ČASOPIS PRO MODERNÍ FILOLOGII 104, 2022, Č. 1, S. 89–109

Hodnoty slovesných morfologických kategorií v korpusu SYN2020 — atribut verbtag[1]

Tomáš Jelínek (Praha) — Vladimír Petkevič (Praha) — Hana Skoumalová (Praha)

VALUES OF VERBAL MORPHOLOGICAL CATEGORIES IN THE SYN2020 CORPUS — THE VERBTAG ATTRIBUTE

The paper describes the *verbtag* attribute, which allows a user to search, in the SYN2020 corpus (and also subsequent corpora, SYNv9 and SYNv10) of contemporary Czech, for all values of morphological categories of verbs, i.e., not only those contained in the *tag* attribute, but also those related mainly to multi-word participial verb predicates, which are prevalent in Czech. The verbtag attribute contains information indicating whether the verb (co-)forming the verbal meaning is either auxiliary or autosemantic, as well as information about the verb mode, diathesis, person, number and tense. The annotation applies both to verb predicates expressed in a single word (e.g., the 1st person indicative present tense: *Čtu rád detektivní příběhy*. 'I like to read detective stories.') and (especially) to verb predicates expressed in multiple words (e.g., the present conditional of the 1st person singular: *Pak bych mu s chutí nabídla výhodnou smlouvu*. 'Then I would gladly offer him a good deal.'). The authors introduce the motivation and the concept of the verbtag annotation, describe relevant morphological categories and their values in detail, and show, via examples, how various multiword structures expressing verbal meaning are annotated in the verbtag attribute. They also offer users a guide to the whole issue of verbal morphosyntax manifested in the verbtag attribute and possibilities for efficient search for and retrieval of morphological/morphosyntactic data. The paper shows which multiple verb complexes are simple in terms of annotation, but also identifies more complex cases (e.g., coordination of participles) which are not easy to automatically annotate, and/or whose annotation is unclear in terms of an adequate theoretical approach. The authors also present the method used for annotating multiword verbal complexes and its current success rate.

KEYWORDS

verbtag attribute, morphology of Czech verbs, morphological categories and values, automatic annotation, SYN2020 corpus

KLÍČOVÁ SLOVA

atribut verbtag, morfologie českých sloves, morfologické kategorie a hodnoty, automatické značkování, korpus SYN2020

DOI

https://doi.org/10.14712/23366591.2022.1.5

1 Článek vznikl během práce na projektu velké infrastruktury reg. č. LM2018137.

1. MOTIVACE

Český národní korpus již řadu let nabízí jazyková data pro výzkum češtiny v podobě jazykových korpusů. Tyto korpusy jsou opatřeny morfologickými značkami, aby uživatelé mohli s korpusy snáze pracovat. Morfologické značky (tagy) poskytují informace o slovním druhu slova, jemuž jsou přiřazeny, a o jeho mluvnických kategoriích, např. tvaru *jdu* je přiřazena značka pro sloveso v přítomném čase v první osobě jednotného čísla. Tyto značky se primárně vztahují k izolovaným slovním tvarům, např. infinitiv *pracovat* je označen jako infinitiv i tehdy, je-li součástí složeného futura *budu pracovat*. Morfologické vlastnosti jednotlivých slovních tvarů (včetně jejich slovního druhu) se určují automaticky na základě kontextu celé věty, protože jedině kontext umožňuje rozpoznat u homonymních tvarů, která z možných interpretací je správná, například zda tvar *se* je předložka, nebo zvratné zájmeno, nebo zda tvar *informací* ve větě *Zdroj informací je ověřený.* je v instrumentálu singuláru, nebo genitivu plurálu.

Donedávna představovalo morfologické značkování vztažené k izolovaným tvarům obtíže pro uživatele, kteří chtěli v korpusu zkoumat složené slovesné tvary, například minulý čas 1. a 2. osoby nebo pasivní či kondicionálové konstrukce. U slovesných příčestí (*koupil, koupen*) byla ve značce sice uvedena osoba (v rozporu se snahou, aby morfologická značka popisovala jen tvar, jemuž je přiřazena), ne však způsob. Čas je u některých tvarů uveden, ale jen vztažený k izolovanému tvaru (tedy minulé příčestí má v tagu přiřazen čas minulý i v případě, kdy je součástí kondicionálu přítomného). Potřeboval-li uživatel tyto údaje, byl nucen je sám dohledat, využívaje složitých dotazů, zahrnujících například pro podmiňovací způsob přítomnost kondicionálového tvaru slovesa *být* nebo spojek *aby* či *kdyby* v téže klauzi (ovšem ne nutně v těsné blízkosti slovesného příčestí). Podobně nebylo snadné rozeznat od sebe tvary pomocného slovesa *být* od tvarů existenčního a sponového *být*.

Pro usnadnění práce se slovesy je slovům v reprezentativním korpusu současné psané češtiny SYN2020 a ve všech nově publikovaných korpusech psané češtiny (tedy zatím také v korpusech SYNv9 a SYNv10) přiřazován nový atribut: **verbtag**. Tento atribut obsahuje na jednom místě všechny důležité gramatické kategorie slovesného tvaru, ať už je složený, nebo prostý. Vyznačuje mimo jiné také pomocná slovesa. Přiřazuje se všem tokenům v korpusu, pouze u sloves (a v malé míře u deverbativních adjektiv) však nabývá jiných hodnot než „------“. Tento nový atribut podrobně popíšeme v následujících částech článku. Původní morfologické značky jsou naopak důsledně zbaveny přenesených kategorií, například se v nich tedy již neuvádí osoba u příčestí. Značkujeme kategorie gramatické (morfosyntaktické), nikoli sémantické; například historický prézens se tak značí jako jakýkoli jiný přítomný tvar.

Atribut **verbtag** byl navržen tak, aby jím bylo možné spolehlivě a srozumitelně automaticky značkovat hodnoty morfologických kategorií v libovolném psaném českém textu. Ve **verbtagu** tedy lze najít například informaci o slovesném způsobu nebo osobě jednoduchých i složených slovesných tvarů. Některé případy, u nichž nelze spolehlivě automaticky rozhodnout, se již v návrhu atributu řeší opatrně. Například u trpných příčestí, jež se nacházejí ve větě bez pomocného slovesa, nelze spolehlivě určit, zda jde o složené pasivum s elidovaným pomocným slovesem, nebo o doplněk;

o takové rozlišení se tedy vůbec nepokoušíme a tato příčestí značkujeme jako samostatně stojící („ostatní" funkce trpného příčestí). Podobně nelze spolehlivě rozpoznat reflexivní pasivum (*Petr se myje.* oproti *Nádobí se myje.*), ve **verbtagu** se tedy takové pasivum vůbec neurčuje (s hodnotou pro reflexivní pasivum se vůbec nepočítá, slovesům se přiřazuje **verbtag** jako pro aktivní tvary).

Kromě zavedení atributu **verbtag** bylo zvoleno i nové řešení pro textová slova, která zahrnují více slov syntaktických, jako je *ses*, *oč* nebo *abychom*. Taková slova se sice i nadále považují za jeden token (jedno slovo, jehož tvar lze v korpusu vyhledat), ale přiřazuje se jim více lemmat, morfologických značek i více hodnot atributu **verbtag**, například tvar *ses* dostává lemma *se* a lemma *být*, značku pro zvratné zájmeno[2] i pro 2. osobu singuláru prézentu slovesa *být*. Dále v textu o nich mluvíme jako o víceslovných tokenech.

Článek je rozvržen do těchto částí. Nejprve stručně pojednáváme o značkování složených slovesných tvarů v některých jiných jazycích a korpusech a také v jiných korpusech češtiny (část 2). Část 3 obsahuje stručný přehled slovesných kategorií zahrnutých ve verbtagu (3.1) a jejich podrobný popis jednotlivě včetně jejich možných hodnot (3.2). Dále je zde několik příkladů náležitého značkování (3.3), vysvětluje se tu také, proč nebyly některé jevy do **verbtagu** zahrnuty (3.4). Část 4 ukazuje na základě souvislostí mezi jednotlivými kategoriemi a jejich hodnotami ve **verbtagu**, jak se v celé problematice snáze orientovat jako uživatel korpusu. V části 5 se probírají složitější případy složených slovesných tvarů, a to jednak z hlediska obtíží a chyb v procesu automatického přiřazení **verbtagu** tokenům v korpusu, jednak z hlediska nejasností v samém chápání morfosyntaktické povahy některých jevů. Část 6 se věnuje automatickému značkování korpusů, především automatickému přiřazování značky **verbtag** slovním tvarům v korpusu, technickým souvislostem zavedení atributu **verbtag** a metodám i úspěšnosti desambiguace zvláště se zaměřením na **verbtag**.

