

ANNOTATION OF ANALYTIC VERB FORMS IN CZECH – COMPLEX CASES

VLADIMÍR PETKEVIČ – HANA SKOUMALOVÁ

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

PETKEVIČ, Vladimír – SKOUMALOVÁ, Hana: Annotation of Analytic Verb Forms in Czech – Complex Cases. *Journal of Linguistics*, 2023, Vol. 74, No 1, pp. 234 – 243.

Abstract: The article deals with complex cases of determining the attribute *verbtage*, which contains the values of morphosyntactic categories of analytic verb forms. The latest corpora of contemporary written Czech from the SYN series are tagged with this attribute. In this paper, we focus on cases where it is difficult to identify values of *verbtage* categories. These include, e.g. the identification of the auxiliary verb *být* ‘to be’, recognition of the mood and tense of coordinated participles, or determining the number in compound forms in which the individual parts have a different morphological number. Some of the problems are of a theoretical nature, since it is not clear what the correct solution should be. Here we have arbitrarily opted for one option that was offered. Other problems are due to imperfections in the algorithms we use for annotation. The solution here is to improve these algorithms.

Keywords: analytic verb forms, *verbtage*, morphological and morphosyntactic annotation

1 INTRODUCTION

Prior to the creation of the SYN2020 corpus (Křen et al. 2020), only orthographic words were morphologically annotated in the SYN corpora of contemporary Czech, e.g. the form *rozhodnuto* ‘decided’ was marked as a passive participle, and there was no indication that this form was part of the passive *bylo rozhodnuto* ‘was decided’. In addition, the tag for participles inappropriately indicated person (while the morphological tag is supposed to describe only the relevant form and participles do not express person). For some forms the tense was given, but related only to the isolated form (the past active participle thus had the past tense assigned in the tag, even though it was a part of the present conditional). For verbs the important morphological category of mood was not given. Information about tense, mood, and voice had to be elicited by the user through complex queries.

In order to be able to conveniently search for values of morphological categories expressed analytically, a new attribute was introduced: *verbtage*, which brings together all important grammatical categories of compound and simple verb forms and also expresses whether the verb form *být* ‘to be’ is auxiliary or main

verb. The attribute is assigned to all tokens in the corpus, but only for verbs and deverbal adjectives ending in *-ný/tý* (but not *-telný*) it takes on values other than “irrelevant”. On the other hand, the original morphological tags are consistently stripped of the transferred categories, e.g. the person is no longer specified for participles.

All corpora of the SYN series since SYN2020, as well as ETALON2020, test and training corpus (Skoumalová 2021), are annotated with the verbtage attribute.

2 MORPHOLOGICAL CATEGORIES AND VALUES IN THE VERBTAG ATTRIBUTE

2.1 Introduction

Although in many languages of the world verb morphological categories are often expressed analytically (e.g. passive voice), in most corpora of these languages only the categories of isolated words in analytic verb complexes are tagged, rather than the properties of these complexes as such. Thus, for Czech, in which the categories of mood, voice, and tense are often expressed analytically, it is desirable to capture the properties of these complexes.

2.2 Morphological categories and their values in the verbtage attribute

In SYN2020 and subsequent corpora each word form is annotated with a 15-position tag and a 6-position verbtage (see Jelínek et al. 2021; Jelínek et al. 2022; Křivan et al. 2022 for details). While the tag expresses the values of grammatical categories of an isolated word form, the verbtage captures information about the grammatical categories of both simple (*napišu* ‘I will write’) and compound (*budeme pracovat* ‘(we) will work’, *nevěděli jsme* ‘(we) did not know’) verb forms. It also distinguishes the auxiliary forms of the verb *být/bývat* ‘to be/used to be’ from the main verb forms including the copula.

