Barbora Štěpánková, Lucie Poláková, Jana Šindlerová, and Michal Novák

WHAT CAN DICTIONARIES TELL US ABOUT PRAGMATIC MARKERS

Building the Lexicon of Epistemic and Evidential Markers in Czech

Abstract In this paper, we explore the possibilities and challenges of lexicographic treatment of pragmatic markers, specifically epistemic and evidential markers in Czech. Our starting point is a detailed comparison of how these expressions are treated in contemporary monolingual Czech dictionaries. Following this, we present the development of the SEEMLex lexicon of Czech epistemic and evidential markers which is based on detailed annotation of selected expressions using data from a Czech-English parallel corpus. We describe the features we annotate when analysing the expressions studied, outline the main aspects that constitute or distinguish their meanings, and emphasise the importance of considering the communicative function in which these expressions are used. Additionally, we highlight the benefits of using a specialised lexical database for the lexicographic processing of pragmatic expressions in general. We demonstrate our approach with a draft of a dictionary entry for the common Czech epistemic marker *asi* 'probably' providing a comprehensive example of our methodology.

Keywords specialized dictionary; epistemic markers; communicative functions; annotation; Czech

1. Introduction

Compiling a monolingual dictionary is a complex task involving many conceptual decisions in order to ensure a comprehensive, yet comprehensible and consistent treatment of word meanings. In the Czech lexicographic tradition, the process of compiling a monolingual dictionary is usually carried out in alphabetical order[1], which in the long run can lead to difficulties in maintaining consistency of treatment due to the change of authors and subsequent changes in the principles of compilation. As a result, expressions belonging to the same semantic group may be treated differently.

This is particularly the case for expressions with a weakened lexical meaning, i.e., grammatical or functional words, and thus also for pragmatic markers, including epistemic and evidential markers (hereafter EEM). There are several

[1] This approach is described, e.g., in the unpublished guidelines for the compilation of the SSJČ dictionary (*Směrnice pro vypracovávání rukopisu Slovníku spisovného jazyka českého*, 1957). There appeared one notable exception to this approach – the comprehensive proposal for the delimitation and treatment of secondary prepositions, including their list and description of their meanings, which was comprised for and applied in the one-volume SSČ dictionary.

This paper is part of the publication: Despot, K. Š., Ostroški Anić, A., & Brač, I. (Eds.). (2024). *Lexicography and Semantics. Proceedings of the XXI EURALEX International Congress.* Institute for the Croatian Language.

reasons for this: 1) monolingual dictionaries focus primarily on content words and leave other expressions on the periphery of interest, or 2) it is only recently that the interest in pragmatic expressions and pragmatic meaning has arisen, thus, there is no standardised lexicographic treatment agreed upon yet, or 3) Czech pragmatic expressions are often homonymous/polysemous and in many cases it is difficult to adequately distinguish the epistemic meanings from other meanings.

The aim of this study is to present a proposal for an architecture of a lexicon of Czech epistemic and evidential markers that would provide a unified and complex description of their meaning.[2] First, we define the group of Czech EEMs and comment on the existing approaches to their description, including lexicographic approaches. Then, the process of building the SEEMLex lexicon based on the underlying annotation of selected EEMs in corpus data is presented. As a case study, we provide the results of the annotation of the most frequent Czech epistemic marker *asi* 'probably', and comment on the relevant semantic and pragmatic features to be captured in its lexicographic description. Also, we present a sample lexicon entry. Finally, we summarise the advantages of a specialised lexical database as a basis for the treatment of a specific class of words in a monolingual dictionary.

2. Epistemic and Evidential Markers

Pragmatic markers, in general, are expressions that encode people's opinions, assumptions and beliefs regarding the propositional content. Their meaning is primarily semantico-pragmatic in nature. In the Czech linguistic tradition, they are classified as *particles*.[3] EEMs are then traditionally referred to as *epistemic particles* (e.g., Komárek et al., 1986; Cvrček et al., 2010).

In the literature, the relationship between epistemic and evidential markers is perceived diversely (cf., e.g., de Haan, 2001; Nuyts, 2001; Plungian, 2001). Our approach assumes a general overlap of both categories in accordance with some theoretical approaches (cf. Hoye, 2008; Carretero & Zamorano-Mansilla, 2013; Komárek et al., 1986; Cvrček et al., 2010) and with the support of our pilot analysis (cf. Šindlerová et al., 2023). Thus, in the context of a project focused on exploring EEMs, we understand them as one broader group containing permeable categories of epistemic markers (such as *možná* 'maybe',[4] *určitě* 'certainly'), evidential markers (*údajně* 'allegedly', *podle všeho* 'to all appearances') and markers confirming/emphasizing the speaker's strong belief in their being right, the so-called confirmatory expressions (cf. Rozumko, 2016) (*ovšem* 'of course').

