POLISH DISCOURSE CORPUS

Corpus Design, ISO-Compliant Annotation, Data Highlights, and Parser Development

LINGUISTIC ENGINEERING GROUP IPI PAN

Maciej Ogrodniczuk, Aleksandra Tomaszewska, Daniel Ziembicki,
Sebastian Żurowski, Ryszard Tuora, Aleksandra Zwierzchowska
Institute of Computer Science, Polish Academy of Sciences; University of Warsaw

LREC-COLING 2024

BACKGROUND

- Diverse corpora annotated with discourse relations have been created by computational and corpus linguists using different schemes.
- Problems: Inconsistent annotation processes, limited comparability and replicability of research.
- ISO 24617-8 provides a core annotation schema for annotating discourse relations.

Discourse relations (DRels), either explicit or implicit, link *situations* in a *discourse*.
ISO 24617-8 provides a standard framework for annotating DRels across languages and genres.
We created an ISO 24617-8-based Polish Discourse Corpus of 1,745 texts, manually annotated with over 17,800 DRels.

CONTRIBUTIONS

- The first ISO 24617-8 compliant DRel corpus for Polish, containing 17,881 identified discourse relations.
- A baseline automatic parsing tool using the sequence-tagging approach to estimate the difficulty of the task.
- A first version of a parser capable of identifying and labelling discourse units, tailored to our corpus.

LITERATURE REVIEW

- ISO 24617-8:2016 standard
- DRIPPS: Discourse Relations in Perfect Participial Sentences (Silvano et al., 2023)
- Hobbs' Theory of Discourse Coherence (Hobbs, 1985)
- Rhetorical Structure Theory (Mann and Thompson, 1988; Taboada and Mann, 2006; Carlson et al., 2002)
- Segmented Discourse Representation Theory (Lascarides and Asher, 2007)
- Penn Discourse Treebank (Prasad et al., 2008)

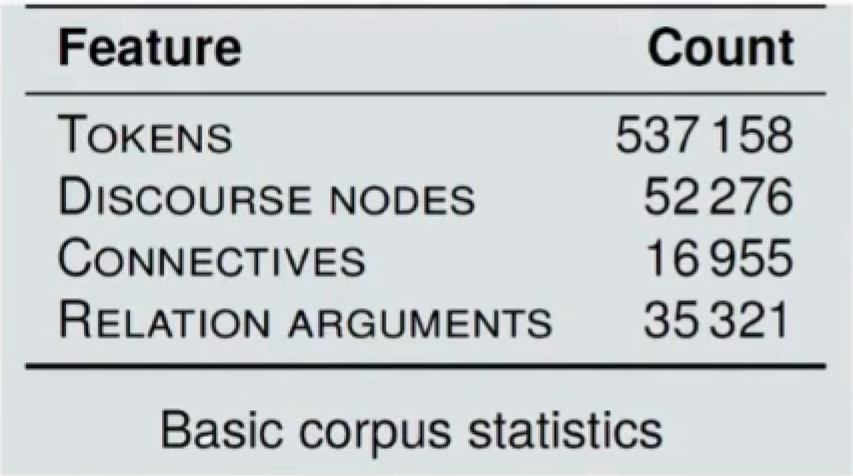
ANNOTATION PROCEDURE

- Annotation team: 3 linguists experienced in annotating discourse relations and one task supervisor.
- Regular meetings to refine guidelines and address annotation challenges.
- Verification and external review of 20% of annotations, providing feedback to annotators.
- Annotation platform: Inforex, a web-based tool for creating and annotating text corpora, manually adapted to the ISO standard.