The 6-position verbtage concentrates the values of the analytic verb complex at the main verb, i.e., it transfers the values of some grammatical categories from the forms of the auxiliary verb to the form of the main verb, exceptionally also from pronouns (if the person is expressed only by a personal pronoun). Modal and phrasal verbs are not marked, they are treated as main verbs. For example, in the sentence

- (1) *Nikdy* *bys* *takový výsledek* *nebyla* *předpokládala*.
 Never be.COND.2SG such result be.NEG.PSTP.SG.F expect.PSTP.SG.F
 ‘You would never have expected such a result.’

we annotate the analytic verb complex *bys nebyla předpokládala* as past active conditional of the 2nd person singular. This information is specified in the verbtage of the main verb form *předpokládala* ‘expect’; the forms *bys*, and *nebyla* are annotated as auxiliaries.

The following values are distinguished at the verbtage positions:

1st position – verb type: auxiliary verb (A) / main verb (V)

2nd position – verb form type: indicative (D) / conditional (C) / infinitive (F) / imperative (I) / transgressive (T) / freestanding passive participle, typically a verbal complement (O)

3rd position – voice: active (A) / passive (P) / deverbal adjective (p)

4th position – person: first (1) / second (2) / third (3)

5th position – number: singular (S) / plural (P) / polite form (v)

6th position – tense: pluperfect (Q) / past tense (R) / present (P) / present or future of biaspectual verbs (B) / future (F).

The advantage of the chosen concept of the verbtage attribute is that it is possible to find synonymous expressions of each tense, mood or voice with a single query, e.g. all forms of the past tense and past conditional can be found in the Corpus Query Language (CQL) by querying: [verbtage=".....R"].

Automatic annotation is performed by the hybrid LanGr+MorphoDiTa system in two phases, where LanGr is a rule-based component (Květoň 2006; Jelinek et al. 2011), and MorphoDiTa is a neural tagger (Straka et al. 2019). First, LanGr removes as many incorrect interpretations as possible, and then MorphoDiTa completes the disambiguation.

3 PROBLEMATIC ANNOTATION CASES

Automatic verbtage annotation has to deal with simple cases (the vast majority) and cases that are theoretically or practically more complex. The theoretical difficulties are mainly related to coordination, where one form of *být* ‘be’ connects a non-verbal (typically adjectival) complement coordinated with a passive participle, and thus it is not clear whether this form should be annotated as main verb (i.e., as a copula with non-verbal complements), or as auxiliary (with passive participles). Another problem is related to the coordination of passive participles: in this case, the second auxiliary form of the verb *být* is elided, but the passive participles do not agree in gender or number.

By practical problems we mean cases where the annotation fails, although it is clear what verbtage values should be assigned to the verb form.

3.1 Theoretically problematic cases

The form of the verb *být/bývat* has both a copulative and an auxiliary function

If the form of the verb *být/bývat* fulfils both a copulative and an auxiliary function, the annotation must cope with the ellipsis of the form of the verb *být/bývat* (generally, no tool for automatic ellipses detection was used for the annotation):

- (2) *Lůžka s nastavitelnou výškou jsou/VDA3PP//wrongly:A----- polohovatelná*
 Height-adjustable beds be.PRES.3PL positionable.ADJ.PL
a elektricky ovládána/VOP-P-//wrongly:VDP3PP
 and electrically control.PASP.PL
 'Height-adjustable beds are positionable and electrically controlled.'
- (3) ... *rostlinná společenstva budou/A----- mozaikovitě*
 ... plant communities be.FUT.3PL mosaically
uspořádána/VDP3PF nebo vzájemně propletená/--p---
 arrange.PASP.PL or mutually intertwine.ADJ.PL
 '... plant communities will be mosaically arranged or intertwined'

In sentence (2), the finite form of the verb *jsou* serves a dual function: (i) the copula in conjunction with the adjectival nominal predicate *jsou polohovatelná*, (ii) as an auxiliary verb in the passive construction *jsou ... ovládána* (due to the ellipsis of the second form of the verb *být*). As it is impossible to assign correct verbtags to all parts of the construction *jsou polohovatelná a ovládána*, we adopted the following rule: The form *jsou* is annotated as main verb, since the nominal part of the predicate (*polohovatelná*) is closer to it than the passive participle *ovládána*.

