sciendo
DOI 10.2478/jazcas-2025-0015

ANNOTATING MOOD, TENSE AND VOICE IN CZECH CORPORA

TOMÁŠ JELÍNEK
Department of Linguistics, Faculty of Arts, Charles University, Prague,
Czech Republic (ORCID: 0000-0002-8521-4715)

JELÍNEK, Tomáš: Annotating Mood, Tense and Voice in Czech Corpora. Journal of Linguistics, 2025, Vol. 76, No 1, pp. 167 – 176.

Abstract: In the corpora of the Czech National Corpus, the verbtag attribute was introduced for annotating the mood, voice, person, and tense of both simple and compound verb forms. This article presents this attribute and the process of its automatic annotation. We then show an example of the use of verbtag in corpus research by comparing five different text genres in terms of verbal categories expressed in this attribute.

Keywords: Czech morphology, verbs, mood, tense, automatic annotation

1 INTRODUCTION

In contemporary linguistic corpora, users usually have access to lemmatization and morphological annotation of words, which facilitates their work with the corpus. However, morphological annotation is typically limited to individual, isolated forms: while the determination of morphological categories occurs within the context of the entire sentence, the tags apply only to single forms. For example, a token might be marked as a past participle of a verb, but the mood, tense, and voice of compound verb forms are not indicated.

For the corpora of the Czech National Corpus (CNC), we decided to address this shortcoming by introducing a new attribute for tagging the morphosyntactic properties of verb forms, both simple and compound, which was named **verbtag**. Verbtag distinguishes whether a verb is auxiliary or not, and for full verbs, mood, voice, person, number, and tense are annotated. The attribute was first implemented for written corpora starting with the SYN2020 corpus; later it was extended to spoken corpora, beginning with the Ortofon_v3 corpus. This article discusses the motivation for introducing verbtag, the process of its automatic annotation, and demonstrates how verbtag can be used in corpus research on statistics of verb properties in various CNC corpora.

2 RELATED WORK

In recent years, several authors have focused on the automatic annotation of mood, tense, and voice, often with the aim of improving performance in subsequent

NLP tasks such as machine translation. Loáiciga et al. (2014) annotated a small parallel English-French subcorpus from the Europarl corpus with tense, aspect, and mode using a syntactic parser and custom rules. They then trained a tense predictor on this subcorpus, achieving improvements in machine translation. Ramm et al. (2017) also annotated mood, tense, and voice using a syntactic parser and subsequent rules to enhance abstract meaning representation. Myers and Palmer (2019) used the same data with a neural-network-based classifier without prior annotation and achieved significantly better results than Ramm. Recent advances in the use of deep learning in NLP have eliminated the need for verb category annotation as an intermediate step in tasks such as machine translation.

There are few corpora with annotations of mood, tense, and other verb properties, and they are usually small. Ramm worked with the English PropBank corpus (Palmer et al. 2005), which is the Penn Treebank corpus enriched in semantic role labeling and verb categories such as tense, but it only has about 180,000 tokens. The tectogrammatical layer of the Prague Dependency Treebank in Czech is larger, with approximately 675,000 tokens, where full verbs are assigned grammatemes corresponding to verb tense, mood, etc.

In the largest current project of multilingual corpora with comparable annotation, Universal Dependencies (Marneffe et al. 2021), properties such as mood, tense, or voice are considered, but they are assigned only to isolated forms. For example, the participle in English, French, or Czech in the phrase *will be saved* / *sera sauvée* / *bude zachráněna* does not have indicative mood or future tense specified in the feats attribute. In English, it is annotated as past participle used in a passive construction, in French as past participle, and in Czech as passive participle. Only for the English version, the annotation contains a feature (voice) derived from the use of an auxiliary verb.

