

THE EPISTEMIC MARKER *URČITĚ* IN THE LIGHT OF CORPUS DATA

BARBORA ŠTĚPÁNKOVÁ¹ – JANA ŠINDLEROVÁ²
– LUCIE POLÁKOVÁ¹

¹ Institute of Formal and Applied Linguistics, Faculty of Mathematics and Physics,
Charles University, Prague, Czech Republic

² Institute of the Czech National Corpus, Faculty of Arts, Charles University,
Prague, Czech Republic

ŠTĚPÁNKOVÁ, Barbora – ŠINDLEROVÁ, Jana – POLÁKOVÁ, Lucie:
The Epistemic Marker *určitě* in the Light of Corpus Data. *Journal of Linguistics*, 2023,
Vol. 74, No 1, pp. 130 – 139.

Abstract: The paper presents a pilot study for a research project on epistemic modality and/or evidentiality markers in Czech. The study focuses on the expression *určitě*. Although, this marker is typically considered to signal high certainty, the dictionary of standard Czech (Slovník spisovné češtiny, SSČ) also offers an alternative meaning of *probability*, indicating a lower degree of certainty. We use parallel data from the InterCorp v15 corpus to determine whether the probability meaning can be identified unequivocally in real language data and whether it correlates with specific translation equivalents, linguistic features, or lexical context. Based on our findings, we propose an alternative method for distinguishing between different shades of meaning based on the communicative functions of the utterances, and we draw conclusions regarding the relevance of individual grammatical and lexical clues in context for future annotations.

Keywords: epistemic markers, Czech, annotation, communicative functions, parallel corpus

1 INTRODUCTION

This paper presents a pilot study for a research project aimed at mapping the semantic space of Czech epistemic and evidential markers. One of the central issues we aim to address is how to approach the description of the meaning of epistemic/evidential markers.

Typically, the meaning¹ of markers expressing epistemic modality² is described in terms of a scale (Grepl – Nekula 2017) that covers different degrees of certainty, ranging from the highest certainty (paraphrased as *it is certain that x*) to the lowest

¹ For epistemic markers, we use the term *meaning* in the sense of the term *relational meaning* (cf. Filipec – Čermák 1985, p. 39).

² I.e., the ways of expressing the degree of certainty about a propositional meaning (cf. Grepl – Šimík 2017).

degree of certainty (e.g. *hardly*), with medium degrees covering a broad area from probability to possibility (cf. Kořenský et al. 1986, p. 233). While few epistemic particles are semantically unambiguous, most are described as “epistemically underspecified”, which means they can be used to signal different degrees of probability, making the interpretation of a text difficult and highly subjective (Ivanová 2017, pp. 242–243).

According to Grepl and Nekula (2017), the Czech word *určitě*³ is ranked third on the scale of certainty, i.e., even higher than other expressions usually considered representative of high certainty, such as *jistě* ‘for sure’, or *nepochybně* ‘undoubtedly’. However, the most recent complete and representative dictionary of standard Czech, SSČ, splits the meaning of *určitě* into two separate meanings that express different degrees of certainty. The first meaning listed expresses a high degree of certainty (synonymous with *jistě*), while the second meaning, marked as *hovor.* (colloquial), expresses probability, i.e., a medium degree of certainty (synonymous with *patrně* ‘apparently’):

SSČ⁴

určitě část. modál. ‘modal particle’

1. *vyj. jistotu, jistě*: zítra určitě přijdu; je to určitě on ‘expresses certainty, definitely: I will definitely come tomorrow; it’s definitely him’

2. *hovor. vyj. pravděpodobnost, patrně*: určitě na to zapomenu, připomeň mi to ‘colloq. expresses probability, likely: I’m likely to forget, do remind me’

The tendency to define epistemic modal markers solely based on the degree of certainty they express is also evident in the most recent attempt for a comprehensive dictionary of standard Czech (ASSČ), cf. *bezpochyby*.

Although, the latest lexicon definition of *určitě* distinguishes two separate meanings that claim a difference in the degree of certainty, the example sentence provided for the probability meaning does not offer enough clues to allow for its identification in real data. Our assumption is that the need to distinguish two (or more) meanings of *určitě* does not reflect only the different degree of certainty perceived, but also other linguistic features of the utterance and its context. These features then might be important for the constitution of individual meanings of the epistemic marker.