Differently – as an auxiliary verb – we annotate the form *budou* in sentence (3), where the passive participle *uspořádána* is closer to it than the deverbal adjective *propletená*. The general criterion is thus the mutual proximity of the respective forms.

The problem of the erroneous annotation in (2) will be solved by creating new, finer disambiguation rules for the LanGr system.

Determining the verb mood in the coordination of past participles

When coordinating past participles, it is often unclear what mood the second conjunct expresses: whether the construction is a coordination of participles expressing the same mood (typically a conditional), or whether the second participle already belongs to an indicative clause, especially if the second clause lacks a nominative subject:

- (4) ... *jako by/A----- také vystupovala/VCA3SP*
 ... as be.COND.3 also ascend.PSTP.SG.F
po Stromu světa a byla/VCA3SP//wrongly:VDA3SR vysoko nad ním.
 on Tree of the World and be.PSTP.SG.F high above it.
 '... as if she too ascended the Tree of the World and was high above it.'

The annotation error in (4) is to be eliminated by creating a new rule in the LanGr system that annotates the coordination, and thus also the agreement of participles in mood (conditional) and tense (present), if the subject of the second clause is omitted.

Person mismatch in coordination of past active participles

In structures with coordination of past active participles, it is often difficult to determine the person of the second participle if the subject is not lexically expressed in the second clause. Here only non-trivial semantic (i.e., difficult to algorithmize) reasoning can help. Compare sentences (5) and (6):

- (5) *Mám jen okamžik na to, abych/A----- se zamyslela/VCA1SP*
I just have a moment for that so that be.COND.1SG REFL think.PSTP.SG.F
a hlavně se zhluboka nadechla/VCA1SP
and mainly REFL deeply inhale.PSTP.SG.F
'I just have a moment to think and, most importantly, take a deep breath.'
- (6) *Na digitalis jsem/A----- samozřejmě myslel/VDA1SR*
On digitalis be.PRES.1SG of course think.PSTP.SG.M
a byl/VDA1SR součástí toxikologického rozboru.
and be.PSTP.SG.M part of toxicological analysis.
'Digitalis was obviously on my mind and was part of the toxicology analysis.'

While in sentence (5) both past participles are part of one clause, in sentence (6) the participles belong to different clauses that differ in person. By analysis of sentences in the corpus, we found that the problematic form is the copula *byl* in the position of the second conjunct. We can therefore set the rule that the coordinated participles of two main verbs always belong to one clause. No rules leading to 100% correct annotation can be formulated for the copula.

If the subject of the second clause is expressed, the situation is clear, and the automatic annotation is not flawed:

- (7) *Když jsem byl malý, něco jsem/A----- napsal/VDA1SR a táta*
When I was small something be.PRES.1SG write.PSTP.SG.M and dad
vtipkoval/VDA3SR a řekl/VDA3SR že to bylo trapné.
joke.PSTP.SG.M and say.PSTP.SG.M that it was embarrassing.
'When I was a kid, I wrote something and my dad joked and said it was embarrassing.'

In the clause *něco jsem napsal* 'I wrote something', the person of the participle (*napsal* 'wrote') is correctly determined as the first one; in the next clause, where the nominative subject (*táta* 'dad') is lexically expressed, the person of the participles (*vtipkoval* 'joked' and *řekl* 'said') is correctly determined as the third one.

