Introducing corpora Hlava Cor and Hlava AD: Human Label Variation in Coreference and Discourse Relations

Anna Nedoluzhko, Šárka Zikánová, Jiří Mírovský, Milan Straka and Eva Hajičová
Charles University, Faculty of Mathematics and Physics, Institute of Formal and Applied Linguistics
Prague, Czech Republic
{nedoluzhko, zikanova, mirovsky, straka, hajicova}@ufal.mff.cuni.cz

Abstract

As previous research on annotator disagreement in discourse phenomena has shown, understanding text coherence varies considerably from one individual to another. To explore this phenomenon, we created two corpora with multiple annotations of Czech texts, accompanied by annotators' explanations of their choices. The first corpus consists of 1,024 contexts annotated in parallel by three annotators. It captures differences in the identification of coreference across various text types and grammatical-semantic categories, including pronouns, full noun phrases, and anaphoric adverbials. The second corpus comprises 512 contexts, annotated in parallel by five annotators, and focuses on identifying discourse relations in attributive and non-attributive constructions. Both corpora achieve a comparable inter-annotator agreement of approximately 60–65%. For coreference annotation, agreement tends to be lower in cases where automatic coreference resolution models disagree, suggesting that when the models disagree, the examples tend to be more difficult or ambiguous for human annotators to interpret. The annotators' comments, both for coreference and discourse relations, further reveal differences in interpretation, varying levels of confidence in text understanding, and individual reading strategies.

Keywords: comprehension of coherence, multiple annotations, commented annotation

1. Introduction

During our long-term development of corpora focused on discourse relations, coreference, and information structure, we observed that certain linguistic phenomena tend to produce lower inter-annotator agreement than others. This observation aligns with a growing body of research on annotation evaluation taking into account different perspectives (Basile et al., 2021) and Human Label Variation (HLV, Plank, 2022), which suggests that disagreement in NLP is often a reflection of inherent linguistic complexity rather than mere noise. It is possible to identify several factors that appear to contribute to this variation, ranging from the polysemy of synsemantic signals to word order patterns (see, e.g., Zikánová, 2024). Furthermore, annotator-related factors, such as the subjective perception of the relative importance of text segments, play a significant role. Such interpretative diversity is well-documented in large-scale projects; for instance, Poesio et al. (2020) and Recasens et al. (2010) have consistently observed that ambiguity in textual meaning naturally reduces agreement in coreference tasks.

In the present paper, we introduce two corpora created to investigate human label variation. **Hlava Cor** (Human Label Variation in Coreference (Nedoluzhko et al., 2026)) contains annotations of coreference and focuses on human label variation related to genericity, subjectivity and underspecification. **Hlava AD** (Human Label Variation in Attribution and Discourse (Šárka Zikánová et al., 2024)) comprises annotations of discourse relations in attributive and non-attributive constructions.

In developing these corpora, we took into account various hypotheses about the possible causes of annotation disagreement, based on our previous annotation experience and analyses. For the coreference corpus, we specifically distinguish generic and specific expressions, as well as different grammatical–semantic categories, including pronouns, full noun phrases, and anaphoric adverbials. We further hypothesize that coreference resolution models may partially predict cases of underspecification, in the sense that where models disagree, human annotators are also likely to find the interpretation of coreference relations problematic (see Section 4.3 and 4.4).

For discourse relations, we assume that in texts with larger segments of direct or indirect speech, it may be more difficult for recipients to recognize which of the upcoming sentences are still related to the direct/indirect speech (see Section 5).

We argue that multiple annotation, and especially the detailed comments accompanying each annotation, provide rich material for investigating the reasons and nuances of human label variation, enhancing our understanding of how people interpret texts differently. We understand our corpora as datasets providing many types of commented structures, and thus serving as a base for future psycholinguistic experiments dealing with special cases. Furthermore, these corpora may serve as reference datasets for evaluating the reliability of previously single-annotated corpora.

Proceedings of the Workshop on Structured Linguistic Data and Evaluation, pages 237–247
May 11, 2026.

2. Related Work

The interpretation of linguistic meaning is never entirely fixed or uniform; it is shaped by context, perspective, and the inherent ambiguity of natural language. This fundamental indeterminacy influences all levels of linguistic analysis, including discourse and coreference annotation, where human judgments often diverge even under detailed annotation guidelines. Researchers in discourse annotation address this challenge in different ways. For example, Marchal et al. (2022) discuss methods evaluating annotation quality across multiple annotators. Building on the "ecologically valid" explanations introduced by Jiang et al. (2023), recent work by Weber-Genzel et al. (2024) utilizes a two-round annotation procedure where participants justify their label choices through natural language explanations. In these studies, annotators classify the relations between clauses as entailment, contradiction, or neutral and provide commentaries that allows researchers to distinguish between interpretative variation and annotation errors. Other studies emphasize the importance of explicitly capturing interpretative plurality (Plank, 2022, Basile et al., 2021, Crible et al., 2019).

New corpora are published capturing multiple understandings of discourse relations, e.g. in the RST approach (Peng et al., 2022, Polakova et al., 2024, Hewett and Stede, 2025), or implicit discourse relations in the PDTB approach (Scholman et al., 2022, Yung et al., 2024). None of these datasets include annotators' comments on their choices.

Comparative research on spoken and written data further demonstrates that modality contributes to interpretative diversity: spoken language tends to involve more implicit and multifunctional relations (Rehbein et al., 2016, Cuenca, 2017, Crible et al., 2019), though its annotation consistency is not necessarily lower than that of written texts (Zufferey and Crible, 2015).

