POLAR: A Corpus of Questions, Responses and Argumentation in Polish Political Radio Discourse

Daniel Ziembicki[1], Aleksandra Zwierzchowska[2], Ewelina Sobol[3], Katarzyna Przerada[3]
[1]University of Warsaw;
[2]Institute of Computer Science, Polish Academy of Sciences;
[3]No affiliation
daniel.ziembicki@uw.edu.pl; aazwierzchowska@gmail.com;
ewelinasobol23@wp.pl; kasia.przerada12@gmail.com

Abstract

In this paper, we present **POLAR**: an experimental dataset designed to investigate question–answer structures in political interviews. The study also aims to integrate this level of annotation with the identification of argumentative structures. The dataset comprises orthographic transcriptions of Polish political radio interviews conducted between December 2023 and March 2024, with a total duration of nearly 10 hours of recordings (119,340 tokens). Manual annotation was performed on three levels: **(a)** identification of questions as speech acts, **(b)** classification of responses to questions, and **(c)** argumentative structures in which interrogative sentences function as premises or conclusions. The results show that not all interrogative sentences function as questions in the sense of requesting information — 23% do not serve this function, while 13% were identified as components of argumentative structures. We also introduce a gold-standard corpus, together with baseline experiments and LLM-based evaluations, demonstrating the usefulness of the resource for both theoretical research and NLP applications.

Keywords: Pragmatics, Question–Answer Pairs, Argument Mining

1. Introduction

The presented study lies at the intersection of discourse analysis and Argument Mining (AM) (Lawrence and Reed, 2020; Li et al., 2025), with a particular focus on question–answer interactions in political discourse. Previous research has shown that questions can play an important role in shaping argumentative exchange and dialogical dynamics (Hautli-Janisz et al., 2022; Hautli-Janisz et al., 2022; Ziembicki, 2025). Building on these insights, our study concentrates on questions and answers as fundamental components of dialogic text, whose formulation is of particular importance in the public sphere, especially in political discourse.

Problem: Analyzing questions and answers in political discourse poses a number of challenges (Clayman and Heritage, 2002). Questions in such interactions rarely serve a purely informational function. They often play rhetorical roles — for example, to accuse, express opinions, or challenge the position — and, at another level, may also function as premises or conclusions within argumentative structures. The answers in political discourse are as diverse as the questions themselves — ranging from direct to evasive or strategically ambiguous (Walton, 1989). Only a small proportion of arguments are explicitly marked by linguistic cues such as discourse connectives (Lawrence and Reed, 2015), while the vast majority depend on pragmatic interpretation. This complexity becomes especially evident in political interviews, where journalists frequently pose confrontational questions and politicians often deviate from classical conversational principles, e.g., Gricean maxims (Grice, 1975), further complicating both theoretical analysis and automatic language processing in NLP.

Gap: Previous research in Argument Mining has largely focused on English and on monological genres of discourse. Despite numerous studies in recent years (Hautli-Janisz et al., 2022; Hautli-Janisz et al., 2022; Kikteva et al., 2022; D'Agostino et al., 2024; Ziembicki, 2025) little is still known about how argumentative structures emerge in dialogue, where questions and answers play a central role. Existing dialogue corpora, such as QT30 (Hautli-Janisz et al., 2022) or US2016 (Hautli-Janisz et al., 2022), are typically limited to basic taxonomies of question types and tend to overlook how implied or suggested meanings are verbalized. They also lack systematic analyses of the relationships between questions and answers. Methodological diversity and the absence of consistent annotation guidelines hinder the comparability of results across studies. For languages other than English — including Polish — there are still no publicly available resources that enable systematic research on the pragmatic functions of questions and answers within argumentative structures.

Solution: To address the challenges outlined

Proceedings of the Fifteenth Language Resources and Evaluation Conference (LREC 2026), pages 5146–5157
11–16 May 2026.

above, we created POLAR — an experimental dataset based on orthographic transcriptions of Polish political radio interviews[1]. The corpus integrates two complementary perspectives: the pragmatic analysis of question–answer structures and the representation of argumentative structures. It includes manual annotation of questions and responses understood as speech acts (Austin, 1975; Searle, 1969), as well as cases where interrogative sentences function as premises or conclusions within reasoning structures. In such instances, annotators verbalized the assertive content implied by the question — the most demanding and, at the same time, the most innovative part of the annotation process. The dataset was manually annotated on two main levels:

(a) dialogic, including questions and answers analyzed as speech acts and categorized according to their pragmatic and semantic properties; and

(b) argumentative, capturing cases where interrogative sentences function as premises or conclusions within reasoning structures.

Consequently, POLAR provides empirical material for analyzing communicative situations in which the act of asking a question serves as a component of argumentation. In such cases, the annotators verbalized the assertive content implied by the question — the most demanding and, at the same time, the most innovative part of the annotation process.

Our contributions are as follows:

(1) Language resource: Preparation of orthographic transcripts of 21 radio interviews with leading Polish politicians, conducted between December 2023 and March 2024 (total: 593 minutes of recordings, 119,340 tokens, 8,980 sentences). These transcripts form the linguistic basis for the annotated dataset.

(2) Annotation guidelines: Development of detailed annotation guidelines covering questions, responses and argument structures, together with a discussion of their applicability and challenges in practice. The annotation scheme also included verbalization of the assertive content of interrogative sentences as premises or conclusions.

(3) Dataset: Creation of a corpus with manual annotations of: (a) questions and responses, (b) argumentative structures in which interrogative sentences serve as components.

(4) Validation: Establishment of a gold standard version of the corpus and reporting of baseline classification experiments as well as LLM-based evaluations, demonstrating the usefulness of the resource for both theoretical research and NLP applications.

