ISSUES OF POS TAGGING OF THE (DIACHRONIC) CORPUS OF CZECH: PREPARING A MORPHOLOGICAL DICTIONARY

ANNA ŘEHOŘKOVÁ
Institute of the Czech National Corpus, Charles University, Prague, Czech Republic

ŘEHOŘKOVÁ, Anna: Issues of POS Tagging of the (Diachronic) Corpus of Czech: Preparing a Morphological Dictionary. Journal of Linguistics, 2017, Vol. 68, No 2, pp. 316 – 325.

Abstract: Many important decisions concerning the part-of-speech categorization remain unexplained in the current practice, only reported in corpus manuals. The aim of this paper is to offer a different perspective on the problems of morphological annotation of corpora – the perspective of mapping and analyzing conceptual problems in the annotation. Focused mainly on function words in Czech, we discuss the possibilities of the POS tagging of the inherently ambiguous category of particles and we introduce criteria for distinguishing particles from interjections.

Keywords: corpus, function words, morphological annotation, Czech

1 INTRODUCTION

The motivation of this paper is to share the experience from the preparation of a new diachronic corpus of Czech, covering the 19th century. Dealing with shifts and changes in the older language, where the lack of native speaker knowledge is perceptible, led us to rethink the principles of morphological annotation, concerning function words in particular, and to seek for inspiration in other corpora (cf. [2]).

Words considered as secondary prepositions, conjunctions, adverbs, particles and interjections, namely all those that have undergone a grammaticalization and conventionalization process, are often difficult to classify. Clues provided by grammars and dictionaries turned out to be insufficient for corpus annotation where every token needs to be tagged. For example, in the Oxford English Dictionary and elsewhere, prepositional, adverbial and conjunctional use of *notwithstanding* is distinguished, the adverbial one according to the meaning ‚nevertheless, all the same' (*he must be told, notwithstanding*). On the contrary, the annotation of the BNC2 corpus is based on contextual features which are recognizable to the automatic tagger, and therefore it is the instances that come after an NP and precede punctuation that are mostly tagged as adverbs:

(1) The author *notwithstanding*, many conclusions can be drawn from this steel-trap of a book [...]

According to the OED, though, (1) is an example of a preposition (used postpositively) meaning ‚in spite of'. Thus, it seems that the adverbial category might have been redefined in the corpus with respect to the formal recognizability of

DOI 10.1515/jazcas-2017-0041

the word in context.[1] Nevertheless, cases like (2), (3) and (4) still can be found where the sentence has the same structure but the word is tagged in three different ways (AV0 - general adverb, PRP – preposition, PRP-CJS: the ambiguity tag for preposition/conjunction):

(2) AV0: *Notwithstanding* all these problems, the bank has kept faith with us [...]
(3) PRP: *Notwithstanding* this promise, the use of road pricing to change travel habits still seems some way off.
(4) PRP-CJS: *Notwithstanding* the re-election of Mrs Thatcher in 1983 and 1987, a clear majority of voters have favoured increased taxes [...]

These examples indicate the complexity of interfaces between various function words. In this article we will focus on the case of particles in Czech.

2 PARTICLES VERSUS OTHER PARTS OF SPEECH

In Czech grammatical theory, particles were not fully recognized as a part of speech until the 1980s [16]. The oldest contemporary grammar [7] introduced a wide and heterogeneous category of adverbs, consisting of content words as well as function words, including idiosyncratic cases like *ne* 'no'. This grammar became a widely used school book and a base for part-of-speech classification in dictionaries of Czech ([8], [14], [19]). Later [13] the definition of adverbs was refined and only clause constituents were considered adverbs, the others being classified as particles (e. g. *snad* 'perhaps' which does not bring any information about the circumstances of the action expressed by a verb and, therefore, unlike other adverbials, can not be used as an answer to any question about the action – how? when? etc.). Interestingly, this criterion was not accepted by Quirk et al. ([20]) who argue that all adverbials (unlike objects, complements etc.) are optional elements to the structure of a clause. Furthermore, in the Czech tradition not only adverbs but also conjunctions, pronouns, nouns, verbs or even phrases have been viewed as particles in cases where they displayed signs of semantic bleaching and/or a shift in their function towards pragmatics of interaction (cf. [6]). Thus, particles, instead of adverbs, became a new heterogeneous category and, in addition, the identification of many of its instances became context-dependent.