Ambiguity in textual meaning naturally reduces inter-annotator agreement (IAA), as has been consistently observed in large-scale coreference annotation projects (Weischedel et al., 2011; Zeldes, 2017; Recasens and Martí, 2010; Poesio, 2020). While analyses of annotation disagreement have occasionally been included (Recasens et al., 2011; Pradhan et al., 2012), such studies are typically limited to small, task-specific subsets of data (cf. Levine and Zeldes, 2025). A few corpora explicitly encode ambiguous or near-identity cases (Uryupina et al., 2020; Bourgonje and Stede, 2020; Ogrodniczuk et al., 2013), but these remain statistically rare, leaving much of the natural variation in human interpretation unexplored.

Poesio et al. (2019) present a preliminary analysis of disagreements in a corpus of anaphoric information in 542 English documents crowdsourced through a game-with-a-purpose. The corpus contains multiple concurrent annotations of about 108 thousand markables, with 20 judgments on average per markable (12 annotations and 8 validations). Expert analysis of a sample of the data shows, however, that genuine ambiguity occurs only in about 9% of the markables and the other cases of annotators' disagreement should be accounted to annotators' errors and to various limitations of the coding scheme and the annotation interface.

In our analysis, we use the Prague Dependency Treebank - Consolidated 2.0 (PDT-C 2.0; (Hajič et al., 2024)) as the primary data source for extracting and then multiply annotating the data.

Namely, we use its subcorpus Prague Dependency Treebank (PDT) for the written mode and Prague Dependency Treebank of Spoken Czech (PDTSC) for the spoken mode. The PDT represents texts from the Czech newspapers from 1990's, while the PDTSC contains transcripts of conversations between interviewers and Holocaust survivors or contemporary witnesses of communism. The PDT-C 2.0 corpus contains manual annotations of coreference and discourse relations, along with multiple linguistic layers ranging from morphology to tectogrammatics. Annotation in PDT-C 2.0 follows the principle of a single correct decision, with approximately 10% of the data annotated in parallel to assess inter-annotator agreement (IAA). The collection comprises both spoken and written subcorpora, enabling comparative analyses of discourse phenomena across modalities.

3. Hlava Cor and Hlava AD: General Settings

To study human label variation, we extracted texts primarily from different parts of the PDT-C 2.0 (see detailed description in Sections 4.3 and 5) and created Hlava Cor and Hlava AD. The overall statistics for the annotated data are listed in Table 1. Hlava Cor contains 1,024 cases annotated by three annotators, while Hlava AD contains 512 cases annotated by five annotators.[1]

Both corpora consist of short Czech text segments containing the relevant linguistic phenomena. Each text segment was multiply annotated by several annotators. The annotations were performed on the linear texts presented in an Excel spreadsheet.

[1]The discrepancy in the size of the two corpora and the number of participating annotators does not stem from a specific theoretical or methodological requirement, but rather reflects the successive stages of their development and the varying availability of resources during each phase of the project.

Table 1: Overall Statistics for Hlava Cor and Hlava AD

Corpus Name	Number of Cases	Number of Annotators
Hlava Cor	1,024	3
Hlava AD	512	5

The annotation team consisted of native Czech speakers, primarily philology students (aged 18–30) with a basic linguistic background but no formal training in specific theoretical frameworks. This lack of prior exposure was intentional to ensure that text comprehension was not biased by theoretical presuppositions. While a core group of five annotators worked on Hlava AD, the Hlava Cor corpus was annotated by six individuals organized into two groups by three annotators. This larger group included the same five participants from the Hlava AD task plus one additional annotator.

The annotation process is designed not only to capture the final annotation decisions but also to elicit the annotators' interpretative considerations. For this reason, all annotations obligatorily include annotators' comments explaining their choices which are crucial for our investigation.

4. Hlava Cor: Human Language Variation in Coreference

The annotation task for Hlava Cor was defined as the identification of coreference, i.e. reference to the same extra-linguistic entity, concept, or situation.

4.1. Hlava Cor: Annotation Process

Annotators were instructed to read the column `Sentence` with a highlighted expression in question and the left-hand context in columns `Adjacent Context` and `Distant Context` (see Table 2)[2] and to find a potential antecedent in the preceding context if any. No potential antecedents have been pre-annotated.[3] Each annotation output contains (i) the antecedent expression if it appeared in the preceding context, (ii) a numeric value indicating the degree to which context was required for interpretation (0 = no distant context needed, 1 = distant context necessary, 2 = even wider context needed), and (iii) a free-text comment explaining the annotator's reasoning or uncertainty.

[2]The column `Distant Context` is not shown in the example for space reasons.

[3]In Table 2, two antecedents are highlighted in bold because the table shows an already completed annotation. The bold highlighting is used only to make the example easier to read in this paper; the annotators did not see these highlights during the annotation process.

The annotation guidelines specified detailed conventions for the form and syntactic scope of antecedents. Annotators were required to record the full nominal phrase as it appears in the text, excluding prepositions unless they constitute part of a named entity or are internal to the referring expression. Dependent modifiers were to be included, while relative clauses were to be excluded. Deviations from these rules had to be explicitly documented in the comment field. Further specifications addressed the treatment of references to clauses or larger textual segments, possessive and other adjectival expressions, coordinated and discontinuous antecedents. When no suitable antecedent was found, annotators entered NE ("no antecedent"), and in cases of complete indecision they recorded a question mark.

