Mikel Iruskieta, Philippe Muller, Chloé Braud, and Sonia Badene. 2021. The DISRPT 2021 shared task on elementary discourse unit segmentation, connective detection, and relation classification. In *Proceedings of the 2nd Shared Task on Discourse Relation Parsing and Treebanking (DISRPT 2021)*, pages 1–12, Punta Cana, Dominican Republic. Association for Computational Linguistics.

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.

7. Language Resource References

Carlson, Lynn and Marcu, Daniel and Okurowski, Mary Ellen. 2002. *RST Discourse Treebank*. Linguistic Data Consortium, LDC2002T07.

Prasad, Rashmi and Dinesh, Nikhil and Lee, Alan and Miltsakaki, Eleni and Robaldo, Livio and Joshi, Aravind K. and Webber, Bonnie L. 2008. *The Penn Discourse TreeBank 2.0*. European Language Resources Association.

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.