We pose the following research questions:

1. Can the medium-certainty meaning be distinguished from the full-certainty meaning in the corpus data:

³ The most common dictionary equivalents of this expression are *surely*, *certainly*, *definitely*, but other translations often appear in specific contexts, see Tab. 1.

⁴ Apart from the word class categorization as a modal particle, SSČ also offers *určitě* as an adverb, nested to the adjectival entry *určitý* ‘certain’.

- a) by comparing the translation equivalents of *určitě* in parallel data? Do translators in any context use medium certainty expressions (e.g. *probably*) as an equivalent to the original data?
 - b) by the mere native speaker annotator judgment provided for the Czech side of the data?
2. What aspects of an utterance support an unequivocal interpretation of the meaning of *určitě* and can thus be considered key points in defining the meaning of *určitě*?
 3. What aspects of an utterance can explain the differences in the meaning of *určitě* and are thus important to be captured in future annotations?

To answer these questions, we have designed a small-scale annotation experiment in which three expert annotators judged the status of *určitě* (or, more precisely, the status of the whole utterance modified by *určitě*) and searched for selected features in its context.

The design of the experiment is described in Section 2. Section 3 discusses the findings with 3.5 presenting an alternative approach to describing the meaning of epistemic markers, namely their communicative functions. Section 4 concludes and offers suggestions for our future work.

2 DATA AND METHODOLOGY

Our corpus study follows the assumption of Aijmer et al. (2006) that comparing data from parallel corpora may be helpful in determining the fine-grained meaning of pragmatic markers, especially because contextual factors become apparent. The interpretation of pragmatic markers is often difficult even for native speakers, due to the natural underspecification of the core meaning in the process of pragmaticalization. The translation process involves looking for functional correspondences within the given context, rather than pure lexical equivalents, providing thus an additional clue for the interpretation.

For this study, we use the Czech (Rosen et al. 2022) and English (Klégr et al. 2022) data of the parallel InterCorp v15 corpus. We use the core part of InterCorp, which consists mainly of fiction. These fictional texts, in contrast to the other parts of InterCorp (subtitles, parliamentary proceedings), represent a suitable basis for our research, as they are presumed to be close to spoken language, e.g. in terms of high frequencies of epistemic markers, a certain degree of subjectivity and elaborate situational context. It also can be assumed that literary translations of fictional texts would be of reliable quality.

For a sample of 300 concordances with Czech as the source language we provided an annotation of the type of meaning of *určitě*. First, we assessed whether the meaning was epistemic or something else (see Section 3). Second, for the epistemic markers, we annotated the degree of certainty perceived (see Section 3.2)

and, additionally, other linguistic features of the utterance that could potentially affect the interpretation of epistemic markers. The repertoire of the annotated features was compiled based on similar previous research (cf. e.g. Pietrandrea 2018, Wiemer – Stathi 2010). The following attributes were captured:

1. lexical environment – predicate lemma, another modal expression in near context (intensifiers, modal markers, modal verbs),
2. grammatical categories of the predicate verb – mood and tense,
3. word order and topic-focus articulation – position within a sentence, scope,
4. annotators’ comments – reflections on/arguments for the decision made plus potential other observations.

Separately,⁵ we manually identified the English translation equivalents and compared whether the translators tend to use epistemic markers with the degree of certainty corresponding to the annotator intuition about the original (see Section 3.1).

3 ANALYSIS

Out of the 300 instances of *určitě* in this annotation task, 238 were agreed on by all three annotators as signalling epistemic meaning. The remaining 62 cases fall into one of the three non-epistemic categories, adverbs, affirmative use, and other (e.g. irony). As we are mainly interested in the epistemic types of use, these are not further analyzed in this study.

3.1 Translation equivalents

For the 238 epistemic instances of *určitě* described earlier, we manually identified English lexical translation equivalents, if present. The distribution of those with frequency 4 and higher is given in Tab. 1.

The equivalents fall mainly among expressions typically indicating a high degree of certainty in English,⁶ translations indicating probability are rare; in 30 cases, there was no equivalent at all. The analysis of the translation equivalents therefore does **not** suggest that there is a discrete meaning with medium certainty in the data.

English equivalent of <i>určitě</i>	Frequency
<i>be bound to</i>	14
<i>be sure</i>	48

⁵ In order not to be influenced by the translation equivalents.

⁶ This claim is supported by an additional annotation of 750 cs > en and 750 en > cs parallel utterances containing *určitě* by a single annotator. Apart from using *probably* as the English translation equivalent (13x) and translating original *probably* to *určitě* (8x) in the opposite direction, there were just 2 instances of other medium-certainty markers (*perhaps*) in the original English data.