[1] POLAR dataset repository

2. Related Works

A key reference in the study of political interviews remains the classical typology of responses developed by (Bull and Mayer, 1993), which was later expanded and adapted in other works (e.g. Clayman and Heritage, 2002; Bull, 2003; Bates et al., 2014). More recent studies (e.g. Ekström, 2009) emphasize that journalist–politician interactions are not only informational, but also ritualistic and persuasive in nature, making them a particularly valuable subject for argumentation research.

The pragmatic and rhetorical functions of interrogative sentences have been widely studied across various text genres. Hyland (2002) analyzed questions in academic writing as rhetorical devices, Webber (1994) examined their strategic role in medical discourse, and Fareh (2008) explored their pragmatic functions in dramatic texts. Other studies, such as Stivers and Enfield (2010), proposed cross-linguistic typologies of questions and answers. Simpler typologies have also been developed for the purposes of argumentation analysis within Argument Mining, most notably by Hautli-Janisz et al. (2022) and Kikteva et al. (2022).

The theoretical framework for this study is the Inference Anchoring Theory (IAT) (Budzynska et al., 2014; Budzynska et al., 2016), which integrates dialogue structure, inferential relations and speech acts into a coherent model for argumentation analysis. Within this framework, several corpora have been developed, including US2016 (Visser et al., 2020) — comprising transcripts of the 2016 U.S. presidential debates — and QT30 (Hautli-Janisz et al., 2022), which includes 30 episodes of the BBC Question Time program. Analyses based on these resources reveal the complexity of illocutionary forces associated with non-canonical questions and the challenges of their automatic classification (Kikteva et al., 2024; Schmidt et al., 2023). Despite the growing number of English-language corpora, there are still very few resources for Slavic languages, making POLAR an step toward cross-linguistic comparison in argumentation and political discourse studies.

Unlike the IAT corpora, POLAR adopts a binary classification of questions — those that contain an informational request component and those that do not. Additionally, in cases where an interrogative sentence forms part of a complex speech act with an argumentative function, annotators verbalized its assertive content and specified its role within the reasoning structure — either as a premise or a conclusion. This approach combines speech act analysis with the pragmatic reconstruction of implicit meanings — an aspect that has so far been absent from the standard practice of IAT.

The understanding of the concept of a question

in the POLAR corpus is rooted in speech act theory (Austin, 1975; Searle, 1969), according to which a question constitutes a speech act aimed at obtaining information. The distinction between canonical and non-canonical questions remains one of the key issues in pragmatic research. Puczyłowski (2022) emphasizes that the way a response meets conversational expectations can reveal the illocutionary type of the question, whereas Farkas (2024), within the framework of Table Theory (Farkas and Bruce, 2010), proposes analogous criteria of canonicity for both questions and assertions. Braun (2011) and Grice (1975) examine how questions can convey implicit meanings through conversational implicatures.

Our typology of responses is rooted in Bull and Mayer's (1993) framework, originally developed for British televised interviews. Given the genre similarity, their typology was particularly relevant to our material. In POLAR, it was simplified in some aspects and expanded in others to better capture the characteristics of Polish radio interviews. This typology also aligns with the interactional and pragmatic studies of Clayman and Heritage (2002), who analyzed journalists' questioning practices in political interviews.

Recent studies have revisited the problem of implicatures in questions (Braun, 2011; Puczyłowski, 2022), yet datasets capturing such phenomena remain scarce. In NLP, this issue has become increasingly prominent — Jeretic (2020) examined scalar implicatures, Lipkin (2023) analyzed model performance on pragmatic inference, and Ruis (2022) developed benchmarks for implicature understanding. These works reveal that current systems still struggle with implicit meaning.

3. Dataset

3.1. Material Selection Criteria

The selection criteria for the linguistic material included: **(a)** the use of contemporary spoken language, **(b)** broad audience reach and accessibility of recordings, and **(c)** public recognizability of both politicians and interviewers.

To meet the first criterion we selected interviews conducted between 2023 and 2024 representing a contemporary sample of Polish political spoken discourse. For the second criterion the interviews were sourced from nationwide radio stations with high audience ratings (including RMF FM[2], Radio Zet[3], Polish Radio[4]) and from their official YouTube channels, ensuring wide reach and public availability of the recordings. The third criterion was met by selecting conversations featuring nationally recognizable politicians—both current and former high-ranking public officials—as well as well-known political journalists.

During the selection process, care was taken to ensure proportional representation of politicians from all major parliamentary parties, reflecting the new political landscape following the October 2023 parliamentary elections, while maintaining a balanced total airtime across interviews. The topics discussed primarily concerned current political affairs, including the formation of the new government, coalition negotiations, and evaluations of decisions made by the previous administration.

3.2. Transcription

As noted by Edwards (2005, pp. 331–332), audio recordings are a valuable resource in discourse research but, on their own, are insufficient for the systematic analysis of interaction. Transcriptions, although inherently selective and interpretative, make it possible to capture the essential features of spoken exchanges in a format suitable for analytical purposes.

In this study, we adopted a minimalist transcription approach, following established conventions for representing spoken language (Sosnowska, 2009). The transcripts cover complete radio interviews, with no omissions or summaries, and are limited to orthographic representation with minimal use of punctuation. Paralinguistic elements such as laughter, sighs, hesitation markers and emotional or intonational cues (e.g., raised voice) were omitted.

A particular challenge involved sentence fragmentation and interruptions, which occur frequently in spontaneous—and especially political—dialogue. To ensure consistency and clarity, the following conventions were applied:

- Incomplete or interrupted utterances were marked with an ellipsis; utterances that were interrupted but later resumed also began with an ellipsis.
- When one speaker interrupted another, an ellipsis was placed at the beginning of the interrupting speaker's utterance.

