SEEM-CZ: Annotation and Classification of Epistemic Markers in Czech

Barbora Štěpánková, Michal Novák, Tomáš Musil, Lucie Poláková
Charles University, Faculty of Mathematics and Physics
Institute of Formal and Applied Linguistics
Malostranské náměstí 25, Prague, Czechia
{stepankova, mnovak, musil, polakova}@ufal.mff.cuni.cz

Abstract

We present a project focused on linguistic description, annotation and automatic classification of the so-called epistemic markers in Czech. These expressions, such as *pravděpodobně* 'probably', *zřejmě* 'apparently' and *určitě* 'certainly', typically operate within the pragmatic domain of language. We introduce a dataset containing manual annotations of the 40 most frequent epistemic markers in Czech, totalling almost 4,000 uses. This annotation was created using parallel InterCorp data (in Czech and English) and the TEITOK tool. We describe the annotation scheme used, the annotation process and data handling. The dataset forms the core of the emerging lexical database of these expressions (SEEMLex). Thanks to the comprehensive manual annotation, the dataset can also serve as a source of further pragmatic information and can be used as a basis for further linguistic research. The proposed annotation scheme can also be used for other languages. To demonstrate the dataset's utility for automatic classification, we trained XLM-RoBERTa classifiers using 10-fold cross-validation, achieving 72.6% accuracy for type of use classification (6 classes) and 54.2% accuracy for degree of certainty classification (4 classes).

Keywords: epistemic markers, annotation, classification, communicative functions, Czech

1. Introduction

Epistemic modality, as one of the three basic types of modality (Palmer, 2001),[1] is a very important linguistic category that encodes people's opinions, assumptions, and beliefs about the content of a statement, specifically by expressing the degree of (un)certainty about a proposition. As such, language devices expressing epistemic modality have vital influence on meaning and (perceived) factuality. In addition to grammatical means such as the future tense and intonation, epistemic modality is typically expressed by various lexical items, including modal auxiliary verbs (*must*, *can*), mental state predicates (*think*) and adverbs (*certainly*, *probably*). The latter, henceforth referred to as "epistemic markers" (EMs),[2] are the focus of our research.

Analyzing the use of epistemic markers can help uncover the illocutions, communication strategies, or ideology of the producer, such as hedging, politeness, persuasion, assurance, etc. which is particularly important for tasks requiring deep understanding of the pragmatic dimension of language.

The aim of this paper is threefold: first, we introduce an annotation project focused on epistemic markers in Czech, with focus on the detailed, newly developed description framework/annotation scheme that can possibly be used across languages. Second, based on the annotated data, we conduct several classification experiments, e.g. we predict the `type of use`, as expressions that typically function as EMs display a large polyfunctionality (see Sec. 3.1.1), and the `degree of certainty`. Third, the annotated data serves as an underlying resource for developing the Czech epistemic markers lexicon SEEMLex. Whereas the annotated dataset, along with the data processing and experiment scripts, is already publicly available in the SEEM-CZ project repository,[3] the lexicon will be published soon.[4]

The paper is organized as follows: After a brief overview of the various approaches to the annotation of pragmatic phenomena, with a particular focus on epistemic modality (Sec. 2), we introduce the SEEM-CZ project (Sec. 3). In Section 4, we describe the data and methodology used in our approach. Section 5 present the results of the annotation process in terms of inter-annotator agreement. Classification experiments based on the annotated data are presented in Section 6. In Section 7, we summarize the results and outline future research steps.

[1] The number of types of modality varies among different authors. We base our classification on Palmer's later division into epistemic, deontic, and dynamic modality.

[2] We consider epistemic markers to be a subtype of Pragmatic markers (cf. Fraser 1996), which broadly correspond to the term Particle used in the Slavic linguistic tradition (cf. Rozumko 2016).

[3] https://github.com/ufal/SEEM-CZ

[4] http://hdl.handle.net/11234/1-6113

Proceedings of the Fifteenth Language Resources and Evaluation Conference (LREC 2026), pages 6856–6869
11–16 May 2026.

2. State of the Art

Despite the fundamental shift towards LLMs, creating annotated datasets continues to play an important role in linguistic research (Opitz et al., 2025). A number of projects are emerging in the field of pragmatics, where many phenomena are still not satisfactorily described; for an overview of the most recent ones, see Ma et al. (2025), for earlier approaches, see e.g. Archer et al. (2008). In the area of lexicography "of function words", traditional dictionaries (cf. Grochowski et al. (2014) for Polish, Helbig (1993) for German) have been followed by attempts to create digital databases, cf. Wiemer and Stathi (2010) or Pietrandrea (2018).

Specifically, research in the field of epistemic modality focuses mainly on recognising lexical or other devices and their annotation in text (Rubin, 2007), on individual markers, their functions and their mutual comparison (Suzuki 2018, Rozumko 2022), as well as the specifics of their use in various genres, registers (cf. González et al. (2017) on a comparison of oral and written discourse in Catalan). However, modality research is also employed for various fundamental linguistic and multidisciplinary tasks: Understanding the linguistic nature of modals helps disentangle the mechanisms of disinformation spread (Abbas, 2022), EMs are to be considered in NLP tasks such as classification and automatic extraction of sentiments, attitudes or negation (Koivunen et al. 2021, Liu et al. 2013, Carrillo-de Albornoz and Plaza 2013, Morante and Sporleder 2012), and they also play a huge role in second language acquisition (Aijmer 2002, Tamaji 2009).

3. SEEM-CZ Project

The SEEM-CZ research project aims at investigation of epistemic markers in Czech based on parallel corpus data and its detailed manual annotation. The main focus is a complex linguistic description of the characteristic features and functions of EMs for the development of the online lexicon of Czech epistemic markers SEEMLex (Štěpánková et al., 2024), as well as for other purposes. This emerging lexicon is the first of its kind in the Czech context. Its uniqueness lies in its systematic account of a pragmatic phenomenon – in other words, the lexicographical work with non-autosemantic expressions.

Regarding the breadth of the epistemic marker category, we follow Czech grammar books, e.g. Komárek et al. (1986), and, in addition to expressions that typically express **epistemic modality**, such as *určitě* 'certainly', *snad* 'perhaps', we also include words expressing **evidentiality**[5] (e.g. *očividně* 'obviously') and expressions confirming/emphasizing the speaker's strong belief in their being right, so called **confirmatory markers** (*opravdu* 'indeed').