4.2. Hlava Cor: Annotation Example

An example of the annotated segment is presented in Table 2. For the expression *vozidlo* 'vehicle' in the column `Sentence`, the annotators were asked to identify an antecedent (if any) and record it in the `ANN_ante` column. In the `ANN_comment` column, the annotators were asked to justify their decision or provide any additional remarks.

This example illustrates the annotation by two annotators out of three who arrived at different conclusions. Annotator 1 (ANN1) selected *vůz* 'car' as the antecedent of *vozidlu* 'vehicle' in `Sentence`, arguing that the expression does not refer to the specifically described model driven by the journalist, but rather to the vehicle in general. Annotator 2 (ANN2), on the other hand, chose *vozu* 'car (genitive)' as the antecedent, noting that both expressions refer to the same entity.

For space reasons, the annotation of the third annotator is not shown; this annotator made the same decision as ANN1 but expressed more uncertainty in the comment. We also omit the annotation of the need for context, which was zero for all three annotators in this case.

4.3. Hlava Cor: Data Description and Statistics

The corpus consists of 1,024 cases, each marked by three annotators. The segments for annotation were selected according to several dimensions of data division. First, we took the same number of spoken and written text segments. Most of the text segments have been excerpted from the PDT-C 2.0, a few segments come from the Czech Academic Corpus 2.0 (CAC) (Hladká et al., 2008). For the written data, we additionally used examples from

Table 2: Example of coreference annotation by two annotators

Adjacent Context	Sentence	ANN1 ante	ANN1 comment	ANN2 ante	ANN2 comment
Firma Renault Česká republika nám k redakčnímu testu poskytla verzi s turbodieselovým motorem s označením RT. Sice jsme v úvodu uvedli, komu lze **vůz** [ANN1] především doporučit. Snad nám nebudou mít případní zájemci z kategorie 'mladá rodina s více dětmi' za zlé, když jim neprozradíme, kde vzít potřebných 969820 korun. V případě testovaného **vozu** [ANN2] dovybaveného dvěma střešními okny, autorádiem s ovládáním pod volantem a ve vlnové metalíze pak ještě přes 66 tisíc korun navrch.	Věnujeme se však raději samotnému **vozidlu**, neboť svezení s ním opravdu stojí za to.	vůz	Neodkazuje na detailně popsaný, konkrétní model vozu, s nímž zrovna jel novinář, ale obecně na vozidlo.	vozu	Oba výrazy označují stejnou skutečnost.

Adjacent Context: The company Renault Czech Republic provided us with a version featuring a turbo diesel engine, designated RT, for an editorial test. Although we mentioned at the beginning who this **car** [ANN1] can primarily be recommended to, we hope that potential buyers from the category of "young families with several children" will not hold it against us if we do not reveal where to get the required 969,820 crowns. In the case of the tested **car** [ANN2], additionally equipped with two sunroofs, a car radio with steering-wheel controls, and metallic paint, the price increases by more than 66,000 crowns.

Sentence: However, let us turn to the **vehicle** itself, since driving it is really worth the experience.
ANN1 ante: *vůz* 'car'
ANN1 comment: Does not refer to the specifically described model that the journalist was driving, but rather to the vehicle in general.
ANN2 ante: *vozu* 'car'(genitive)
ANN2 comment: Both expressions refer to the same entity.

the archive of iRozhlas.[4] Table 3 shows the data distribution from these resources.

Table 3: Structure of Hlava Cor: written vs. spoken modes

Mode	Source Corpus	Number of Cases
Spoken	PDT-C 2.0	505
	Czech Academic Corpus 2.0 (CAC)	7
	Total Spoken	512
Written	PDT-C 2.0	461
	iRozhlas	40
	Czech Academic Corpus 2.0 (CAC)	11
	Total Written	512
Grand Total		**1,024**

We also considered the referential status of nominal groups for all categories, and divided the segments according to the specific and generic reference of the nominal groups and pronouns in coreferential relations (see Table 4). The division of the specific/generic reference did not apply for the cases with the local pronominal adverb *tam* 'there'.

Table 4: Structure of Hlava Cor: specific vs. generic reference

Reference Category	Number of Cases
full NPs + to 'it' with specific reference	384
full NPs + to 'it' with generic reference	384
tam 'there' (division is not relevant)	256
Total Cases	**1,024**

Another dimension was the division of our data into coreference types where the anaphoric expression is a full nominal group (noun, noun + adjective, etc.), pronominal coreference with *to* 'it', or local coreference with the pronominal adverb *tam* 'there' (see Table 5).

Table 5: Structure of Hlava Cor: nominal and pronominal anaphors

Anaphor Type	Number of Cases
full NPs	512
pronominal coreference with to 'it'	256
local coreference with tam 'there'	256
Total Segments	**1,024**

Being interested in exploring multiple readings of coreference, we aimed to extract examples that exhibit ambiguous or otherwise vague instances of coreference, and at the same time to allow for comparison with cases where coreference appears to be easily identifiable. In an effort to assess such a distinction in potential examples automatically, we have first passed the examples to five different Transformer-based coreference resolution models.[5] All potential examples were pre-selected to be theoretically ambiguous, i.e., based on morphological properties, at least two possible antecedents were present within a five-sentence context, making

[4] the internet archive of the Czech news outlet iRozhlas, https://www.irozhlas.cz/zpravy-archiv

[5] We trained 14 models on CorefUD 1.2 (Popel et al., 2024) using the coreference resolution system CorPipe (Straka, 2024), the winner of the CRAC 2024 shared task (Novák et al., 2024), and selected 5 achieving the best results on the two Czech CorefUD datasets.

disagreement possible. Based on the output of the five models, the potential examples were divided into two groups: (i) examples where all five models agreed on the antecedent, and (ii) examples where at least two models disagreed, with preference for examples with larger number of disagreements. Final examples for annotation were selected from these two groups and included in Hlava Cor in a 1:2 ratio. The ratio was uneven intentionally, as we assumed that cases of models' disagreement would provide more informative material for analysis. Table 6 presents the data distribution according to the models' agreement and disagreement.