The initial transcription was automatically generated using the Whisper (large model) (Radford et al., 2023) and subsequently manually verified and corrected by two independent annotators[5]. Manual post-editing included removing segmentation errors, normalizing orthographic forms, and standardizing the representation of proper names. The Whisper model was used exclusively for automatic speech-to-text transcription, and not for any

[2] https://www.rmf24.pl/
[3] https://www.radiozet.pl/
[4] https://www.polskieradio.pl/
[5] https://huggingface.co/openai/whisper-large-v3

subsequent linguistic processing or annotation.

3.3. Text Segmentation

The orthographic transcriptions were automatically converted into a `.json` format, in which each sentence represents the basic unit of annotation. Each record in the file corresponds to a single sentence, with boundaries determined during the transcription process. For every sentence, a unique identifier was assigned, along with a set of metadata including: **(a)** the speaker's full name, **(b)** their role in the conversation (journalist or politician), **(c)** party affiliation, and **(d)** the radio station and program, both accompanied by direct source links to the original recordings. This structure makes the corpus useful not only for research in AM but also for broader analyses of political discourse and public communication.

3.4. Manual Annotation Process

3.4.1. Annotators

Manual annotation was conducted by two experienced annotators: a graduate in Polish philology and a graduate in English philology specializing in NLP. Each text was annotated twice, with the annotators working independently. Before starting the main annotation task, they completed approximately 20 hours of training using sample interviews that were not included in the final dataset.

The annotation was performed using the Inforex web-based platform[6] (Marcińczuk and Oleksy, 2019), while the super-annotation was carried out using the *Tag Me Maybe* tool (developed by the authors; under review). The text was annotated on three levels: **(a)** identification of questions as speech acts, **(b)** classification of responses, and **(c)** marking of interrogative sentences functioning as premises or conclusions within argumentative structures. In such cases, the annotators reconstructed the entire argumentative structure. Importantly, they also verbalized the assertive content implied by the questions, which constituted the most challenging—and at the same time the most innovative and experimental—aspect of the annotation process. The methodological foundation for this task draws on the work of Ziembicki (2025).

The decision to annotate only interrogative sentences at the argumentative level was an intentional experimental design choice, aimed at exploring how this kind of sentences participate in reasoning structures within political discourse. This stage represents a natural point of departure and does not preclude extending the annotation to other sentence types in future phases of the project.

[6] https://clarin-pl.eu/tools/inforex

A gold standard version of the corpus was also created, based on reconciled annotations from both annotators. This version serves as a reference for model evaluation, including experiments with simple classifiers and large language models (LLMs).

3.4.2. Taxonomy

Information Question: The first annotation task was to identify sentences that function as questions in a pragmatic sense. We adopted the classical approach, according to which a question is defined as a speech act (Austin, 1975; Searle, 1969). However, this definition alone does not resolve the theoretical and practical challenges associated with the notion of a question: there is no universal agreement on how questions should be defined, and a single utterance may perform more than one pragmatic function simultaneously. See, for instance, the challenges involved in distinguishing canonical from non-canonical questions (Puczyłowski, 2022; Farkas, 2024).

In line with this pragmatic perspective, our goal was to determine whether a sentence that takes an interrogative form at the purely syntactic level actually performs the illocutionary act of requesting information. To this end, we developed a semantic–syntactic test that enabled annotators to determine whether a given sentence could be transformed into reported speech using a main predicate with a semantic core associated with questioning. The annotation results show that the most frequently used predicates of this type were ***zapytać*** (*ask*; *inquire*), ***dopytać*** (*ask further*; *follow up on*), ***dociekać*** (*probe*; *investigate further*; *inquire persistently*), and ***upewniać się*** (*verify; make sure*; *confirm*). For example:

> **(1)** *Minister, do you believe that Poland will receive six trillion two hundred billion in war reparations from Germany?*
> **(1a)** *He asked whether the minister believed that Poland would receive six trillion two hundred billion in war reparations from Germany.*

If the transformation of sentence (1) into (1a) is possible, this indicates that sentence (1) should be classified as a question. In this specific case, both annotators independently agreed that such a transformation was valid. However, it is worth noting that the speaker (the journalist) simultaneously expresses doubt about the proposition in the subordinate clause. This example illustrates that, despite performing the speech act of expressing doubt, the utterance still functions as a question in the pragmatic sense.

Although the semantic–syntactic test proved to be a useful tool for ensuring annotation consistency, its proper application required a high level of semantic intuition and sufficient time for reflection. It was

not conclusive on its own — it had to be used in parallel with pragmatic interpretation, which involved world knowledge and awareness of rhetorical strategies typical of political interviews. During annotator meetings, the practical effectiveness of the test was frequently discussed and refined: although its outcome ultimately depended on the annotator's linguistic intuition, the test provided a repeatable and structured interpretative framework that helped to systematize this intuition.

In the present procedure, the search for questions was limited exclusively to interrogative sentences, even though questions can also be formulated through declarative forms — particularly in the interview genre. This decision was a deliberate methodological choice aimed at maintaining annotation consistency and focusing on clear, unambiguous realizations of the questioning act. In future stages of the project, the analysis will be extended to other sentence types. The identification of interrogative sentences and the distinction between them and utterances containing question tags were performed entirely manually.

Responses: The second annotation task involved identifying types of responses to questions. We adopted a modified version of the typology developed by Bull and Mayer (1993), adapted to the specific characteristics of our dataset — simplified in certain areas, revised in others, and expanded with new categories. The typology used in our study is based on two mutually exclusive categories: Explicit Direct Answer and Explicit Non-Direct Answer. Below we present the labels for the Explicit Direct Answer category along with brief descriptions.