3.1. Annotation Scheme

The design of the annotation scheme is one of the key components of this project. The set of annotated features was compiled based on our pilot studies (Šindlerová et al., 2023, Štěpánková et al., 2023) and on similar previous research (e.g. Pietrandrea, 2018, Wiemer and Stathi, 2010). The annotation captures two fundamental features that define the use of a given expression: the `type of use` and the `communicative function`. A third feature, which is somewhat traditional, is the `degree of certainty`. Furthermore, together with the epistemic marker, we annotated contextual linguistic phenomena that can influence the overall interpretation of the meaning of an individual use (for details see Sec. 3.1.4).

In the project, we also adopt the proposal of Aijmer et al. (2006) that parallel corpora can be valuable resources of information on the meaning and functions of pragmatic markers, because translations reflect their contextual dimensions and other broad factors of meaning correspondence. We therefore used a parallel corpus as our data source, and `translation equivalents` were one of the annotated features.[6] The list of annotated categories, including their values, is presented in Table 1; see also the table of annotated examples in the Appendix A.

3.1.1. Type of Use

Expressions that function as epistemic markers are often subject to processes referred to as grammaticalization[7] (cf. Traugott, 1995). In addition to their epistemic function, they are also used in other functions (cf. Carretero and Zamorano-Mansilla, 2019, Rozumko, 2022, Poláková et al., 2025). As the border between these functions can be fluid and sometimes unclear (cf. Diewald, 2013), we use the term 'epistemic markers' for all uses. Let us exemplify the variety of functions on the expression *určitě*. Czech epistemic markers often originate

[5] The transition between the categories of epistemicity and evidentiality is a much-discussed topic (cf. Nuyts, 1999, De Haan, 2001). In Czech, there is significant overlap between expressions belonging to these two categories, which is why we examine them together in our research.

[6] Thanks to the TEITOK tool mentioned below, we were able to use Czech and English parallel texts simultaneously.

[7] Following Diewald (2011) and Degand and Evers-Vermeul (2015), we use the term as an umbrella term for other expressions used with similar meanings, such as pragmaticalization and (inter)subjectivization.

Annotated feature	Values
Type of use	epistemic (evidential, confirmatory), response, autosemantic, other
Degree of certainty	high, higher medium, medium, low
Type of comm. function	assertive, directive, interrogative, commissive, (dis)approval, expressive
Specific comm. function	e.g. assumption, recommendation, wish
Scope	clause, member, ellipsis
Predicate verb	lemma, morphological tag
Position in a sentence	first, last, other
Negation	Y/N
In a contrastive pattern	Y/N
Other modal expression	e.g. intensifier, modal marker, modal verb
Type of evidence	sensory, hearsay, reasoning, inference
Translation equivalent	choice from the parallel English sentence

Table 1: Annotated categories in SEEM-CZ

from adverbs (see Example 1). Apart from their primary role as epistemic markers (see Example 2), these expressions can also perform further pragmatic, discourse-related functions, see Example 3 demonstrating a response function. Example 4 then illustrates the fully blended, idiomatic use of the expression *určitě* 'certainly'. Therefore, our primary focus was on determining whether the examined expression exhibits an epistemic[8] or other function.

(1) *Petr se vyjadřuje až příliš **určitě** a důrazně*
'Peter is too **outspoken** and emphatic' [Lit: Petr speaks too much **confidently** and emphatically]

(2) *Jeho drápy vydávaly na dřevě slabý dutý zvuk, takže pod ním **určitě** byla voda.*
'His claws made a faint hollow sound on the wood, so there **must** be water underneath.'

(3) *„I dnes odpoledne to budou hlásit?" „**Určitě**."*
"Will there be an announcement this afternoon as well?' '**Sure**."

(4) *„Já vám ho seženu." „Platí. Ale **určitě**!"* "I'll get you one.' Right. **See that you do.**" [Lit: 'But **surely**!']

3.1.2. Degree of Certainty

As mentioned above, EMs are characterized by the degree of certainty they express, which is usually described using a scale ranging from the highest certainty (paraphrased as *it is certain that x*) to the lowest degree of certainty, doubt (paraphrased *I doubt that x*), with medium degrees covering a broad area from probability to possibility. The number of degrees of certainty varies in different approaches (cf. Hoye, 2014). Based on our pilot studies, we used a 4-point scale, see Table 2.

high certainty	higher medium certainty	medium certainty	low certainty / doubt
určitě 'certainly'	*nejspíš* 'probably'	*možná* 'maybe'	*těžko* 'hardly'

Table 2: Degrees of certainty

3.1.3. Communicative Functions

Based on pilot studies, essential information regarding the use of epistemic markers was considered to be the speaker's intention. We annotated this information using two features: `communicative function` (six possible values: *assertive*, *directive/contact*, *interrogative*, *commissive*, *(dis)approval*, *expressive*)[9] and a more fine-grained `specific communicative function` (open-ended category, with possible values such as 'statement', 'promise', 'assurance').

According to Grepl (2017), epistemic modality is an integral part of assertive communicative function such as 'announcement', 'statement', 'assertion'. In these cases, the degree of certainty conveyed by the marker corresponds to its position on the certainty scale (see Example 2 above). However, when a marker is used in a different communicative function, the degree of certainty may change and its epistemicity may weaken (cf. Míguez, 2022, Štěpánková et al., 2023). As an example, let us mention the so-called *directive assumption*, i.e. a situation in which the speaker firmly assumes a particular (agreeing or disagreeing) stance of their communication partner, but the acknowledgement or denial from the partner is not expected at all (see Example 5).

[8] Annotators could optionally annotate whether it is an epistemic, evidential or confirmatory subtype.

[9] Communicative function (Grepl, 2017), a term inspired by the speech act theory, is defined as a meaning of an utterance resulting from the intention with which the utterance is produced by the speaker towards the addressee in that particular communicative situation.

(5) *A koho tady ještě máme - s Arturem Weasleym se **určitě** znáte, že?* 'And let's see who else - you know Arthur Weasley, I **daresay**?'

