# Analytic Morphology – Merging the Paradigmatic and Syntagmatic Perspective in a Treebank

**Vladimír Petkevič Alexandr Rosen Hana Skoumalová Přemysl Vítovec**
Charles University in Prague, Faculty of Arts
first_name.surname@ff.cuni.cz premysl.vitovec@gmail.com

## Abstract

We present an account of analytic verb forms in a treebank of Czech texts. According to the Czech linguistic tradition, description of periphrastic constructions is a task for morphology. On the other hand, their components cannot be analyzed separately from syntax. We show how the paradigmatic and syntagmatic views can be represented within a single framework.

## 1 Introduction

Analytic verb forms (henceforth AVFs) consist of one or more auxiliaries and a content verb. The auxiliaries can be seen either as marking the content verb with morphological categories or as being part of a multi-word expression, to which the categories are assigned. This is the perspective taken by all standard grammar books of Czech, which treat AVFs as a morphological rather than a syntactic phenomenon. AVFs are listed in conjugation paradigms quite like synthetic forms for a good reason: from a meaning-based view, whether a certain category in a certain language happens to be expressed by a single word or a string of words is an epiphenomenon.

From a different perspective, each of the components has its role in satisfying a syntactic grammaticality constraint and in making a contribution to the lexical, grammatical or semantic meaning of the whole. This approach is common in both corpus and generative linguistics (including theories such as LFG or HPSG),[1] where each form is treated as a syntactic word and AVFs belong to the domain of syntax. As a result, morphological categories are not assigned to units spanning word boundaries. This is for several reasons: (i) an AVF does not emerge as a single orthographical word (often even phonological word); (ii) AVFs may be expressed by a potentially discontinuous string of a content verb and multiple auxiliaries, sometimes in an order determined by information structure rather than by rules of morphology or syntax proper; (iii) some auxiliary forms share properties with some content words – like weak pronouns, the past tense auxiliary is a 2nd position clitic.

Our claim is that the two views are compatible, complementary and amenable to formalization within a single framework, combining the traditional paradigmatic view with a syntagmatic view. This reconciliatory effort is part of a more general goal: a choice of different interpretations of annotated corpus data, depending on the preferences of a user or an application.

AVFs are assigned a syntactic structure: the (finite) auxiliary is treated as the surface head, governing the rest of the form – the deep head.[2]

In Czech, AVFs are used to express the verbal categories of mood, tense and voice in periphrastic passive (all moods and tenses), in periphrastic future, in 1st and 2nd person past tense, in pluperfect and in present and past conditional. In all these forms the auxiliary is *být* 'to be'. Here we focus on past tense and conditional forms, including pluperfect and past conditional, but the solution works for all the above AVFs, and covers also negation of some components of the AVFs by the prefix *ne-* and can be extended to some other kinds of function words, such as prepositions and conjunctions. In (1)–(4) below we show some properties of the past and conditional forms. The finite auxiliary is marked for person, number and mood, while the *l*-participle[3] is marked for gender and number. Past tense (1) consists of the auxiliary in the present tense and the *l*-participle of

[1] See, e.g., Webelhuth (1995), Dalrymple (1999), Pollard and Sag (1994).

[2] We use the term *government* in the sense of "subcategorization" or "imposition of valency requirements."

[3] We avoid the frequently used term *past participle* because the same form is also used in *present* conditional.



*Proceedings of the 5th Workshop on Balto-Slavic Natural Language Processing*, pages 9–16,
Hissar, Bulgaria, 10–11 September 2015.

the content verb. Present conditional (2) consists of the conditional auxiliary and the content verb's *l*-participle. Past conditional (3) includes an additional *l*-participle of the auxiliary. In (4) we show that other words can be inserted, the auxiliary *l*-participle can be repeated, and any *l*-participle can be negated.

(1) *Já jsem přišel*
I be.PRS.1SG come.PTCP.M.SG
'I have come.'

(2) *Já bych přišel*
I be.COND.1SG come.PTCP.M.SG
'I would come.'

(3) *Já bych byl přišel*
I be.COND.1SG be.PTCP.M.SG come.PTCP.M.SG
'I would have come.'

(4) ***Kdybys*** *tenkrát* ***nebyl***
If-be.COND.2SG back then be.PTCP.M.SG.NEG
***býval*** *tak duchapřítomně*
be.PTCP.M.SG.ITER so readily
***zasáhl***...
intervene.PTCP.M.SG
'If you haven't intervened so readily back then...'