English equivalent of <i>určitě</i>	Frequency
<i>bet</i>	6
<i>certainly (almost certainly)</i>	39 (5)
<i>definitely</i>	31
<i>for sure</i>	7
<i>must</i>	13
<i>no doubt</i>	6
<i>probably</i>	4
<i>surely</i>	11
no equivalent	30

Tab. 1. The most frequent English translation equivalents of *určitě*

3.2 Annotation of the degrees of certainty

For the subsequent annotation tasks, the subset of 238 concordances detected as epistemic readings of *určitě* served as a basis. In this annotation task, the degree of certainty conveyed by the word *určitě* was assessed. We annotated the data according to the meaning split in SSČ (see Section 1), as either expressing high certainty or medium certainty, with a third possibility being “unknown”. Given the assumption regarding the degree of certainty of *určitě* articulated above, one might expect frequent disagreements in this respect. The resulting figures (Tab. 2) show that the annotators preferred the high certainty interpretation; the 103 cases agreed on by all three annotators stand for 43% of the data. In contrast, there was almost no agreement (3 cases) on the medium certainty interpretation. The annotations therefore suggest that, in line with expectations, *určitě* is largely perceived as a marker of the high degree of certainty, especially if also the high agreement of just two annotators is considered.

Type of agreement	Count
high certainty (3 agreed)	103
high certainty (2 of 3 agreed)	94
medium certainty (3 agreed)	3
medium certainty (2 of 3 agreed)	3
no agreement	35
total	238

Tab. 2. Agreement of three annotators on the degrees of certainty in the epistemic subset of *určitě*

This claim is further corroborated when we compare the annotators’ decision and the interpretation of the translators. As Tab. 3 demonstrates, out of the 103 cases

agreed upon by all three annotators as high-certainty markers, the chosen translation equivalents are in the vast majority (90 cases) also expressions typically perceived as high-certainty markers.⁷

Label	3 annotators agreed	3 annotators and translations agreed	Disagreed with translations	No translation equivalent
high certainty	103	90	3	11
medium certainty	3	2	0	1

Tab. 3. Agreement of 3 annotators and the translators’ choice of expression on the degrees of certainty in the epistemic subset of *určitě* in cs > en parallel data

On the other hand, the overall 66% annotator disagreement, with 35 cases even exhibiting complete disagreement (among all three annotators), indicates that there must be other perspectives to consider for the semantic interpretation of these markers (see Section 3.5).

3.3 Relevant interpretation features

The high level of disagreement on the degree of certainty for epistemic uses of *určitě* indicates a general high inconsistency in the perception of features relevant to a given annotation decision. In individual cases, however, specific context patterns or collocations can be discerned, which facilitate the interpretation of a given use.

For this analysis, we used the annotation of the attributes described in Section 2 in order to identify distinctive patterns and clues. The following linguistic features appear to be the most conclusive in our data (instances agreed on by all three annotators):

a) Typically, *určitě* can be considered a high-certainty marker in a **contrastive** relation to another epistemic marker with a different (lower) degree of certainty (1).

(1) *Vyplňuje celou čtvrtou komoru, **určitě** utlačuje mozeček a **možná** i kmen.*

‘It’s filled the whole fourth ventricle, **certainly** it’s pressing on the cerebellum and **probably** on the stem too.’

b) High-certainty interpretation is strengthened by the use of **another epistemic signal** with a high degree of certainty, typically the verb *muset* ‘must’. In (2), the original Czech utterance combines two high-certainty markers, which is mostly a redundancy. The English translation drops one of them.⁸

⁷ See footnote 6 and Tab. 1 for details on English high-certainty markers.

⁸ These cases are not to be confused with a co-occurrence of *určitě* with an objective meaning of *must* (necessity, duty): *Důstojně odchází, **určitě** se musí převléci a umýt.* ‘He strides off in a dignified manner, but he **must be in need** of a wash and a change of clothes.’

(2) *Určitě se musela podívat, že je celkem čistá.*

'She **must**'ve been amazed how clean she was.'

In contrast, a co-occurring epistemic signal of a medium certainty weakens the degree of certainty (e.g. *asi* 'might'). In the English translation of (3), by analogy, a medium-certainty marker ('might') is used.