3 THE VERBTAG ATTRIBUTE IN CNC

3.1 Motivation

Verbs in Czech form both simple (e.g. *jdu* ‘I'm going’) and compound forms (e.g. *byl bych šel* ‘I would have gone’). The full-verb part of a compound form can include active participles (e.g. *přišel* ‘came’), passive participles (e.g. *zachráněn* ‘saved’), as well as the infinitive in the compound future tense (e.g. *chodit* in *budu chodit* ‘I will walk’). In compound forms, some morphosyntactic features of verbs are carried by auxiliary verbs (e.g. person), others by the main verb forms, and still others by the choice of the specific compound form. For instance, in the form *přišla byste* ‘you would come’, which is a polite form of 2nd person singular of the present conditional, mood and voice (active conditional) follow from the entire form (i.e., the conditional form of the verb *být* ‘to be’ and the past participle), the tense from the absence of another auxiliary verb in the past tense, the number and gender from the participle form, and the politeness from the number of the auxiliary verb (which differs from the number of the participle). Given the relatively free word order and frequent use of embedded subordinate clauses in written Czech, the individual parts of a compound form can be far apart: it is possible to find comprehensible Czech sentences where thirty other tokens stand between the auxiliary verb and the participle. Without annotation focused on verb categories in compound verb forms, it would be difficult for corpus users to determine these properties using corpus queries alone. Feedback from users indicated a demand for this information. Therefore, a new attribute was conceived which supplements the existing morphological tag with information derived from the entire verb form: it has been named **verbtag**.

3.2 The verbtag attribute

The verbtag attribute has been described elsewhere, e.g. (Jelínek et al. 2021), but for understanding this article, it is necessary to be familiar with it, so we will briefly summarize its properties here. The verbtag attribute is a six-position tag that supplements the original fifteen-position morphological tag. It is relevant only for verbs; for other parts of speech, all positions are empty.[1]

3.2.1 Full verb or auxiliary

The first position of the verbtag specifies whether the verb is auxiliary (A) or full (V). Only forms of the verb *být* 'to be' are considered auxiliary, rather than, e.g. *mít* 'to have'. For auxiliary verbs, all other positions of the verbtag are empty. The auxiliary verb *být* appears in combinations like *četl jsem* 'I have read', *budu číst* 'I will read', *byl bych četl* 'I would have read'; the verb *být* is considered a full verb both as existential (e.g. *Bůh je*. 'God is.') and copular (e.g. *Opak je pravdou.* 'The opposite is the truth').

3.2.2 Mood

The second position distinguishes mood: indicative (D), conditional (C), imperative (I), infinitive (F), transgressive (T), and passive participle not forming a compound verb form (O). The character O stands for "other uses" of the passive participle, such as cases when it stands alone, typically as a predicative complement (e.g. *Hořce* ***zklamán*** *se vrací do ateliéru.* 'Bitterly disappointed, he returns to his studio.'), in sentence segments without a predicate (*Cyklisté* ***vítáni****.* 'Cyclists welcome'), and often in sentences with verbs *mít* 'to have' and *zůstat* 'to stay' (*V této oblasti* ***máme rozpracováno*** *několik iniciativ.* 'We have several initiatives underway in this area.').

3.2.3 Voice

The third position indicates voice: active (A) or passive (P), with passive referring only to the periphrastic passive (*důraz je kladen* 'emphasis is placed'), not to the reflexive passive (*důraz se klade* 'emphasis is placed').

[1] Adjectives derived from passive participles are assigned 'p' on the third position of the verbtag.

3.2.4 Person

The fourth position distinguishes person (1, 2, 3, -). In some rare cases, a syntactic construction may cause a conflict between the person expressed in the morphological tag and the person in the verbtag, such as in the phrase *Bůh suď!* 'God be the judge!' where the imperative has the 2nd person in the tag, whereas the 3rd person in verbtag.

3.2.5 Number

The fifth position indicates number: singular (S), plural (P), and the form of politeness (v); the form of politeness is identified only in the combination of an auxiliary verb with a past or passive participle (e.g. *řekla jste* 'you said'), where the number of the auxiliary verb differs from the number of the participle.

3.2.6 Tense

The last, sixth position indicates tense: pluperfect (Q), past (R), present (P), future (F), and present or future of biaspectual verbs (B). The pluperfect, as found in sentences such as *Víno, jemuž dávno byl odvykl, uvolnilo nyní cenzuru myšlenek i slov.* 'The wine he had given up a long time ago now loosened the censorship of thoughts and words.' is rarely used in Czech. In the texts we annotated automatically, cases where this tense is incorrectly determined due to a text error (such as a typo or a missing comma) are much more frequent than the correct ones.

3.3 Automatic annotation with the verbtag attribute

To allow users access to verbal categories contained in verbtag, it was first necessary to integrate the annotation of verbtag into our annotation process. The annotation process used for both written and spoken corpora of the CNC in the SYN2020 corpus standard has been described for written corpora (Jelínek et al. 2021) and spoken corpora (Jelínek 2023). Here, we focus only on the automatic annotation of verbtag. Extending the original annotation to include verbtag required adding verbtag to the training data (both written and spoken). For written text, where we use a rule-based module for disambiguation, it was necessary to design and test several disambiguation rules. For both written and spoken text, neural tagger models were trained.