3.1.4. Contextual Features

While the previous three features focus on the semantic-pragmatic properties and functions of EMs and their annotation require an overall understanding of the sentence in context, the following features are explicitly present in the text. As mentioned above, these phenomena contribute to signaling the communicative function of the sentence and help interpret the use of the EM. Some relate to the examined marker itself (e.g. `position in the sentence, scope`), others identify and describe phenomena in its close context (e.g. `predicate verb` in the sentence, `evidentiality, negation` and `contrast`).

Values of these features are annotated in the following ways: selection from a list of options (e.g., type of evidence, presence of negation); marking one or more words directly in the text (e.g., predicate, other modal expression, translation equivalent, see Fig. 1); automatic tag-based annotation (e.g. grammatical features of the predicate), cf. Sec. 4.

Note that, due to the project's focus, only markers in epistemic use were annotated for these features. Also note that not all features have to be present in every sentence, so not all features have to be annotated.

4. Data

The first step in creating the dataset involved manually compiling a list of expressions classified as EMs documented in Czech grammars, monolingual dictionaries and the PDT-C corpus (Hajič et al., 2020). The list contained 140 items in total. From this list, we have selected the 40 most frequent expressions according to their frequency in the InterCorp corpus (Čermák and Rosen, 2012).

The Czech and English subparts of the parallel InterCorp corpus were then chosen as the main data source, specifically the *Core* part, which mainly consists of fiction (421 books). These fictional texts are presumed to be close to spoken language (e.g. they have a high frequency of EMs, a certain degree of subjectivity), and at the same time they also provide sufficient situational context. Additionally, translations of fiction tend to be of higher quality because publishers carry out proofreading and editing of manuscripts.

As for the corpus in general, in addition to its parallel nature, its advantages also include detailed morphological tagging and the presence of sentence alignment. Furthermore, we provided the parallel texts with automatic word alignment using AWESOME aligner (Dou and Neubig, 2021).

For each marker, we compiled a list of all its occurrences in the corpus using simple lemma-based queries. The only exception was the word *fakt* 'really', for which we applied part-of-speech filtering to exclude its nominal use (meaning 'piece of information'). For a distribution of selected EMs across books in the corpus, see Appendix B.

For manual annotation, we selected a random sample of 100 sentences for each of the 40 selected markers: 50 of these examples were from original Czech texts provided with an English translation (Rosen et al.), the other 50 were originally written in English and translated into Czech (Klégr et al.). The total number of annotated uses was 3,968, because two markers did not reach 50 occurrences in the data,[10] covering 74 originally Czech and 256 originally English books.

The search results were divided into 80 files, with each file containing 50 corpus contexts of a given expression from either the Czech or English part of the corpus. In other words, the data was not mixed in terms of individual markers or source language. One goal is to build a lexicon capturing each word's typical behavior and this setting enables to observe the differences.

4.1. Tool

For annotation we used the TEITOK web-based platform (Janssen, 2016), which was modified for working with a parallel corpus for the purposes of our project. For each sampled marker, the annotators can see the sentence in which it appears and can explore an arbitrarily large context of the sentence. This is essential for the annotation of modality. Moreover, the corresponding English sentence and its context can be displayed (see Figure 1).

4.2. Annotation

As mentioned above, a new annotation scheme was developed for the project's purposes. Several rounds of pilot annotations (both individual and parallel) were conducted by a team of three experienced linguists during its creation. The annotation process itself consists of two stages: (i) In the first stage, translation equivalents were annotated; (ii) in the second stage, comprehensive annotation was performed, i.e. detailed annotation of all features in the Czech part.

(i) Although sentence and word alignment was applied in the texts, it was necessary to perform **translation equivalents manual post-editing**.

[10] The Czech data contained only 29 occurrences of *evidentně* and 39 of *podle všeho*.

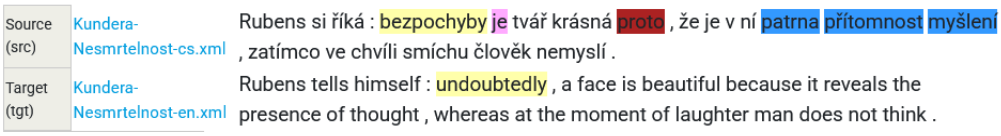

Figure 1: Illustration of text annotation in the TEITOK annotation tool. The individual categories are indicated by the following colours: yellow - `translation equivalents`; pink - `predicate verb`; brown: `other modal expression`; blue: `type of evidence` (in this case, reasoning).

Feature	All markers		Epistemic only	
	Simple (%)	α	Simple (%)	α
Type of use	83.6	*0.673*	N/A	N/A
Degree of certainty	57.6	0.540	52.4	0.533
Type of CF	81.0	*0.681*	84.4	0.391
Scope	80.0	0.638	90.9	*0.669*
Position in a sent.	78.7	*0.668*	88.8	*0.760*
Negation	92.8	**0.844**	96.1	**0.837**
In a contrast. pattern	74.4	0.518	81.5	0.209

Table 3: Inter-annotator agreement on EM attributes. Left: all parallel annotations; right: pairs with epistemic-use agreement. Reported are simple agreement (%) and Krippendorff's α (chance-corrected). $\alpha \geq 0.8$ in **bold**, $\alpha \geq 0.667$ in *italics*, marking definitive and tentative reliability thresholds (Krippendorff, 2013).

The translations of the word *určitě* (see Examples 1-5) demonstrate the variety of translation possibilities. In particular Examples 4 and 5 highlight the challenges of automatic alignment. The idioms in Example 4 are not fully established as translation equivalents, and example 5 shows that the word *daresay* is occasionally translated as a combination of the epistemic marker *určitě* 'certainly' and the interrogative particle *že* 'right'. Example 6 demonstrates that epistemic modality can be expressed differently across languages. In the Czech sentence, for instance, it is expressed twice (by the epistemic marker *asi* 'probably' and by the modal verb *museli* 'must', whereas in the English sentence there is only the modal verb *must* and no corresponding epistemic marker.

(6) ***museli*** *vodněkad* ***asi*** *dovízt ňáký herce*
'they **must** have brought in actors' [Lit: they-must-have from-somewhere **probably** to-import some actors]

Ten philology students participated in these annotations. The translation equivalents annotation was performed independently from the comprehensive annotation (ii), so that the translations would not influence the annotators' interpretation of a given marker's functions.