(3) *To právě myslela ty moje potíže v hlavě a ty já teda asi určitě mám.*

'She was thinking about the troubles in my head and they **might** well be something that I do have.'⁹

c) The presence of an **amplifier**, e.g. *zcela, docela* (fully, completely, quite) usually confirms the high-certainty interpretation (4).

(4) *Musím si prokousnout ret, protože tentokrát bych se už docela určitě rozbřečela.*

'I had to bite my lip because this time I would **definitely** have burst into tears.'

Lit: '...I would **quite surely** have burst...'

On the contrary, the presence of a **downtoner** weakens the degree of certainty (5).

(5) *Hovořil s holohlavým suchým mužikem, skoro určitě to byl její otec, ani toho jsem živého nikdy neviděl...*

'He was talking to a bald-headed, wizened old man, **almost certainly** her father, whom I hadn't met either.'

Apart from the above mentioned contextual clues, grammatical features (such as sentence-initial position or future tense) also seem to show some interesting tendencies, nevertheless, larger data would be needed to make reliable claims about their impact.

3.4 Analysis of annotators' comments

Based on the annotators' notes, a number of findings have emerged that significantly affect the interpretation of texts but have not yet been systematically addressed here. These findings relate primarily to the areas of semantics and pragmatics.

a) There were several utterances in the data in which the combination of a predicate verb's meaning, past tense, and situational context elicited the reader's expectation of the absence of epistemic marking or the presence of an epistemic marker indicating lower certainty. For example, in (6) the use of a high certainty marker sounds illogical and raises doubts about the interpretation of the utterance.

(6) *Určitě jsem se do ní zamiloval. 'I'm sure I fell in love with her.'*

b) The annotators experienced difficulties when the perspectives of the author and the character in the text were blended. For the judgment, it was crucial to identify unequivocally who was evaluating the certainty of the statement, see (7). Note that the English translation interprets the perspective clearly.

⁹ One of the three examples in total uniformly annotated as medium certainty.

(7) ...byla ráda, že ten den nemusí vůbec nic důležitého nebo potřebného dělat, protože ten den by **určitě** zkazila cokoli, co by vyžadovalo být i jen kapku soustředění.
'...she was glad she had nothing important she needed to do that day, since **she was sure** she would have made a mess of anything that required even a drop of concentration.'

3.5 Communicative functions

The annotation of the degree of certainty did not appear conclusive considering the task of differentiating between distinct uses of the epistemic marker *určitě*. Nevertheless, the annotators were able to intuitively group similar uses on a different basis. It appears that the distinctive feature for describing the differences in use is the intention with which the speaker produces the utterance, i.e., its communicative function. This is not quite surprising since communicative functions are frequently mentioned in connection to epistemic modality (cf. Pietradrea 2018).

For *určitě*, the following types of communicative functions seem relevant¹⁰:

– assertive functions which can be specified as **assumptions**, regarding past, present or future events.

(8) *Ano, určitě tu je daleko víc lidí mladších než já.*

'Yes, definitely there's a lot more people younger than me around.'

Here the epistemic marker signals that the utterance is not presented as an objective fact, but as a (strong) personal belief of the speaker.

(self-)encouragement

(9) *Určitě se vše vyjasní.* 'I'm sure it'll all get cleared up.'

By using *určitě*, the speaker aims at convincing him/herself or the audience about a positive outcome of the situation.

worry

(10) *Sdělila otci, že Věrušku ušpiněná ručka stále svědí, stále si ji umývá a určitě dostane nějaký ekzém.*

'She told Father that her darling little Vera's soiled hand wouldn't stop itching, she kept washing it and was bound to get eczema.'

Here the author expresses negative predictions about the future with the secondary intention to make the audience prevent them.

accusations

(11) *Mláďi v hajzlu, revoluce v prdeli, rektor furt v Americe a ty mi určitě s někým zahejbáš!*

'Youth down the tubes, revolution up shit's creek, the dean still in America, and you are definitely cheating on me with someone else!'

– commissive functions which can be specified as **promises**.

(12) *Až to začne jít, dám ti určitě vědět.*

'As soon as it starts to go anywhere, I'll be sure to let you know.'

¹⁰ Here we follow the terminology and classification of Grepl et al. 1995, p. 585ff.

The speaker expresses his (strong) intentions to carry on some action in the future. *Určitě* serves for emphasis on the promise here.

- directive functions which can be specified as **persuasion**.

(13) *Jen se neostýchejte, určitě si chcete sochu podrobně prohlédnout.*

‘There’s no need to be shy. I am sure you would like to look at the statue more closely.’