2. ANOTACE SLOŽENÝCH SLOVESNÝCH TVARŮ V KORPUSECH

Složené slovesné tvary se vyskytují snad ve všech indoevropských jazycích, experimenty s automatickou anotací slovesného způsobu, rodu nebo času byly prováděny například pro angličtinu, němčinu nebo francouzštinu, viz např. (Ramm et al., 2017), (Dönicke, 2020). Tyto experimenty používají syntaktickou anotaci (tedy automatické určení souvztažnosti mezi jednotlivými částmi slovesného tvaru).

Ve většině velkých, veřejně dostupných korpusů indoevropských jazyků (například Polský národní korpus, Britský národní korpus, Referenční korpus současné španělštiny[3] aj.) se jednotlivé části složených slovesných tvarů značkují izolovaně,

2 U reflexiva *se* nerozlišujeme z důvodů technických (obtížnost automatické anotace) i teoretických (nejasná hranice mezi jevy) mezi reflexivním zájmenem (*Myli se*), tvarotvorným prostředkem *se* (*Auta se vyrábějí*) a slovotvorným prostředkem *se* (*Usmála se na mě*).

3 Polský národní korpus: www.nkjp.pl; Britský národní korpus: https://www.english-corpora.org/bnc/; Referenční korpus současné španělštiny: https://www.rae.es/banco-de-datos/crea.

tedy minulému příčestí se přiřadí pouze značka pro minulé příčestí bez ohledu na to, jakého složeného tvaru je slovo součástí. V anotačním standardu Universal Dependencies[4] možnost anotovat složené tvary sloves existuje, ale v korpusech tímto standardem anotovaných se vesměs také značkují jednotlivé části slovesných tvarů izolovaně (min. příčestí: Tense=Past|VerbForm=Part).

Anotace složených slovesných tvarů se objevuje v některých syntakticky anotovaných korpusech (treebancích), které však nejsou tak rozsáhlé jako korpusy textové. Je to například závislostní treebank češtiny PDT 2.5,[5] ve kterém je na tektogramatické rovině u sloves vyznačeno, zda jsou součástí složeného tvaru a jaký mají způsob. V jiném, složkovém treebanku (Petkevič et al., 2015) jsou u sémantické hlavy složeného slovesného tvaru uvedeny charakteristiky celého složeného tvaru. Slovesný způsob a čas jsou podobně uvedeny v polském treebanku Składnica (Patejuk — Przepiórkowski, 2014).[6] V jiných treebancích, např. v ruském Syntaktickém korpusu (Апресян et al., 2005),[7] se dá informace o způsobu a času získat dotazem na plnovýznamové sloveso a na něm závislá pomocná slovesa, ale nikde není uvedena explicitně.

3. MORFOLOGICKÉ KATEGORIE A HODNOTY V ATRIBUTU *VERBTAG* A JEJICH ZNAČKOVÁNÍ

V této části nejprve podrobně popíšeme, které morfologické kategorie a jejich hodnoty ve **verbtagu** zachycujeme a kterým slovním tvarům jsou tyto údaje připisovány (3.1 a 3.2). Poté předvádíme na příkladech, jak se struktury se složenými slovesnými tvary značkují v korpusech SYN2020 a SYNv9 (3.3), a v části 3.4 uvádíme vlastnosti, které ve **verbtagu** nevyjadřujeme.

3.1 MORFOLOGICKÉ KATEGORIE A HODNOTY V ATRIBUTU *VERBTAG*

Slovní tvary jsou v korpusu SYN2020 (a také v korpusu SYNv9 a SYNv10) morfologicky značkovány dvěma atributy: základním 15pozičním **tagem** a 6pozičním atributem **verbtag**. Atribut **tag** obsahuje gramatické kategorie vztažené přímo k izolovanému slovnímu tvaru, u sloves tedy především: typ tvaru (imperativ, minulé příčestí, kondicionálový tvar), číslo, čas, osobu (u tvarů přímo osobu vyjadřujících), jmenný rod (u příčestí a přechodníků), slovesný rod, vid.

Atribut **verbtag** obsahuje informace o gramatických kategoriích slovesa, a to jak u tvarů jednoduchých (*napíšu*, *slyšet*, *jsem*, *hovořil*), tak u tvarů složených (*budu psát*, *slyšeli jste*, *byli bychom připravili*, *být spasen*). Odlišuje také pomocné tvary slovesa *být* a *bývat* od tvarů plnovýznamových a od spony u verbonominálního predikátu, která

4 Anotační standard, který používá v současné době nejvíce různých jazykových korpusů: www.universaldependencies.org.

5 http://hdl.handle.net/11858/00-097C-0000-0006-DB11-8.

6 https://clarino.uib.no/iness/page.

7 https://ruscorpora.ru/new/search-syntax.html.

se rovněž chápe jako plnovýznamová. Atribut **verbtag** tvoří šest pozic. Přiřazuje se všem tokenům v korpusu (tedy včetně například substantiv či interpunkce), jen u sloves a u deverbativních adjektiv zakončených na *-ný*, *-tý* (viz popis 3. pozice níže) však nabývá jiných hodnot než hodnoty „------“ (irelevantní).

Atribut **verbtag** tedy přenáší hodnoty některých gramatických kategorií od tvarů pomocného slovesa *být* ke slovesu plnovýznamovému, výjimečně i od zájmen (například tam, kde je osoba vyjádřena osobním zájmenem, nikoli slovesným tvarem: *já nemohl spát*). Nepřenáší však hodnoty gramatických kategorií od jiných sloves, například od sloves modálních či fázových, např. infinitiv po modálním slovese tedy bude mít pro osobu a číslo hodnotu „-“ (irelevantní).

Při určování gramatických kategorií daného tvaru atribut **verbtag** více využívá okolního kontextu a je interpretativnější než základní 15poziční morfologický tag. Například u příčestí neobsahuje tag kategorii osoby (samo izolované příčestí totiž osobu nevyjadřuje, typ *slyšels* je pojednán jako víceslovný token, viz výše), zato **verbtag** osobu určuje. Imperativ může mít v tagu pouze první či druhou osobu (*chraňme*, *chraň*, *chraňte*), ve **verbtagu** může ale v určitých kontextech nabýt i osoby třetí (*chraň Bůh*). Rozpor mezi číslem u pomocného slovesa a u příčestí v případě vykání je zachycen hodnotou **v** pro gramatickou kategorii čísla ve **verbtagu**: například ve spojení *abyste nepřišla* bude mít příčestí *nepřišla* číslo v tagu hodnotu **S** (singulár), ve **verbtagu** hodnotu **v** (vykání).

Nyní uvedeme nejprve přehledově a poté podrobně obsah atributu **verbtag**. Tvoří jej šest pozic s těmito hodnotami:

— 1. pozice — **druh slovesa**: pomocné sloveso / plnovýznamové sloveso
— 2. pozice — **typ slovesného tvaru**: indikativ / kondicionál / infinitiv / imperativ / přechodník; a navíc samostatně stojící trpné příčestí
— 3. pozice — **diateze**: aktivní / pasivní
— 4. pozice — **osoba**: první / druhá / třetí
— 5. pozice — **číslo**: singulár / plurál / vykání
— 6. pozice — **čas**: předminulý čas / minulý čas / prézens / futurum

3.2 PODROBNÝ POPIS JEDNOTLIVÝCH POZIC *VERBTAGU*

1. pozice — **druh slovesa**: pomocné/**A,** plnovýznamové/**V,** irelevantní/-
Na této pozici mají plnovýznamová slovesa v jakémkoli tvaru a sponové *být, bývat* hodnotu **V** (viz př. 1–5 níže); pomocná slovesa *být*, *bývat* včetně klitického *s* a kondicionálových tvarů samostatných i spojených se spojkami *aby, kdyby* mají hodnotu **A** (viz př. 3–5); všechny ostatní slovní druhy včetně deverbativních substantiv a adjektiv mají hodnotu - (irelevantní). U pomocných sloves obsahuje celý **verbtag** právě jen hodnotu A, další hodnoty se neuvádějí (A-----). Sloveso *mít* za pomocné nepovažujeme ani v případech jako *mít uvařeno*.[8]

8 Více o rezultativních a dalších konstrukcích v oddíle 3.4.

(1) *Přišel*/V *pozdě.*

(2) *Jeho matka a bratr tu budou*/V *každou chvíli.*

(3) *Proč se****s***/A *s ním hádala*/V*?*

(4) *Měl*/V *by*/A *se pokusit*/V *překonat*/V *tu starou rivalitu.*

(5) *Vzorce musí*/V *být*/A *použity*/V *přesně.*

Obecně platí, že jednoduchý plnovýznamový slovesný tvar nebo složený slovesný tvar obsahuje právě jeden jednoduchý slovesný tvar s hodnotou **V** na první pozici **verbtagu**.