[2] For EEMs, we use the term *meaning* in the sense of the term *relational meaning* (cf. Filipec & Čermák, 1985, p. 39).

[3] In Czech linguistics, *particles* are defined as expressions with a predominant attitudinal function; they are generally accepted as one of ten parts of speech. For a detailed analysis of the concept of *particles* in Slavic languages compared to anglophone linguistics using the example of epistemic adverbs, see Rozumko (2016). For a description of the behaviour of these expressions, cf. also Volkova (2017), or Grochowski et al. (2014).

[4] Here are the most common translation equivalents. For specific examples, equivalents appropriate to the context are used.

2.1 Dictionary Treatment of the EEMs

As mentioned above, in Czech monolingual dictionaries, EEMs are usually treated as function words. As such, they have received limited attention so far. The definitions of EEMs mostly comprise only a vague description of the degree of certainty (1).

(1) *dozajista*: 'vyjadřuje nejvyšší míru jistoty, přesvědčení mluvčího o platnosti tvrzení' (*certainly*: 'expresses the maximum degree of certainty, speaker's belief in the validity of the proposition') (ASSČ)

Only rarely and inconsistently do we come across attempts to capture other aspects (2). In some grammars, their role in expressing negativity (e.g., Daneš et al., 1987) or their function in communication (Hoffmannová et al., 2019) is mentioned.

(2) *zajisté*: '2. vyjadřuje zdůrazněně souhlas; 3. vyjadřuje subjektivní přesvědčení o něčem' (*certainly*: '2. emphatically expresses agreement, 3. expresses a subjective belief about something') (SSJČ)[5]

This does not correspond to the definitions in monolingual dictionaries of languages other than Czech which often provide more detailed explanations (cf. selected definitions of *maybe* (3)), commenting specifically on the possible intentions of the speaker or giving hints on the degree of politeness.

(3) *maybe:* 'used to politely suggest or ask for something; used to avoid giving a clear or certain answer to a question' (CLD)

Therefore, in general, Czech lexicographic treatments of EEMs typically contain only a basic semantic feature, namely the aforementioned specification of the degree of certainty, which most closely resembles the treatment of content words. However, as "words with a primarily pragmatic nature and usage"[6] (Čermák, 1992, p. 257), the expressions under examination require a more specific approach, based on the pragmatic component of their meaning. Our pilot study (Štěpánková et al., 2023) suggested that the degree of certainty may be weakened in some uses and that the **communicative function** (CF) of an utterance can serve as a crucial component to follow. According to Grepl (2017), we understand CF as "the 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."[7] Based on our experience with empirical data so far, we hypothesise that an annotation of CF is crucial not only for the description of basic EEM meanings but also for distinguishing other meanings of the studied expressions. A large-scale CF annotation can provide evidence for the repeated and regular use of some of these meanings and thus for their lexicalization.

[5] The SSJČ and SSČ dictionaries are cited as https://prirucka.ujc.cas.cz/

[6] "…slova s primárně pragmatickou povahou a územ…" (Čermák, 1992, p. 257)

[7] "…smysl výpovědi vyplývající ze záměru, s jakým je nějaká výpověď mluvčím vůči adresátovi v dané konkrétní komunikační situaci produkována…" (Grepl, 2017)

3. Methodology

The specialised SEEMLex lexicon is planned as one of the outputs of the project researching EEMs. In recent years, electronic lexicons describing functionally defined groups of expressions have often been created in diverse areas, e.g., in the area of subjectivity (Veselovská & Bojar, 2013), or valency (Lopatková et al., 2016); in the area of discourse markers and connectives, see e.g., Mírovský et al. (2017) for Czech, or Stede (2002) for German; a multilingual database of discourse markers is available online (Stede et al., 2019)[8]; in the area of epistemic and evidential markers, see e.g., proposal for a database by Wiemer & Stathi (2010). Such lexicons are usually created on the basis of large corpus data and their specific annotation, which is also the case of the SEEMLex.

In our project, we use the parallel InterCorp v15 corpus (Čermák & Rosen, 2012), specifically the core part of its Czech and English sections, containing mainly fiction, as the underlying language resources for lexicon development. The fiction data were chosen as suitable for the study of EEMs for the following reasons:

- presumed closeness to spoken language (e.g., in terms of high frequencies of epistemic markers, a certain degree of subjectivity);
- in contrast to originally spoken texts, the fiction writing also mostly offers elaborate situational context which is helpful for EEM interpretation;[9]
- unlike often in journalistic texts, the origin of a fiction text (the author and the original language) is known. Our primary focus is on original Czech texts, and, secondly, original English texts translated into Czech;
- reliable translation quality of the texts is another advantage of using a parallel corpus. It allows us to use translation equivalents to clarify the meaning of the Czech markers in context (cf. Aijmer et al., 2006).