Table 6: Structure of Hlava Cor: model agreement vs. model disagreement

Coreference Status	Number of Cases
model agreement	352
model disagreement	672
Total Cases	**1,024**

Table 7 shows how various categories and dimensions in Tables 3 to 6 relate to each other. For example, the first line states that there are 128 cases with an NP type of the anaphor in spoken data with generic coreference (in 84 of them the coreference resolvers disagreed, in 44 cases they were in agreement). The features can be combined to larger groups of equal size, for example 'to' spoken (128 cases) vs. 'to' written (128 cases), or 'tam' (256 cases) vs. 'to' (256 cases), or spoken (512 cases) vs. written (512 cases).

Table 7: Numbers of cases for detailed combinations of features in Hlava Cor

Feature Combination	Models Disagreed	Models Agreed	Total
NP spoken GEN	84	44	128
NP spoken SPEC	84	44	128
NP written GEN	84	44	128
NP written SPEC	84	44	128
tam spoken	84	44	128
tam written	84	44	128
to spoken GEN	42	22	64
to spoken SPEC	42	22	64
to written GEN	42	22	64
to written SPEC	42	22	64
Total	**672**	**352**	**1,024**

4.4. Hlava Cor: Inter-Annotator Agreement

For each category described in Section 4.3 (and summarized in Tables 3, 4, 5, and 6) we calculate the inter-annotator agreement of our three human annotators for setting the coreferential antecedent.[6] Table 8 presents the IAA results for coreference annotation in Hlava Cor. Overall, the agreement of all three annotators reached 49%, while partial agreement (at least two out of three annotators) reached 83%. The average pairwise agreement for Hlava Cor reaches 60.4%.[7]

Table 8: Hlava Cor: Inter-Annotator Agreement Overview

Category	All 3	At Least 2
Text Mode (from Table 3)		
written data	45%	81%
spoken data	53%	86%
Referential Status (from Table 4)		
generic coreference (GEN)	45%	82%
specific coreference (SPEC)	48%	82%
Grammatical Form (from Table 5)		
full NPs	51%	86%
pronominal (to 'it')	38%	75%
local (tam 'there')	57%	88%
Model Status (from Table 6)		
model agreement	66%	91%
model disagreement	39%	79%
All data	**49%**	**83%**

The inter-annotator agreement results reflect the inherent complexity of coreference annotation, particularly in naturally occurring data and in cases involving vague or underspecified referential relations. When broken down by referential status, agreement for generic coreference (45%) and specific coreference (48%) is relatively similar, suggesting that both types of reference pose comparable challenges for human annotation. This contradicts our initial observations and is possibly related to the lexical composition of generic nominal phrases in Hlava Cor. Regarding grammatical form, the pronominal coreference with *to* 'it' showed the lowest (38%) agreement as compared to full nominal groups (51%) and the local adverbial *tam* 'there' (57%), confirming that non-nominal and underspecified *to* 'it' may be more ambiguous. A more detailed analysis of these patterns will be the subject of further research.

Another relevant observation supports our hypothesis that instances where the models dis-

[6] Exact match was required. In order to avoid typo-related errors, the annotators were asked to copy/paste the segments from the original text; also, the disagreements were manually checked by an expert to find and fix cases of insignificant omissions (for example, a comma at the end of the coppied text, a missing character etc.).

[7] This number is not included in Table 8.

agreed tend to be more difficult for human annotators as well. Human annotation on the model-disagreement subset reached only 39% inter-annotator agreement (IAA), whereas cases where the models agreed showed considerably higher consistency, with an IAA of 66%. Notably, an IAA of 66% is still relatively low, especially given that the models themselves were in agreement. This fact clearly deserves special attention and can be partially explained by the inherent implicitness and ambiguity of coreference relations in our data. Model agreement in such cases likely reflects a combination of chance and biases induced by the training data. Overall, these findings suggest that model disagreement may serve as an indicator of interpretative complexity in the data.

Taken from a slightly different perspective, we can also examine how many distinct annotation choices were made for a single case, see Table 9. The table presents the absolute numbers of unique coreference choices per context annotated by the three annotators. The specific reasons for inter-annotator disagreement are not analyzed in detail here; however, in general, they often stem from differences in the level of detail or interpretation of the referential scope of the entities.

Table 9: Hlava Cor: Distribution of Coreference Choices by Annotator Agreement (absolute numbers).

Choice Type	Agreement Level	Cases
One choice	Full (3/3)	503
Two different antecedent	Partial (2/3)	351
Three different antecedents	No agreement (1/3)	170
Total Annotated Choices		**1,024**

5. Hlava AD: Human Label Variation in Attribution and Discourse

The corpus contains 512 short Czech text segments with explicit inter-sentential discourse relations, multiply annotated by five annotators. All texts in Hlava AD come from PDT-C 2.0.

Similarly to Hlava Cor, we separately considered the written and spoken modalities.