2.1 Identification of Particles

To our knowledge, there is no universal criterion for defining particles, except the negative one (a non-declined word which is not a conjunction, an adverb, a preposition nor an interjection). In an attempt to define this category on a functional basis, several sets of subcategories have already been proposed and the research remains ongoing (see [16] for an extensive enumeration). Bearing in mind a practical goal of the delimitation of particles, we chose a bottom-up approach: the first step was to compile a list of particle candidates based on example words obtained from grammars and related works ([22], [5], [3], [10], [13], [9]) and on lists of words tagged as particles (CNC – SYN2015, SNK – prim-7.0) or as similar classes (ATT,

[1] The Collins COBUILD Dictionary, based also on a corpus, probably introduced such an interpretation for the first time.

CM and MOD functor in the Prague Dependency Treebank 3.0) in corpora. In the next step, the items were sorted approximately according to prominent features they had in common, in relation to their function. Inspired by previously suggested subclasses (namely by [13], [10]), we built a generalized system which integrates commonly used perspectives. With many overlaps between the groups, we identified particles:

1. structuring discourse and/or information in an utterance (sentence adverbials, restrictive particles):
   mimochodem 'by the way', *obzvlášť* 'particularly, especially', *ostatně* 'anyway', *také* 'also'
2. indicating sentence mood/type or its illocutionary function (questions, wishes, appeals, threats etc.), often adding expressivity:
   ***Kéž** bych měla dítě* '**If only** I had a child' (CNC – InterCorp v9)
   ***Běda, jestli** za to můžeš ty* ‚This had **better not** be your fault' (CNC – InterCorp v9)
3. implying a presupposition:
   ***ještě** větší* '**even** bigger' (assuming smaller)
   *to je **teprve** začátek* 'that's **just** the start' (despite the assumption that nothing more is to come, CNC – InterCorp v9)
4. commenting on a proposition and its wording, in terms of modality, emotions or attitude (hedges, amplifiers, emphatics):
   asi 'perhaps', *jaksi* 'somehow', *naštěstí* 'fortunately', *naprosto* 'absolutely', *opravdu* 'really'
5. expressing affirmation and negation:
   ***Pravda**, ale nemáme na vybranou*. '**True**, but we have no choice.' (CNC – InterCorp v9)
   *žádné plachty, **kdepak*** 'no oars, **nay**' (CNC – InterCorp v9)
6. serving as fillers:
   tentononc 'whatsit', *jako* 'like (colloquial)'

The list of particle candidates was further refined. Firstly, since our goal is to tag texts from the 19th century, we checked the items against the first modern dictionary of Czech ([19], 1935–1957), which captures the language of classic writers of the period in question, and removed words that started to be used as pragmatic devices only later (e. g. *prakticky* 'practically, basically') and also foreign words (e. g. *apropos*) due to the unknown degree of their integration into Czech vocabulary. Secondly, we extracted older derived forms, variants and synonyms from the dictionary using the categories obtained from the list.

The main decisions made throughout the whole procedure concerned the extent to which we should adhere to the criterion of function. This criterion goes across established boundaries of parts of speech, and when there is no additional feature distinguishing particles from the other classes, as mentioned above, the decision about the inclusion or exclusion of particular words can be made only on the basis of convention and with respect to a practical purpose. For example, we did not include many of the words which specify the intensity of particular action or quality (degree