In contrast to several studies devoted to epistemic modality (cf. the Modal Corpus described in Pietrandrea (2018)), SEEMLex does not use a corpus-driven approach to identify EEMs, instead, we annotate a predetermined list of them.

This headword list has been compiled by a manual selection of lemmas and forms from various Czech grammars, complemented by a selective list of markers annotated as modal or attitudinal in the Prague Dependency Treebanks. While grammars mostly list typical (very frequent) expressions representing the epistemic group and usually do not provide any context, the corpora used to compile the headword list (PDT 3.5, Hajič et al., 2018, and PDTSC, Mikulová et al., 2017) capture various expressions in their (syntactic) contexts, and additionally enrich the list with less frequent items. The comprehensive list contains approximately 140 entries. It includes both single-word (*pravděpodobně* ‘probably’) and multi-word (*s jistotou* ‘certainly’, lit. ‘with certainty’) markers. We do

[8] http://connective-lex.info/

[9] On the other hand, the written texts have the disadvantage of not including intonation, which often serves as a very useful interpretative device in the case of particles.

not include modal verbs – their properties are somewhat different, and, in contrast to EEMs, they are well researched in Czech (cf. Grepl, 1979; Ivanová, 2017, etc.).

3.1 Tool

Given the headword list and the parallel corpus, we collect all occurrences of the expressions in the corpus and select samples of them for manual annotation. The manual annotation of the EEMs is conducted using the TEITOK web-based platform (Janssen, 2016). For each sampled expression, the annotators can see the sentence it appears in, including the possibility to explore an arbitrarily large context of the sentence, which may be crucial for the annotation of modality.

Furthermore, the annotation environment can display the English equivalent of the sentence and highlight the expression's counterpart in the sentence. While the sentence alignment is an integral part of the InterCorp corpus, the counterpart of the expression is obtained automatically by running the AWESOME aligner (Dou & Neubig, 2021). We fine-tuned the default model based on multilingual BERT (Devlin et al., 2019) first on the parallel Czech-English data from PCEDT 2.0 (Hajič et al., 2012) in the unsupervised setting, followed by Czech-English manual word alignments (Mareček, 2008) in the supervised setting. While annotating the modality features, the annotators are also asked to fix potential errors in the automatic alignment of the expressions.

The annotators may also decide to label an annotated sentence as a candidate for dictionary exemplification. Out of these candidate sentences, we manually select the most suitable examples to be shown in the SEEMLex lexicon.

3.2 Annotation of EEM Features

Each instance of the selected markers in the corpus data is annotated for a set of features. This repertoire of features was compiled based on the available state-of-the-art studies (e.g., Wiemer & Stathi, 2010; Lavid et al., 2016; Pietrandrea, 2018) and was confirmed convenient through test annotation (Štěpánková et al., 2023).

Although our project focuses on expressions with epistemic and evidential meanings, we consider it necessary to annotate also other meanings of the given expressions at least in a basic manner. In this way, we are able to document their widely polysemous/homonymous nature and map their overall use. For example, for the expression *jistě* 'certainly', the meaning of a manner adverb – 'walk surely' (4) or response particle 'sure' (5) are annotated.

(4) *Kráčela lehce a **jistě**.*[10] 'She walked lightly and **surely**.'

(5) *"Chodil jste do kostela?" "**Jistě**. Každé Vánoce a Velikonoce."* '"Did you go to church?" "**Sure**. Every Christmas and Easter."'

[10] All examples used come from the InterCorp corpus.

In agreement with Carretero & Zamorano-Mansilla (2019), we also consider as EEMs those uses where an additional feature is present in combination with the epistemic meaning, e.g., expressing an attitude (6).

(6) *"Znáte **jistě** tuto scénu z desítek špatných filmů: hoch a dívka se drží za ruce a běží krásnou jarní (eventuelně letní) přírodou.* 'You **certainly** remember this scene from dozens of bad films: a boy and a girl are running hand in hand in a beautiful spring (or summer) landscape.'

The annotated features (see Table 1) relate to the expression itself (e.g., position in the sentence), describe phenomena in its close context (grammatical features of the predicate, presence of evidentiality, negation, contrast, etc.), or comment on the utterance as a whole (CF).

Table 1: List of annotated features

Annotated feature	**Values**
Type of use	epistemic, evidential, confirmatory, response, other, autosemantic
Degree of certainty	high, higher medium, medium, low
Type of CF	assertive, directive/contact, interrogative, commissive, (dis)approval, expressive
Specific CF	e.g., assumption, recommendation, wish
Scope	clause/member
Predicate verb	verb 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

4. Case Study – *asi*

Asi is the most frequent epistemic marker on our list, therefore we have selected it as an example lexicon entry. Moreover, it is one of the expressions that have already been processed in the most recent (unfinished) monolingual dictionary, the ASSČ[11], which itself is proclaimed to be based on corpus data and its concept explicitly mentions a shift towards reflecting pragmatics in the processing of entries (cf. Kochová & Opavská, 2016).