Based on our previous findings of frequent cases of inter-annotator disagreement on discourse relations (Zikánová, 2024), we hypothesized that a higher measure of disagreement in inter-sentential relations might be related to the presence of attributive constructions (sentences with verbs of thinking or saying). The supposed reason is that it may be difficult for recipients to recognize whether the upcoming sentence is related to the governing clause (author's speech) or to the dependent direct or indirect speech.

Table 10: Structure of the Hlava AD (number of items)

	Spoken data	Written data	**Total**
Attribution	96	125	**221**
Other	173	118	**291**
Total	**269**	**243**	**512**

To analyze the influence of attributive constructions on interpretive variation, we divided the Hlava AD data into two subsets: one containing attributive constructions with verbs of thinking and saying (including both direct and indirect speech) and a control subset with non-attributive constructions. Table 10 summarizes the structure of Hlava AD concerning the spoken/written and attribution/non-attribution dimensions.

The text segments were chosen from the original source (PDT-C 2.0) based on the syntactic and discourse structure. We searched for explicit inter-sentential discourse relations with a complex sentence in the role of the left argument. This complex sentence contained at least one dependent clause and its governing verb was either a verb of thinking or saying (attributive constructions) or not (non-attributive constructions). In non-attributive items, we filtered out all the text segments containing verbs of thinking or saying in any other role.

5.1. Hlava AD: Annotation Process

The annotators were shown pairs of sentences which had been related by an explicit discourse relation in the previous annotation of PDT-C 2.0. The `Right sentence` column contains a discourse connective (marked green, see Figure 1). The `Left sentence` column consists of several clauses; in the attributive group of items, the left sentence contains a governing verb of saying / thinking, whereas in the control group, another semantic type of verb is used instead. For better understanding of the text, annotators were additionally given the three preceding sentences in the `Preceding context` column.

The annotators' initial task was to identify the clause in the left sentence, to which the discourse connective relates the right sentence (they searched for the so-called "target of the relation").

To technically simplify the task for the annotators, we pre-marked the finite verbs (or parts of verbal forms) in clauses in the `Left sentence` in red, see Figure 1. Thus, the verbs served as identifiers of the whole clauses; annotators entered the chosen verb into the corresponding column.

In case of coordination between two verbal forms (clauses), the entire construction could be annotated as the target of the relation. Coordination was

Figure 1: Annotators' prompt in the Hlava AD annotation task

Preceding context	Left sentence	Right sentence
Španělská sekce prestižní Asociace evropských novinářů pozvala již po šesté různorodé spektrum středoevropských politiků, publicistů a filosofů, aby početným posluchačům letního universitního semináře s názvem Střední Evropa mezi Bruselem a Moskvou přednesli své úvahy o tom, co se uprostřed kontinentu děje. Pozornost od počátku budila polská účast. Vítěz nedávných voleb Aleksander Kwasniewski a premiérka poražené vlády Hanna Suchocká, každý ze svého úhlu, ale stejně přesvědčivě popsali, co se stalo v Polsku.	Kwasniewski opakovaně **zdůraznil**, že z cesty zásadních proměn země **nelze** sejít: pravice a levice se budou **přít** se středem o tempo změn, **ale** o základní vzorec reforem **není** sporu.	Co **však** je vážné, je nevelký zájem veřejnosti o věci veřejné.

[Preceding context: The Spanish section of the prestigious Association of European Journalists has invited for the sixth time a diverse spectrum of Central European politicians, publicists and philosophers to present their reflections on what is happening in the middle of the continent to the numerous audience of the summer university seminar entitled Central Europe between Brussels and Moscow. The Polish participation attracted attention from the beginning. The winner of the recent elections, Aleksander Kwasniewski, and the Prime Minister of the defeated government, Hanna Suchocka, each from their own perspective, but equally convincingly described what happened in Poland.

Left sentence: Kwasniewski has repeatedly **emphasized** that the country **cannot** be diverted from the path of fundamental changes: the right and left will **argue** with the center about the pace of change, **but** there **is** no dispute about the basic formula of reforms.

Right sentence: What is serious, **however**, is the public's lack of interest in public affairs.]

identified by the pre-marked (blue) conjunctions or punctuation marks.

Thus, the annotators could mark red verbal forms or blue conjunctive phrases/marks as targets of the relation expressed by the green discourse connective. Additionally, they had a possibility to use an exclamation mark (!) to indicate that the target was in the preceding or far left context, and they used a question mark (?) to signal that they could not find any target in the given text (e.g., they did not understand the text segment as a whole).

The second annotation task was to mark to what degree the annotator needed the previous context (before the left sentence) to understand the discourse relation expressed by the discourse connective. There were three possible answers: 0 (I did not need the previous context); 1 (I needed the previous context to understand and it helped me); and 2 (I needed the previous context to understand, but it did not help me).

The third annotation task required annotators to explain their choice of the target of the relation by entering comments in a separate plain text field.

5.2. Hlava AD: Annotation Example

Figure 1 presents an example with an attributive construction in the left sentence (*Kwasniewski **emphasized** that...*). Annotators interpret this segment in two different ways. Some annotators identify the main clause with the governing verb *emphasized* as the target of the discourse relation expressed by the discourse connective *however*. Their comments explain that Kwasniewski emphasized political happenings. On the other hand, the annotators stress that what is truly important is what is going on among ordinary people. This group of annotators does not see the right sentence as a part of the Kwasniewski's speech.