adverbs in Czech school tradition, e. g. *velmi* 'very', ***moc** hezký* '**pretty** good', ***strašně** dobře* '**awful** good') into the fourth subclass because such intensifiers are largely metonymy- and metaphor-based and therefore still productive. The subclass would thus be unpredictably extensive (cf. **hodinářsky** přesná práce, lit. 'watchmaker.ADV accurate work', 'very accurate'). We chose only the words explicitly expressing the highest/lowest grade of intensity, which also function as rheme indicators in an utterance (e. g. *maximálně* 'maximally, a maximum of'). Similarly, we distinguished between two types of "commenting words" (*hlavně* 'mainly' vs *většinou* 'mostly') according to the difference between "limit" and "degree". When a borderline case occurred (e. g. *nadmíru* 'above the line', 'extraordinarily'), we tended to make a decision according to the semantics of the word (*nadmíru* refers to the usualness rather than to the highest extent, and therefore we classified it as an adverb). The overall aim thus was not to come up with the one and only right set of principles to identify particles but to keep them as a category „for the remaining cases" while understanding what makes them different (and which cases can be still counted as less typical representatives of other parts of speech).

2.2 The Estimate of Particle Ambiguity

Having adjusted the compiled list to 19th century language, we arrived at a final list (further referred to as P-list and P-words) consisting of more than 500 items (available at `https://trnka.korpus.cz/~zitova/`). This number was quite surprising given that the list obtained from the CNC – SYN2015 contains 214 items (excluding words with hyphens that were incorrectly tagged as particles) and even the more extensive list from the SNK – prim-7.0 comprises 374 items.[2] We would also expect more particles identified in newer texts than in the older ones given a general shift towards oral discourse during the time (cf. [21]: 254). Our assumption is that the class of particles is intentionally maintained rather small to leave out words with multiple morphological interpretation. Therefore, to estimate the ambiguity rate of the items in the P-list and to map the approach to tagging particles in the corpus of present-day Czech, we tested the P-list against the CNC – SYN2015 corpus.

We used a multi-level frequency distribution function of the KonText interface to get a list of matched words and their tags. Despite the adjustments of the P-list to the older language, the vast majority of words was found in the corpus (442 items of the original 512). Words with more than one part-of-speech tag were counted as ambiguous.

	particles	%	non-particles	%
ambiguous	67	48.55	35	11.55
unambiguous	71	51.45	268	88.45
total	138	100	303	100

Tab. 1. Part of speech assigned to the words from the P-list in the CNC – SYN2015 corpus

[2] We found also 395 particles in the CNC - Prague Spoken Corpus but the list largely consists of phonological variants of a limited set of words, preserved in the transcription.

As can be seen from Table 1, roughly a half of the P-words tagged as particles in the CNC – SYN2015 is, according to the tagging scheme, homonymous with representatives of another part of speech. On the contrary, almost 90% of the P-words not tagged as particles do not need to be disambiguated in the context. The tendency to somewhat avoid particles in the POS tagging is thus understandable given its ambiguity rate. In most cases, particles are homonymous with adverbs: 66% of ambiguous particles (44 out of 67) also have an adverbial interpretation and adverbs represent 68% of non-particles in this analysis (183 out of 268, the rest is accounted for by 8 other POS).

It is precisely the difference between particles and adverbs that is most difficult to recognize. Examples 5 and 6 show one of the cases that are fairly impossible to distinguish for an automatic tagger (stochastic or rule-based), example 7 poses a problem even for a human:

(5) *"Uzavřeme sázku," řekl Lukáš. [...] "**Dobře**," řekl nakonec [Richard]*. "Let's make a bet," said Lucas. - "**Alright** then," said Richard. (CNC – SYN2015, affirmative particle in Czech)

(6) *"Jak se ti vede?" - "**Dobře**."* "How are you?" "I'm **fine**." (CNC – SYN2015, adverb in Czech)

(7) *hebrejština se normálně píše zprava doleva, ale átbaš můžeme **jednoduše** použít i takto* (CNC – InterCorp v9)

‚Hebrew is normally written in the opposite direction, but we can just as **easily** use Atbash this way' (adverb in Czech)

‚Hebrew is normally written in the opposite direction, but **in short**, we can use Atbash this way' (alternative interpretation; discourse-structuring particle in Czech)

Thus it seems recommendable not to integrate particles into the morphological tagging scheme unless there is a possibility of their manual disambiguation (and even in that case only with certain restrictions, see section 2.4). Standard dictionaries of Czech, containing example sentences or phrases, continue the tradition of treating such words as adverbs probably for similar reasons. Another option is to introduce ambiguity tags with information about the probable accuracy in large corpora which, however, presupposes at least the identification of the typical cases in their contexts.