Disagreement in the coordination of passive participles

In this type, two passive participles are connected by one form of the auxiliary *být*, the other form is elided, which makes it difficult to determine the values of the *verbtg* attribute. In the sentence:

- (8) *V 16. a 17. století* byl/A----- *hrad silně opevněn/VDP3SR*
 In the 16th and 17th centuries be.PSTP.SG.M castle heavily fortify.PASP.SG.M
a vybudovány/VDP3PR další pevnostní objekty.
 and build.PASP.PL.M other fortress objects
 ‘In the 16th and 17th centuries the castle was heavily fortified with other fortress structures built.’

The passive participles *opevněn* ‘fortified’ and *vybudovány* ‘built’ do not match in number because they belong to different clauses – the auxiliary form *byly* ‘were’ before the participle *vybudovány* is elided.

A similar mismatch occurs in constructions with the passive infinitive as in the sentence:

- (9) *Použité zdroje musí být/A----- citovány/VFP--- a*
 Used sources must.PRES.SG/PL be.INF cite.PASP.PL.M and
odkazováno/VFP---//wrongly:VOP-S- na ně podle normy ISO 690.
 reference.PASP.SG.N to them according to norm ISO 690
 ‘Used sources must be cited and referenced in accordance with ISO 690.’

Sentence (9) contains the coordination of two clauses with a passive infinitive, the second lacking the auxiliary form of the verb *být* ‘be’, whose presence would have facilitated the annotation of the passive participle *odkazováno* ‘referenced’. This participle does not agree with the first participle *citovány* ‘cited’ and is therefore incorrectly annotated as a complement (VOP-S-). However, the sentence is grammatical, albeit odd – the repetition of the verbs *musí být* ‘must be’ in the second clause (with changes in the word order) would help the clarity of the sentence.

The difficulties with reconstructing the ellipsis of the passive participle and the ellipsis of the auxiliary verb *být* are encountered in the sentence:

- (10) *podrobný popis účelu, na který je/A-----//wrongly:VDA3SP půjčka*
 detailed description of purpose for which be.PRES.SG loan.SG
nebo při možném souběhu půjčky požadovány/VDP3PP//wrongly:VOP-P-
 or in possible concurrence loan.PL request.PASP.PL
 ‘a detailed description of the purpose for which the loan or, in the case of possible overlapping loans, the loans are requested’

Here, the first clause lacks a singular passive participle (*požadována* ‘requested’) that would agree with the expressed auxiliary form *je* ‘is’ in number; the second clause, on the other hand, lacks an auxiliary form (*jsou* ‘are’) that would agree with the expressed participle *požadovány* ‘requested’. The form *je* is incorrectly annotated as the main verb, the form of the participle *požadovány* is incorrectly annotated as a verbal complement, because the corresponding auxiliary form of the verb *být* (*jsou*) is not found.

The form of the auxiliary verb *být* in the future tense and the infinitive of the imperfective/biaspectual verb

In Czech, the future tense of imperfective or biaspectual verbs is formed analytically: by combining the auxiliary future form of the verb *být* and infinitive of an imperfective or biaspectual main verb, while perfective verbs express future tense simply by the present forms. The construction of analytic future tense is morphologically and syntactically ambiguous. In the sentence

- (11) *Všiml si, že se chystá něco podotknout, ale pak se zřejmě rozhodla,*
 He noticed that she is going to make a comment but then she perhaps decided
že nejchytřejší bude/VDA3SF/A----- mlčet/VFA---/VDA3SF
 that smartest.NOM be.FUT.SG be-silent.INF
 ‘He noticed that she was about to make a comment, but then she apparently decided that the smartest thing to do was to keep quiet.’
 ‘He noticed that she was about to make a comment, but then she apparently decided that the smartest one would keep quiet.’

it is semantically very likely that the subject is the infinitive of the imperfective verb *mlčet* ‘keep quiet’, the form *bude* ‘will be’ is a copula, and the adjective *nejchytřejší* ‘smartest’ is a nominal part of the predicate; syntactically, however, the form *nejchytřejší* may have the function of a nominal subject and the construction *bude mlčet* ‘will keep quiet’ would have the function of the analytic future.