Conversely, other annotators mark the clause *that the country **cannot** be diverted...* as the target. They claim in their comments that Kwasniewski knows about the upcoming changes and is concerned about the public's lack of interest in public affairs. Thus, they understand the right sentence as a part of the Kwasniewski's speech.

5.3. Hlava AD: Inter-Annotator Agreement

We calculate the inter-annotator agreement on setting the target of a discoure relation, expressed by the discourse connective in the right sentence. We can obtain theoretically 1-5 solutions from 5 annotators; nevertheless, some syntactic structures in the left sentences are simpler, they do not provide up to 5 potential targets of the discourse relation. On the other hand, the disagreement can be increased by the possibility to use the mark "?" (not understandable) and "!" (relation to a distant left

Table 11: IAA on the target of a discourse relation in Hlava AD (percentage points)

	Agreement (1 solution from 5 annotators)	Disagreement (2-5 solutions from 5 annotators)
In general	38%	62%
Simple structures (2 possible targets)	48%	52%
Complex structures (3-10 pos. targets)	27%	73%

context). The average pairwise agreement disregarding the syntactic complexity reached 64.9%. The results for the IAA regarding the syntactic complexity of the left sentence (and, subsequently, a different number of potential targets) is presented in Table 11.

A more detailed analysis of the IAA on the target of a discourse relation can be found in Zikánová et al. (2025), where the multiple annotation serves as a test dataset for the default golden data. According to the authors, the measure of the IAA in Hlava AD is strongly affected by the syntactic complexity of the left sentence; however, it is influenced neither by the presence of the attribution, nor by the text mode (spoken or written).

6. Observations across Hlava Cor and Hlava AD

Given that Hlava Cor covers a dataset twice the size of Hlava AD, and that Hlava AD was annotated by five annotators while Hlava Cor involved only three parallel annotations, the results are not fully comparable. However, it is possible to compare the average pairwise agreement on target identification, which neutralizes the difference in the number of annotators. For Hlava AD, the average pairwise agreement reaches 64.9%, while for Hlava Cor it is 60.4%.

In both corpora, annotators marked the need for context (see Sections 4.1 and 5.1). The analysis of the annotations reveals substantial variation across annotators, while the overall tendencies remain similar in both corpora. What we observe likely reflects, to some extent, the annotators' degree of confidence in their interpretation of the text, as well as individual differences in annotation style. Some annotators tend to consult a broader preceding context to verify their interpretation, while others consider their immediate reading sufficient without referring to additional context. To a certain degree, this also mirrors the annotator's thoroughness and attention to detail.

Particular attention should be paid to the annotators' comments, which, in our view, represent the most interesting part of our annotation. Comments were mandatory in both corpora, and they certainly deserve a separate study beyond the scope of this paper. They capture the annotators' doubts about their chosen decisions, indicate possible alternative interpretations, and often describe in detail the strategies underlying their choices. Especially intriguing is the overlap between the annotators' comments and their decisions in other tasks. In some cases, the antecedents were chosen identically (inter-annotator agreement), yet the comments still mentioned alternative options. Conversely, in certain cases of inter-annotator disagreement, the analysis shows that the annotators reasoned in a very similar way but ultimately arrived at different decisions.

7. Conclusions

In this paper, we presented and described two datasets created to study human label variation: Hlava Cor (Nedoluzhko et al., 2026), focused on coreference, and Hlava AD (Šárka Zikánová et al., 2024), focused on attribution and discourse relations. Hlava Cor explores coreference with respect to the reference status of anaphoric expressions (specific or generic), their form (pronouns, full noun phrases, and anaphoric adverbials), and the extent to which coreference resolution models agree on identifying these relations. Hlava AD pays special attention to attribution and non-attributive constructions. Both corpora include spoken and written data and consider both registers equally.

The introduced corpora offer space for a more detailed analysis of label variation in human annotation. Preliminary observations suggest individual differences between annotators and the presence of personal annotation styles. Of particular interest are the annotators' comments, which often explain annotation choices, reflect subjective judgments, and point to interpretative ambiguity. These and related aspects will be the focus of our future work based on the data obtained from the corpora.

Acknowledgements

The authors gratefully acknowledge support from the Grant Agency of Czech Republic, project 24-11132S, as well as the OP JAK project CZ.02.01.01/00/23_020/0008518 of the Ministry of Education, Youth and Sports of the Czech Republic.

The research reported in the present contribution has been using language resources developed, stored and distributed by the LINDAT/CLARIAH-CZ project of the Ministry of Education, Youth and Sports of the Czech Republic (LM2023062).

8. Bibliographical References

V. Basile, B. Plank, D. Hovy, P. Van Der Lee, L. Van Der Plas, and M. Poesio. 2021. We need to consider disagreement in evaluation. In *Proceedings of 1st Workshop on Benchmarking, ACL*, pages 15–21.

Peter Bourgonje and Manfred Stede. 2020. The Potsdam commentary corpus 2.2: Extending annotations for shallow discourse parsing. In *Proceedings of the 12th Language Resources and Evaluation Conference*, pages 1061–1066, Marseille, France. European Language Resources Association.

L. Crible, M. J. Cuenca, L. Degand, C. Fuentes, and J. Vandelanotte. 2019. Functions and translations of underspecified discourse markers in TED talks: A parallel corpus study on five languages. *Journal of Pragmatics*, 142:139–155.

Maria Josep Cuenca. 2017. Discourse markers in speech: Distinctive features and corpus annotation. *Dialogue & Discourse*.