In this section, we first compare and critically evaluate three different ways of lexicographic treatment of the selected marker in Czech dictionaries, then we present our draft for the EEM lexicon entry, and, finally, we discuss the underlying principles in detail.

[11] Currently, entries starting with the first ten letters of the Czech alphabet (A–CH) have been published, i.e., approx. 20,000 entries.

The expression *asi,* roughly translatable as ‘probably’, is considered an epistemic marker situated approximately in the middle of the certainty scale (in agreement with e.g., Komárek et al., 1986; or Grepl, 2017). This basic semantic characteristic is also evident in dictionary treatments. In older monolingual dictionaries (SSJČ (7), and SSČ (8)), the word is described by means of two meanings, or shades of meaning, both expressing a lower degree of certainty. In the SSČ, the only meaning included applies the lower degree of certainty also to the meaning of approximation (paraphrased by *přibližně*).

(7) SSJČ: **1.** *přibližně:* a. před týdnem; a. pět knih; a. čtyři lidé; **2.** *jak se zdá; pravděpodobně, snad, možná, patrně:* to a. nepůjde; a. to tak je; a. někde pršelo; a. to přinesu[12]

(8) SSČ: *vyj. menší míru jistoty, pravděpodobně, možná, snad 1, patrně:* asi bude pršet; asi tak před týdnem *přibližně*[13]

In the most recent ASSČ dictionary (9), *asi* is divided into three separate meanings: Meaning 1 contains uses expressing at least a medium degree of certainty; Meaning 2 expresses the adverbial meaning of measure (see (7), Meaning 1 above), and Meaning 3 records other uses expressing various sentiments of the speaker towards the proposition. It must be noted that in the last type of meaning, the original certainty meaning is no longer manifested and the attitudinal function prevails. To sum up, in the ASSČ dictionary, we can see a noticeable attempt to separate the certainty meaning from other types of meaning.

(9) ASSČ: **1.** *vyjadřuje střední nebo vyšší míru jistoty syn. pravděpodobně:* Soupeři se ho asi bojí. Sluneční hodiny zná asi každý. Krupobití a vichru se asi nevyhneme. *ve funkci citoslovce* Pojedete na dovolenou? – Asi. *asi ano*; **2.** *vyjadřuje přibližnost míry, množství, délky trvání, zprav. před číselným výrazem syn. přibližně:* Seskočil z výšky asi jednoho a půl metru. Na letišti čekalo asi pět set lidí. Pracovala tam asi rok. Zpoždění vlaku bude asi šedesát minut. **3.** *zdůrazňuje citový postoj mluvčího k situaci •(v otázce) zvědavost, nevědomost:* Kdepak mám asi lístek? Budeme mít novou paní učitelku. Jaká asi bude? *•expresivní rozhořčení, nesouhlas, často při odmítání předchozí výpovědi***:** Ty ses asi zbláznil! Byl tady Tom! – Kde by se tady asi vzal? Jestli toho necháš, šeredně na to doplatíš. – A co mám asi dělat?[14]

[12] ‘**1.** *approximately:* a. a week ago; a. five books; a. four people; **2.** *it seems, probably, perhaps, maybe, apparently:* it seems impossible; it seems so; it has probably rained somewhere; I may bring it with me’)’

[13] ‘*expressing lower degree of certainty, probably, maybe, perhaps 1, apparently:* it may rain; approximately a week ago *approximately*’

[14] ‘**1.** *expressing medium or higher degree of certainty, synonymous with probably*: His enemies probably fear him. Sundial is known by probably everyone. We probably won’t avoid hail and wind. *in the function of an interjection*: Are you going on vacation? – Probably. *Probably we are.* **2.** *expressing inexactness of the measure, quantity, duration, usually in front of a numerical expression, synonymous with about*: He jumped from a height of about one and a half meters. There were about 500 people waiting at the airport. She has been working there for about a year. The train will be delayed by about sixty minutes. **3.** *emphasizing the emotional attitude of the speaker to the situation• (in questions) curiosity, ignorance***:** Where might my ticket be? We will have a new teacher. What will she be like? •*expressive indignation, disagreement, often when rejecting previous statement***:** You must be crazy! Tom was here! – How would he have gotten here? If you don’t stop, you’ll pay dearly for it. – And what am I supposed to do?’