3.3.1 Adding verbtag to training data

For training the neural tagger and testing the entire annotation process, we use the training data created in the CNC named Etalon corpus, which consist of approximately 2.25 million tokens of written text and additional 200,000 tokens of spoken text. The written Etalon comprises a balanced selection of texts from the three main genre types of the SYN2020 corpus: fiction, non-fiction (academic and professional literature), and newspapers and magazines. The spoken Etalon was selected from the Ortofon corpus, which consists of transcriptions of spontaneous speech into phonetic and orthographic levels, with the orthographic level used for tagging. These data were previously manually annotated for lemmas and morphological tags by two annotators. With the introduction of verbtag, verbtags were assigned to tokens in the data. For verb forms that are ambiguous in terms of verbtag (e.g. all participles, imperfective infinitives, forms of the verb *být* 'to be'), a set of all potential verbtags was added, from which two annotators independently selected the correct verbtag based on the context of the sentence.

3.3.2 Adding verbtag to data during annotation

In the automatic annotation process, first a set of all possible combinations of lemmas and tags for a given token is assigned to each token. This set is then refined in subsequent disambiguation steps until only one (presumably correct) combination remains. We assign this set based on a version of the MorfFlex dictionary modified for the purposes of the CNC. However, verbtag is not included in this dictionary, as it would multiply the number of dictionary entries and slow down its operation. Instead, an additional step has been included in the processing which expands the set of lemmas and tags (on average, 5.56 tags per token) with verbtag (on average, 7.73 verbtags per verb). The subsequent disambiguation steps then select from the set of lemma-tag-verbtag triplets.

3.3.3 Expanding the linguistic rule module for the verbtag disambiguation

For written texts, a combination of a rule-based module and a neural tagger is used for automatic annotation: we refer to this process as hybrid disambiguation. First, the rule-based module is applied, and for tokens that the rule-based module cannot fully disambiguate, the final combination of lemma and tag is selected by the neural tagger.

With the introduction of verbtag, it was necessary to expand the rule-based module with disambiguation rules focused on verbtag. Generally, the rules work by gradually removing tags, verbtags and lemmas from individual tokens that are not correct in a given context. The rules are applied repeatedly, so the action of one rule can enable the later application of another one. One such verbtag rule is the removal of the conditional interpretation from a participle in a sentence where the conditional form of the verb *být* 'to be', e.g. *bych* 'I would' does not appear, and conversely, the removal of all interpretations except conditional for a past participle located in the same clause with a conditional form of the auxiliary verb. The conditional form can be separated from the participle by an embedded clause; in the case of a subordinate clause, the presence of the conditional form is processed independently. For example, in the sentence *Řekl bych, že moc peněz nevydělal.* 'I would say that he did not earn much money,' the first participle *Řekl* 'said' is undoubtedly a conditional, while the second participle *nevydělal* 'did not earn' is an indicative. Approximately 80 rules focused on verbtag were added.

3.3.4 Training neural tagger models

For disambiguation in spoken corpora and for the second phase of disambiguation in written corpora, a deep-learning-based tagger is used. This is an unpublished, beta version of a tagger, developed as part of the MorphoDiTa family of NLP tools, we call this version MorphoDiTa-research. Its properties are described in (Straka et al. 2019).

After adding verbtag, a model for this tagger for written text was independently trained based on the Etalon corpus data of written language, and another tagger model for spoken text was trained based on the combined data of the Etalon corpus of both written and spoken language, as there is not enough data to train on spoken language alone.

Adding verbtag to the training data increased ambiguity in the text by 39% (the average number of lemma-tag combinations per token in written data is 4.03, the average number of lemma-tag-verbtag combinations per token is 5.60). The accuracy of the tagger during training on the written corpus decreased by only about 0.2% (from 97.69% to 97.47%), indicating that the neural tagger handled the more complex data very well.

3.3.5 Disambiguation accuracy

We measured disambiguation accuracy using the method of ten-fold cross-validation. In the case of spoken data, the tagger was trained on both written and spoken data, but testing was conducted only on spoken data.

Tab. 1 shows disambiguation accuracy. The first column indicates the accuracy of assigning the correct verbtag calculated only for verbs, the second column shows the accuracy of morphological tags calculated for all tokens, the third one shows the accuracy of both tag and verbtag, and the fourth one the accuracy of the combination of lemma, tag, and verbtag (i.e. all attributes). The first row shows the accuracy of tagging written texts using the process used by the CNC for its written corpora, i.e., a combination of linguistic rules and the neural tagger. The second row shows the accuracy of tagging spoken data using the neural tagger, MorphoDiTa-research.