In the persuasive contexts, *určitě* is a means of expressing polite urge on someone to act according to the intentions of the speaker.

As can be seen from this list, the degree of certainty is a dominant element of meaning in case of assumptions, while in other cases, it is overshadowed by other functions, e.g. with adding emphasis (fears, promises) or politeness (persuasion) to the statement. We believe that it is exactly the presence of another dominant aspect in the meaning that led the SSČ authors to add the medium certainty meaning. Their example *Určitě na to zapomenu, připomeň mi to* is a manifestation of worry with an explicit request to the partner to prevent the negative outcome.

4 CONCLUSION

Based on the annotations and subsequent analyses, the addressed research questions can be answered as follows:

Ad 1) We were able to document the postulated medium degree of certainty of *určitě* only very rarely in the studied dataset. Similarly, only two translation equivalents with medium certainty in the parallel English translations have been detected (*probably* and *perhaps*) both with low frequencies. The absolute majority of translation equivalents of the epistemic marker *určitě* express a high degree of certainty. In line with this, three annotators agreed on high-certainty meaning of *určitě* in 103 corpus instances, but only on 3 instances on medium-certainty meaning.

Ad 2) As for the aspects of an utterance facilitating the interpretation of the meaning of *určitě*, a distinctive role has been observed in the co-occurrence of *určitě* and other epistemic markers or intensifiers within an utterance – these expressions, in compliance with their own semantics, affect the degree of certainty in the given utterance. Another impactful factor are contrastive or comparative syntactic patterns with epistemic markers where different degrees of certainty are juxtaposed. Further relevant clues tend to appear in combinations, e.g. sentence-initial position and future tense. An analysis of such combinations is subject to more elaborate research, for which we intend to apply statistical methods and tools.

Ad 3) As suggested by the findings in 1 above, the core of the lexical meaning of *určitě* is high certainty. There is not enough evidence in the studied data to support the hypothesis on the existence of an individual meaning of medium certainty.

The analysis of the degree of certainty of the epistemic marker *určitě* has proven to be a less appropriate tool for a deeper description of the meanings and

various fine-grained functions of this expression in a text. It turns out that a more informative method of description, mapping the actual breadth of its use, is an analysis of the communicative functions that this expression may have in different contexts. However, the reliability of such a pragmatic approach also remains to be verified by empirical corpus experiments (i.e. annotation and annotator agreement).

ACKNOWLEDGEMENTS

The research has been supported by the Czech Science Foundation under the project GA23-05240S.

References

- Aijmer, K. et al. (2006). Pragmatic markers in translation: a methodological proposal. In K. Fischer (ed.): *Approaches to discourse particles*. Oxford: Elsevier, pages 101–114.
- Akademický slovník současné češtiny (2017–2023). [online]. Praha: ÚJČ AV ČR (ASSČ).
- Filipec, J., and Čermák, F. (1985). *Česká lexikologie*. Praha: Academia, 284 p.
- Filipec, J. et al. (2003). *Slovník spisovné češtiny pro školu a veřejnost*. Praha: Academia. (SSČ).
- Grepl, M. et al. (1995). *Příruční mluvnice češtiny*. Praha: Nakladatelství Lidové noviny.
- Grepl, M., and Nekula, M. (2017). Epistémická částice. In *CzechEncy – NESČ*.
- Grepl, M., and Šimík, R. (2017). Epistémická modalita. In *CzechEncy – NESČ*.
- Ivanová, M. (2017). *Modálnosť a modálne verbá v slovenčine*. Prešov: Filozofická fakulta Prešovskej univerzity, 310 p.
- Klégr, A. et al. (2022). *Korpus InterCorp – angličtina*, verze 14 z 31. 1. 2022. Praha: Ústav Českého národního korpusu FF UK. Accessible at: <http://www.korpus.cz>.
- Kořenský, J. et al. (1986). *Mluvnice češtiny 2*. Praha: Academia, 536 p.
- Pietrandrea, P. (2018). Epistemic constructions at work. A corpus study on spoken Italian dialogues. *Journal of Pragmatics*, 129, pages 171–191.
- Rosen, A. et al. (2022). *Korpus InterCorp – čeština*, verze 14 ze 31. 1. 2022. Praha: Ústav Českého národního korpusu FF UK. Accessible at: <http://www.korpus.cz>.
- Wiemer, B., and 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, pages 275–289.