ANNOTATION CHALLENGES

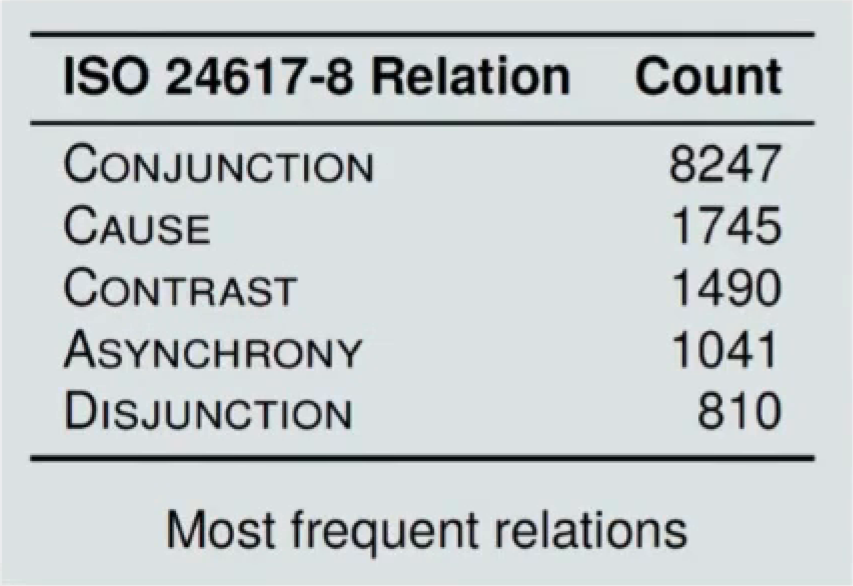
- Certain discourse relations (e.g: *Negative Condition* or *Feedback Dependence*) are underrepresented.
- Difficulty in distinguishing between *Expansion* and *Evaluation* in text samples.
- Problems with identifying implicit relations based on intuition, leading to omissions.

DATASET STATISTICS



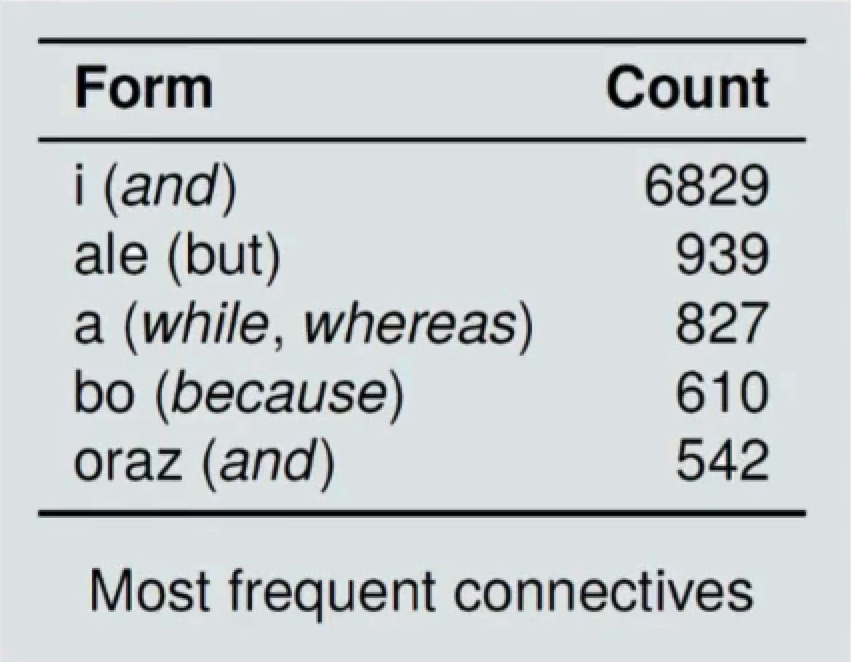
Feature	Count
Tokens	537 158
Discourse nodes	52 276
Connectives	16 955
Relation arguments	35 321

Basic corpus statistics



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

Most frequent relations



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

Most frequent connectives

DISCOURSE PARSER

- Discourse structures are very rich, and in most cases only limited aspects of them are handled by parsing. We use a sequence-tagging architecture to identify and label discourse units.
- A number of simplified tasks are considered, and in most cases training on the more robust task yields better results on the limited task.
- Identifying the precise boundaries of EDUs is a challenge.