	Verbtag (verbs)	Tag (all tokens)	Tag+Verbtag (all tokens)	All (all tokens)
Written: hybrid approach	99.08	97.76	97.70	97.62
Spoken: neural tagger	96.95	92.95	92.56	92.34

Tab. 1. Disambiguation accuracy

The accuracy of tagging spoken corpora is significantly lower in all measured parameters compared to the accuracy of tagging written corpora. This is primarily due to two reasons: firstly, disambiguation of spoken language is more challenging due to its characteristics (non-standard syntax, word repetition, unfinished statements, etc.), and secondly, we have only a relatively small amount of training data available, with most of the data used to train the model coming from written text.

In written text, verb tag annotation is reliable, with the hybrid process incorrectly assigning verbtags to less than one percent of verbs. The highest error rate is in tag annotation, mainly due to the challenges of case ambiguity in Czech.

4 STATISTICS OF VERB FORMS BASED ON VERBTAG

4.1 Corpora

We provide statistics for three basic text genres of the SYN2020 corpus, a representative corpus of contemporary written Czech: newspapers and magazines (NMG), non-fiction (NFC), and fiction (FIC), and for two corpora of contemporary spoken Czech: the Ortofon_v3 corpus (ORT), consisting of transcripts of spontaneous informal spoken Czech, and the Orator_v3 corpus (ORA), consisting of transcripts of formal, prepared monologic speeches.

4.2 Proportion of auxiliary verbs

Tab. 2 shows the proportion of auxiliary (A) and full verbs (V) in the total number of verbs in the corpus.

	NMG	**NFC**	**FIC**	**ORA**	**ORT**
V	89.47	87.75	87.43	87.88	84.14
A	10.53	12.25	12.57	12.22	15.86
Total	**100**	**100**	**100**	**100**	**100**

Tab. 2. Proportion of auxiliary verbs

The higher proportion of auxiliary verbs in the Ortofon corpus corresponds to a significantly higher proportion of the first person past tense indicative mood in this corpus.

4.3 Proportion of mood

Tab. 3 presents the proportion of mood among full verbs: indicative (D), conditional (C), imperative (I), infinitive (F), transgressive (T) and other uses of passive participle (O).

	NMG	**NFC**	**FIC**	**ORA**	**ORT**
D	80.68	78.22	80.93	80.52	83.82
C	4.34	4.31	5.59	4.67	4.89
I	1.28	1.81	2.01	1.73	2.32

	NMG	NFC	FIC	ORA	ORT
F	13.38	15.15	11.10	12.80	8.88
T	0.04	0.09	0.11	0.02	0.01
O	0.29	0.43	0.26	0.27	0.07
Total	100	100	100	100	100

Tab. 3. Proportion of mood

In all corpora, the indicative mood (D) prevails. In non-fiction, the proportion of the infinitive (F) is noticeably higher, because of a greater representation of modality (modal verbs, modal nouns, the adverb *lze* 'may' etc.) and more complex sentence constructions in this subcorpus. In fiction, the proportion of the conditional (C) is slightly higher.

4.4 Proportion of voice

Tab. 4 presents the proportion of voice: active (A) and periphrastic passive (P) among full verbs.

	NMG	NFC	FIC	ORA	ORT
A	97.18	93.93	98.84	97.98	99.83
P	2.82	6.07	1.16	2.02	0.17
Total	**100**	**100**	**100**	**100**	**100**

Tab. 4. Proportion of voice

The proportion of the periphrastic passive in non-fiction is significantly higher than in other corpora, whereas in the Ortofon corpus, its proportion is negligible. It is noteworthy that in the Orator corpus, which is a corpus of formal spoken discourse, the proportion of the periphrastic passive is higher than in the written corpus of fiction.

4.5 Proportion of person

Tab. 5 shows the proportion of person among full verbs, ignoring cases when person is not expressed (infinitive, transgressive).

	NMG	NFC	FIC	ORA	ORT
1	11.73	12.16	17.37	22.43	29.29
2	4.36	4.73	7.94	9.13	13.99
3	83.91	83.11	74.69	68.44	56.73
Total	**100**	**100**	**100**	**100**	**100**

Tab. 5. Proportion of person

The highest proportion of the first and second person is in the corpus of spontaneous spoken language Ortofon, and the lowest in the subcorpus of newspapers.