On the other hand, the authors' lack of experience or methodological lexicographic support in the processing of expressions with a strong pragmatic component of meaning is evident here, as well as insufficient comparative approach that would take into account also other expressions from this functional-semantic group.

Meaning 1 also includes examples that are defined by their communicative function rather than by the degree of certainty, e.g., the conversational premise *Sluneční hodiny zná asi každý.* 'Sundial is known by probably everyone.' or the response particle *Pojedete na dovolenou? – Asi.* 'Are you going on vacation? – Probably.'

What is more, Meaning 3 merges different types of attitudes which imply different synonymous alternatives, e.g., in the example *Ty ses asi zbláznil* 'You must be crazy' [lit. 'You have gone probably crazy'], *asi* can be replaced by a number of other certainty markers with varying degrees of certainty (*určitě* 'definitely', *nejspíš* 'probably'); on the other hand, *Kde by se tady asi vzal?* 'How would he have [lit. 'probably'] gotten here?' is a use in which no such substitution is possible. It is therefore worth considering whether these examples should be grouped together under a single meaning.

Within the case study, we performed a parallel annotation of 200 occurrences of *asi* (100 from original Czech texts, 100 from the Czech translations of English originals). Randomly selected samples were annotated in parallel by three annotators, native speakers of Czech with a linguistic background. Inter-annotator comparisons were made on the annotated data, focusing mainly on the basic type of use categories (epistemic – pragmaticalized – autosemantic). Inconsistencies – in most cases – included annotator's misinterpretation of the annotation guidelines, the treatment of expressive usages with a more vague solution in the guidelines, multiple possible interpretations of a sentence. The first two types of inconsistencies should be improved in future annotations, on the other hand, different interpretations are unavoidable. Our draft of an EEM lexicon entry is based on the analysis of the annotation results as well as on the dictionary comparison above. When creating a dictionary entry, we are primarily guided by two principles: 1. the presence or absence of epistemic modality or its weakening, and 2. the communicative function in which the expression is used.

In our draft, the basic, unmarked use of *asi* is the epistemic meaning with a medium certainty or higher medium certainty degree. Within this major use, several communicative functions can be distinguished. The strongest one is the assertive CF which includes several subtypes of CFs mostly distinguishable thanks to the lexical or syntactic contexts. The default assertive CF is the assumption (I.a). Further, if the context contains strong evidence, the CF of the utterance is explanation (I.b); different verb tenses and moods imply the CFs of future guess (I.c). Within the epistemic type of use, we have also documented a directive/contact CF subtype which is primarily used to express a recommendation (II.a), but is also further used as a conversational formula/assumption (II.b). Meaning III shows a specific type of assertive CF with a weakened degree of certainty – introspection. In Meanings IV–VII, *asi* appears in functions of attitudinal pragmatic markers. Meaning VIII describes various uses of approximation.

Draft of a SEEMLex entry *asi*[15]

Epistemický význam

I. Vyjádření střední či vyšší střední jistoty vzhledem k propozici

I.a Domněnka o tom, co se stalo nebo děje: Šlo ***asi*** *o import z Dálného východu.*

I.b Vysvětlení, s přítomností evidence: ***Asi*** *jsem na chvíli usnul, protože když jsem otevřel oči, byl jsem ve třídě s provaleným stropem sám.*

I.c Odhad toho, co se bude dít: *Ti vojáci tu* ***asi*** *taky nebudou věčně.*

II. Direktivní/kontaktový význam, epistemičnost částečně oslabená

II.a Doporučení: *„****Asi*** *byste měl," řekl.*

II.b Direktivní předpoklad, předjímání názoru komunikačního partnera: ***Asi*** *se tomu divíš, takhle přímo jsi to ode mne neslyšel.*

III. Introspekce, epistemičnost částečně oslabená: *Já se v tý chvíli* ***asi*** *pomát.*; ***Asi*** *bych si od něho měla něco přečíst.*

Bez epistemičnosti, pragmatikalizovaný význam

IV. Tázací: *Jaké poruchy by se* ***asi*** *jevily u profesora Devrienta, kdybych mu odňal pravý čelní mozkový lalok?*

V. (Slabý) souhlas/nesouhlas: *„To je všechno, co máme? Fazole?" „****Asi.****"*

VI. Hedging, konverzační formule: *„Je mi hrozně líto, ale to* ***asi*** *nepůjde."*

VII. Expresivní, umocňuje postoj: *To bude to středisko* ***asi*** *pěkně vypadat.*

Bez epistemičnosti, plnovýznamové

VIII. Přibližnostní význam

VIII.a Přibližné množství, před číselným výrazem nebo srovnáním: *večer* ***asi*** *v pět nebo v šest*; *Byl velký* ***asi*** *jako menší město.*

VIII.b Přibližnost, podobnost: *Ale říkal* ***asi*** *tohle.*

‘Epistemic modality

I. expression of medium and higher medium certainty regarding a proposition

I.a Assumption about past or present events: *This* ***probably*** *referred to silk imported from the Far East.*

I.b Explanation: *I* ***must*** *have nodded off, because when I opened my eyes I was alone in the classroom with the collapsed ceiling.*

I.c Estimation of what will happen: *Those soldiers* ***might*** *not be here forever either.*

II. Directive/contact meaning, epistemic meaning partially weakened

[15] The SEEMLex lexicon will be provided both in Czech and English language version.