Slovesný tvar mající hodnotu **V** na první pozici **verbtagu** může mít v závislosti na svém typu vyplněny všechny hodnoty na 2. až 6. pozici (viz níže).

2. pozice — **typ slovesného tvaru**: indikativ/**D**, kondicionál/**C**, imperativ/**I**, infinitiv/**F**, přechodník/**T**, ostatní funkce trpného příčestí/**O** (doplněk nebo predikát bez spony), irelevantní/-
Písmeny označené hodnoty může mít jen slovesný tvar, který má na první pozici **V**.

Hodnota	Význam	Příklady
D	indikativ	***odešla***, ***odejdu***; ***byl*** *doma*; *byl jsem* ***přirovnáván***; *bude* ***pršet***
C	kondicionál	***odešel*** *bych*; *byla bys bývala* ***chycena***; *abych/kdybych byla* ***potrestána***
I	imperativ	***odejdi***; *buďte* ***chváleny***
F	infinitiv	*chci* ***vědět***; ***být*** *úspěšný*; *být* ***přirovnáván***;
T	přechodník	*chutně* ***vaříce***; *jsouc* ***vystavována*** *velké zátěži*
O	ostatní funkce trpného příčestí	*poslouchal mne* ***znuděn***; *mám jídlo* ***uvařeno***; *bankovní loupež* ***zmařena***

TABULKA 1. Typy slovesných tvarů

— Hodnotu **D** (indikativ) na 2. pozici tedy mají aktivní tvary indikativu sloves (*odcházím*/D, *odejdu*/D), a to včetně sloves *být*, *bývat* v nepomocné funkci, tj. existenční a sponové (*ďábel je*/D; *byl*/D *doma*), trpná příčestí v indikativu pasiva (*byl jsem přirovnáván*/D) a tvary infinitivu v aktivním futuru nedokonavých sloves (*budu tady sedět*/D).
— Hodnotu **C** (kondicionál) na 2. pozici mají minulá příčestí významových sloves a sponového či existenčního *být* v aktivní kondicionálové konstrukci (*odešel*/C *bych*; *byl bych (býval) odešel*/C; *byla*/C *bych smutná*) a trpná příčestí plnovýznamových sloves v pasivní kondicionálové konstrukci (*bylo by znárodněno*/C; *abyste nebyli překvapeni*/C).

— Hodnotu **I** (imperativ) na 2. pozici mají aktivní tvary imperativu (*odejdi*/I, *buď*/I *připravený*) a trpná příčestí plnovýznamových sloves v konstrukci s pasivním imperativem (*buď připraven*/I; *buď pozdravena*/I, *Maria*).
— Hodnotu **F** (infinitiv) na 2. pozici mají všechny tvary infinitivu plnovýznamových sloves mimo opisné konstrukce pro tvoření aktivního futura nedokonavých sloves (tam mají hodnotu D): *sedět*/F *na zahradě se mu líbilo*; *chci tady sedět*/F; *možnost odejít*/F *do zahraničí ji lákala*; *opak může býti*/F *pravdou* a trpná příčestí plnovýznamových sloves v konstrukci s pasivním infinitivem: *a tak chci být připraven*/F *na všecko*; *nebývat osočován*/F.
— Hodnotu **T** (přechodník/transgresiv) na 2. pozici mají aktivní tvary přechodníku plnovýznamových sloves i nepomocného *být*: *předpokládaje*/T, *že budeme skeptičtí*; *a nevyčkav*/T *odpovědi odešel si pro zákusek*; *sám jsa*/T *hříchu neschopen* a trpná příčestí plnovýznamových sloves v pasivní přechodníkové konstrukci: *odvětila dívka, jsouc dojata*/T *takovým náporem úcty a lásky*.
— Hodnotu **O** (ostatní funkce trpného příčestí) na 2. pozici mají trpná příčestí ve funkci doplňku, v rezultativních konstrukcích a ve funkci jmenného přísudku s elidovanou sponou: *poslouchal mne znuděn*/O; *mám jídlo uvařeno*/O; *na výsost znechucen*/O *špatným dojmem, který Clevinger vyvolával*; *koncert zrušen*/O *bez náhrady*. Oproti tomu ve strukturách se sponou se trpná příčestí značkují podle typu slovesného tvaru, jaký pasivní konstrukce vyjadřuje, např. indikativ v klauzi: *byl jsem zahrnován*/D *holdy jako milionář*.

3. pozice — slovesný rod: aktivum/**A**, pasivum/**P**, deverbativní adjektivum odvozené z pasivního příčestí/**p**, irelevantní/**-**

Hodnota	Význam	Příklady
A	aktivum	***odešla***; ***odešel*** *bych*; ***byl*** *doma*; ***popojdu***; *budou* ***sedět***
P	pasivum	*byla bys* ***chycena***; *buďte* ***pochváleny***
p	deverbativní adjektiva odvozená od trpného příčestí	*dílo bylo* ***vyvážené***; *ležel* ***zabitý*** *na trávníku*; *býval* ***přirovnávaný*** *k Mozartovi*; ***zabitá*** *neděle*

TABULKA 2. Slovesný rod

Na 3. pozici mají hodnotu **A** (aktivum) aktivní slovesné tvary, hodnotu **P** (pasivum) pasivní slovesné tvary. Tyto hodnoty jsou ve **verbtagu** totožné s příslušnými hodnotami v tagu (12. pozice). Hodnotu **A** nebo **P** může mít jen slovesný tvar, který má na první pozici **V**. Hodnotu **p** má pouze deverbativní adjektivum zakončené na *-ný* nebo *-tý* a toto adjektivum nemá žádné další hodnoty, jeho **verbtag** je tudíž vždy roven --p---. Tato hodnota umožňuje odlišit deverbativní adjektiva odvozená od pasiva od nedeverbativních adjektiv; adjektiva odvozená od přechodníků jsou rozlišena již v tagu (AG.*, AM.*: adjektivum od přítomného, resp. minulého přechodníku). V konstrukci se sponou a deverbativním adjektivem (např. *byl unavený*) je spona na rozdíl od pasiva značena jako plnovýznamové sloveso.

Ve **verbtagu** nezachycujeme reflexivní diatezi (medium), tj. například ve větě: *učešu si vlasy* není u slovesného tvaru *učešu* uveden reflexivní význam (lexikálně signalizovaný reflexivem *si*). Ani v něm nezachycujeme reflexivní pasivum, značkujeme pouze pasivum opisné. Reflexivní pasivum totiž není snadné spolehlivě značkovat vzhledem k velmi obtížně desambiguovatelné homonymii reflexivních částic *se*, *si* i vzhledem k neshodám v teoretickém pohledu na tento jev.

4. pozice — **osoba**: **1** / **2** / **3** / irelevantní/**-**
Hodnoty označené čísly může mít jen slovesný tvar, který má na první pozici V.

Hodnota	Význam	Příklady
1	první osoba	***přišel*** *jsem*; ***vylézáme***; ***připravili*** *bychom to*; *budeme* ***zainteresováni***; *buďme* ***odhodláni***
2	druhá osoba	***odejdete***; ***odejděte***; *byla bys* ***chycena***; *budete* ***zastrašeni***; *buďte* ***připraveny***; *hodně ses* ***snažila***
3	třetí osoba	***přišel*** *včas*; *byl* ***ztracen***; *bude* ***připravovat***; *je* ***odhodlán***; ***chraň*** *Bůh*

TABULKA 3. Osoba

Ve **verbtagu** se osoba (1/2/3) uvádí u všech nepomocných tvarů morfologického prézentu (*přinášíme*, *přinesou*), futura (*bude*, *půjdu*), tvaru imperativního (*vyskoč*) a u minulého (l-ového) příčestí, pasivního příčestí a infinitivu ve finitní konstrukci: *otec zemřel*/3; *byli bychom připravili*/1; *byl byste zatčen*/2; *nebudeme to odkládat*/1. Osoba se neuvádí u nefinitních konstrukcí infinitivních: *být přinesen*/-; *být*/- *přinesený* (ani když lze osobu dovodit například z modálního slovesa), přechodníkových: *nechtě*/-; *jsouc přinášeno*/- a u trpných příčestí ve funkci doplňku: *silně rozrušen*/-, *odešel*.