1. Explicit Direct Answer

The label Explicit Direct Answer indicates whether the respondent has provided a strictly direct answer to the question in its exact sense. It was assigned to sentences that explicitly contain the Content of a Question (Puczyłowski, 2022): *A proposition p is part of the semantic content of a question Q iff there is a sentence S such that (1) S expresses p, and (2) S is a proper answer to Q.*

Determining whether a response includes the Content of a Question is a strictly semantic task, requiring the annotator to assess whether the proposition expressed in the response corresponds to the informational gap established by the question. If the response includes the Content of a Question but, for instance, does so in an elliptical manner — that is, without stating it explicitly — it was assigned a different label (see Elliptical Direct Answer below and other labels within this group). Example (from an actual interview):

(1) *What is this lady's name?*
(2) *Ms. Grażyna, a Polish teacher.*

To clearly define the sought information in question (1), it can be reformulated as: "This lady's name is xxx". This makes it evident that response (2) directly provides the missing content without requiring further interpretation.

2. Non-Factive Direct Answer

This label applies to sentences in which the Content of a Question falls within the scope of non-factive epistemic expressions. In the case of non-factive epistemic verbs (*think that...*, *be sure that...*, *be convinced that...*, *believe that...*, *suppose that...*, etc.), such verbs do not imply the truth of the subordinate clause, in contrast to factive verbs (*know that...*, *realise that...*, *be aware that...*), which presuppose its truth (see Ziembicki et al., 2024). In this type of response, the questioned content is preserved but presented through a speaker's epistemic stance rather than as a factual assertion. Example:

(1) *Would you like to run for office?*
(2) *I think so. Politics is my passion.*

3. Contextual Direct Answer

This type of response contains the Content of a Question, but determining it requires additional contextual information that is not included in the response itself. Example:

(1) *Minister, how many Ukrainian children are enrolled in Polish schools?*
(2) *Seven thousand fewer than on June 24*.

The response in (2) does not directly provide the numerical value requested by the journalist. To determine the actual number, the listener must know the reference point (the number as of June 24). Full understanding thus depends on external contextual knowledge.

4. Conditional Direct Answer

This category includes responses in which the Content of a Question is expressed within a conditional construction. In this type of response, the requested information depends on the fulfillment of a specific condition. The value may appear either in the protasis (if-clause) or the apodosis (then-clause) of the sentence. Example:

(1) *Would a referendum be binding for the Polish People's Party?*
(2) *Yes, if the result were representative — then of course, yes.*

5. Elliptical Direct Answer

This category includes responses in which the Content of a Question is present but not explicitly stated. The information is implied through elliptical structure. Example:

(1) *Do you know where [Germany] had colonial possessions?*
(2) *Namibia, for example...*

In the response “Namibia, for example…”, there is an implicit yes, which functions as confirmation and provides the missing informational content. The value is not stated explicitly but inferred from the elliptical structure.

6. Declared Ignorance Answer

This category includes responses in which the speaker explicitly states their lack of knowledge, thus indicating the inability to provide the requested information. Example:

> **(1)** *Will Jacek Kurski run on your lists for the European Parliament?*
> **(2)** *I don't know anything about that.*

7. Scalar Implicature Answer

This label applies to utterances in which the speaker conveys information implicitly by means of a scalar implicature (see: Jeretic et al., 2020). Example:

> **(1)** *Were you satisfied with the way public television looked?*
> **(2)** *I wasn't satisfied with everything either — that's obvious.*

The seven labels above constitute the Explicit Direct Answer subcategory. The complementary category, Explicit Non-Direct Answer, is defined negatively: if a response does not fit into any of the seven Direct Answer subcategories, annotators labeled it as Explicit Non-Direct Answer.

The final stage of annotation involved identifying argumentative structures — but only those in which an interrogative sentence appeared at the syntactic level. This restriction was a deliberate experimental choice aimed at examining how interrogative sentences — and, at the level of speech acts, questions — can function as components of reasoning in political discourse. Below is an example of an annotated structure:

> **[1]** *I have a printed payslip in front of me, showing the salary of Ms. Marta, a mathematics teacher at a high school in a small town.*
> **[2]** *Ms. Marta takes home two thousand five hundred thirty zlotys and twenty-three groszes.*
> **[3]** *I have a question for you, quite seriously: how can a teacher afford to buy an apartment?*

Argumentative structure:

[PREMISE] [1] → Ms. Marta is a mathematics teacher at a high school in a small town.

[PREMISE] [2] → Ms. Marta takes home two thousand five hundred thirty zlotys and twenty-three groszes.

[CONCLUSION] [3] → A teacher cannot afford to buy an apartment.

In the three-sentence excerpt above, the first two sentences were annotated as premises, while the interrogative sentence functions as a conclusion. For each premise and conclusion, annotators reconstructed the propositional content by verbalizing its assertive meaning, thereby making implicit reasoning explicit.

The process of verbalization followed two methodological rules described in (2025):

R1: Preserve as much lexical similarity to the original sentence as possible.

R2: Produce a sentence that is both unambiguous and natural-sounding in Polish.

The proportion of interrogative sentences functioning as elements of argumentative structures (13%) may appear moderate, yet it is crucial for understanding the dynamics of political interviews — it is precisely these cases that reveal how questions co-construct lines of reasoning and shape the development of argumentation. We emphasize that this annotation layer was treated as experimental — its results are intended primarily to provide empirical material for theoretical analysis.

3.5. Dataset in Numbers

The basic numerical data describing the presented dataset is provided in Table 1. Compared to other datasets used in AM, its size can be classified as medium (119,340 tokens). Table 2 and Table 3 , in turn, summarizes the annotation results obtained for the dataset.

Category	Value
Number of texts	21
Duration (min)	593
Tokens	119,340
Total number of sentences	8,980
Total number of speakers	28
Journalists	7
Politicians	21

Table 1: Dataset in Numbers.