Freya Hewett and Manfred Stede. 2025. Disagreements in analyses of rhetorical text structure: A new dataset and first analyses. In *Proceedings of the 19th Linguistic Annotation Workshop (LAW-XIX-2025)*, pages 35–47, Vienna, Austria. Association for Computational Linguistics.

Nan-Jiang Jiang, Chenhao Tan, and Marie-Catherine de Marneffe. 2023. Ecologically valid explanations for label variation in NLI. In *Findings of the Association for Computational Linguistics: EMNLP 2023*, pages 10622–10633, Singapore. Association for Computational Linguistics.

Lauren Levine and Amir Zeldes. 2025. Subjectivity in the annotation of bridging anaphora. In *Proceedings of the 19th Linguistic Annotation Workshop (LAW-XIX-2025)*, pages 48–59, Vienna, Austria. Association for Computational Linguistics.

M. Marchal, B. Hladká, V. Lojda, and J. Jírka. 2022. Establishing annotation quality in multi-label annotations. In *Proceedings of COLING*, pages 3659–3668.

Michal Novák, Barbora Dohnalová, Miloslav Konopik, Anna Nedoluzhko, Martin Popel, Ondrej Prazak, Jakub Sido, Milan Straka, Zdeněk Žabokrtský, and Daniel Zeman. 2024. Findings of the third shared task on multilingual coreference resolution. In *Proceedings of the Seventh Workshop on Computational Models of Reference, Anaphora and Coreference*, pages 78–96, Miami. Association for Computational Linguistics.

Maciej Ogrodniczuk, Katarzyna Glowinska, Maciej Kopec, Agata Savary, and Magda Zawisławska. 2013. Polish coreference corpus. In *Human Language Technology. Challenges for Computer Science and Linguistics - 6th Language and Technology Conference, LTC 2013, Poznan, Poland, December 7-9, 2013. Revised Selected Papers*, volume 9561 of *Lecture Notes in Computer Science*, pages 215–226. Springer.

Siyao Peng, Yang Janet Liu, and Amir Zeldes. 2022. GCDT: A Chinese RST treebank for multigenre and multilingual discourse parsing. In *Proceedings of the 2nd Conference of the Asia-Pacific Chapter of the Association for Computational Linguistics and the 12th International Joint Conference on Natural Language Processing (Volume 2: Short Papers)*, pages 382–391, Online only. Association for Computational Linguistics.

Barbara Plank. 2022. The "problem" of human label variation: On ground truth in data, modeling and evaluation. In *Proceedings of EMNLP*, pages 10671–10682.

Massimo Poesio. 2020. Ambiguity. In D. Gutzmann, L. Matthewson, C. Meier, H. Rullmann, and T. Zimmermann, editors, *The Wiley Blackwell Companion to Semantics*. Wiley.

Massimo Poesio, Jon Chamberlain, Silviu Paun, Juntao Yu, Alexandra Uma, and Udo Kruschwitz. 2019. A crowdsourced corpus of multiple judgments and disagreement on anaphoric interpretation. In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers)*, pages 1778–1789, Minneapolis, Minnesota. Association for Computational Linguistics.

Lucie Polakova, Jiří Mírovský, Šárka Zikánová, and Eva Hajicova. 2024. Developing a Rhetorical Structure Theory treebank for Czech. In *Proceedings of the 2024 Joint International Conference on Computational Linguistics, Language Resources and Evaluation (LREC-COLING 2024)*, pages 4802–4810, Torino, Italia. ELRA and ICCL.

Sameer Pradhan, Alessandro Moschitti, Nianwen Xue, Olga Uryupina, and Yuchen Zhang. 2012. CoNLL2012 shared task: Modeling multilingual unrestricted coreference in OntoNotes. In *Proceedings of the Sixteenth Conference on Computational Natural Language Learning (CoNLL 2012)*, Jeju, Korea.

Marta Recasens, Eduard Hovy, and M. Antònia Martí. 2011. Identity, non-identity, and near-

identity: Addressing the complexity of coreference. *Lingua*, 121:1138–1152.

Marta Recasens and M. Antònia Martí. 2010. AnCora-CO: Coreferentially annotated corpora for spanish and catalan. *Language Resources and Evaluation*, 44:315–345.

I. Rehbein, M. Reznicek, R. Schüller, H. Schüppert, and M. Stede. 2016. Annotating discourse relations in spoken language: A comparison of the PDTB and CCR frameworks. In *Proceedings of LREC'16*, pages 1039–1046, Portorož.

Merel Scholman, Tianai Dong, Frances Yung, and Vera Demberg. 2022. DiscoGeM: A crowdsourced corpus of genre-mixed implicit discourse relations. In *Proceedings of the Thirteenth Language Resources and Evaluation Conference*, pages 3281–3290, Marseille, France. European Language Resources Association.

Milan Straka. 2024. CorPipe at CRAC 2024: Predicting zero mentions from raw text. In *Proceedings of the Seventh Workshop on Computational Models of Reference, Anaphora and Coreference*, pages 97–106, Miami. Association for Computational Linguistics.

Olga Uryupina, Ron Artstein, Andrea Bristot, Federico Cavicchio, Fabio Delogu, Kenia J. Rodriguez, and Massimo Poesio. 2020. Annotating a broad range of anaphoric phenomena, in a variety of genres: the ARRAU corpus. *Natural Language Engineering*, 26:95–128.