Na 4. pozici **verbtagu** zachycujeme také sémanticky třetí osobu imperativu: *chraň*/3 *Bůh*; *pozdrav*/3 *pánbůh*; *vem*/3 *to čert*, která je morfologicky vyjádřena druhou osobou (*chraň*/2, *pozdrav*/2, *vem*/2) na 8. pozici tagu, neboť čeština morfologicky třetí osobu imperativu nevyjadřuje.

5. pozice — **číslo**: singulár/**S**, plurál/**P**, vykání/**v**, irelevantní/**-**
Hodnoty označené písmeny může mít jen slovesný tvar, který má na první pozici V.

Hodnota	Význam	Příklady
S	singulár	***přišel*** *jsem*; *bude je* ***škádlit***; *budeš* ***sedět***; ***připravil*** *bych to*; *bude* ***zabit***; *buď* ***odhodlána***
P	plurál	***odejdete***; ***vylézáme***; *nebyly by* ***prodány***; ***čtěme***; *nebuďte* ***překvapeni***; *budou* ***číst***
v	vykání jednotlivci	***nedostavil*** *jste se*; ***měla*** *byste, paní účetní, vědět*

TABULKA 4. Číslo

Hodnotu **v** (vykání jednotlivci) má pouze minulé příčestí a trpné příčestí plnovýznamového slovesa v singuláru (v přechodníkových konstrukcích se v dnešní češtině nevyskytuje) v konstrukci s vykáním jednotlivci, kde je rozpor v čísle mezi plurálem pomocných sloves a singulárem plnovýznamového slovesa: *vy byste to zařídil*/v oproti *vy byste to zařídili*/P. Neznačkuje se vykání v plurálu (tedy vykání skupině osob) vzhledem k tomu, že homonymii mezi vykáním a nevykáním v plurálu nelze patrně nikdy pouze na základě textu desambiguovat.

U plnovýznamových sloves je hodnota **S** (singulár) nebo **P** (plurál) na 5. pozici **verbtagu** totožná s hodnotou čísla (**S** nebo **P**) na 4. pozici tagu až na

— minulá a trpná příčestí v aktivních a pasivních konstrukcích s vykáním jednotlivci (*to byste viděl*: tag obsahuje **S** (singulár), **verbtag** obsahuje **v** (vykání jednotlivci))
— infinitiv v aktivních konstrukcích tvořících opisné futurum (*budeme pracovat*: tag obsahuje - (irelevantní), **verbtag** obsahuje **P** (plurál))
— trpné příčestí v pasivních infinitivních konstrukcích (*být stíhán*: tag obsahuje **S** (singulár), **verbtag** obsahuje - (irelevantní)).

6. pozice — čas: prézens/**P**, futurum/**F**, prézens/futurum obouvidých sloves/**B**, minulý čas/**R**, předminulý čas/**Q**, irelevantní/-
Hodnota atributu **verbtag** na této pozici vyjadřuje čas gramatický (tedy například přítomný čas u kondicionálu přítomného), ne čas sémantický (jako je třeba historický prézens nebo futurum vyjádřené přítomným časem). Hodnoty označené písmeny může mít jen slovesný tvar, který má na první pozici V.

Hodnota	Význam	Příklady
P	prézens	***sedím***; ***jsem*** *doma*; *jsem* ***přesvědčen***; *jsouc* ***vychovávána*** *jen italsky*; ***udělal*** *bych* (kondicionál přítomný)
F	futurum	*budu* ***pracovat***; *budu* ***chtít*** *spát*; ***poběží***; ***odešle***; *zítra* ***budu*** *v práci*; *bude* ***potrestán***
B	prézens/futurum obouvidých sloves	***rezignuji***; ***věnuje***[9]
R	minulý čas	***přišel***; *byl* ***chycen***; *byl bych (býval)* ***seděl***; *kdo by byl býval* ***zarmoucen*** (kondicionál minulý); *byvši* ***vyzvána***
Q	předminulý čas	*pak jsem byl* ***odešel***; *byl býval* ***chytán***
-	hodnota nevyplněna (irelevantní)	*být* ***zraněn***; *buď* ***připraven***; ***znechucen*** *odešel*

TABULKA 5. Slovesný čas

9 Slovesa *věnovat* a *věnovat se* jsou obouvidá: podle slovníků SSČ, SSJČ, PSJČ je sloveso s valencí *věnovat něco někomu* dokonavé, řídčeji nedokonavé, a reflexivní *věnovat se někomu/něčemu* je nedokonavé, řídčeji dokonavé. Sémantiku obou sloves nerozlišujeme, obě však mají v prézentu význam přítomnosti nebo budoucnosti.

Hodnotu **P** mají

— aktivní prézentní tvary plnovýznamových sloves: *sedím*/P; *jsem*/P *doma*
— aktivní přechodníky plnovýznamových sloves: *vytvářeje*/P
— trpné příčestí plnovýznamového slovesa v pasivní konstrukci indikativu/přechodníku přítomného: *jsa považován*/P *za trpaslíka*
— minulé příčestí v konstrukci s kondicionálem přítomným: *já bych to nepřežila*/P.

Hodnotu **F** má

— infinitiv opisného futura nedokonavých sloves: *budu pracovat*/F
— synteticky tvořené tvary významového futura sloves pohybu: *půjdu*/F; *pojedeme*/F; tvary dokonavých sloves v morfologickém prézentu, ale s významem futura: *pochválí*/F; *pojde*/F
— nepomocné (existenční, sponové) tvary futura slovesa *být*: *zítra budu*/F *v lese*; *budete*/F *ovlivnění*
— pasivní příčestí v pasivu futura: *budu chycen*/F

Verbtag tedy umožňuje nalézt synonymní vyjádření futura jediným dotazem: například všechny tvary futura lze v dotazovacím jazyce CQL nalézt dotazem [verbtag=".....F"].

Hodnotu **B** mají tvary obouvidých sloves (viz i pozn. 9), u nichž není jasné, zda čas je prézens, nebo futurum: *abdikuji*/B; *rezignuju*/B; *věnuji*/B.

Hodnotu **R** mají

— aktivní tvary minulého příčestí plnovýznamového slovesa spolutvořící indikativ minulého času včetně spony: *pak jsem odešel*/R, *byl*/R *jsem doma* a minulý kondicionál: *byl bych* (*býval*) *seděl*/R
— tvary trpného příčestí plnovýznamového slovesa spolutvořící pasivum indikativu minulého času: *byl chycen*/R; pasivum přechodníku minulého času: *byvši vyzkoušena*/R a pasivum kondicionálu minulého: *byli by bývali vyzváni*/R *zřeknout se islámu*

Hodnotu **Q** mají

— aktivní tvary minulého příčestí plnovýznamového slovesa spolutvořící předminulý čas: *pak jsem byl odešel*/Q
— tvary trpného příčestí plnovýznamového slovesa spolutvořící předminulý čas: *on byl býval podepsán*/Q. Takováto struktura je však z hlediska úzu ryze teoretická: v korpusu SYNv9 (cca 5,6 mld. tokenů) jsme ji nenalezli.

Čas se neuvádí (jeho hodnota je tedy: -):

— u infinitivu, který nespolutvoří opisné futurum: *studovat*/- *práva je něco speciálního*
— u imperativu: *pracuj*/-, a to ani u sponového *být*: *buď*/- *zdráv*

— u trpných příčestí v imperativní konstrukci: *buď připraven*/- a v infinitivní konstrukci: *netěší mne být nalezen*/-

3.3 PŘÍKLADY ZNAČKOVÁNÍ ATRIBUTEM *VERBTAG*

Aby se čtenář v popisu víceslovných predikátů, který jsme právě představili, lépe orientoval, uvádíme v této části tři ilustrativní příklady vět z korpusů SYN2020 a SYNv9.