Category	Count	(%)
Interrogative Sentence	1,267	14.1
Information Question	1,007	11.2
Explicit Direct Answer	334	3.7
Explicit Non-Direct Answer	530	5.9
Non-Factive Direct Answer	32	0.4
Declared Ignorance Answer	29	0.3
Elliptical Direct Answer	29	0.3
Premise	239	2.7
Conclusion	150	1.7
Total sentences	**8,980**	**100.0**

Table 2: Annotation statistics for the gold-standard version of the POLAR dataset.

Category	Count	(%)
IS as Premise or Conclusion	167	13.2
IQ as Premise or Conclusion	140	11.0

Table 3: Frequency of interrogative sentences and Information questions functioning as elements of argumentative structures.

4. Inter Annotator Agreement (IAA)

To assess IAA, Cohen's Kappa coefficient was calculated separately for each level of annotation. The results are presented in Table 4.

Category	Cohen's κ
Information Question	0.86
Explicit Direct Answer and Explicit Non-Direct Answer	0.64
Types of Direct Answer	0.62
IS as Premise or Conclusion	0.40

Table 4: Cohen's κ values for individual annotation categories.

The high level of agreement in the use of the Information Question label indicates that the criteria for applying it were clear and unambiguous, and that the task itself did not raise major disagreements. Nevertheless, as mentioned earlier, feedback gathered from annotators during regular meetings suggests that the reported-speech test used in this task is not fully conclusive. Moreover, the genre itself — the radio interview — may have influenced the identification of speech acts as questions, since asking questions in the sense of requesting information constitutes the journalist's canonical communicative role.

The results for the distinction between Explicit Direct Answer and Explicit Non-Direct Answer were somewhat lower but still relatively high, considering the pragmatic–semantic nature of the task. A marked decrease in agreement was observed in the classification of Direct Answer subtypes (moderate agreement), which is not surprising given the seven detailed categories and their uneven frequency in the dataset.

The lowest Kappa values were obtained for the category Interrogative sentence as part of an argumentative structure, reflecting the highly interpretive nature of this annotation layer. It should be emphasized, however, that this part of the annotation was still experimental at the current stage of the project. The obtained IAA values do not differ significantly from those reported in comparable studies in the field of AM.

5. Experiments

A series of experiments was conducted to evaluate the performance of selected models in two main areas: the classification of utterances into predefined semantic categories and the verbalization of implicit content within argumentative structures.

5.1. Tasks

Q-A Classification Two classification scenarios were defined for this part of the study:

(1) (Answers): This task involved classifying utterances as either Explicit Direct Answer or Explicit Non-Direct Answer. Only samples labeled with one of these two categories were included. Each utterance was provided to the model along with a dialogue context consisting of two preceding utterance blocks. An utterance block is understood as a sequence of utterances representing a segment of interaction between consecutive contributions from the same speaker.

(2) (Information Question): The goal of this task was to determine whether a given utterance constitutes an Information Question. Only utterances containing a question mark were considered. Each input sample included both the preceding and the following dialogue context, represented by two utterance blocks on each side.

For the Answer classification task, the dataset comprised 961 annotated utterances. A stratified 70/30 split was applied, resulting in 672 training and 289 test instances. The test set contained 131 Explicit Direct Answer and 158 Explicit Non-Direct Answer instances. For the Information Question task 1,321 utterances were used. Applying the same stratified 70/30 split resulted in 923 training and 398 test instances, including 303 Information Question and 95 Non-Information Question instances in the test set.

Experiments were carried out using both discriminative and generative models. For the discriminative models, `bert-base-multilingual-cased` (Devlin et al., 2018), `herbert-base-cased` and `herbert-large-cased` (Mroczkowski et al., 2021) were employed. These models were fine-tuned in a standard classification setting, employing a linear layer and a cross-entropy loss function. Training was performed for 3 epochs with a learning rate of 2e-5 and a batch size of 8.

Additionally, large language models (LLMs) were assessed using a generative approach. In this setup, models dedicated to the Polish language were employed: `speakleash/Bielik-11B-v3.0-Instruct` (Ociepa et al., 2025) and `pllum-12b-nc-chat-250715` (Consortium, 2025). Three configurations were tested: zero-shot,

Category		Model	EA	IQ
Discriminative		bert-base-multilingual-cased	73.07	64.38
		herbert-base-cased	79.72	65.14
		herbert-large-cased	**80.89**	**78.21**
Generative	Zero-shot	Bielik-11B-v3.0-Instruct	58.88	64.69
		pllum-12b-nc-chat-250715	59.23	53.18
	Few-shot	Bielik-11B-v3.0-Instruct	60.31	63.03
		pllum-12b-nc-chat-250715	55.43	50.34
	Fine-tuned	Bielik-11B-v3.0-Instruct	**81.46**	**83.58**
		pllum-12b-nc-chat-250715	75.70	75.88

Table 5: Evaluation results of classification tasks (F1 scores).

Model	Premise	Conclusion
Bielik-11B-v3.0-Instruct	**87.20**	**79.84**
pllum-12b-nc-chat-250715	73.18	69.78

Table 6: Evaluation results of implicatures generation task (cosine similarities).

few-shot and fine-tuned. In the fine-tuned configuration, the models were adapted to the task using the LoRA method, trained for 3 epochs with a learning rate of 1e-4 and a batch size of 1.

Generating implicit argumentative content The second task aimed to assess the models' ability to verbalize implied components of argumentative structures based on dialogue context. The models were provided with utterances labeled as Premise or Conclusion, along with two preceding utterance blocks, and were prompted to generate the corresponding content. The evaluation was performed on a test set of 341 instances. In total, 239 instances were annotated with a Premise implicature and 150 with a Conclusion implicature.