2.3 The Current State of the Tagging of Particles in the CNC – SYN2015

Concerning the original set (214 items after refinement), 42% of particles have more than one tag which is less than in the case of the P-list. Nevertheless, we have found certain inconsistency in the tagging scheme. The original set contains also salutations and swear words (e. g. *ahoj* 'Hi!', *kčertu* 'Damn!', *ježíši* 'Jesus!') which are traditionally regarded as interjections (cf. [1]). It is to be said, though, that the difference between interjections and particles is not always clear (cf. category names like "particles of contact" and "particles of emotions" [13]). We will focus on this issue in section 3.

Overall, there does not seem to be any function-based conception of particles behind the CNC - SYN2015. Candidate words have thus been probably assessed independently, as can be seen from the different tagging of close variants and synonyms, e. g. *nejspíš* and *nejspíše*, both meaning 'probably' (1. adverb or particle,

distinguished without any obvious contextual clue by the stochastic module of the tagger; 2. adverb only), *opravdu* and *doopravdy*, both 'really, truly' (the same case) or *bezesporu* and *nepochybně*, both 'undoubtely, certainly' (1. particle, 2. adverb). Although there certainly exist some different features of contexts of these words, they are rather subtle or their importance for the POS categorization is questionable (e. g. there are 60,22 i.p.m. of *opravdu* before an adjective, whereas only 2,58 i.p.m. of *doopravdy* in the same position in the corpus, however, this has not been recognized as an important feature yet).

Another consequence of the lack of conception is the uncertain boundary between particles, adverbs and conjunctions. For example, *vždyť* 'after all; because', *však* 'well; however, though' and *přece* 'surely, after all; though' are all able to express a syntactic relationship as well as a pragmatic meaning but they are tagged differently (1. conjunction only, 2. conjunction or particle, 3. adverb or particle). We deal with this issue in the next section.

2.4 Particle as a Functional Attribute

Trying to avoid loss of information about the pragmatics of texts (which comes with using adverbial tags only) on the one hand and unreliability of tagging on the other, we suggest to follow a morphological criterion first (almost every particle is morphologically an adverb, having similar affixes etc.), as the dictionaries usually do, and then to optionally add information about the function of such an adverb, which can be not only pragmatic but also syntactic (connective), as mentioned above. As examples 8 and 9 show, the same word can have different functions and none of them is typical for adverbs (primarily used to denote circumstances) to which it points with its formation (the suffix *-ak* occurs also in *tak*, *jinak* and a few other adverbs).

(8) ***Však** víte*. '**Well**, you know.' (CNC – InterCorp v9, pragmatic)

(9) *...první večer padla volba na ni. Nazítří ráno **však** došlo ke změně* ‚...for the first evening she was his settled choice. The next morning, **however**, made an alteration...' (CNC – InterCorp v9, syntactic)

Tagging the first case as an adverb serving as a particle due to its pragmatic function (ADV + PART) and the second case as an adverb with a connective function (ADV + CONJ) allows us to avoid the difficult clear-cut decision whether the word *však* is still a particle when it connects two adjacent utterances (should we conceive it as a discourse-structuring particle, to keep the interpretation close to its other usage, as an adverbial connector or as a conjunction?). This manner of annotation also enables us to capture the connective function of traditional adverbs like *přesto* (lit. 'over it', 'yet, still, however'), *proto* (lit. 'for it', 'therefore') etc. which can not only modify a conjunction but also substitute it, so they are partially grammaticalized as connective devices.