(6) *Zřejmě se z toho budu*/A----- *muset*/VDA1SF *dostat*/VFA--- *sám.*

Ve větě (6) je opisné futurum nedokonavého plnovýznamového (**V**) slovesa *muset*, které je v indikativu (**D**) aktiva (**A**) 1. osoby singuláru (**S**) futura (**F**), přičemž slovesný tvar *budu* je pomocný (A-----). Dále se v ní nachází infinitivní (**F**) aktivní (**A**) tvar jednoslovného plnovýznamového slovesa *dostat*. Atribut **verbtag** dostávají i neslovesné tvary ve větě, ovšem s prázdnými hodnotami všech kategorií: ------.

(7) *Nakopal*/VCA1SP *bych*/A----- *si, že jsem*/A----- *s tím vším začal*/VDA1SR.

V této větě nacházíme (i) víceslovný slovesný tvar *nakopal bych*, což je aktivní (**A**) konstrukce přítomného (**P**) kondicionálu (**C**) v 1. osobě singuláru (**S**); (ii) víceslovný slovesný tvar *jsem … začal*, což je aktivní (**A**) konstrukce 1. osoby singuláru (**S**) indikativu (**D**) minulého času (**R**).

(8) *Půjdu*/VDA1SF, *až se mi bude*/A----- *chtít*/VDA3SF *a až budu*/A----- *připraven*/VDP1SF.

Ve větě (8) je *půjdu* tvarem syntetického futura (**F**) indikativu (**D**) aktiva (**A**) slovesa *jít* v 1. osobě singuláru (**S**). Dále tu jsou dvě konstrukce s indikativem (**D**) budoucího (**F**) času: (i) aktivní (**A**) opisné futurum *bude chtít* s pomocným slovesným tvarem *bude* (**A**) a plnovýznamovým infinitivním tvarem *chtít*, jemuž jsou přiřazeny všechny relevantní hodnoty atributu **verbtag**; (ii) indikativní (**D**) pasivní (**P**) konstrukce opisného futura (**F**) 1. osoby singuláru (**S**) *budu připraven*, kde tvar *budu* je pomocný a *připraven* plnovýznamový.

3.4 MORFOSYNTAKTICKÉ VLASTNOSTI SLOVES NEZAHRNUTÉ VE *VERBTAGU*

Koncept **verbtagu** nemůže zahrnout všechny morfosyntaktické vlastnosti sloves, a to jednak proto, že cílem **verbtagu** je mít jednoduchou, přehlednou značku, jednak proto, že některé jevy není možné spolehlivě značkovat a na dalších se neshodují lingvistické teorie. Zde stručně představíme dva jevy, které ve **verbtagu** zahrnuté nejsou, přestože by je uživatelé mohli v této značce očekávat.

Prvním z nich je afirmace/negace víceslovného komplexu jako celku (afirmace/negace jednotlivých tvarů se určuje na 11. pozici v tagu). Patrně by automaticky bylo možné zachytit negaci v těchto větách:

(9) *Nikdy bych*/A----- ***ne****byl*/A----- *věřil*/VCA1SR, *že jsem*/VDA1SP *schopen takhle někoho kopnout*/VFA---.

(10) *V živoucím městě bych*/A----- *byl*/A----- *nikdy* ***ne****mohl*/VCA1SR *tolik pozorovat*/VFA---.,

kde celý víceslovný slovesný komplex vyjadřuje negaci, přičemž ve větě (9) se neguje pomocné sloveso *nebyl*, zatímco ve větě (10) naopak plnovýznamový tvar *nemohl* — obě v podstatě plně synonymní varianty čeština připouští. Problém však může nastat u jevu dvojí negace.

(11) *Nikterak však nejsem*/A----- *nepřipraven*/VDP1SP *na těžké časy*.

Dvojí negace vyjadřuje oslabenou afirmaci (litotes) a je otázka, zda tuto vlastnost komplexu připsat, a pokud ano, jak by to bylo spolehlivé. Uživatel je tedy při svých rešerších odkázán na kombinaci vlastností obsažených v tagu (kde je kromě negace i vid a další morfologické vlastnosti izolovaného slovesného tvaru) a v morfosyntaktickém **verbtagu**. Další problém nastává u konstrukcí, ve kterých můžeme negaci použít na různých místech, přičemž se význam mírně změní, jako v příkladech (12) a (13):

(12) *Nebyl bych spal*/VCA1SR.

(13) *Byl bych nespal*/VCA1SR.

Druhým jevem, který ve **verbtagu** není zpracován, jsou struktury s modálními a fázovými slovesy a struktury rezultativní. Hlavním důvodem je to, že u těchto struktur nepanuje mezi lingvisty shoda, které konstrukce mezi ně patří. U modálních konstrukcí jde například o konstrukce se slovesy *nutit*, *chtít* nebo *hodlat*, která se morfologicky i syntakticky liší od nesporných modálních sloves, jako je *muset* nebo *smět*. Podobně u rezultativních konstrukcí není neshoda pouze v tom, která slovesa (např. *mít*, *dostat*, *zůstat*) tvoří rezultativní konstrukce, ale i v tom, zda za rezultativní konstrukce považovat i konstrukce s objektem typu *mít dort upečen*. Kromě toho jsou v některých pojetích za rezultativní považovány i konstrukce se slovesem *být*, jako třeba *bylo uklizeno*. Tyto konstrukce se ale dají jen těžko odlišit od trpného rodu, a bylo by tak obtížné je správně označkovat.

4. JAK SE VE *VERBTAGU* NEZTRATIT ANEB RYCHLÁ NAVIGACE PRO UŽIVATELE

Pro uživatele korpusu může být zpočátku nesnadné orientovat se ve významu a pozicích jednotlivých hodnot ve **verbtagu**. Proto v této kapitole sepíšeme návod, jak snadno **verbtag** číst. V prvé řadě je třeba si uvědomit, že každý (složený) slovesný tvar má právě jeden člen (plnovýznamové sloveso), který nese **verbtag** začínající pís-

menem **V**. Ostatní slovesa ve složeném slovesném tvaru jsou pomocná a jejich **verbtag** začíná písmenem **A**. Mezi složené tvary ale (jak je již řečeno výše) nepočítáme konstrukce s fázovými a modálními slovesy, konstrukce rezultativní se slovesem *mít* nebo *dostat*, ani konstrukce s významovými slovesy, která vyžadují jako své doplnění infinitiv. Pomocné sloveso tedy bude vždy tvar slovesa *být*, *bývat*, avšak opačné tvrzení neplatí — ne každé sloveso *být* nebo *bývat* je pomocné. Kromě slovesa *být* v existenčním významu (*strašidla nejsou*), je sloveso *být* plnovýznamové i tam, kde slouží jako spona, kromě opisného pasiva.

Různé tvary slovesa mají vyplněné různé pozice ve **verbtagu**, přičemž tři první (**V**, tvar, slovesný rod) jsou vyplněné vždy, kdežto u druhé trojice (osoba, číslo, čas) mohou některé hodnoty chybět. V následující tabulce ukážeme, které pozice jsou vyplněné pro které tvary. Zaplněné pozice značíme tečkou, nezaplněné znakem **-**.

aktivní indikativ	VDA...
pasivní indikativ	VDP...
aktivní kondicionál	VCA...
pasivní kondicionál	VCP...
aktivní imperativ	VIA..-
pasivní imperativ	VIP..-
aktivní infinitiv	VFA---
pasivní infinitiv	VFP---
aktivní přechodník	VTA-..
pasivní přechodník	VTP-..
doplněk/volné pasivní příčestí	VOP-.-

TABULKA 6. Obsazení pozic ve verbtagu plnovýznamového slovesa podle tvarů

Některé hodnoty **verbtagu** se vyskytují pouze u jednoduchých slovesných tvarů, některé pouze u složených slovesných tvarů (tzn., že ve větě musí být zároveň alespoň jedno pomocné sloveso) a některé mohou být přítomny jak u jednoduchých, tak u složených tvarů.

Hodnoty **verbtagu**, které se vyskytují pouze u jednoduchých tvarů:

— aktivní imperativ (**VIA**: *přijď*, *pojďme*)
— aktivní infinitiv (**VFA**: *dělat*)
— aktivní přechodník (**VTA**: *jda*, *přišedši*)
— volné pasivní příčestí (**VOP**: *zarmoucen*).

Hodnoty atributu **verbtag**, které se vyskytují u plnovýznamového členu ve složených tvarech (ostatní pasivní tvary a aktivní kondicionál):

— pasivní indikativ (**VDP**: *jsem pozván*, *byla potěšena*)
— aktivní kondicionál (**VCA**: *přišel bych*)

— pasivní kondicionál (**VCP**: *kdybych byl pozván*)
— pasivní imperativ (**VIP**: *buď vítán*)
— pasivní infinitiv (**VFP**: *být vyvolán*)
— pasivní přechodník (**VTP**: *jsa připravován, byvši potěšena*).