5.2. Results

The results of the conducted classification tasks are presented in Table 5. Macro F1 score was adopted as the primary metric for evaluating the quality of the models' predictions. Among the discriminative models, herbert-large-cased achieved the highest performance, reaching F1 scores of 80.89 for EA and 78.21 for IQ. For the generative models, the few-shot approach resulted in only a minor improvement over the zero-shot setting in one case, while in most cases it led to lower performance. A substantial increase in effectiveness was observed only after applying fine-tuning. In particular, the fine-tuned Bielik-11B-v3.0-Instruct model achieved the best overall results, with F1 scores exceeding 80 across both tasks. These results suggest that generative models require additional adaptation to the specific characteristics of the tasks.

To evaluate the performance of the models in the task of generating implicit content, semantic similarity between the generated and reference content was employed. This similarity was computed using the sdadas/stella-pl-retrieval-8k (Dadas et al., 2024) model with the cosine similarity metric. Evaluation results are presented in Table 6, reported separately for Premise and Conclusion implicature. It can be observed that the scores are higher for Premises, while overall both Premise and Conclusion attain reasonably good performance.

6. Qualitative Error Analysis

Our qualitative analysis focuses only on the most recurrent and methodologically informative patterns. In particular, we examine cases in which all three models made the same classification error.

Explicit Direct Answer vs Explicit Non-Direct Answer Incorrect predictions occurred both in highly context-dependent examples and in relatively simple utterances, which indicates that contextual complexity alone does not explain all errors. The example below illustrates one such type of case.

> **(1)** *A: Who will be in it, how many people will it include, and what exactly will it do?*
> **(2)** *B:* (2a) *It will focus on work in more than a dozen areas. For each of those areas, I will invite many people, both politicians and experts.* **(2b)** ***I want to bring in as many specialists as possible, including local government people and entrepreneurs, into this team for public affairs.***
> **Gold:** Exp. direct answ. | **Models:** Exp. non-direct answ.

Question (1) contains more than one content of a question, which in itself makes it a problematic type of example. Answer (2a) addresses only one of these content. The rationale behind the gold

standard was that, although the utterance does not provide a full resolution, it partially closes the informational gap concerning the composition of the team. Cases of this kind do not fit neatly into a simple distinction between Explicit Direct and Non-Direct Answer; in this instance, (2a) was judged to be closer to answering the question than to avoiding it. Examples of this kind require particular attention in future research.

Information Question vs Non-Information Question The analysis of results for questions reveals a clear predominance of cases in which the gold annotation labeled an utterance as a Non-Information Question, while all models classified it as an Information Question. These examples suggest that models are particularly prone to errors when a question carries a strong assertive component, that is, when an utterance has interrogative form but simultaneously functions as a suggestion, pressure, disbelief, or reinforcement of a previously expressed stance; the examples below illustrate this pattern.

(1) *A: You don't read comments from...*
(2) *B: ... Not at all...*
(3) *A: ... internet users?*
(4) *B:* (4a) *Not at all.* **(4b) *But why would I, then?*** (4c) *I'd rather...*

(1) *A: ...* **(1a) *You are obliging yourselves to do something?*** (1b) *Then change the law.*

Sentence (4b) does not serve to obtain information, but rather conveys that reading comments is pointless. In the second example, (1a) likewise does not establish a genuine informational gap, but instead expresses surprise and functions as rhetorical pressure.

Cosine Similarity The analysis of implicit content generation results suggests that the CS measure alone may, in some cases, overestimate prediction quality. Although the model's output may be semantically close to the GOLD version, it may still omit crucial lexical elements, such as negation. This is illustrated by the premise verbalizations in the GOLD version and in the Bielik output below:

GOLD: Sienkiewicz was not demoted.
Bielik: Sienkiewicz stopped being a colonel.

These sentences stand in a relation of contradiction; nevertheless, the CS score for this example is 0.71. It should be noted, however, that such cases are relatively rare.

7. Conclusion and Future Work

This paper presents POLAR, an experimental corpus of 21 Polish political radio interviews. The aim of this annotation was to collect material for analyzing the relationship between the form and function of questions and the types of responses they elicit. On a theoretical level, the project was experimental in nature: it combined the pragmatic analysis of dialogue with the modeling of argumentative structures, representing a step toward the creation of AM resources that integrate multiple levels of analysis.

The analysis of the corpus shows that not all interrogative sentences perform the function of a question understood as a request for information — approximately one quarter of utterances ending with a question mark were not labeled as questions. Moreover, 8–11% of interrogative sentences were annotated as elements of argumentative structures, functioning as premises or conclusions. These findings confirm that both questions and other speech acts realized syntactically through interrogative sentences can participate in argumentative processes and should be taken into account in computational models within AM.

The conducted experiments showed that, although the obtained results can be considered satisfactory, fully capturing the complex pragmatic dependencies in dialogue remains a challenge for contemporary language models. This stems both from the need to account for multi-turn conversational context and from the subtle, often ambiguous relationships between discourse elements.

Although the POLAR corpus is of medium size (119,340 tokens), it stands out due to its manual annotation covering three complementary analytical dimensions — questions, answers and argumentative structures. As a resource in Polish, a language rarely represented in discourse and argument mining studies, POLAR fills an important gap and complements existing predominantly English-language datasets. Owing to its rich metadata, the corpus can also support broader analyses of political communication and public discourse.

In the next stages of the project, the annotation scheme will be extended to include declarative sentences functioning as information-seeking questions. A further step will be to apply the question typology proposed by Hautli-Janisz (2022) or related frameworks, ensuring greater methodological consistency and broader cross-study comparability.

A more fine-grained classification of interrogative speech acts beyond the binary distinction adopted here would also be desirable. One promising direction is Nielsen's (2020) extension of Searle's taxonomy based on so-called preparatory conditions.