The introduction of multiple tags, however, also presupposes clear rules for their application. For example, when there is a collision between pragmatic and syntactic function (e. g. *vždyť* indicating reproach and marking an explicative relationship at the same time in some cases), there are at least two possible solutions: 1. the pragmatic function (ADV + PART) will be given precedence for the relationship between the two utterances is implied by their propositions and does not

need to be expressed overtly (explication is based on a partial reformulation of the previous proposition; more on the nature of such relationships in [18]); 2. a new tag (e. g. ADV + MIX1) will be introduced to denote this combination (to avoid a triplet of tags), which seems to capture the nature of the problem more accurately. Nevertheless, despite the difficulties with setting rules, this system allows more space to deal with problematic cases than a single-tag solution and well documented rules will be informative both for the users of the corpus and for an automatic tagger.

2.5 The Interface Between Particles and Interjections

Words, that can be found included either in the category of particles or interjections, are especially response words, *ano* 'yes' and *ne* 'no'. As opposed to our view in 2.1 (also e. g. [10]), which conforms to the school tradition, some Czech papers ([5], [23]) argue that *ano*, *ne* are interjections due to the criterion of forming independent non-elliptical utterances (cf. [1], [20]). Cvrček et al. ([5]) mention *ne* along with content words used in rejections (cf. example 10 and 11). As interjections are supposed to be closer to content words than particles, the analogy with content words of rejection would support the view that *ne* is an interjection.

(10) *Jseš na flámu, bejby? – **Hovno**, já jsem na flámu pořád.* ‚You been partyin', baby? **Shit**, I been partyin' all the time.' (CNC – InterCorp v9)

(11) *Mrzí mě to. – Ale, **houby** se stalo.* 'I'm sorry about that. – Hey, **shit** happens.' (lit. 'mushrooms', CNC – InterCorp v9)

On the other hand, interjections are also supposed to express a rudimentary proposition which should be paraphrasable (e. g. *Ouch, it hurts!*) and it is hard to imagine how to paraphrase *ne* otherwise than by repeating the previous utterance (usually a question), only with the negative polarity. The non-elliptical nature of *ne* is thus questionable.

Example 11 is further complicated by the fact that *houby* 'shit', originally a noun, is a clause constituent which is untypical both for particles and interjections. Although Komárek et al. ([13]) and Kleňhová ([11]) argue that interjections can perform a function of any other part of speech in the clause structure (with an implicit reference to their primarily independent use), the concept of secondary interjection in its secondary function, which would be the case here, seems too complex. As in section 2.4, we prefer to tag the word according to its morphology first (NOUN) and its function second (PART). It is obvious that the word was reanalysed as uninflected thanks to the homonymy of its inflectional suffix *-y* with an adverbial suffix *-y* (*hovn-o* is the same case).

Overall, it seems that the devices of negation and affirmation should be conceived as particles rather than as interjections. Besides the reanalysed cases mentioned above, there may certainly be a problem with disambiguation of sound-like words like *hm* (does it express a response, hesitation or something else?) and even with *ano*, *ne* 'yes, no' expressing emotional reaction to an event (success, loss etc., cf. below). The most appropriate solution seems to be to tag them as borderline cases between particles and interjections, though it indicates a need for another type of multiple tag: the OR tag (different from the AND tag suggested in section 2.4), denoting two competing interpretations.

Particles expressing emotional comment on the formulation of an utterance are closely related to interjections. In an attempt to distinguish between them, Vondráček ([23]) proposed to follow the criterion of syntactic independency (examples are taken from [23]):

(12) ***Bohužel*** *se ještě nevyjádřila* 'She **unfortunately** has not commented on it yet' (PART)

(13) ***Bohužel****, ještě se nevyjádřila.* '**Unfortunately**, she has not commented on it yet' (INTJ)

Unlike the English equivalent of *bohužel* 'unfortunately', the Czech expression can be either intervowen with the structure of a clause through a change in the word order of enclitics (e. g. *se* above) or separated by a comma as an independent element. When the word is separated, Vondráček draws a parallel with interjections and their paraphrases. However, it remains unclear what to do with clauses without such a change in the word order (in 14, the enclitic *tam* 'there' stays in Wackernagel's position):