(ii) The **comprehensive annotation** covers all other features described in Sec. 3.1 and in Table 1. The annotators followed the rules of the annotation manual (which was created based on pilot studies and supplemented with examples).[11] In addition, they had the opportunity to use the comments section to justify their decisions or provide additional information, e.g., regarding alternative annotations of certain features. The annotation was performed by five experienced linguists and eight trained students.

During the annotation process, it was ensured that the same annotator would not annotate files containing both Czech and English examples of the same word. In other words, each word was annotated by at least two annotators, albeit on different examples. Additionally, some files underwent parallel annotation (see Sec. 5).

Once the annotations had been completed, the following final stages were carried out:

- A compiled file was created based on the parallelly annotated files. (Two experienced linguists always created the final, third version.)
- Semi-automatic checks were performed. (A set of checks was created to detect annotation omissions; e.g., mandatory fields for a given use were not filled in. Errors found were corrected manually or automatically).

5. Inter-Annotator Agreement

To estimate the reliability of the proposed annotation scheme, we collected multiple annotations of the same set of occurrences and measured the

[11] The manual will be published as a technical report.

extent of agreement among annotators across various features.

In total, 690 occurrences of eight selected epistemic markers were annotated multiple times. Of these, 86% were annotated by two annotators and 14% by three annotators. Altogether, ten annotators participated in the analysis of inter-annotator agreement (IAA).

We computed both *simple agreement*, which measures, for a given feature, the proportion of annotation pairs in which the annotators agreed.[12] To account for chance agreement, we also calculated IAA using *Krippendorff's* α (Krippendorff, 2013). We decided for this metric because it accommodates more than two annotators and naturally handles missing data. Furthermore, for the `degree of certainty` feature, it is important that α can weight disagreements in ordinal variables according to the distance between their values. Antoine et al. (2014) also demonstrate that this coefficient is far less sensitive to an inappropriate choice of granularity level.

5.1. Results and Discussion

Table 3 presents both simple agreement and Krippendorff's α for most of the features listed in Table 1. Since the epistemic use of markers is central to our analysis, we also report IAA calculated only on pairs where both annotators agreed on the epistemic use.

As expected, the highest agreement was observed for the feature `negation`, which is explicitly either present or absent. We also found good agreement ($\alpha \geq 0.667$) for three other features, including `type of use`.[13]

In contrast, the lowest agreement across all EMs was observed for two features: `degree of certainty` and `in a contrastive pattern`. Regarding the former, as discussed in Section 3.1.3, the `degree of certainty` may diverge from the original lexical meaning of a marker when the marker is used with a communicative function other than *assertive*. Analysis of parallel annotations shows that, in such cases, individual annotator approaches are evident: some focus more on the original meaning of the expression, while others give greater weight to communicative functions and context. For example, both annotators labeled Example 7 as *expressive*, but one assigned it the value *high certainty*, while the other chose *higher medium certainty*.

(7) *Ty ses* ***určitě*** *zbláznila!*
'You've gone mad' [Lit: you've **definitely** gone mad]

Nevertheless, identifying a specific communication function within the non-assertive group seems to be a much harder task than distinguishing between the assertive and non-assertive communicative function. Since the assertive function is indeed prevailing for the epistemic use, it results in high simple agreement for markers in epistemic use, whereas the corresponding value of Krippendorff's α, which compensates for chance agreement, remains much lower (see the *Epistemic only* column in Table 3).

The low agreement for the feature `in a contrastive pattern` is likely due to an insufficient description in the annotation manual. When focusing only on the epistemic use, the simple agreement for this feature increases, but α decreases even further. This disproportion again reflects a skewed distribution of feature values, with the 'No contrast' category being strongly dominant.

6. Experiments

We conducted experiments to automatically classify epistemic markers in Czech using transformer-based models. We trained separate classifiers for two related tasks: (1) identifying the `type of use`, and (2) determining the `degree of certainty` for markers in epistemic uses.

Data and Tasks We trained classifiers on the complete annotated dataset (see Section 4), which comprises 3,968 instances across 40 different EMs. The classification tasks were defined as follows:

- **`Type of use` classification (6 classes):** Classify all annotated instances into one of six[14] categories: *epistemic*, *evidential*, *confirmatory*, *response*, *autosemantic*, and *other* (other pragmatic or discourse-related uses).

- **`Degree of certainty` classification (4 classes):** For instances annotated as epistemic use, classify their degree of certainty into four levels: *high*, *higher medium*, *medium*, and *low certainty/doubt*. The *no* certainty class was excluded due to insufficient data (only 3 annotated examples). This task uses 1,320 instances.

[12] Note that there is one annotation pair for occurrences labeled by two annotators and three pairs for those labeled by three annotators.

[13] For measuring IAA on the `type of use` feature, we do not distinguish among the subtypes of the epistemic category (see Section 3.1.1). If we did, Krippendorff's α would decrease to 0.618.

[14] Unlike in the IAA analysis, here we distinguish between subtypes of the epistemic use.

Task	Classes	Accuracy	F1 (weighted)	Samples
Type of Use	6	0.726 ± 0.020	0.727 ± 0.021	3,968
Degree of Certainty	4	0.542 ± 0.068	0.540 ± 0.067	1,320

Table 4: Results for epistemic markers classification using XLM-RoBERTa-base on Czech data. Both models were trained on the complete dataset (Czech and English source sentences) using 10-fold cross-validation. Values shown as mean ± standard deviation across folds.

Class	Precision	Recall	F1-score	Support
response	0.67	0.84	0.75	139
epistemic	0.80	0.69	0.74	1,323
confirmatory	0.61	0.74	0.67	586
autosemantic	0.77	0.71	0.74	841
evidential	0.77	0.85	0.80	633
other	0.61	0.65	0.63	446
Weighted avg	**0.73**	**0.73**	**0.73**	**3,968**

Table 5: Per-class results for `type of use` classification (6 classes). Results aggregated across all 10 folds (3,968 total predictions). Support indicates the number of instances for each class.

Model and Training We used XLM-RoBERTa-base (Conneau et al., 2020), a multilingual transformer model pretrained on 100 languages including Czech. The model processes the contextual window (50 words on each side of the target expression) and produces a classification for the marker.