We exemplify the solution using a treebank of Czech. The framework is based on the HPSG.[4] The annotation, originally produced by a stochastic dependency parser, is checked by a formal grammar, using a valency lexicon and implemented in Trale.[5] Trees complying with grammatical and lexical constraints are augmented with information derived from the lexicon and any annotation provided by a stochastic parser.

## 2 Previous Work

Grammars of Czech take a paradigmatic perspective, treating AVFs as an exclusively morphological phenomenon (Karlík et al., 1995; Cvrček et al., 2010; Komárek et al., 1986), glossed over without describing their syntagmatic and word-order properties. In Komárek et al. (1986), components of AVFs are assigned a particular grammatical meaning (person, number, tense, mood, voice) but their syntactic status is not specified.

The syntagmatic approach has been introduced to Czech by Veselovská (2003) and Veselovská and Karlík (2004), who analyze past tense and periphrastic passive within the Minimalist Program. A non-transformational account was pursued by Karel Oliva in an HPSG-inspired prototype grammar checker of Czech (Avgustinova et al., 1995). HPSG and LFG have been used to account for similar phenomena in closely related Polish, where the border between morphology and syntax is even less apparent than in Czech: all forms of the past tense and conditional auxiliaries are floating suffixes, attached either directly to the *l*-participle, or to some other preceding word. In the following, we briefly review several proposals for Polish, with an extension to Czech.

Based on the analysis of similar phenomena in West European languages, Borsley (1999) proposes two structures for modelling Polish AVFs: (i) classic VP complementation where the auxiliary is a subject-raising verb selecting a phrasal complement headed by an *l*-participle (Fig. 1), and (ii) flat structures where the auxiliary subcategorizes for an *l*-participle and its complements (Fig. 2).[6] The former is used for future tense while the latter for present conditional and past tense. This distinction is motivated by the ability or inability of the auxiliary to be preceded by the associated *l*-participle and its complements: while the future auxiliary allows for VP-preposing, the other auxiliaries are prohibitive in this respect.

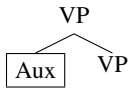


Figure 1: VP complementation.

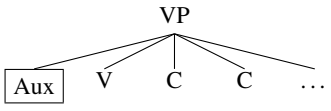


Figure 2: Flat structure.

Kupść (2000) follows Borsley (1999) but rejects the flat structure for past tense and present conditional as it makes incorrect predictions with respect to clitic climbing. Instead, she assumes VP complementation for all AVFs.

Kupść and Tseng (2005) argue against the unified treatment of AVFs. Only the future tense auxiliary behaves like a full syntactic word. In contrast, the forms of conditional auxiliary, albeit syn-

[4] See, e.g., Pollard and Sag (1994).

[5] See http://www.sfs.uni-tuebingen.de/hpsg/archive/projects/trale/. For more details see Jelínek et al. (2014).

[6] Heads are denoted by boxed nodes in the figures.

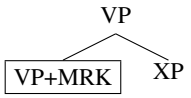


Figure 3: Local agreement marker.

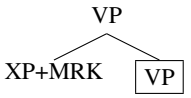


Figure 4: Nonlocal agreement marker.

tactic words, are clitics and thus subject to specific word order constraints (dependencies on various clitic hosts). Past tense is viewed as a simple tense and the past tense agreement markings are treated as inflectional elements, even if they are not attached to the *l*-participle. This analysis builds on (i) an observation that agreement markings are much more closely bound to the preceding word than conditional clitics, and (ii) the fact that there are no agreement markings used in the third person. As a result, the *l*-participle becomes the head of the whole structure. In order to ensure that the agreement marking appears somewhere in the structure the head acts as its trigger, carried by an agreement marker, either the head itself (Fig. 3), or some other preceding element (Fig. 4).