II.a Recommendation: *“**Perhaps** you should,” he said.*

II.b Directive assumption, anticipating the partner’s opinion: ***Maybe** that comes to you as a bit of a shock, you’ve never heard it like that, straight from me.-*

III. Introspection, epistemic meaning partially weakened: *And at that very instant I **must** have gone mad.*; ***Maybe** I should read one of his things.*

Without epistemicity, pragmaticalized meaning, politeness, hedging

IV. Interrogative: *What kind of disturbances would appear in Professor Devrient if I removed his right frontal lobe?*[16]

V. (Weak) agreement/disagreement: *“Is that all we have? Beans?” “**Could be**.”*

VI. Hedging, some form of weakening, conversational figure: *‘Oh, I’m terribly sorry, but I **don’t think** that’s possible.*

VII. Expressive, emphasizes an attitude: *What an awful place that re-education centre **must** be!*

Without epistemicity, autosemantic

VIII. Approximation

VIII.a Approximate quantity preceding a numerical expression or comparison: *in the evenings, **around** five or six; It was the size of a small city.*

VIII.b Approximation, similarity: *But what he said was **roughly** this.’*

5. Conclusion

By comparing traditional lexicographic approaches to pragmatic markers, we have demonstrated that contemporary Czech dictionaries typically handle them using principles more suited to content words, i.e., the pragmatic component is often disregarded. The example of the relatively detailed treatment of *asi* in the ASSČ dictionary demonstrates the current lack of theoretical lexicographic support for dealing with such expressions.

Hence, the primary objective of this study was to propose a more comprehensive lexicographic approach to EEMs in a specialized lexicon based on an in-depth annotation of the studied expressions in context using a parallel corpus. In the annotation process, two guiding principles for the lexicographic treatment of EEMs proved relevant: the first one is the degree of certainty conveyed by a given expression in a specific context, the second one is the communicative function of the utterance containing the expression. While the degree of certainty (epistemicity) may be weakened or completely emptied for some uses, the communicative function remains a strong lead to distinguish among various attitudinal meanings of the examined expressions.

[16] In the Meanings IV, VI, and VIII.a, *asi* is implied or paraphrased by quite different linguistic features in the English InterCorp translation (e.g., in VI lit. *perhaps not possible = I don’t think that’s possible*).

The future SEEMLex lexicon will offer a comprehensive overview of the selected items. Each entry will primarily focus on describing the epistemic and evidential features of a given marker but it will also capture other functions (e.g., response functions, expressing politeness, etc.), as well as intrinsic lexical meanings (e.g., expressing approximation, manner). Apart from ensuring greater consistency, this approach facilitates an in-depth analysis of the universal and specific features of the individual expressions within the given group including their mutual relations (such as synonymy, antonymy, etc.).

References

Aijmer, K., Foolen, A., & Simon-Vandenbergen, A-M. (2006). Pragmatic markers in translation: a methodological proposal. In K. Fischer (Ed.), *Approaches to discourse particles.* (pp. 101–114). Elsevier.

ASSČ: *Akademický slovník současné češtiny.* Retrieved May 26, 2024, from https://slovnikcestiny.cz/

CLD: *Cambridge Learner's Dictionary.* Retrieved January 24, 2024, from https://dictionary.cambridge.org/dictionary/english/maybe

Carretero, M., & Zamorano-Mansilla, J. R. (2013). Annotating English adverbials for the categories of epistemic modality and evidentiality. In J. Marín Arrese, M. Carretero, J. Arús, & J. van der Auwera (Eds.), *English Modality: Core, Periphery and Evidentiality. Topics in English Linguistics, 81.* (pp. 317–355). Mouton De Gruyter.

Carretero, M., & Zamorano-Mansilla, J. R. (2019). Disentangling epistemic modality, neighbouring categories and pragmatic uses: the case of English epistemic modal adverbs. In K. Filippi-Deswelle (Ed.), *Quinze études de cas sur les modalités linguistiques. / Fifteen case studies on types of linguistic modalities* (pp. 131–157). Publications Électroniques de l'ERIAC, Epilogos 6.