U aktivního indikativu nastávají všechny tři zmíněné možnosti v závislosti na slovesném čase, vidu a osobě, jak ilustruje následující tabulka.

VDA..F	*půjdu, poplavete, koupím*	*budu dělat, budou číst*
VDA..B	*abdikuje*	
VDA..P	*čteš, pracují*	
VDA..R	*přišel, vyhráli, (já) spal*	*četl jsem, prohrál*s,[10] *přišel jest*[11]
VDA..Q		*(jak) jsem byl řekl*

TABULKA 7. Jednoduché a složené tvary u aktivního indikativu

Čtvrtá pozice ve **verbtagu**, osoba, se shoduje s osmou pozicí v tagu pomocného slovesa (v případě složeného tvaru), anebo plnovýznamového slovesa (v případě jednoduchého tvaru) až na případ tzv. imperativu ve 3. osobě (např. *čert to vem*), kdy slovesný tvar *vem* má v tagu druhou osobu, ale ve **verbtagu** třetí.[12]

Pátá pozice ve **verbtagu**, číslo, se shoduje se čtvrtou pozicí v tagu pomocného slovesa složeného slovesného tvaru až na případ vykání jednotlivci, potažmo dalších konstrukcí, ve kterých je finitní tvar pomocného slovesa v plurálu, ale příčestí v singuláru. Hodnota **v** ve **verbtagu** vyjadřuje nesoulad mezi číslem v tagu u finitního pomocného slovesa a číslem v tagu ostatních částí složeného slovesného tvaru.

Šestá pozice atributu **verbtag**, čas, je určena na základě tvaru plnovýznamového slovesa a kombinace tvarů všech pomocných sloves, která náleží do složeného slovesného tvaru.

Při práci v rozhraní KonText lze **verbtag** použít pro pokročilé vyhledávání CQL, a to jak samostatně, tak v kombinaci s jinými atributy (např. dotaz [verbtag="VD...F" & lemma=".*(běžet|běhat|bíhat|běhnout)"] najde všechny výskyty indikativu futura sloves se slovním základem *běžet/běhat*). Lze ho využít pro vytváření frekvenčních seznamů (Frekvence — Vlastní — Atribut **verbtag**). Dá se také zobrazit u vyhledaných slov přímo v konkordanci (Zobrazení — Korpusová nastavení — Poziční atributy, zaškrtnout **verbtag**).

10 Tvary indikativu v 2. osobě singuláru s připojenou klitikou *-s* jsou graficky jedním slovem, ale ve skutečnosti jde o víceslovný token (agregát) dvou tvarů; dostane proto dva tagy i dva verbtagy.

11 Tento tvar je sice archaický, přesto se vyskytuje v korpusech řady SYN; většinou jde o citáty ze starší literatury. Jinak je minulý čas indikativu ve 3. osobě v současné češtině vyjádřen tvarem jednoduchým.

12 Takové případy (*Pozdrav Pán Bůh*), ponejvíce ustálené výrazy, určujeme podle seznamů těchto výrazů a také podle ne/přítomnosti substantivního subjektu u tvaru imperativu.

5. SLOŽITĚJŠÍ PŘÍPADY

Při určování, jaký **verbtag** se má přiřadit tomu či onomu slovesu, narážíme jak na jednoduché, tak na složitější případy. V této kapitole uvádíme, jaké typy konstrukcí činí při značkování potíže, ať už teoretické nebo praktické. Teoretické potíže pramení obvykle z toho, že dojde ke koordinaci adjektiva a trpného příčestí, takže není jasné, jestli má být spona označena jako pomocné, nebo jako plnovýznamové sloveso. Další podobné problémy vznikají, je-li v koordinaci druhá spona vypuštěna, ale trpná příčestí se neshodují ve všech kategoriích.

K praktickým problémům (tedy k problémům, které nastávají v procesu automatického přiřazení atributu **verbtag** tokenům v textech) řadíme takové, kdy je teoreticky jasné, jak mají být slovesa značkována, ale z nějakých důvodů nejsme schopni správné značky přiřadit. Následuje několik příkladů typů složitějších jevů; pokud je u takového jevu zjištěna chybně automaticky přiřazená hodnota atributu **verbtag** v korpusu, uvádíme níže jak chybnou značku, tak značku, která by v daném místě měla být správně přiřazena. Podrobnější popis automatického značkování se nachází v části 6. Publikované korpusy jsou referenční, tedy neměnné, nicméně identifikované typy chyb se vždy snažíme v dalších, nově publikovaných korpusech odstraňovat.

a. koordinace příčestí:

(14) *Mám jen okamžik na to, abych se zamyslela*/VCA1SP *a hlavně se zhluboka nadechla*/VCA1SP/**chybně**:VDA3SR.[13] (korpus SYN2020)

(15) *Před padesáti lety byl*/A----- *cikánský tábor v Březince zlikvidován*/VDP3SR *a zbývající věz*ni *zavražděn***i**/VDP3PR/**chybně**:VOP-P-. (SYNv9)

(16) *Použité zdroje musí být citovány*/VFP--- *a odkazováno*/VFP---/**chybně**:VOP-S-*na ně podle normy ISO 690*. (SYN2020)

(17) *Na digitalis jsem samozřejmě myslel*/VDA1SR *a byl*/VDA3SR *součástí toxikologického rozboru*. (SYNv9)

Ve větách (14) až (16) jsou vždy koordinována dvě příčestí (buď minulá, nebo trpná). Pro účely náležitého značkování tu schází další pomocné sloveso, jehož vlastnosti by nástrojům automatického značkování pomohly určit morfosyntaktické hodnoty druhého členu koordinace: ve větě (14) schází vyjádření kondicionálu 1. osoby singuláru ve víceslovném tokenu *aby**ch**: … a* [*aby**ch***] *se hlavně zhluboka nadechla…*. Ve větě (15) chybí sponové ***byli***: … [***byli***] *zavražděni*… a funkce trpného příčestí *zavražděni* se mylně interpretuje jako doplněk, navíc tu není shoda s předchozím příčestím *zlikvidován* v čísle. Ve větě (16) schází infinitiv opisného pasiva ***být***: … *a* [***být***] *odkazováno*,

13 Nejprve uvádíme náležitou značku, zde VCA1SP, a poté značku chybnou, která je v příslušném korpusu, zde VDA3SR. Podobně dále.

a funkce trpného příčestí *odkazováno* se tak nesprávně chápe jako doplněk; opět zde není shoda s předchozím příčestím *citovány* v čísle. Ve větě (17) naopak nejde o koordinaci dvou příčestí, nýbrž o koordinaci dvou klauzí v rámci souvětí.

b. sloveso *být* se sponovou a pomocnou funkcí zároveň

(18) *Vítěz bude*/VDA3SF *znám a vyhlášen*/VOP-S- *v 18 hodin.* (SYN2020)

(19) *Do února 1988 byl*/A----- *celý mechanismus připraven*/VDP3SR *a hotov k rozjezdu.*

Ve větě (18) vystupuje finitní tvar futura *bude* ve dvojí funkci — jednak jako spona, jednak jako pomocné sloveso v pasivní konstrukci. Sloveso *být* je označeno jako plnovýznamové proto, že jmenná část přísudku stojí blíže[14] než trpné příčestí *připraven*. Samo trpné příčestí je označeno jako „ostatní funkce trpného příčestí" (**O**).