4.6 Proportion of number

Tab. 6 shows the proportion of singular (S), plural (P), and form of politeness (v) in the corpora studied.

	NMG	NFC	FIC	ORA	ORT
S	70.23	68.57	81.79	64.66	81.02
P	29.43	31.29	17.70	35.26	18.92
v	0.35	0.14	0.52	0.08	0.07
Total	**100**	**100**	**100**	**100**	**100**

Tab. 6. Proportion of number

The differences in the proportion of singular and plural among the corpora are the largest among the observed attributes. In the Orator corpus, plural is used approximately twice as much compared to the fiction subcorpus (the most frequent in the Orator is the 3rd person plural present, followed by the 1st person plural present, which is largely pluralis modestiae). The fiction subcorpus has a similar proportion of number as the Ortofon corpus, and the non-fiction literature subcorpus is similar to the subcorpus of newspapers.

4.7 Proportion of tense

Tab. 7 presents the proportion of past (R), present (P), and future (F) tense and undifferentiated present or future tense of biaspectual verbs (B). We disregard pluperfect (Q) because its annotation is not reliable.

	NMG	NFC	FIC	ORA	ORT
R	37.87	33.72	55.36	23.28	31.00
P	49.65	55.57	34.92	62.99	55.83
F	12.04	10.12	9.58	13.13	12.94
B	0.43	0.59	0.14	0.61	0.23
Total	**100**	**100**	**100**	**100**	**100**

Tab. 7. Proportion of tense

4.8 Comparison of genre types

The above comparison of five broadly defined genre types is rough; more detailed work with both verbtag and tag values (which exceeds the scope of this article) would better show the differences between genres in the use of verb forms. However, we can draw several conclusions.

It cannot be said that written texts behave in one way and spoken texts in another in terms of verbal categories. In the case of person, the written-text subcorpora of newspapers and of non-fiction stand on one side, and the spoken corpora and the subcorpus of fiction on the other. In terms of number, non-fiction and the corpus of formal spoken language behave similarly, while written fiction is close to the corpus of informal spoken language, with the subcorpus of newspapers standing between them. For mood and voice, the situation is similar; non-fiction and newspapers subcorpora have similar proportions along with the corpus of formal spoken language, while the fiction subcorpus and the corpus of informal spoken language differ from them, being similar to each other.

5 CONCLUSION

The verbtag attribute, which has recently been introduced into the annotation of both written and spoken corpora of the CNC, is a useful tool for searching verb forms in the corpus, regardless of whether these forms are compound or simple. The annotation of verbtag in both written and spoken corpora is relatively reliable. Currently, only a small portion of CNC users utilize verbtag, so the aim of this article was also to raise awareness about it.

ACKNOWLEDGEMENTS

The research has been supported by the Ministry of Education of the Czech Republic through the project Czech National Corpus (LM2023044), and by Charles University through the Cooperatio programme (Linguistics).

References

Jelínek, T., Křivan, J., Petkevič, V., Skoumalová, H., and Šindlerová, J. (2021). SYN2020: A New Corpus of Czech with an Innovated Annotation. Proceedings of TSD 2021, Springer, pp. 48–59.

Jelínek, T. (2023). Morphological Tagging and Lemmatization of Spoken Corpora of Czech. Proceedings of TSD 2023, Springer, pp. 154–163.

Marneffe, M. C., Manning, C., Nivre, J., and Zeman, D. (2021). Universal Dependencies. Computational Linguistics, 47(2), pp. 255–308.

Myers, S., and Palmer, M. (2019, August). ClearTAC: Verb Tense, Aspect, and Form Classification Using Neural Nets. Proceedings of the First International Workshop on Designing Meaning Representations, pp. 136–140.

Palmer, M., Gildea, D., and Kingsbury, P. (2005). The Proposition Bank: An Annotated Corpus of Semantic Roles. Computational Linguistics, 31(1), pp. 71–106.

Ramm, A., Loáiciga, S., Friedrich, A., and Fraser, A. (2017). Annotating tense, mood and voice for English, French and German. Proceedings of ACL 2017, System Demonstrations, pp. 1–6.

Straka, M., Straková, J., and Hajič, J. (2019). Czech Text Processing with Contextual Embeddings: POS Tagging, Lemmatization, Parsing and NER. Proceedings of TSD 2019, Springer, pp. 137–150.