Training was performed on an NVIDIA A100 GPU with the following configuration: batch size of 8 with gradient accumulation over 8 steps (effective batch size 64), learning rate of 2×10^{-5} with 10% warmup, and mixed-precision training (bfloat16). We applied class-balanced weighting in the loss function to handle class imbalance. Early stopping with patience of 7 epochs was used to prevent overfitting, monitoring weighted F1 score on the validation set.

For `type of use` classification, we trained for up to 30 epochs and evaluated every epoch. For `degree of certainty` classification, given the smaller dataset size, we trained for up to 40 epochs with a lower learning rate (1×10^{-5}) and smaller batch size (4) for more stable learning.

We used 10-fold stratified cross-validation to obtain robust performance estimates. In each fold, data was split into training (90%) and validation (10%) sets while maintaining class distributions.

Results Table 4 presents the overall cross-validation results for both classification tasks. The `type of use` classifier achieved 72.6% ± 2.0% accuracy (mean ± standard deviation across 10 folds) with an F1 score of 0.727 ± 0.021, demonstrating strong and consistent performance across all six categories. The `degree of certainty` classifier achieved 54.2% ± 6.8% accuracy with an F1 score of 0.540 ± 0.067, showing greater variance due to the smaller dataset and class imbalance.

Table 5 shows per-class performance for `type of use` classification, aggregated across all folds. The model performed particularly well on *evidential* (F1=0.80), *response* (F1=0.75), and *epistemic*/*autosemantic* (both F1=0.74). The model also achieved good performance on *confirmatory* uses (F1=0.67). The most challenging class was *other* (F1=0.63), likely due to its heterogeneous nature encompassing various pragmatic functions. The low standard deviation (2.0%) indicates high consistency across folds.

Table 6 presents results for `degree of certainty` classification. The model achieved moderate performance on *high certainty* (F1=0.66) and *medium certainty* (F1=0.50), but struggled significantly with the rare *low certainty / doubt* class (F1=0.16), which had only 45 instances total across all folds. The higher variance (6.8%) reflects the challenging nature of this task with limited training data.

7. Conclusion

We presented a manually annotated dataset of **epistemic markers** in Czech, which is based on the parallel InterCorp corpus. Thanks to the complex manual annotation, including annotations of English translation equivalents, the dataset captures the polyfunctionality of these expressions and their use in various pragmatic and communicative functions. The dataset represents a vital underlying resource for the development of the lexicon of epistemic markers in Czech (currently in progress). A further valuable outcome is the newly developed annotation scheme that can potentially also be used for annotating EMs in other languages.

Part of the data was annotated in parallel, and

Class	Precision	Recall	F1-score	Support
high certainty	0.68	0.64	0.66	617
higher medium certainty	0.42	0.39	0.40	306
medium certainty	0.46	0.55	0.50	352
low certainty / doubt	0.20	0.13	0.16	45
Weighted avg	**0.54**	**0.54**	**0.54**	**1,320**

Table 6: Per-class results for `degree of certainty` classification (4 classes). Results aggregated across all 10 folds (1,320 total predictions). Support indicates the number of instances for each class. The *no* certainty class was excluded due to insufficient data (3 examples).

inter-annotator agreement was calculated for several attributes. The results confirm the importance of adopting a complex pragmatic approach to the analysis of these expressions. On the one hand, they show that the annotators achieved satisfactory agreement on the `type of use`; on the other, they highlight the problematic and subjective nature of the `degree of certainty` category.

We presented automatic classification experiments for Czech epistemic markers using XLM-RoBERTa-base trained and evaluated on the annotated data in a cross-validation fashion. The results demonstrate that `type of use` classification achieves strong performance (72.6% ± 2.0% accuracy across 6 classes), with particularly good results for evidential and epistemic uses. `Degree of certainty` classification proved more challenging (54.2% ± 6.8% accuracy across 4 classes), primarily due to limited training data and extreme class imbalance, particularly for rare certainty levels. These results establish a baseline for automatic epistemic marker classification in Czech and highlight the need for additional annotated data to improve fine-grained certainty distinctions.

The annotated dataset and the scripts to reproduce the experiments are available in the SEEM-CZ project repository.

8. Limitations and Ethical Considerations

A possible **limitation** of the project is that we only used fiction texts, meaning the results may differ for other genres. We plan to test other genres in the future. Furthermore, the research was conducted in one language only, with the second language serving mainly for lexical purposes, such as determining the semantic field. However, pilot studies comparing English and Swedish were conducted, which provided an indication of the relevant categories and the broader applicability of the approach.

Although the IAA analysis involved most of the annotators, it covered only 8 out of 40 expressions. Reliability of IAA analysis would be higher if the examples selected for parallel annotation were distributed across a broader range of expressions. IAA analysis indicates that some annotated features (e.g., `in a contrastive pattern`) would have benefited from clearer guidelines.

Ethical considerations The annotators were duly employed and paid. Copyright regarding the resources was handled in accordance with applicable legislation.

9. Acknowledgements

The research has been supported by the Czech Science Foundation under the project GA23-05240S, by Ministry of Education, Youth and Sports of the Czech Republic under the LINDAT/CLARIAH-CZ project (LM2023062), and by Charles University Research Centre program No. 24/SSH/009. We gratefully thank Maarten Janssen for adjustments of his TEITOK tool, the rest of the team for their collaboration, and three anonymous reviewers for their comments.

10. Bibliographical References

Ali Haif Abbas. 2022. Politicizing COVID-19 vaccines in the press: A critical discourse analysis. *International Journal for the Semiotics of Law - Revue internationale de Sémiotique juridique*, 35(3):1167–1185.

Karin Aijmer. 2002. *Modality in advanced Swedish learners' written interlanguage*, pages 55–76. John Benjamins Publishing Company.

Karin Aijmer, Ad Foolen, and Anne-Marie Simon-Vandenbergen. 2006. Pragmatic markers in translation: a methodological proposal. In *Approaches to Discourse Particles*, pages 101–114. Brill.

Jean-Yves Antoine, Jeanne Villaneau, and Anaïs Lefeuvre. 2014. Weighted Krippendorff's alpha is a more reliable metrics for multi-coders ordinal annotations: Experimental studies on emotion,

opinion and coreference annotation. In *Proceedings of the 14th Conference of the European Chapter of the Association for Computational Linguistics*, pages 550–559, Gothenburg, Sweden. Association for Computational Linguistics.