Odlišně se značkuje tvar *byl* a trpné příčestí *připraven* ve větě (19), tvoří totiž těsnější strukturu s tvarem *byl*, než je spojení *byl … hotov*.

c. pomocné sloveso *být* ve futuru a infinitiv nedokonavého slovesa

(20) *Ideální bude*/VDA3SF/**chybně**:A----- *psát*/VFA---/**chybně**:VDA3SF *knížku ve dvou.* (SYNv9)

(21) *Nejdůležitější bude*/VDA3SF/**chybně**:A----- *umět*/VFA---/**chybně**:VDA3SF *nepracovat*/VFA---. (SYN2020)

Žádná z vět (20) a (21) neobsahuje konstrukci s opisným futurem: *bude* je tu vždy spona spojující podmět v infinitivu (nedokonavé sloveso *jít*, resp. *umět*). Slovesa v obou větách jsou značkována nesprávně — konstrukce s infinitivem nedokonavého slovesa se anotují obtížně vzhledem k možné homonymii.

d. složené tvary, kde se číslo pomocného slovesa neshoduje s číslem slovesa plnovýznamového

(22) *Nešel*/VCA2vP *byste*/A----- *někdo na oběd?* (SYNv9)

(23) *…, že jsme*/A----- *každý zaměstnán*/VDP1vP/**chybně**:VDP1SP *jinde.* (SYNv9)

Ve větě (22) nejde nutně o vykání, i když slovesné tvary tak vypadají. V současné době se tyto konstrukce označují hodnotou **v** na pozici čísla a my předpokládáme, že rozšíříme definici této hodnoty. Nově bude zahrnovat případy neshody v čísle u finitního

14 V případech, kdy si u jednoho tvaru plnovýznamového slovesa konkurují dvě pomocná slovesa či jeden tvar slovesa *být* může být zároveň sponovým i pomocným vždy rozhoduje vzdálenost mezi těmito tvary. Jiné řešení by vyžadovalo, aby celý systém atributu **verbtag** byl kvůli okrajovému jevu výrazně složitější.

tvaru pomocného slovesa a příčestí, kdy pomocné sloveso je v plurálu a příčestí v singuláru. I ve větě (23) tak bude hodnota čísla ve **verbtagu** rovna **v**.

Uvedli jsme několik typů složených slovesných tvarů, u nichž se hodnota atributu **verbtag** určuje nesnadno. S výjimkou koordinace příčestí se tyto typy (a jejich kombinace) vyskytují v korpusech dost výjimečně, to však neznamená, že by se jim v zájmu co nejlepšího značkování neměla v budoucnu věnovat pozornost.

6. AUTOMATICKÉ ZNAČKOVÁNÍ A JEHO ÚSPĚŠNOST

Vzhledem k rozsahu korpusu se jeho lemmatizace a značkování provádějí plně automaticky a to se samozřejmě týká i atributu **verbtag**. Obvyklý postup při značkování textů je takový, že každému slovu se při morfologické analýze přiřadí všechny jeho možné interpretace, tedy kombinace lemmat a morfologických značek (například slovu *při* se přiřadí tyto údaje: (i) lemma *při* a značka pro předložku s lokálem, (ii) lemma *pře* a značky pro substantivum ženského rodu v dativu, akuzativu a lokálu singuláru a (iii) lemma *přít* (*se*) a značka pro sloveso ve 2. osobě singuláru imperativu) a potom se v procesu zvaném desambiguace vybírá z množiny těchto interpretací jedna v daném kontextu správná. Desambiguaci lze provádět jednak programy založenými na ručně vytvořených (lingvistických) pravidlech, která odstraňují nevhodné interpretace a ponechávají pouze interpretace žádoucí (například tvar *se*, po němž následuje sloveso, jistě není vokalizovaná předložka *s*), jednak programy používajícími strojové učení, tzv. taggery, které se na menším souboru kvalitně, ručně označkovaných dat „naučí", jak má být cílový jazyk značkován, a vytvořený jazykový model se pak aplikuje na texty jiné, nedesambiguované. Pro značkování korpusů řady SYN Českého národního korpusu (ČNK) se používají metody obě, tzv. hybridní značkování: nejprve se uplatňuje systém lingvistických desambiguačních pravidel (zvláště syntaktických, ale i sémantických a fonetických), zvaný LanGr,[15] a poté desambiguaci dokončí tagger, konkrétně „neuronový" tagger MorphoDiTa,[16] který desambiguuje případy, jež pravidla nedokázala při desambiguaci rozhodnout. Podrobnější popis postupu značkování je popsán jinde, srov. Květoň (2006), Jelínek et al. (2011), Petkevič (2014), Jelínek et al. (2021); zde se zaměříme jen na atribut **verbtag**.

Pro přehlednější práci při přípravě korpusu je **verbtag** během zpracování textů spojen s původním tagem, dohromady tak vytvářejí komplexní 21poziční značku. Jako nově vytvořený atribut nebyl **verbtag** dosud obsažen v morfologických slovnících, díky nimž je možné přiřadit slovům jejich možné interpretace, nepracovalo se s ním v desambiguačních pravidlech a nebyl ani zahrnut v ručně značkovaných „trénovacích" datech pro tagger. Všechny tyto otázky bylo nutné v souvislosti se zavedením **verbtagu** řešit.

15 LanGr je nástroj vytvořený pro desambiguaci založenou na lingvistických pravidlech v ÚTKL FF UK, viz (Květoň, 2006).

16 https://ufal.mff.cuni.cz/morphodita.

6.1 ÚPRAVA MORFOLOGICKÉHO SLOVNÍKU PRO PRÁCI S *VERBTAGEM*

Morfologický slovník (Štěpánková et al., 2020) byl automaticky upraven tak, že počet původních pozic v morfologické značce se rozšířil z 15 na 21, konkrétně u sloves tak, že zároveň mohlo dojít k navýšení počtu jejich značek. Například k jednoznačné interpretaci tvaru *pracovat* v tagu (infinitiv nedokonavého slovesa) přibylo sedm možností ve **verbtagu**: jeden **verbtag** pro aktivní infinitiv (VFA---) a šest různých hodnot atributu **verbtag** pro indikativ složeného futura v různých osobách a číslech (například VDA2PF, tj. indikativ 2. osoby plurálu futura pro *budete pracovat*, uvedený u plnovýznamového slovesa *pracovat*). U minulých příčestí se tak původní počet tagů zvýšil 14krát, u trpných příčestí dokonce 28krát. Průměrný počet různých tagů na token před desambiguací tak obecně vzrostl z původního 4,06 na 5,67, u sloves pak z 2,49 na 11,83 tagu na token (měřeno na korpusu SYN2020).

6.2 ÚPRAVA PRAVIDLOVÉHO DESAMBIGUAČNÍHO SYSTÉMU V SOUVISLOSTI S *VERBTAGEM*

Desambiguační pravidla systému LanGr byla se zavedením **verbtagu** rozšířena o pravidla specificky zaměřená na **verbtag**, a to jak pravidla jednoduchá (například stojí-li pozitivní tvar slovesa *být* v indikativu prézentu 1. nebo 2. osoby těsně po minulém příčestí se shodným číslem, např. *přišla jsem*, je sloveso *být* pomocné a minulé příčestí je značkováno jako tvar indikativu), tak pravidla složitější, zahrnující více slovesných tvarů v jedné klauzi nebo složené slovesné tvary, jež tvoří jeden celek, oddělené jinými slovy. Pravidla zaměřená na **verbtag** poměrně málo chybují, nezvládají však dosud některé složitější případy, jako je pomocné sloveso *být* oddělené od příčestí vloženou vedlejší větou, např. *Zkrátka jsem se, co se věku týče, utrhl.*, a v mnoha případech desambiguaci nedokončí, a ponechají tak částečně desambiguovaný text k dokončení taggeru. Při vývoji pravidel zaměřených na **verbtag** byla v některých případech pravidla chybně formulována, například u po sobě stojících infinitivů modálního a významového nedokonavého slovesa byl i u druhého slovesa ponechán jako jediná možná interpretace indikativ futura a v následné fázi už tagger nemohl tuto chybu napravit; často se tedy v korpusu SYN2020 vyskytuje chybná konstrukce jako ve větě (24). V korpusu SYNv10 je tato chyba již z větší části opravena.

(24) *Teď budeme*/A----- *muset*/VDA1PF *čekat*/VFA---/**chybně**:VDA1PF.

Během desambiguace se počet tagů na token snižuje z 5,67 na 1,76 (cílový stav je právě jeden tag na token).

6.3 ÚPRAVA TRÉNOVACÍCH DAT PRO NEURONOVÝ TAGGER

Tagger potřebuje ke své práci nejprve vhodná „trénovací data“. Po zavedení **verbtagu** byla původní trénovací data — korpus Etalon — manuálně doplněna o **verbtag** (možné **verbtagy** byly doplněny automaticky a dva anotátoři pak v každém víceznač-

ném případě vybírali v daném kontextu správnou značku; tam, kde se anotátoři neshodli, ověřil výsledek odborník). Nově upravená trénovací a testovací data byla zveřejněna (Skoumalová, 2021).[17]

Tagger MorphoDiTa (Straka et al., 2019), založený na tzv. hlubokém učení (či neuronových sítích), byl potom na těchto datech natrénován a otestován. Tagger má poměrně vysokou úspěšnost, tedy poměrně málo chybuje, a to jak obecně, tak při značkování **verbtagu** (viz níže), zřídka se však stává, že se dopouští i „hloupých" chyb v (pro člověka) jednoznačných kontextech, například u těsně sousedícího pomocného slovesa *být* a minulého příčestí neurčí správně sloveso *být* jako pomocné (ovšem poté, co ani pravidlový systém LanGr z důvodu přílišné opatrnosti nebo opomenutí tuto desambiguaci neprovede).