Another important avenue for future work is the adoption of widely used discourse annotation standards, such as the PDTB framework (Prasad et al., 2004; Prasad et al., 2008) and ISO standards — ISO 24617-8 (ISO 24617-8:2016, 2016), which concerns discourse relations and ISO 24617-2 (Bunt et al., 2012), designed for dialogue act annotation.

8. Limitations

The first limitation concerns the number of annotators. Although the corpus was double-annotated and the gold standard version was established through consensus between annotators under the supervision of a super-annotator, the interpretation of implicit meanings and illocutionary force inevitably involves a degree of subjectivity. In the future, it would therefore be advisable to develop a dataset annotated by a larger number of annotators.

Another limitation — and at the same time an important direction for future development — concerns the absence of prosodic annotation. Previous studies have shown that prosody, particularly intonational contours, plays a crucial role in signaling the illocutionary force of questions and, for instance, in distinguishing rhetorical from information-seeking questions (Dehé and Braun, 2020; Cresti and Moneglia, 2023). These phenomena, however, are highly language-dependent and require specialized acoustic annotation. For Polish, systematic prosodic analyses of rhetorical and non-rhetorical questions are still lacking, which makes this an important direction for further research.

In a broader perspective, it is worth noting that the use of prosodic—and more broadly, multimodal—data in discourse analysis, including in Argument Mining research, is only beginning to develop (Mancini et al., 2024). Yet it is precisely within these communicative channels that essential information about the structure and dynamics of discourse is encoded.

For this reason, the absence of a prosodic layer should not be viewed merely as a limitation but rather as an indication of a broader methodological gap: the fields of linguistic prosody research and argumentation analysis remain largely separate. Integrating these two areas—particularly through multimodal models of discourse analysis—appears essential for achieving a more comprehensive account of argumentative communication in natural language.

Another limitation concerns the experimental nature of the argumentative structure annotation. This layer was primarily designed to explore the feasibility of identifying reasoning patterns involving interrogative sentences, rather than to serve as a mature benchmark for NLP evaluation. While it provides valuable qualitative insights into how questions can function as premises or conclusions, the current sample size and level of granularity are insufficient for training or reliably testing computational models. Future iterations of the corpus will therefore focus on expanding and standardizing this annotation layer so that it can be incorporated into machine learning workflows.

Acknowledgements

This research was funded by the National Science Centre, Poland, in the framework of the PRELUDIUM project 2025/59/D/HS2/01238.

9. Bibliographical References

John Langshaw Austin. 1975. *How to do things with words*. Harvard university press.

Stephen R Bates, Peter Kerr, Christopher Byrne, and Liam Stanley. 2014. Questions to the prime minister: A comparative study of pmqs from thatcher to cameron. *Parliamentary Affairs*, 67(2):253–280.

David Braun. 2011. Implicating questions. *Mind & Language*, 26(5):574–595.

Katarzyna Budzynska, Mathilde Janier, Chris Reed, and Patrick Saint-Dizier. 2016. Theoretical foundations for illocutionary structure parsing. *Argument & Computation*, 7(1):91–108.

Peter Bull. 2003. *The microanalysis of political communication: Claptrap and ambiguity*. Routledge.

Peter Bull and Kate Mayer. 1993. How not to answer questions in political interviews. *Political psychology*, pages 651–666.

Steven Clayman and John Heritage. 2002. *The news interview: Journalists and public figures on the air*. Cambridge University Press.

HIVE AI Consortium. 2025. Pllum: A family of polish large language models.

Emanuela Cresti and Massimo Moneglia. 2023. The role of prosody for the expression of illocutionary types. the prosodic system of questions in spoken italian and french according to language into act theory. *Frontiers in Communication*, 8:1124513.

Slawomir Dadas, Michał Perełkiewicz, and Rafał Poświata. 2024. Pirb: A comprehensive benchmark of polish dense and hybrid text retrieval methods. In *Proceedings of the 2024 Joint International Conference on Computational Linguistics, Language Resources and Evaluation (LREC-COLING 2024)*, pages 12761–12774.

Nicole Dehé and Bettina Braun. 2020. The prosody of rhetorical questions in english. *English Language & Linguistics*, 24(4):607–635.

Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 2018. BERT: pre-training of deep bidirectional transformers for language understanding. *CoRR*, abs/1810.04805.

Giulia D'Agostino, Ella Schad, Eimear Maguire, Costanza Lucchini, Andrea Rocci, and Chris Reed. 2024. Superquestions and some ways to answer them. *Journal of Argumentation in Context*, 13(3):319–372.

Jane A Edwards. 2005. The transcription of discourse. *The handbook of discourse analysis*, pages 321–348.

Mats Ekström. 2009. Power and affiliation in presidential press conferences: A study on interruptions, jokes and laughter. *Journal of Language and Politics*, 8(3):386–415.

Shehdeh Fareh. 2008. Pragmatic functions of interrogative sentences in english: A corpus-based study. *International Journal of Arabic-English Studies*, 9(1):145–164.

Donka Farkas. 2024. Canonical and non-canonical questions in discourse.

Donka F Farkas and Kim B Bruce. 2010. On reacting to assertions and polar questions. *Journal of semantics*, 27(1):81–118.

Herbert P Grice. 1975. Logic and conversation. In *Speech acts*, pages 41–58. Brill.

Annette Hautli-Janisz, Katarzyna Budzynska, Conor McKillop, Brian Plüss, Valentin Gold, and Chris Reed. 2022. Questions in argumentative dialogue. *Journal of Pragmatics*, 188:56–79.

Ken Hyland. 2002. What do they mean? questions in academic writing. *Text & Talk*, 22(4):529–557.

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.

Zlata Kikteva, Kamila Gorska, Wassiliki Siskou, Annette Hautli, and Chris Reed. 2022. The keystone role played by questions in debate. In *Proceedings of the 3rd Workshop on Computational Approaches to Discourse*, pages 54–63.