Dawn Archer, Jonathan Culpeper, and Matthew Davies. 2008. Pragmatic annotation. In Anke Lüdeling and Merja Kytö, editors, *Corpus Linguistics: An International Handbook*, pages 613–641. Mouton de Gruyter.

Marta Carretero and Juan Rafael Zamorano-Mansilla. 2019. Disentangling epistemic modality, neighbouring categories and pragmatic uses: the case of english epistemic modal adverbs. *InQuinze études de cas sur les modalités linguistiques/Fifteen case studies on types of linguistic modalities*, pages 131–157.

Jorge Carrillo-de Albornoz and Laura Plaza. 2013. An emotion-based model of negation, intensifiers, and modality for polarity and intensity classification. *Journal of the American Society for Information Science and Technology*, 64(8):1618–1633.

Alexis Conneau, Kartikay Khandelwal, Naman Goyal, Vishrav Chaudhary, Guillaume Wenzek, Francisco Guzmán, Edouard Grave, Myle Ott, Luke Zettlemoyer, and Veselin Stoyanov. 2020. Unsupervised cross-lingual representation learning at scale. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 8440–8451, Online. Association for Computational Linguistics.

František Čermák and Alexandr Rosen. 2012. The case of InterCorp, a multilingual parallel corpus. *International Journal of Corpus Linguistics*, 13(3):411–427.

Ferdinand De Haan. 2001. The Relation between Modality and Evidentiality. *Linguistische Berichte*, 9:201–216.

Liesbeth Degand and Jacqueline Evers-Vermeul. 2015. Grammaticalization or pragmaticalization of discourse markers? more than a terminological issue. *Journal of Historical Pragmatics*, 16(1):59–85.

Gabriele Diewald. 2011. Pragmaticalization (defined) as grammaticalization of discourse functions. *Linguistics*, 49(2):365–390.

Gabriele Diewald. 2013. *"Same same but different" – Modal particles, discourse markers and the art (and purpose) of categorization*, pages 19–46. John Benjamins Publishing Company.

Zi-Yi Dou and Graham Neubig. 2021. Word alignment by fine-tuning embeddings on parallel corpora. In *Conference of the European Chapter of the Association for Computational Linguistics (EACL)*.

Bruce Fraser. 1996. Pragmatic markers. *Pragmatics; Vol 6, No 2 (1996)*, 6.

Montserrat González, Paolo Roseano, Joan Borràs-Comes, and Pilar Prieto. 2017. Epistemic and evidential marking in discourse: Effects of register and debatability. *Lingua*, 186-187:68–87. Essays on evidentiality.

Miloslav Grepl and Radek Šimík. 2017. Epistemická modalita. In *CzechEncy - Nový encyklopedický slovník češtiny.*

Miroslav Grepl. 2017. Komunikační funkce výpovědi. In Petr Karlík, Marek Nekula, and Jana Pleskalová, editors, *CzechEncy – Nový encyklopedický slovník češtiny*. CzechEncy. Poslední přístup: 16. 10. 2025.

Maciej Grochowski, Anna Kisiel, and Magdalena Żabowska. 2014. *Słownik gniazdowy partykuł polskich*. Polska Akademia Umiejętności, Kraków.

Jan Hajič, Eduard Bejček, Jaroslava Hlaváčová, Marie Mikulová, Milan Straka, Jan Štěpánek, and Barbora Štěpánková. 2020. Prague Dependency Treebank - Consolidated 1.0. In *Proceedings of the Twelfth Language Resources and Evaluation Conference*, pages 5208–5218, Marseille, France. European Language Resources Association.

Agnes Helbig, Gerhard Helbig. 1993. *Lexikon deutscher Modalwörter*, 2., durchges. aufl. edition. Langenscheidt, Leipzig , Berlin [u.a.].

Leo Hoye. 2014. *Adverbs and modality in English*. Routledge.

Maarten Janssen. 2016. TEITOK: Text-faithful annotated corpora. In *Proceedings of the Tenth International Conference on Language Resources and Evaluation (LREC'16)*, pages 4037–4043, Portorož, Slovenia. European Language Resources Association (ELRA).

Anu Koivunen, Anu Kanner, Maciej Janicki, Auli Harju, Jussi Hokkanen, and Essi Mäkelä. 2021. Emotive, evaluative, epistemic: A linguistic analysis of affectivity in news journalism. *Journalism*, 22(5):1190–1206.

Miroslav Komárek, Jan Kořenský, and Jan Petr. 1986. *Mluvnice češtiny 2.* Academia.

Klaus Krippendorff. 2013. *Content Analysis: An Introduction to Its Methodology (third edition)*. Sage Publications, Thousand Oaks, CA.

Yang Liu, Xiaohui Yu, Zhongshuai Chen, and Bing Liu. 2013. Sentiment analysis of sentences with modalities. UnstructureNLP '13, page 39–44, New York, NY, USA. Association for Computing Machinery.

Bolei Ma, Yuting Li, Wei Zhou, Ziwei Gong, Yang Janet Liu, Katja Jasinskaja, Annemarie Friedrich, Julia Hirschberg, Frauke Kreuter, and Barbara Plank. 2025. Pragmatics in the era of large language models: A survey on datasets, evaluation, opportunities and challenges. In *Proceedings of the 63rd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 8679–8696, Vienna, Austria. Association for Computational Linguistics.

Roser Morante and Caroline Sporleder. 2012. Modality and negation: An introduction to the special issue. *Computational Linguistics*, 38(2):223–260.

Vítor Míguez. 2022. On epistemic modality and discourse strategy: Evidence from galician adverbs. *Journal of Pragmatics*, 201:32–42.

Jan Nuyts. 1999. *Epistemic modality, language, and conceptualization*. John Benjamins.

Juri Opitz, Shira Wein, and Nathan Schneider. 2025. Natural language processing relies on linguistics. *Computational Linguistics*, 51(3):1009–1032.

Frank Robert Palmer. 2001. *Mood and Modality*. Cambridge University Press.

Paola Pietrandrea. 2018. Epistemic constructions at work: A corpus study on spoken italian dialogues. *Journal of Pragmatics*, 129:171–191.