Leon Weber-Genzel, Siyao Peng, Marie-Catherine De Marneffe, and Barbara Plank. 2024. VariErr NLI: Separating annotation error from human label variation. In *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 2256–2269, Bangkok, Thailand. Association for Computational Linguistics.

Ralph Weischedel, Eduard Hovy, Mitchell Marcus, Martha Palmer, Robert Belvin, Sameer Pradhan, Lance Ramshaw, and Nianwen Xue. 2011. Ontonotes: A large training corpus for enhanced processing. In *Handbook of Natural Language Processing and Machine Translation: DARPA Global Autonomous Language Exploitation*, pages 54–63. Springer-Verlag, New York.

Frances Yung, Merel Scholman, Sarka Zikanova, and Vera Demberg. 2024. DiscoGeM 2.0: A parallel corpus of English, German, French and Czech implicit discourse relations. In *Proceedings of the 2024 Joint International Conference on Computational Linguistics, Language Resources and Evaluation (LREC-COLING 2024)*, pages 4940–4956, Torino, Italia. ELRA and ICCL.

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

Šárka Zikánová. 2024. Text structure and its ambiguities: Corpus annotation as a helpful guide. In *Proceedings of the 24th Conference Information Technologies – Applications and Theory (ITAT 2024)*, pages 2–12, Košice, Slovakia. Univerzita Pavla Jozefa Šafárika v Košiciach, Košice, Slovakia, CEUR-WS.org.

Šárka Zikánová, Anna Nedoluzhko, Jiří Mírovský, and Eva Hajičová. 2025. Gold data and multiple understanding of discourse relations. In *28th International Conference on Text, Speech and Dialogue (Part II)*, volume 16030 of *Lecture Notes in Computer Science*, pages 250–262, Cham, Switzerland. Friedrich-Alexander-Universität Erlangen/Nürnberg, Springer.

Sandrine Zufferey and Ludivine Crible. 2015. Using a unified taxonomy to annotate discourse markers in speech and writing. In *Proceedings of the 11th Joint ACL - ISO Workshop on Interoperable Semantic Annotation*, London.

9. Language Resource References

Jan Hajič, Eduard Bejček, Alevtina Bémová, Eva Buráňová, Eva Fučíková, Eva Hajičová, Jiří Havelka, Jaroslava Hlaváčová, Petr Homola, Pavel Ircing, Jiří Kárník, Václava Kettnerová, Natalia Klyueva, Veronika Kolářová, Lucie Kučová, Markéta Lopatková, David Mareček, Marie Mikulová, Jiří Mírovský, Anna Nedoluzhko, Michal Novák, Petr Pajas, Jarmila Panevová, Nino Peterek, Lucie Poláková, Martin Popel, Jan Popelka, Jan Rompoltl, Magdaléna Rysová, Jiří Semecký, Petr Sgall, Johanka Spoustová, Milan Straka, Pavel Straňák, Pavlína Synková, Magda Ševčíková, Jana Šindlerová, Jan Štěpánek, Barbora Štěpánková, Josef Toman, Zdeňka Urešová, Barbora Vidová Hladká, Daniel Zeman, Šárka Zikánová, and Zdeněk Žabokrtský. 2024. Prague Dependency Treebank - Consolidated 2.0 (PDT-C 2.0).

Barbora Hladká, Jan Hajič, Jirka Hana, Jaroslava Hlaváčová, Jiří Mírovský, and Jan Raab. 2008. The czech academic corpus 2.0 guide. *The Prague Bulletin of Mathematical Linguistics*, 89:41–96.

Anna Nedoluzhko, Jiří Mírovský, Šárka Zikánová, Eva Hajičová, Bianca Chuffartová, Šárka

Dohnalová, Lucie Hartmanová, Eliška Nodlová, Dominik Teska, and Františka Zikánová. 2026. Human Label Variation in Coreference (Hlava Cor). Data/software, ÚFAL MFF UK, Prague, Czech Republic, LINDAT/CLARIAH-CZ: http://hdl.handle.net/11234/1-6131.

Martin Popel, Michal Novák, Zdeněk Žabokrtský, Daniel Zeman, Anna Nedoluzhko, Kutay Acar, David Bamman, Peter Bourgonje, Silvie Cinková, Hanne Eckhoff, Gülşen Cebiroğlu Eryiğit, Jan Hajič, Christian Hardmeier, Dag Haug, Tollef Jørgensen, Andre Kåsen, Pauline Krielke, Frédéric Landragin, Ekaterina Lapshinova-Koltunski, Petter Mæhlum, M. Antònia Martí, Marie Mikulová, Anders Nøklestad, Maciej Ogrodniczuk, Lilja Øvrelid, Tuğba Pamay Arslan, Marta Recasens, Per Erik Solberg, Manfred Stede, Milan Straka, Daniel Swanson, Svetlana Toldova, Noémi Vadász, Erik Velldal, Veronika Vincze, Amir Zeldes, and Voldemaras Žitkus. 2024. Coreference in universal dependencies 1.2 (CorefUD 1.2). LINDAT/CLARIAH-CZ digital library at the Institute of Formal and Applied Linguistics (ÚFAL).

Šárka Zikánová, Jiří Mírovský, Anna Nedoluzhko, Eva Hajičová, Šárka Dohnalová, Anna Kmječová, Eliška Nodlová, and Dominik Teska. 2024. Human Label Variation in Attribution and Discourse (Hlava AD). Data/software, ÚFAL MFF UK, Prague, Czech Republic, LINDAT/CLARIAH-CZ: http://hdl.handle.net/11234/1-5819.