(14) *Dámy tam,* ***bohužel****, přístup nemají* 'Ladies, **unfortunately**, are not allowed to enter there' (CNC – SYN2015)

Furthermore, graphically separated occurrences of *bohužel* are quite infrequent and may thus be the result of a stylistic rather than a functional variation. Examining the frequency of such occurrences in related particles of emotions (*naštěstí* 'fortunately'; *naneštěstí* 'unfortunately'; *díkybohu*, *bohudíky*, *bohudík*, *chválabohu*, *zaplaťpánbůh* 'thank God'), however, we found substantial differences between particular words indicating that relying purely on the analogy with one of them could be misleading.

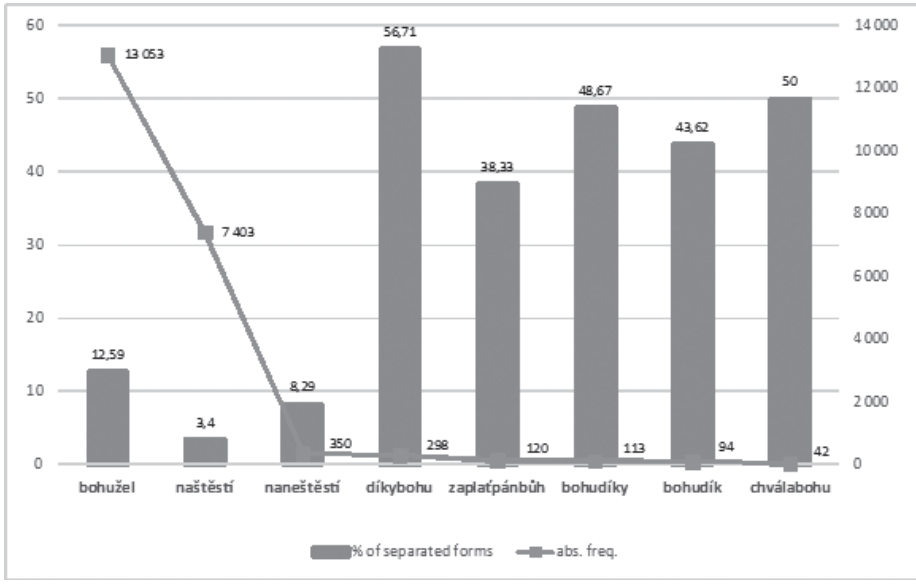

Fig. 1. The percentage of graphically separated particles and their absolute frequency in the CNC – SYN2015

As can be seen from Graph 1, a group of compound words with the element *-bůh*, *-bohu* ('God') besides another noun or verbal element tend to be separated

more often than the others, unless they are too frequent (as is the case of *bohužel*). On the other hand, *naneštěstí* (lit. 'to unhapiness', with a prepositional element), though rather infrequent, is mostly accepted to a clause structure. Word formation and frequency thus have an impact on whether a word is perceived as an integral part of a clause (and therefore should not constitute a truly non-elliptical utterance) or still as a parenthesis. Given that various stages of conventionalization are visible even in contemporary language, let alone the older periods, when the word is graphically separated, we suggest to tag it 1. as adverb due to the compound form, 2. both interjection and particle (e. g. ADV + MIX2).

3 CONCLUSION

Showing problematic cases of function words, we aimed to draw attention to theoretical backgrounds of morphological annotation of texts in corpora. The analysis of the corpus of present-day Czech allowed us to considered the complexity of including the category of particles into a tagging scheme and we arrived at a recommendation not to apply this category to large and automatically tagged corpora because of a high rate of ambiguity of respective words. Inspired by the BNC2 and Czech dictionaries, we recommend rather the extensive use of the category of adverbs and the application of ambiguity tags. This seem to be reasonable also for the diachronic corpus of Czech in preparation because of the language change that affects this pragmatic means considerably. The basic interpretation of word forms should lean on formal morphology and word formation and then attributes of particular function should be added if the word is listed in a list of functionally-conventionally defined particles. When such a word has also a clause-linking function, it should be given also a tag for conjunction. Multiple tags and tags with attributes seems to be the right mean to tackle the problem of categorization of scalar phenomena like those of language.