In light of diachronic and comparative considerations, Tseng and Kupść (2006) and Tseng (2009) extend the analysis of Kupść and Tseng (2005) to other Slavic languages, including Czech. The Czech past auxiliary forms are at the same time syntactic words and clitics with a restricted distribution (2nd position, cannot be negated). Moreover, the 2nd person singular clitic *-s* is similar to the Polish floating suffixes, suggesting that the head is the *l*-participle. As a result, the analysis of the Polish past tense can be applied to Czech with only a slight modification: the agreement markings are carried (mostly) by syntactic rather than morphological elements. No changes are proposed for the analysis of the Czech conditional either, where the only complication is the separable ending *-s* in the 2nd person singular (*bys*). However, the extremely restricted distribution of this phenomenon (only in combination with the *si* and *se* reflexives, resulting in the *sis* and *ses* forms) does not motivate treating Czech conditional structures like Polish past tense structures. The authors admit that the analysis based on the standard VP complementation is equally possible.

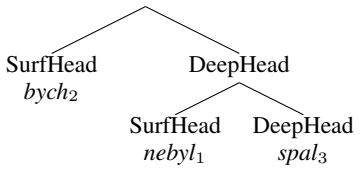


Figure 5: Structure of *nebyl bych spal* 'I wouldn't have slept'.

## 3 Our Approach

### 3.1 Two Types of Heads: Surface and Deep

In addition to strictly linguistic criteria for an optimal analysis of AVFs, our choice of the core representation format was influenced by the treebank design, which should allow for the derivation of syntactic structure and categorial labels of various shapes and flavours to be used in queries, responses and exported data. Adopting a uniform analysis for all AVFs simplifies the task. Each AVF is represented as a syntactic phrase with two constituents: a surface head daughter representing the auxiliary, and a deep head daughter representing the auxiliary's VP complement, which includes the content verb.[7] Multiple auxiliaries within a single AVF are surface heads within recursively embedded deep heads (see Fig. 5).[8]

### 3.2 Modifications of the HPSG Signature

HPSG represents linguistic data as typed feature structures. Words and phrases are subtypes of *sign*, a structure representing their form, meaning and combinatorial properties. Fig. 6 shows a simplified representation of an English sentence *dogs bark*. Types are in italics, attributes in upright capitals, boxed numbers indicate identity of values.

Each word consists of two parts: PHONOLOGY for the analyzed string and SS (SYNSEM) for its paradigmatic analysis. Phrases have two additional attributes: SD (SUBJECT-DAUGHTER) and HD (HEAD-DAUGHTER). The value of SS has L (LOCAL) as its single attribute; its NONLOCAL counterpart, used for discontinuous constituents, is not relevant for our example. The CAT (CATEGORY) attribute specifies (i) morphosyntactic properties of the expression as its HEAD features and (ii) its VALENCY. The CONT (CONTENT) attribute is responsible for semantic interpretation.

[7] Cf. Przepiórkowski (2007) for an equivalent distinction between syntactic and semantic heads.

[8] The node labels in Fig. 5 are actually feature structure attributes modelling phrasal daughters, abbreviated as SH and DH in Fig. 7.

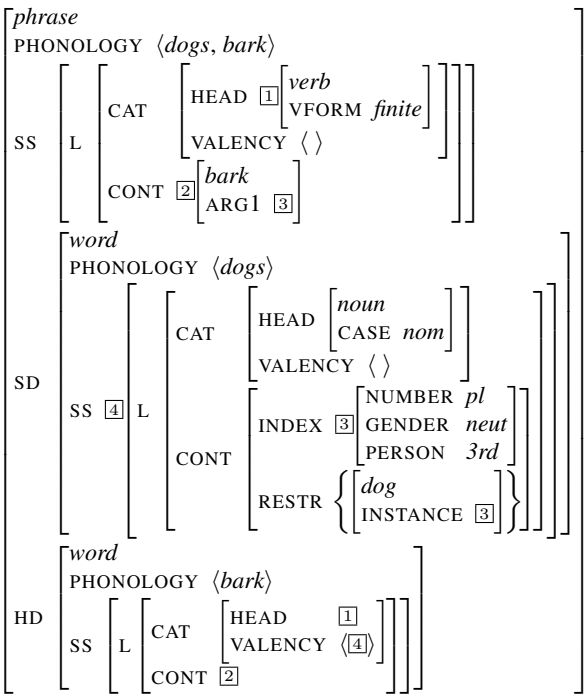


Figure 6: An HPSG representation of a sentence.

Some values are shared due to Head Feature Principle – HFP, projecting features of the head daughter to its phrasal mother, Valency Principle – ValP, a general valency satisfaction mechanism, and Semantics Principle. Morphosyntactic categories of the noun relevant for pronominal reference are the properties of CONTENT's INDEX, while those of the verb relevant for agreement are specified indirectly, as properties of its subject. For languages with rich morphology, NP-internal agreement and null subjects, such as Czech, other arrangements of morphosyntactic and valency features have been proposed.