	EVALUATED TASK				
	TRAINED TASK	REDUCED TASK			
TRAINING 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

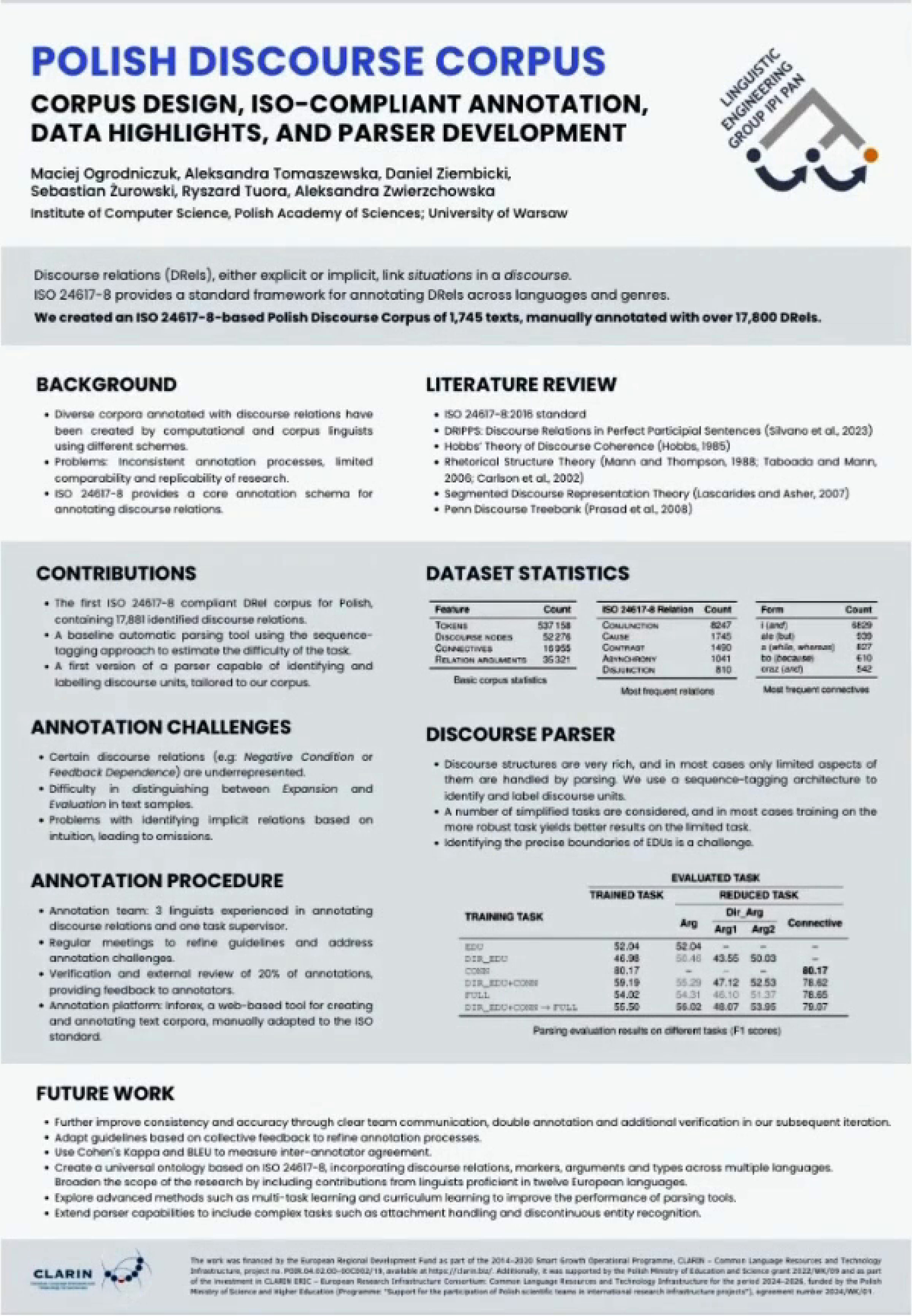
Table 3: Parsing evaluation results on different tasks (F1 scores).

FUTURE WORK

- Further improve consistency and accuracy through clear team communication, double annotation and additional verification in our subsequent iteration.
- Adapt guidelines based on collective feedback to refine annotation processes.
- Use Cohen's Kappa and BLEU to measure inter-annotator agreement.
- Create a universal ontology based on ISO 24617-8, incorporating discourse relations, markers, arguments and types across multiple languages. Broaden the scope of the research by including contributions from linguists proficient in twelve European languages.
- Explore advanced methods such as multi-task learning and curriculum learning to improve the performance of parsing tools.
- Extend parser capabilities to include complex tasks such as attachment handling and discontinuous entity recognition.

THANK YOU

The work was financed by the European Regional Development Fund as part of the 2014–2020 Smart Growth Operational Programme, CLARIN – Common Language Resources and Technology Infrastructure, project no. POIR.04.02.00–00C002/19, available at https://clarin.biz/. Additionally, it was supported by the Polish Ministry of Education and Science grant 2022/WK/09 and as part of the investment in CLARIN ERIC – European Research Infrastructure Consortium: Common Language Resources and Technology Infrastructure for the 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.