The disambiguation of these structures by linguistic rules can be improved by enlarging a group of adjectives that predominantly have the function of a nominal part of predicate in these constructions (e.g. *ideální* ‘ideal’, *lepší* ‘better’, *nejlepší* ‘best’, etc.).

Structures of addressing individuals politely

In Czech, there are analytic verb constructions expressing polite addressing of individual persons where there is a mismatch in number – the form of the auxiliary verb *být* is in plural and the form of the past or passive participle of the main verb is in singular. For example, in the sentence

- (12) *Jak dobře jste/A----- ho znala/VDA2vR*
 How well be.PRES.2PL him know.PSTP.SG.F
 ‘How well did you know him?’

where the auxiliary form *jste* is in plural (tag=VB.P), the past participle *znala* in feminine singular (tag=VpFS), and the polite form is expressed by the value *v* in the *verbtg*.

Syntactically, sentence (12) is ambiguous, since the past participle *znala* ‘knew’ can also be the form of the neuter plural, which, however, agrees with the plural auxiliary *jste*.

Analytic forms where the number of the auxiliary verb does not match the number of the participle of the main verb

There are other types of constructions in Czech where the plural auxiliary form does not agree with the past active participle or the passive participle in singular and there is no polite addressing of individuals.

- (13) *Poslouchal/VDA2vR* *jste/A-----* *někdo* *ráno rádio?*
Listen.PSTP.SG.M be.PRES.2PL somebody.SG morning radio
'Did anyone listen to the radio this morning?'
- (14) ... *vždyt' jsme/A-----* *každý* *vytížen/VDP1vP* *svými aktivitami.*
... after all be.PRES.1PL everyone.SG busy.PASP.SG.M own activities.
'I mean, we're all busy with our own activities.'

We annotate also these constructions as polite forms, but thanks to the mismatch between the subject and the auxiliary form in number, the annotation rules are not too complex. However, the pronominal subject (*někdo* 'someone', *každý* 'everyone') must be identified in the clause.

3.2 Practically problematic cases

Here we present cases where it is obvious how to annotate analytic forms, but automatic disambiguation fails.

Wrong POS disambiguation

Among the forms difficult to annotate there are three forms of the verb *být*:

- *si* – the non-standard form of *jsi* (2nd person singular present tense of the verb *být*) / reflexive particle
- *je* – 3rd person singular present tense of the verb *být*/personal pronoun in accusative singular or plural
- *bud'* – 2nd person singular imperative of the verb *být*/conjunction 'either'.

An error in determining the POS will affect, of course, the selection of the correct verbtage. Disambiguation errors occur especially in more complex contexts where the parts of the analytic verb form are separated and there is syntactically complex linguistic material (punctuation, verbs, etc.) between them. The LanGr system therefore contains many sub-rules for disambiguating these forms.

Incorrect number and gender of participles

It often happens that the POS of ambiguous forms are disambiguated correctly, but their morphological properties are not. This happens, for example, in the case of systemic ambiguity in past and passive participles between feminine singular and neuter plural (*odešla* 'she/they.N left'). The sentence context is sometimes insufficient, as in

- (15) *Bylo jich málo a spíše dávno, ale byla/VDA3PR//wrongly:VDA3SR*
 be.PSTP.SG them few and rather long-ago but be.PSTP.PL.N//wrongly:PSTP.SG.F
 ‘They were few and rather ancient, but they were.’

where the form *byla* ‘were’ is incorrectly identified as the feminine singular form.

If there is no subject in the clause, the disambiguation rules and MorphoDiTa fail, as they only use the context of the given sentence. There is no satisfactory solution here.