In addition to the introduction of surface and deep heads, the standard HPSG signature has been modified in two main aspects: (i) at least in the current version, attributes such as LOCAL and HEAD are missing to simplify annotation of extensive data – discontinuities are treated as word order variations and head features are the value of CATEGORY; and (ii) the signature is extended by introducing a cross-clasification of morphological and morphosyntactic categories along three dimensions: morphological (inflectional), syntactic and semantic (lexical).[9] This is useful especially for word classes where classification criteria in the three dimensions do not coincide, such as numerals and pronouns. Their standard definitions are based on semantic criteria, but otherwise cardinal numerals and personal pronouns behave like nouns, whereas ordinal numerals and possessive pronouns behave like adjectives. The cross-classification can also be used to model some regular derivational relations, e.g., deverbal nouns and adjectives (inflectional classes) are derived from verbs (lexical class).

[9]See Rosen (2014) for more details.

### 3.3 Representing Analytic Categories

To accommodate AVFs, the 3D classification has been extended by an *analytic* dimension. The AC attribute specifies categories appropriate to the AVF as a whole. A verbal AC includes three basic properties: TENSE, MOOD and VOICE. Their values are encoded in the lexical specifications of function words, including the deep head's contribution, which is mediated through the valency frame of the auxiliary, including the content verb's ALEMMA. The rest is the task of ValP and HFP. More specifically, the surface head and its mother share their head features, including the analytic categories, and their deep valency frames – the deep structure is thus available in the phrasal category. Since AC is a head feature, tense, mood and voice are projected from the auxiliary as the surface head of the AVF.

### 3.4 An Example

The mechanism is illustrated in Figs. 5 and 7, using the past conditional form of the verb *spát* 'to sleep' (5).[10]

(5) *nebyl* *bych* *spal*
be.PTCP.M.SG.NEG be.COND.1SG sleep.PTCP.M.SG
'I wouldn't have slept'

Past conditional consists of the finite conditional auxiliary (*bych*), the *l*-participle form of the 'to be' auxiliary (*nebyl*) and the *l*-participle of the content verb (*spal*).[11]

[10]Fig. 5 ignores word order, which is specified within the PHON list of the phrase.

[11]Past conditional may include additional *l*-participle auxiliaries with the meaning unchanged: an iterative and a plain form (6). Passive past conditional, where two *l*-participle auxiliaries are obligatory (7), shows that the iterative is used to avoid two identical *l*-participles.

(6) *nebyl* *bych* *býval*
be.PTCP.M.SG.NEG be.COND.1SG be.PTCP.M.SG.ITER
(*byl*) *spal*
(be.PTCP.M.SG) sleep.PTCP.M.SG
'I wouldn't have slept'

The auxiliary *bych* is the surface head of the entire structure (see Fig. 5). Its sister phrase, i.e., *nebyl spal*, is the deep head, consisting of *nebyl* and *spal* as the surface and the deep head daughter. The auxiliary *bych* takes a single *l*-participle, unspecified as a content verb or auxiliary (see Fig. 8 below). This distinction, related to the interpretation of tense and/or voice of the AVF, is handled by grammar – see (8)–(12) below. In our example, the conditional auxiliary takes an auxiliary form *nebyl*, which in turn can take another *l*-participle as part of (i) indicative pluperfect (as in *byl spal*, 'he had slept'), or (ii) past conditional, as in our example. It is the presence or absence of the conditional auxiliary that identifies the structure as conditional or indicative. The substructure *nebyl spal* determines its tense as past, the resulting phrase is thus identified as past conditional.

The binary tree shown in Fig. 5 is represented as a feature structure of the *sdheaded* type (i.e., a surface/deep-headed phrase) in Fig. 7. The structure is similar to that in Fig. 6, except for the additions and some abbreviations: PH stands for PHONOLOGY, SH for the surface head daughter, DH for the deep head daughter, C for CATEGORY and COMPS for non-subject valency.[12]

The C attribute consists of three parts, representing three aspects of the category: analytic in AC, inflectional in IC and lexical in LC.[13] The AC attribute includes the lemma of the content verb (ALEMMA, shared as 2 with the lexical deep head and all its deep head projections), its mood, polarity (*minus* due to the negated auxiliary *nebyl*), tense and voice (*actv* for active). As in ALEMMA, 3 shows that the lexical deep head shares AVOICE with its projections. AMOOD and ATENSE are unspecified, because they can be determined only when the AVF is evaluated as a whole. E.g., the embedded DH phrase *nebyl spal* can be either part of indicative pluperfect or past conditional and the content verb participle *spal* can be part of past or pluperfect indicative, pluperfect indicative, present conditional or past conditional.