POLISH DISCOURSE CORPUS

CORPUS DESIGN, ISO-COMPLIANT ANNOTATION, DATA HIGHLIGHTS, AND PARSER DEVELOPMENT

Maciej Ogrodniczuk, Aleksandra Tomaszewska, Daniel Ziembicki,
Sebastian Żurowski, Ryszard Tuora, Aleksandra Zwierzchowska
Institute of Computer Science, Polish Academy of Sciences; University of Warsaw

Discourse relations (DRels), either explicit or implicit, link *situations* in a *discourse*.
ISO 24617-8 provides a standard framework for annotating DRels across languages and genres.
We created an ISO 24617-8-based Polish Discourse Corpus of 1,745 texts, manually annotated with over 17,800 DRels.

BACKGROUND

- Diverse corpora annotated with discourse relations have been created by computational and corpus linguists using different schemes.
- Problems: Inconsistent annotation processes, limited comparability and replicability of research.
- ISO 24617-8 provides a core annotation schema for annotating discourse relations.

LITERATURE REVIEW

- ISO 24617-8:2016 standard
- DRIPPS: Discourse Relations in Perfect Participial Sentences (Silvano et al., 2023)
- Hobbs' Theory of Discourse Coherence (Hobbs, 1985)
- Rhetorical Structure Theory (Mann and Thompson, 1988; Taboada and Mann, 2006; Carlson et al., 2002)
- Segmented Discourse Representation Theory (Lascarides and Asher, 2007)
- Penn Discourse Treebank (Prasad et al., 2008)

CONTRIBUTIONS

- The first ISO 24617-8 compliant DRel corpus for Polish, containing 17,881 identified discourse relations.
- A baseline automatic parsing tool using the sequence-tagging approach to estimate the difficulty of the task.
- A first version of a parser capable of identifying and labelling discourse units, tailored to our corpus.

ANNOTATION CHALLENGES

- Certain discourse relations (e.g. *Negative Condition* or *Feedback Dependence*) are underrepresented.
- Difficulty in distinguishing between *Expansion* and *Evaluation* in text samples.
- Problems with identifying implicit relations based on intuition, leading to omissions.

ANNOTATION PROCEDURE

- Annotation team: 3 linguists experienced in annotating discourse relations and one task supervisor.
- Regular meetings to refine guidelines and address annotation challenges.
- Verification and external review of 20% of annotations, providing feedback to annotators.
- Annotation platform: Inforex, a web-based tool for creating and annotating text corpora, manually adapted to the ISO standard.

DATASET STATISTICS

Feature	Count
Tokens	537 158
Discourse nodes	52 276
Connectives	16 955
Relation arguments	35 321

Basic corpus statistics

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

Most frequent relations

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

Most frequent connectives

DISCOURSE PARSER

- Discourse structures are very rich, and in most cases only limited aspects of them are handled by parsing. We use a sequence-tagging architecture to identify and label discourse units.
- A number of simplified tasks are considered, and in most cases training on the more robust task yields better results on the limited task.
- Identifying the precise boundaries of EDUs is a challenge.

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

Parsing evaluation results on different tasks (F1 scores)

FUTURE WORK

- Further improve consistency and accuracy through clear team communication, double annotation and additional verification in our subsequent iteration.
- Adapt guidelines based on collective feedback to refine annotation processes.
- Use Cohen's Kappa and BLEU to measure inter-annotator agreement.
- Create a universal ontology based on ISO 24617-8, incorporating discourse relations, markers, arguments and types across multiple languages. Broaden the scope of the research by including contributions from linguists proficient in twelve European languages.
- Explore advanced methods such as multi-task learning and curriculum learning to improve the performance of parsing tools.
- Extend parser capabilities to include complex tasks such as attachment handling and discontinuous entity recognition.

The work was financed by the European Regional Development Fund as part of the 2014–2020 Smart Growth Operational Programme, CLARIN – Common Language Resources and Technology Infrastructure, project no. POIR.04.02.00-00C002/19, available at https://clarin.biz/. Additionally, it was supported by the Polish Ministry of Education and Science grant 2022/WK/09 and as part of the investment in CLARIN ERIC – European Research Infrastructure Consortium: Common Language Resources and Technology Infrastructure for the 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.