Lucie Poláková, Jana Šindlerová, and Barbora Štěpánková. 2025. No doubt, but... on connective functions of epistemic markers. *Jazykovedný časopis / Journal of Linguistics*, 76(1):75–84.

Agata Rozumko. 2016. Linguistic concepts across languages: The category of epistemic adverbs in english and polish. *Yearbook of the Poznan Linguistic Meeting*, 2.

Agata Rozumko. 2022. Textual functions of low confidence adverbs: The case of perhaps. *Lingua*, 268:103191.

Victoria L. Rubin. 2007. Stating with certainty or stating with doubt: Intercoder reliability results for manual annotation of epistemically modalized statements. In *Human Language Technologies 2007: The Conference of the North American Chapter of the Association for Computational Linguistics; Companion Volume, Short Papers*, pages 141–144, Rochester, New York. Association for Computational Linguistics.

Daisuke Suzuki. 2018. Variation between modal adverbs in british english: The cases of maybe and perhaps. *Functions of Language*, 25:392–412.

Jana Šindlerová, Barbora Štěpánková, and Ingrid Andrén. 2023. Epistemická částice zřejmě pohledem paralelního korpusu. *Korpus – gramatika – axiologie*, (27):37–52.

Barbora Štěpánková, Lucie Poláková, Jana Šindlerová, and Michal Novák. 2024. What can dictionaries tell us about pragmatic markers – building the lexicon of epistemic and evidential markers in czech. In *Lexicography and Semantics. Proceedings of the XXI EURALEX International Congress.*, pages 728–741, Zagreb, Croatia. Institute for the Croatian Language, Institut za hrvatski jezik.

Barbora Štěpánková, Jana Šindlerová, and Lucie Poláková. 2023. The epistemic marker určitě in the light of corpus data. *Jazykovedný časopis / Journal of Linguistics*, 74(1):130–139.

Mizuho Tamaji. 2009. What L2 processing strategy reveals about the prototypical relationship: A case of Japanese modal markers of possibility. In *Proceedings of the 23rd Pacific Asia Conference on Language, Information and Computation, Volume 2*, pages 540–548, Hong Kong. City University of Hong Kong.

Elizabeth Traugott. 1995. The role of discourse markers in a theory of grammaticalization.

Björn Wiemer and Katerina Stathi. 2010. The database of evidential markers in european languages: A bird's eye view of the conception of the database. *Language Typology and Universals*, 63:275–289.

11. Language Resource References

Klégr, Aleš and Kubánek, Michal and Malá, Markéta. *Korpus InterCorp – angličtina*. PID http://www.korpus.cz.

Rosen, Alexandr and Vavřín, Martin and Zasina, Adrian. *Korpus InterCorp – čeština*. PID http://www.korpus.cz.

A. List of Annotated Features

Here we show the annotated features in SEEM-CZ, which are supplemented with typical Czech examples and their English translation equivalents. The list of features is divided into two tables. Table 7 contains fundamental features: `Type of use`, `Type of degree of certainty`, and `Type of communicative function`. Table 8 contains contextual features: `Scope of the marker`, `Predicate verb`, `Position in a sentence`, `Negation`, `Contrast`, `Other modal expression`, and `Type of evidence`.

B. Distribution of Epistemic Markers

Figure 2 shows the distribution of the 40 selected expressions across books originally written in Czech. For each book, the ratios are normalized by book length.

The figure demonstrates that both the variety and frequency of EM usage differ substantially across titles. We present these results primarily to illustrate the nature of the source data; nevertheless, they suggest potential applicability for stylometric analysis. Works by Václav Havel exhibit a relatively high frequency of EM usage compared to most other authors. Moreover, Josef Škvorecký and Michal Viewegh display relatively consistent EM usage in terms of both frequency and distribution across their works. In contrast, books by Jáchym Topol vary considerably in both overall frequency and the distribution of EMs.