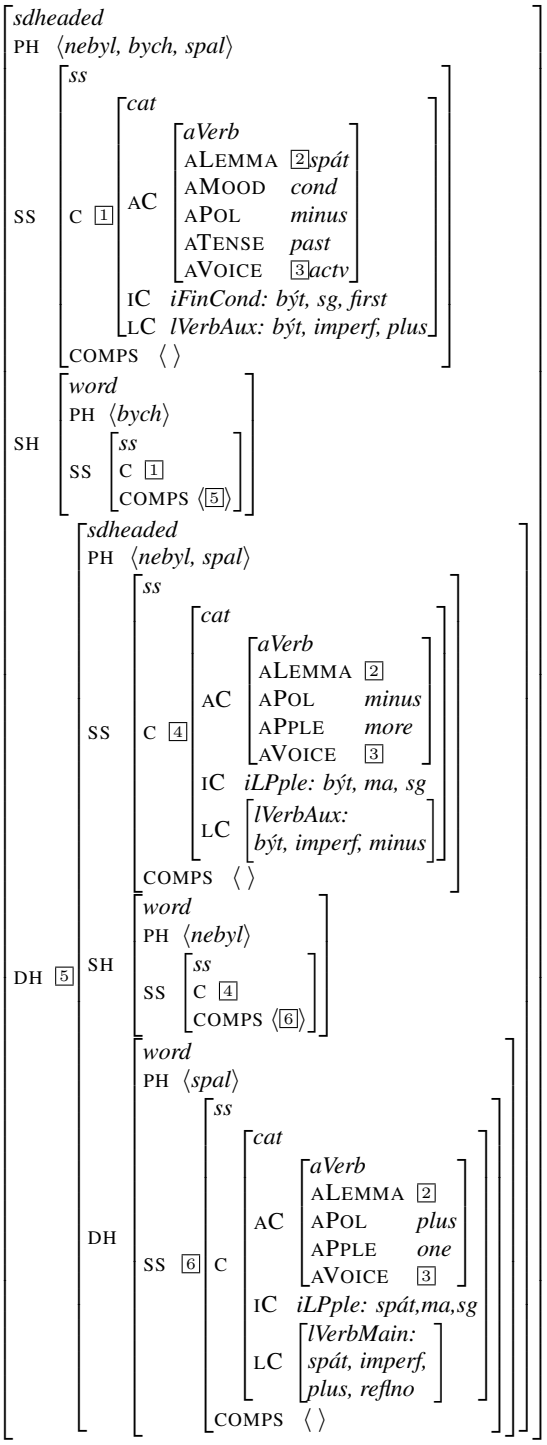


Figure 7: Analysis of *nebyl bych spal*.

(7) *byl* *bych* *býval* /
be.PTCP.M.SG be.COND.1SG be.PTCP.M.SG.ITER /
*?byl* *dopaden*
be.PTCP.M.SG catch.PASS.M.SG
'I would have been caught'

[12] Subject valency, specified by a separate attribute, SUBJ, is not shown in Fig. 7.

[13] sC for the syntactic aspect is omitted for brevity.

Values of the other two attributes IC and LC are abbreviated: the categorial type is followed by a list of attribute values. More importantly, they refer only to the surface head of the phrase. They are obtained from the input parse. The grammar checks that some of them (person, number, also gender) agree with corresponding values in the

rest of the predicate and/or in the subject.

Inflectional properties of the surface head as the finite conditional auxiliary (*iFinCond*) are *first* person *s*ingular of the lemma *být*. As the value of LC shows, *bych* is an *imperf*ective positive (*plus*) form of the verb *být*.