Zlata Kikteva, Alexander Trautsch, Steffen Herbold, and Annette Hautli. 2024. Question type prediction in natural debate. In *Proceedings of the 25th Annual Meeting of the Special Interest Group on Discourse and Dialogue*, pages 624–630.

John Lawrence and Chris Reed. 2015. Combining argument mining techniques. In *Proceedings of the 2nd Workshop on Argumentation Mining*, pages 127–136.

John Lawrence and Chris Reed. 2020. Argument mining: A survey. *Computational Linguistics*, 45(4):765–818.

Hao Li, Viktor Schlegel, Yizheng Sun, Riza Batista-Navarro, and Goran Nenadic. 2025. Large language models in argument mining: A survey. *arXiv preprint arXiv:2506.16383*.

Benjamin Lipkin, Lionel Wong, Gabriel Grand, and Joshua B Tenenbaum. 2023. Evaluating statistical language models as pragmatic reasoners. *arXiv preprint arXiv:2305.01020*.

Robert Mroczkowski, Piotr Rybak, Alina Wr 'oblewska, 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.

Niels Møller Nielsen. 2020. Expanding searle's analysis of interrogative speech acts: A systematic classification based on preparatory conditions. *Scandinavian Studies in Language*, 11(1):7–19.

Krzysztof Ociepa, Łukasz Flis, Remigiusz Kinas, Krzysztof Wróbel, and Adrian Gwoździej. 2025. Bielik 11b v3: Multilingual large language model for european languages.

Tomasz Puczyłowski. 2022. A taxonomy of non-canonical uses of interrogatives. *Axiomathes*, 32(3):505–527.

Laura Ruis, Akbir Khan, Stella Biderman, Sara Hooker, Tim Rocktäschel, and Edward Grefenstette. 2022. Large language models are not zero-shot communicators. *arXiv preprint arXiv:2210.14986*.

Klaus Schmidt, Andreas Niekler, Cathleen Kantner, and Manuel Burghardt. 2023. Classifying speech acts in political communication: A transformer-based approach with weak supervision and active learning. In *2023 18th Conference on Computer Science and Intelligence Systems (FedCSIS)*, pages 739–748. IEEE.

John R Searle. 1969. *Speech acts: An essay in the philosophy of language*. Cambridge university press.

Natalia Sosnowska. 2009. Oznaczanie segmentacji w zapisie tekstu mówionego. *Roczniki Humanistyczne*, 57(06):189–200.

Tanya Stivers and Nick J Enfield. 2010. A coding scheme for question-response sequences in conversation. *Journal of pragmatics*, 42:2620–2626.

Douglas Neil Walton. 1989. Question-reply argumentation.

Pauline Webber. 1994. The function of questions in different medical journal genres. *English for Specific Purposes*, 13(3):257–268.

Daniel Ziembicki. 2025. Questions as elements of argumentation in political debates. *Argumentation*, 39(4):601–634.

Daniel Ziembicki, Karolina Seweryn, and Anna Wróblewska. 2024. Polish natural language inference and factivity: An expert-based dataset and benchmarks. *Natural Language Engineering*, 30(2):385–416.

10. Language Resource References

Katarzyna Budzynska, Mathilde Janier, Chris Reed, Patrick Saint-Dizier, Manfred Stede, and Olena Yaskorska. 2014. A model for processing illocutionary structures and argumentation in debates. In *LREC 2014: Ninth International Conference on Language Resources and Evaluation*, pages 917–924. European Language Resources Association.

Harry Bunt, Jan Alexandersson, Jae-Woong Choe, Alex Chengyu Fang, Koiti Hasida, Volha Petukhova, Andrei Popescu-Belis, and David R Traum. 2012. Iso 24617-2: A semantically-based standard for dialogue annotation. In *LREC*, pages 430–437.

Hautli-Janisz, Annette and Kikteva, Zlata and Siskou, Wassiliki and Gorska, Kamila and Becker, Ray and Reed, Chris. 2022. *Qt30: A corpus of argument and conflict in broadcast debate*. European Language Resources Association (ELRA).

Jeretic, Paloma and Warstadt, Alex and Bhooshan, Suvrat and Williams, Adina. 2020. *Are Natural Language Inference Models IMPPRESsive? Learning IMPlicature and PRESupposition*. Association for Computational Linguistics. PID https://github.com/alexwarstadt/data$_{g}eneration$.

Eleonora Mancini, Federico Ruggeri, Stefano Colamonaco, Andrea Zecca, Samuele Marro, and Paolo Torroni. 2024. Mamkit: A comprehensive multimodal argument mining toolkit. In *Proceedings of the 11th Workshop on Argument Mining (ArgMining 2024)*, pages 69–82.

Michał Marcińczuk and Marcin Oleksy. 2019. Inforex — a collaborative systemfor 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.

Rashmi Prasad, Nikhil Dinesh, Alan Lee, Eleni Miltsakaki, Livio Robaldo, Aravind K Joshi, Bonnie L Webber, et al. 2008. The penn discourse treebank 2.0. In *LREC*.

Rashmi Prasad, Eleni Miltsakaki, Aravind Joshi, and Bonnie Webber. 2004. Annotation and data mining of the penn discourse treebank. In *Proceedings of the Workshop on Discourse Annotation*, pages 88–95.

Alec Radford, Jong Wook Kim, Tao Xu, Greg Brockman, Christine McLeavey, and Ilya Sutskever. 2023. Robust speech recognition via large-scale weak supervision. In *International conference on machine learning*, pages 28492–28518. PMLR.

Jacky Visser, Barbara Konat, Rory Duthie, Marcin Koszowy, Katarzyna Budzynska, and Chris Reed. 2020. Argumentation in the 2016 us presidential elections: annotated corpora of television debates and social media reaction. *Language Resources and Evaluation*, 54(1):123–154.