Cvrček, V., Kodýtek, V., Kopřivová, M., Kováříková, D., Sgall, P., Šulc, M., Táborský, J., Volín, J., & Waclawičová, M. (2010). *Mluvnice současné češtiny I.* Karolinum.

Čermák, F. (1992). Paradigmatika a syntagmatika slovníku: problémy a možnosti. *Slovo a slovesnost, 53*(4), 249–264.

Čermák, F., & Rosen, A. (2012). The case of InterCorp, a multilingual parallel corpus. *International Journal of Corpus Linguistics*, *17*(3), 411–427.

Daneš, F., Hlavsa, Z., & Grepl, M. (Eds.). (1987). *Mluvnice češtiny 3.* Academia.

De Haan, F. (2001). The Relation between Modality and Evidentiality. *Linguistische Berichte*, *9*, 201–216.

Devlin, J., Chang, M. W, Lee, K., & Toutanova, K. (2019). BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding. In J. Burstein, Ch. Doran, & Th. Solorio (Eds.), *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) (pp. 4171–4186). Association for Computational Linguistics.

Dou, Z., & Neubig, G. (2021). Word Alignment by Fine-tuning Embeddings on Parallel Corpora. In P. Merlo, J. Tiedemann, & R. Tsarfaty (Eds.), *Proceedings of the 16th Conference of the European Chapter of the Association for Computational Linguistics* (pp. 2112–2128). Association for Computational Linguistics.

Filipec, J., & Čermák, F. (1985). *Česká lexikologie.* Academia.

Grepl, M. (2017). Komunikační funkce výpovědi. In P. Karlík, M. Nekula, & J. Pleskalová (Eds.), *CzechEncy – Nový encyklopedický slovník češtiny.* Retrieved May 24, 2024, from https://www.czechency.org/slovnik/KOMUNIKA%C4%8CN%C3%8D%20FUNKCE%20V%C3%9DPOV%C4%9ADI

Grepl, M. (1979). Úvodní poznámky k tzv. jistotní modalitě. *Slovo a slovesnost*, 40(2), 81–87.

Grochowski, M., Kisiel, A., & Żabowska, M. (2014). *Słownik gniazdowy partykuł polskich.* Polska Akademia Umiejętności.

Hajič, J., Hajičová, E., Panevová, J., Sgall, P., Bojar, O., Cinková, S., Fučíková, E., Mikulová, M., Pajas, P., Popelka, J., Semecký, J., Šindlerová, J., Štěpánek, J., Toman, J., Urešová, Z., & Žabokrtský, Z. (2012). Announcing Prague Czech-English Dependency Treebank 2.0. In N. Calzolari, K. Choukri, Th. Declerck, Mehmet U. Doğan, B. Maegaard, J. Mariani, A. Moreno, J. Odijk, & S. Piperidis (Eds.), *Proceedings of the Eighth International Conference on Language Resources and Evaluation (LREC'12)* (pp. 3153–3160). European Language Resources Association (ELRA).

Hajič, J., Bejček, E., Bémová, A., Buráňová, E., Hajičová, E., Havelka, J., Homola, P., Kárník, J., Kettnerová, V., Klyueva, N., Kolářová, V., Kučová, L., Lopatková, M., Mikulová, M., Mírovský, M., Nedoluzhko, A., Pajas., P., Panevová, J., Poláková, L., Rysová, M., Sgall, P., Spoustová, J., Straňák, P., Ševčíková, M., Štěpánek, J., Urešová, Z., Vidová Hladká, B., Zeman, D., Zikánová, Š., & Žabokrtský, Z. (2018). *Prague Dependency Treebank 3.5.* ÚFAL, LINDAT/CLARIN, Charles University. http://hdl.handle.net/11234/1-2621

Helcl, M, Sochová, Z., & Kozlová, K. (Eds.). (1957). *Směrnice pro vypracovávání rukopisu Slovníku spisovného jazyka českého.* Interní tisk. ÚJČ.

Hoffmannová, J., Homoláč, J., & Mrázková, K. (Eds.). (2019). *Syntax mluvené češtiny.* Academia.

Hoye, L. F. (2008). Evidentiality in discourse: A pragmatic and empirical account. In J. Romero-Trillo (Ed.), *Pragmatics and Corpus Linguistics* (pp. 151–174). De Gruyter Mouton.

Ivanová, M. (2017). *Modálnosť a modálne verbá ve slovenčině.* FF Prešovskej univerzity.

Janssen, M. (2016). TEITOK: Text-Faithful Annotated Corpora. In N. Calzolari, K. Choukri, Th. Declerck, S. Goggi, M. Grobelnik, B. Maegaard, J. Mariani, H. Mazo, A. Moreno, J. Odijk, & S. Piperidis (Eds.), *Proceedings of the Tenth International Conference on Language Resources and Evaluation* (pp. 4037–4043). ELRA.