References

[1] Ameka, F. (2006). Interjections. In Brown, K., editor, *Encyclopaedia of Language and Linguistics*, pages 743–746, Elsevier, Amsterdam.

[2] Atwell, E. S. (2008). Development of tag sets for part-of-speech tagging. In Ludeling, A. and Kyto, M., editors, *Corpus Linguistics: An International Handbook*, Volume 1, pages 501–526, Walter de Gruyter.

[3] Bedřichová, Z. (2008). Částice implikující presupozici jako podstatná složka větného významu. *Čeština doma a ve světě* 3–4:119–126.

[4] *Collins COBUILD Dictionary.* Accessible at: `http://collinsdictionary.com`, retrieved 2017-03-20.

[5] Cvrček, V. et al. (2010). *Mluvnice současné češtiny: Jak se píše a jak se mluví*, Karolinum, Praha.

[6] Grepl, M. (1989). Partikulizace v češtině. *Jazykovědné aktuality*, 26:95–100.

[7] Havránek, B. and Jedlička, A. (1960). *Česká mluvnice*. Státní pedagogické nakladatelství, Praha.

[8] Havránek, B. et al., editor (1989). *Slovník spisovného jazyka českého*. Academia, Praha.

[9] Hoffmannová, J. (1983). *Sémantické a pragmatické aspekty koherence textu.* Ústav pro jazyk český ČSAV, Praha.

[10] Karlík, P., Nekula, M., and Rusínová, Z. (1995). *Příruční mluvnice češtiny*. Nakladatelství Lidové noviny, Praha.

[11] Kleňhová, E. (2011). Pojetí interjekcí v některých českých mluvnicích. *Naše řeč*, 94(5):242–255.
[12] Kleňhová, E. (2012). Postavení a užívání interjekcí v současné češtině. *Naše řeč*, 95(5):238–254.
[13] Komárek, M. et al. (1986). *Mluvnice češtiny: vysokoškolská učebnice pro studenty filozofických a pedagogických fakult, aprobace český jazyk. [Díl] 2*. *Tvarosloví*. Academia, Praha.
[14] Kroupová, L. et al. (eds.) *Slovník spisovné češtiny pro školu a veřejnost: s Dodatkem Ministerstva školství, mládeže a tělovýchovy České republiky*. Academia, Praha.
[15] Milička, J. (2013). Bootstrapper [software]. Accessible at: `http://milicka.cz/en/bootstrapper`.
[16] Nekula, M. (2017). Částice. In Karlík, P., Nekula, M., and Pleskalová, J., editors, *Nový encyklopedický slovník češtiny*. Accessible at: `https://www.czechency.org/slovnik/ČÁSTICE`, retrieved 2017-03-25.
[17] *Oxford English Dictionary.* Accessible at: `http://www.oed.com`, retrieved 2017-03-20.
[18] Poláková L. et al. (2012). *Manual for Annotation of Discourse Relations in the Prague Dependency Treebank.* Technical Report No. 47, ÚFAL, Charles University, Prague. Accessible at: `http://ufal.mff.cuni.cz/discourse/publications`.
[19] *Příruční slovník jazyka českého* (1935-1957). Státní nakladatelství, Praha.
[20] Quirk, R., Greenbaum, S., Leech, G., and Svartvik, J. (1985). A Comprehensive Grammar of the English language. Longman, London and New York.
[21] Reppen, R., Fitzmaurice, S. M., and Biber, D., editors (2002). *Using corpora to explore linguistic variation* (Vol. 9). John Benjamins Publishing.
[22] Štícha, F. et al. (2013). *Akademická gramatika spisovné češtiny.* Academia, Praha.
[23] Vondráček, M. (1998). Citoslovce a částice – hranice slovního druhu. *Naše řeč*, 81(1):29–37.