6.4 ÚSPĚŠNOST ZNAČKOVÁNÍ VČETNĚ *VERBTAGU*

Následující tabulka ukazuje úspěšnost hybridního automatického značkování (podíl správně určených hodnot daného atributu podle testovacích dat). Uvádíme celkovou úspěšnost pro všechny atributy (tedy podíl tokenů, u nichž je správně určeno lemma, tag i **verbtag**), dále úspěšnost určení slovního druhu, celé morfologické značky a **verbtagu** u sloves — měřeno na datech korpusu Etalon (10násobná křížová validace).

Atribut	Úspěšnost
Lemma+tag+verbtag	97,36 %
POS	99,56 %
Tag	97,60 %
Lemma	99,67 %
Verbtag	99,77 %
Verbtag u sloves	98,57 %

TABULKA 8. Úspěšnost značkování hybridním systémem

Celková úspěšnost lemmatizace a morfologické anotace korpusu SYN2020 je 97,36 %, tj. 2,64 % tokenu má chybně určené lemma, tag nebo **verbtag**. Úspěšnost určení **verbtagu** je zdánlivě velmi vysoká, 99,77 %, je to ale dáno tím, že u všech slovních druhů kromě sloves je **verbtag** vždy jednoznačný. Úspěšnost značkování **verbtagu** měřená pouze na slovesech je 98,57 %, chybně určený **verbtag** má tedy 1,43 % sloves.

7. ZÁVĚR

Nově zavedený atribut **verbtag** může být pro pokročilejší uživatele korpusů ČNK vhodným pomocníkem pro práci se slovesy v synchronních korpusech řady SYN. V přehledném 6pozičním atributu může uživatel na jednom místě zjistit morfosyntaktické vlast-

17 https://wiki.korpus.cz/doku.php/cnk:etalon.

nosti složeného slovesného tvaru, může vyhledávat slovesa podle způsobu, času, osoby atp., aniž by musel brát v úvahu specifika jednotlivých složených tvarů.

Seznámit se s atributem **verbtag** představuje pro uživatele ČNK užitečnou investici i proto, že se do budoucna počítá s používáním tohoto značkování u všech dalších synchronních korpusů; plánuje se i rozšíření stejného značkovacího standardu na mluvené korpusy ČNK a výhledově i na korpusy historické (přinejmenším na korpusy 19. a 20. stol.).

U atributu **verbtag** nedojde v dohledné době ke změnám koncepce, věnujeme však úsilí zejména:

— vyjasnění některých teoretických východisek
— dalšímu zlepšení kvality značkování, tedy odstranění současných nedostatků a chyb
— adekvátnějšímu popisu a značkování typů neshody pomocného tvaru slovesa *být* v plurálu a l-ového a trpného příčestí v singuláru, případně dalších zjištěných zvláštností.

POUŽITÉ KORPUSY

Křen, M. — Cvrček, V. — Henyš, J. — Hnátková, M. — Jelínek, T. — Kocek, J. — Kováříková, D. — Křivan, J. — Milička, J. — Petkevič, V. — Procházka, P. — Skoumalová, H. — Šindlerová, J. — Škrabal, M. (2020): *SYN2020: reprezentativní korpus psané češtiny*. Praha: Ústav Českého národního korpusu FF UK. https://www.korpus.cz

Křen, M. — Cvrček, V. — Henyš, J. — Hnátková, M. — Jelínek, T. — Kocek, J. — Kováříková, D. — Křivan, J. — Milička, J. — Petkevič, V. — Procházka, P. — Skoumalová, H. — Šindlerová, J. — Škrabal, M. (2021): *Korpus SYN, verze 9 z 5. 12. 2021*. Praha: Ústav Českého národního korpusu FF UK. https://www.korpus.cz

LITERATURA

Апресян, Ю., Д. — Богуславский, И., М. — Иомдин, Б., Л., и др. (2005): Синтаксически и семантически аннотированный корпус русского языка: современное состояние и перспективы // Национальный корпус русского языка: 2003—2005. М.: Индрик, s. 193–214. https://ruscorpora.ru/new/sbornik2005/12apresyan.pdf

Bejček, E. et al. (2011): Prague Dependency Treebank 2.5, LINDAT/CLARIAH-CZ digital library at the Institute of Formal and Applied Linguistics (ÚFAL). Prague: Faculty of Mathematics and Physics, Charles University, http://hdl.handle.net/11858/00-097C-0000-0006-DB11-8.

Bejček, E. — Panevová, J. — Popelka, J. — Straňák, P. — Ševčíková, M. — Štěpánek, J. — Žabokrtský, Z. (2012): Prague Dependency Treebank 2.5 — a revisited version of PDT 2.0. In: *Proceedings of the 24th International Conference on Computational Linguistics (Coling 2012)*. Mumbai, India: Coling 2012 Organizing Committee, s. 231–246.

Dönicke, T. (2020): Clause-Level Tense, Mood, Voice and Modality Tagging for German. *TLT*. Göttingen: University of Göttingen, Centre for Digital Humanities Papendiek 16, 37073. https://aclanthology.org/2020.tlt-1.1.pdf

Jelínek, T. — 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, s. 154–170.

Jelínek, T. — Křivan, J. — Petkevič, V. — Skoumalová, H. — Šindlerová, J. (2021): SYN2020: A New Corpus of Czech With an Innovated Annotation. In: K. Ekštein — F. Pártl — 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, s. 48–59. https://doi.org/10.1007/978-3-030-83527-9

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

Patejuk, A. — Przepiórkowski, A. (2014): Synergistic development of grammatical resources: A valence dictionary, an LFG grammar, and an LFG structure bank for Polish. In: V. Henrich — E. Hinrichs — D. de Kok — P. Osenova — A. Przepiórkowski (eds.), *Proceedings of the Thirteenth International Workshop on Treebanks and Linguistic Theories (TLT 13)*. Tübingen: Department of Linguistics (SfS), University of Tübingen, s. 113–126.

Petkevič, V. (2014): Problémy automatické morfologické disambiguace češtiny. *Naše řeč*, 97, 4–5, s. 194–207.

Petkevič, V. — Rosen, A. — Skoumalová, H. — Vítovec, P. (2015): Analytic Morphology — Merging the Paradigmatic and Syntagmatic Perspective in a Treebank. In: J. Piskorski — L. Pivovarova — J. Šnajder — H. Tanev — R. Yangarber (eds.), The 5th Workshop on Balto-Slavic Natural Language Processing (BSNLP 2015). Hissar, Bulgaria, s. 9–16. http://bsnlp-2015.cs.helsinki.fi/.

Ramm, A. — Loáiciga, S. — Friedrich, A. — Fraser, A. (2017): Annotating tense, mood and voice for English, French and German. *Proceedings of ACL 2017, System Demonstrations*. Vancouver: Association for Computational Linguistics. https://www.cis.uni-muenchen.de/~fraser/pubs/ramm_acldemo2017.pdf

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

Straka, M. — Straková, J. — 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, s. 137–150.

Štěpánková, B. — Mikulová, M. — Hajič, J. (2020): The MorfFlex Dictionary of Czech as a Source of Linguistic Data. In: *Euralex XIX Proceedings Book: Lexicography for inclusion*. European Association for Lexicography, s. 387–391.

Tomáš Jelínek | Ústav teoretické a komputační lingvistiky, Filozofická fakulta Univerzity Karlovy | Celetná 13, 110 00 Praha 1
ORCID ID**:** 0000-0002-8521-4715
tomas.jelinek@ff.cuni.cz

Vladimír Petkevič | Ústav teoretické a komputační lingvistiky, Filozofická fakulta Univerzity Karlovy | Celetná 13, 110 00 Praha 1
ORCID ID: 0000-0003-0468-4158
vladimir.petkevic@ff.cuni.cz

Hana Skoumalová | Ústav teoretické a komputační lingvistiky, Filozofická fakulta Univerzity Karlovy | Celetná 13, 110 00 Praha 1
ORCID ID**:** 0000-0002-3519-0233
hana.skoumalova@ff.cuni.cz