Kochová, P., & Opavská, Z. (Eds.). (2016). *Kapitoly z koncepce Akademického slovníku současné češtiny*. Ústav pro jazyk český AV ČR.

Komárek, M., Kořenský, J., Petr, J., & Veselková, J. (Eds.). (1986). *Mluvnice češtiny 2*. Academia.

Lavid, J., Carretero, M., & Zamorano-Mansilla, J. R. (2016). A linguistically-motivated annotation model of modality in English and Spanish: Insights from MULTINOT. *Linguistic Issues in Language Technology*, 14, 1–35. https://aclanthology.org/2016.lilt-14.4/

Lopatková, M., Kettnerová, V., Bejček, E., Vernerová, A., & Žabokrtský, Z. (2016). *Valenční slovník českých sloves VALLEX*. Karolinum.

Mareček, D. (2008). *Automatic Alignment of Tectogrammatical Trees from Czech-English Parallel Corpus*. [Master thesis. Charles University, MFF].

Mikulová, M., Bémová, A., Hajič, J., Hajičová, E., Ircing, P., Kolářová, V., Lopatková, M., Mareček, D., Mírovský, J., Nedoluzhko, A., Pajas, P., Panevová, J., Peterek, N., Romportl, J., Sgall, P., Ševčíková, M., Štěpánek, J., Urešová, Z., & Žabokrtský, Z. (2017). *Prague Dependency Treebank of Spoken Czech 2.0*. ÚFAL, LINDAT/CLARIN, Charles University. http://hdl.handle.net/11234/1-3189.

Mírovský, J., Synková, P., Rysová, K., & Poláková, L. (2017). CzeDLex – A Lexicon of Czech Discourse Connectives. *The Prague Bulletin of Mathematical Linguistics*, *109*, 61–91.

Nuyts, J. (2001). *Epistemic Modality, Language, and Conceptualization: A cognitive-pragmatic perspective*, Vol. 5. John Benjamins Publishing.

Pietrandrea, P. (2018). Epistemic constructions at work. A corpus study on spoken Italian dialogues. *Journal of Pragmatics, 128*(15), 171–191.

Plungian, V. (2001). The place of evidentiality within the universal grammatical space. *Journal of Pragmatic*s, *33*, 349–357.

Rozumko, A. (2016). Linguistic Concepts across Languages: The Category of Epistemic Adverbs in English and Polish. *Yearbook of the Poznan Linguistic Meeting*, *2*(1), 195–214.

Stede, M. (2002). DiMLex: A Lexical Approach to Discourse Markers. In A. Lenci & V. Di Tomaso (Eds.), *Exploring the Lexicon - Theory and Computation*. Edizioni dell'Orso.

Stede, M., Scheffler, F., & Mendes, A. (2019). Connective-lex: A web-based multilingual lexical resource for connectives. *Discours*, 24. DOI: 10.4000/discours.10098

SSČ: Filipec, J., Daneš, F., Machač, J., Kroupová, L., Mejstřík, V., Poštolková, B., & Sochová, Z. (Eds.). (2003). *Slovník spisovné češtiny pro školu a veřejnost*. Academia.

SSJČ: Havránek, B., Bělič, J., Helcl, M. Křístek, V., & Trávníček, F. (Eds.). (1960–1971). *Slovník spisovného jazyka českého*. ČSAV.

Šindlerová, J., Štěpánková, B., & Andrén, I. L. (2023). Epistemická částice *zřejmě* pohledem paralelního korpusu. *Korpus – gramatika – axiologie, 27*, 37–52.

Štěpánková, B., Šindlerová, J., & Poláková, L. (2023). The Epistemic Marker *určitě* in the Light of Corpus Data. *Jazykovedný časopis, 74*, 130–139.

Veselovská, K., & Bojar, O. (2013). *Czech SubLex 1.0.* LINDAT/CLARIAH-CZ, ÚFAL.

Volkova, L. (2017). Pragmatic markers in dialogical discourse. *Lege artis. Language Yesterday, Today, Tomorrow*, *2*(1), 379–427.

Wiemer, B., & Stathi, K. (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.

Acknowledgements

The work has been supported by Czech Science Foundation under the project GA23-05240S and by Charles University Research Centre program No. 24/SSH/009. We gratefully thank Maarten Janssen, the main developer of TEITOK, for adapting the tool to meet the needs of this research.

Contact information

Barbora Štěpánková
Charles University, Prague
stepankova@ufal.mff.cuni.cz

Lucie Poláková
Charles University, Prague
polakova@ufal.mff.cuni.cz

Jana Šindlerová
Charles University, Prague
jana.sindlerova@ff.cuni.cz

Michal Novák
Charles University, Prague
mnovak@ufal.mff.cuni.cz