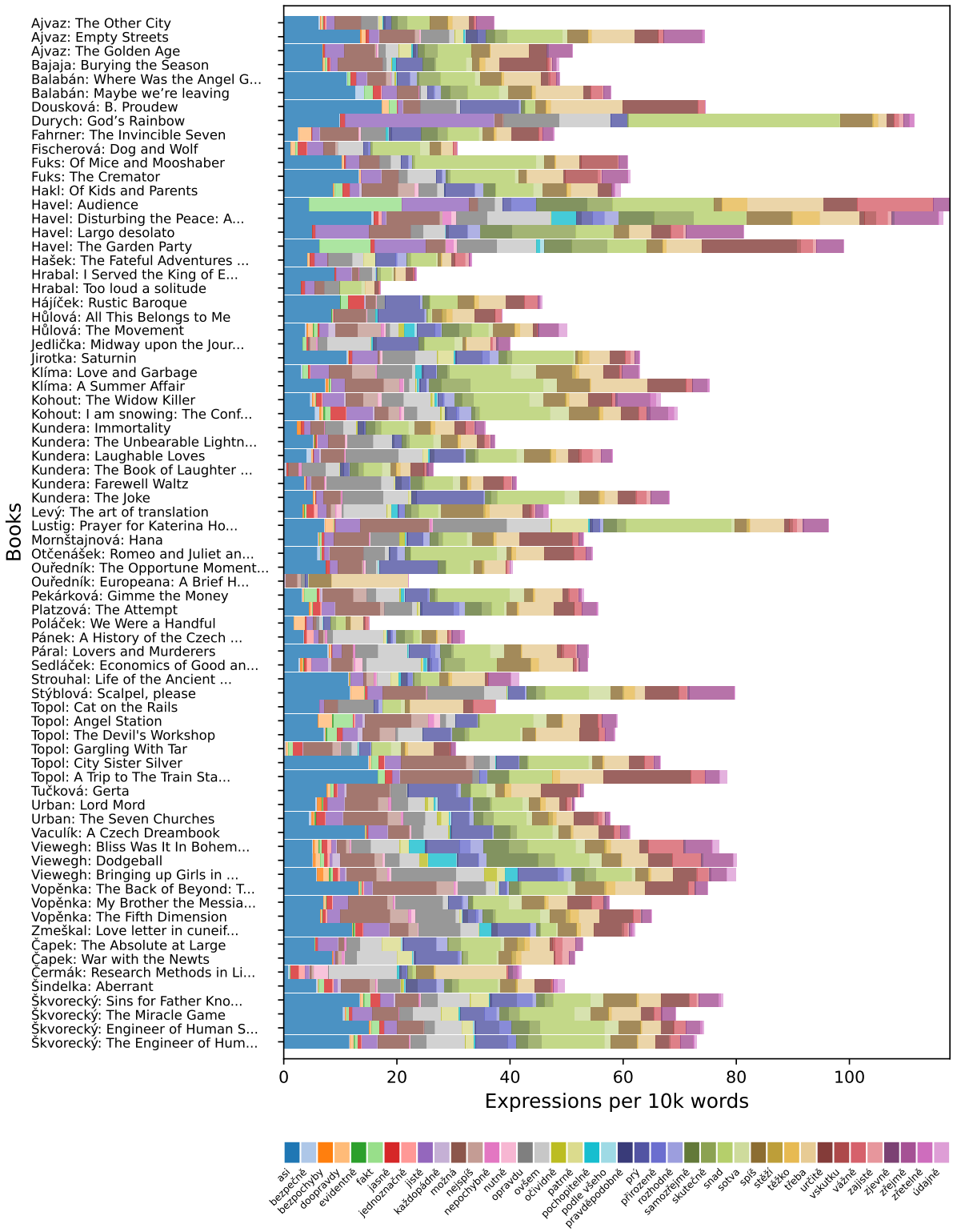

Figure 2: Distribution of expressions selected for annotation across books from InterCorp (Core part) originally written in Czech.

Annotated feature	Values
Type of use	**epistemic** *V dešti vypadala křehce, křehká však **nejspíš** nebyla.* She looked fragile in the rain, but she **probably** wasn't fragile.
	evidential *Tohle je **evidentně** normální kosmická loď, i když je obrovská*. This is **obviously** a conventional spacecraft, although very large.
	confirmatory *Malej je ale **fakt** zloděj.* Shorty **really** is a thief, though.
	response *Myslíš, že to maloval von? **Jasně**, ten na to vypadá* Think he painted all that? **Obviously**, he looks it.
	autosemantic *Dávala **jasně** najevo, že to chce.* But she 'd made it **clear** that she wanted it.
	other *„Jaký **asi**," odsekla babička.* "**What do you mean**, what notices?" Grandma snapped back.
Degree of certainty	**high** *Já ti **rozhodně** nic nevyčítám.* Well I'm **certainly** not berating you for anything.
	higher medium *...dostal šéfku, kerá **nejspíš** prodává sirotkům španělský mušky* ...got a boss **probly** sellin Spanish fly to orphan
	medium *...psal, že **možná** ještě jednoho dne půjde na vojnu.* he said he **might** go to the army himself one day.
	low *...bílou ručkou (v níž **sotva** kdy držela něco těžšího než cigaretu)* with her white hand (in which she'd **hardly** ever held anything heavier than a cigarette)
Type of comm. function *(Specific comm. function)*	**assertive** *Z těch tří jsem na tom **bezpochyby** nejlíp.* Of the three of us I am **without doubt** the best off. *(assumption)* *To se **pochopitelně** nelíbilo redaktorovi.* **Of course**, that wasn't to the liking of the editor. *(confirmation)*
	directive *To nám **bezpochyby** ještě vyjasníte, inspektore.* You will **doubtless** enquire into that. *(directive assumption)*
	interrogative ***Opravdu** se to jmenuje Symfonie milénia?* Is it **really** called the Millennial Symphony? *(request for confirmation)*
	commissive *Až to začne jít, dám ti **určitě** vědět.* As soon as it starts to go anywhere, I'll be **sure** to let you know. *(assurance)*
	(dis)approval *„**Nejspíš** ano," přikývl Harry, „proč by ale chtěl Rona otrávit?"* "**Probably**," said Harry, "but why would Slughorn want to poison Ron?" *(admission)*
	expressive *„Je to **zjevně** důležité," přisadila si jedna kamenná obluda.* "It's urgent, **apparently**," said the second gargoyle snidely. *(irony)*

Table 7: List of annotated fundamental features with examples

Annotated feature	Values
Scope of the marker	**whole clause** *To **opravdu** nevím.* I **really** do n't know
	member (narrow scope) Je lékařka s dobrou pověstí a s **nepochybně** tučným příjmem. She's a surgeon with a good reputation and **no doubt** a six-figure income.
	ellipsis *„A neměli bychom vědět?" zeptala se Beth. „**Možná** ne."* "Well, shouldn't we know?" Beth said. "Not **necessarily**."
Predicate verb *lemma, morphological tag:*	*To**s** už asi **zjistil** sám.* You'**ve** probably already **found** that **out** yourself. být zjistit, VB-S—2P-AAI-1 VpMS—-R-AAP–
Position in a sentence	**first** ***Možná** u něj nejste zapsán tak dobře jako já, Traversi.* ***Perhaps** your credit is not as good with him as mine is, Travers.*
	last *Jestli mladej, nevím, ale dědeček **určitě.*** Young I don't know, but grand-daddy **for sure**.
	other *Cestovali **přirozeně** na koních.* They had been on horseback, **naturally**.
Negation (Y/N)	*Záhirovi to podle všeho **ne**vadilo.* This did**n't** seem to bother Zahir.
In a contrastive pattern (Y/N)	*Všechno je **snad** v pořádku, oficiálně **určitě**.* Everything is [**hopefully**] in order, **at least** officially speaking.
Other modal expression (e.g. intensifier, modal verb)	*Ale tělo s tím **mohlo jen** těžko vystačit...* A body **could** hardly make ends meet...
Type of evidence *(if it is present)*	**sensory** ***Kluk zčervenal**, jasně se bál dalšího debaklu a výsměchu.* **The boy turned red**, clearly he was afraid of any further humiliation
	hearsay ***ti mu řekli**, že je na něm určitě dcera majitele vily ..* **they were fairly certain** the young woman was the daughter of the villa's owner
	reasoning *Dlouhý denní spánek ji zřejmě osvěžil; **teď je veselá a milá jako jindy.*** Her long sleep all day have refresh her, **for now she is all sweet as ever.**
	inference – [evidence is inferred, not present in the text]

Table 8: List of annotated contextual features with examples