Misidentification of the auxiliary vs. main verb *být* in structurally complex sentences

It may happen that the parts of analytic verb forms are distant from one another and the sequence of words between them is complex. For example, in the sentence

- (16) *Primátor Richard Svoboda byl/A----//wrongly:VDA3SR sdělením,*
 Mayor Richard Svoboda be.PSTP.SG.M by-announcement
že jméno jeho podřízeného figuruje v registru svazků, nemile
 that name of his subordinate appears in the registers unpleasantly
překvapen/VDP3SR//wrongly:VOP-S-
 surprised
 ‘Mayor Richard Svoboda was unpleasantly surprised by the announcement that his subordinate’s name appeared in the registers.’

The disambiguation system does not see the pair *byl ... překvapen* ‘was ... surprised’ as a unit, since the two parts are separated by the nested clause *že ... svazků*. Proper annotation can be assured by creating additional LanGr rules.

4 CONCLUSION

Although we have shown a number of theoretical and practical problems associated with the annotation of analytic verb forms in this paper, the overall annotation accuracy (measured on the ETALON2020 corpus) is quite high. For verbtags computed over the whole corpus, it is 99.77%; if we compute the accuracy only on verbs, it is 98.57% (Jelínek et al. 2022).

In the future, in the interest of improved annotation, we will focus on

- more linguistically adequate solutions to the theoretical problems,
- further improving the annotation (disambiguation of the listed complex cases),
- identifying analytic verb forms as such, i.e., marking which components of an analytic verb form make the relevant unit (the software annotation system does not yet allow this).

ACKNOWLEDGEMENTS

The research was supported by the grants LM2023044 and GA-23-05240S.

References

Jelínek, T., and Petkevič, V. (2011). Systém jazykového značkování současné psané češtiny. In *Korpusová lingvistika Praha 2011*, sv. 3. Gramatika a značkování korpusů. Praha: Nakladatelství Lidové noviny / Ústav českého národního korpusu, pages 154–170.

Jelínek, T., Křivan, J., Petkevič, V., Skoumalová, H., and Šindlerová, J. (2021). SYN2020: A New Corpus of Czech With an Innovated Annotation. In K. Ekštejn – F. Párt – M. Konopík (eds.): *Proceedings of the Text, Speech and Dialogue 24th International conference TSD 2021*. Olomouc, Czech Republic, September 6–9, 2021. LNAI 12848. Springer Nature Switzerland AG 2021, pages 48–59. Accessible at: <https://doi.org/10.1007/978-3-030-83527-9>.

Jelínek, T., Petkevič, V., and Skoumalová, H. (2022). Hodnoty slovesných morfologických kategorií v korpusu SYN2020 – atribut verbtage. *Časopis pro moderní filologii* 104(1), pages 89–109.

Křen, M., Cvrček, V., Henyš, J., Hnátková, M., Jelínek, T., Kocěk, J., Kovářiková, D., Křivan, J., Milička, J., Petkevič, V., Procházka, P., Skoumalová, H., Šindlerová, J., and Škrabal, M. (2020). SYN2020: reprezentativní korpus psané češtiny. Praha: Ústav Českého národního korpusu FF UK. Accessible at: <https://www.korpus.cz>.

Křivan, J., and Šindlerová, J. (2022). Změny v morfologické anotaci korpusů řady SYN: nové možnosti zkoumání české gramatiky a lexikonu. *Slovo a slovesnost*, 83(2), pages 122–145.

Květoň, P. (2006). Rule-based Morphological Disambiguation. Ph.D. thesis. Praha: MFF UK.

Skoumalová, H. (2021). Etalon: manuálně anotovaný synchronní korpus českých textů. Praha: Ústav Českého národního korpusu FF UK. Accessible at: <https://www.korpus.cz> and <http://hdl.handle.net/11234/1-3698>.

Straka, M., Straková, J., and Hajič, J. (2019). Czech Text Processing with Contextual Embeddings: POS Tagging, Lemmatization, Parsing and NER. In *International Conference on Text, Speech, and Dialogue*, Ljubljana: Springer, pages 137–150.