The top-level SS part, concerning the entire phrase, is followed by two daughters: (i) SH *bych*, whose categorial properties are identified with those of the whole phrase (1, due to HFP), and whose single valency is identified with the SS part of its deep head sister (5); and (ii) DH *nebyl spal*. As for AC, the negative polarity *minus* as the value of APOL is due to the negative form *nebyl*. The rather technical APPLE attribute specifies the number of *l*-participles (*more*) and helps to determine AMOOD and ATENSE. The attributes IC and LC refer only to the *l*-participle (*iLPple*) *nebyl* as the surface head of the embedded phrase in singular masculine animate. LC (the lexical category) states that *nebyl* is a negative form of the *imperf*ective auxiliary *být*.

The COMPS (non-subject valency) list is empty – the phrase *nebyl spal* is saturated. It is made up of DH, the content verb *spal*, and SH, the auxiliary participle *nebyl*, whose C is shared with that its mother's C (4) and whose single item on the COMPS list (6) is identified with its deep head sister, the content verb's SS. The categorial features of the content verb are specified in SS|C. The form is positive (APOL *plus*) and the phrase *spal* consists of the single form (APPLE *one*). As above, the values of the IC and LC attributes refer to the form *spát* itself: lemma = *spát*, masculine animate form (*ma*), *imperf*ective voice, polarity positive (*plus*), non-reflexive (*reflno*) content verb (*lVerbMain*). The intransitive content verb has no non-subject valency – the COMPS list is empty.

Representations of AVFs are built from: (i) skeletal phrase structures, converted from dependency trees produced by the parser, including morphosyntactic information about the terminals, and (ii) valency of auxiliaries (except for subject, valency of content verbs are irrelevant for analytic predicates).

### 3.5 Lexical Entries for the Auxiliaries

The forms *bych* and *nebyl*, used in Fig. 7, are derived from lexical entries shown in Figs. 8 and 9. The entries stand for all forms of the conditional auxiliary and *l*-participle.

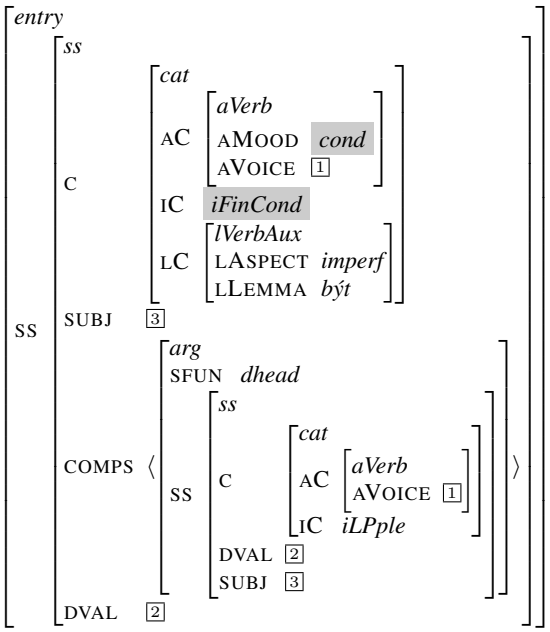


Figure 8: Lexical entry for the conditional auxiliary (e.g., *bych*).

Fig. 8 describes the conditional auxiliary irrespective of person or number, i.e., including *bych*. The value of C determines the *cond*itional mood (in AMOOD) for the whole AVF. Its voice is the same as the voice of its *l*-participle complement and of the entire structure. Inflectional category is finite conditional and lexically a form of the imperfective auxiliary *být* (*lVerbAux*). The valency (COMPS) specifies an *l*-participle (*iLPple*) whose deep valency is shared with that of *bych* itself. The subject of *bych* is also shared with that of its complement, including a potentially null subject. The lexical entry for the form *nebyl* in Fig. 9 differs from the entry for *bych* in the following respects: (i) no value of mood is present (there is no AMOOD in the AC attribute), (ii) the type of the IC attribute is *iLPle*, i.e., the form *nebyl* is an *l*-participle, and (iii) there is an SC (syntactic category) attribute whose *sLPple* value states that the form is a syntactic participle rather than *iFinPlain*, reserved for 3rd person *l*-participles.

### 3.6 Constraints for the Analytic Categories

Additional specifications are due to constraints of the grammar. Deep and surface heads share their ALEMMAS (8), deep head shares AVOICE with its auxiliary (9), *l*-participle surface head is marked as APPLE:*more* if the deep head is also an *l*-participle (10), tense is determined by the mood and num-

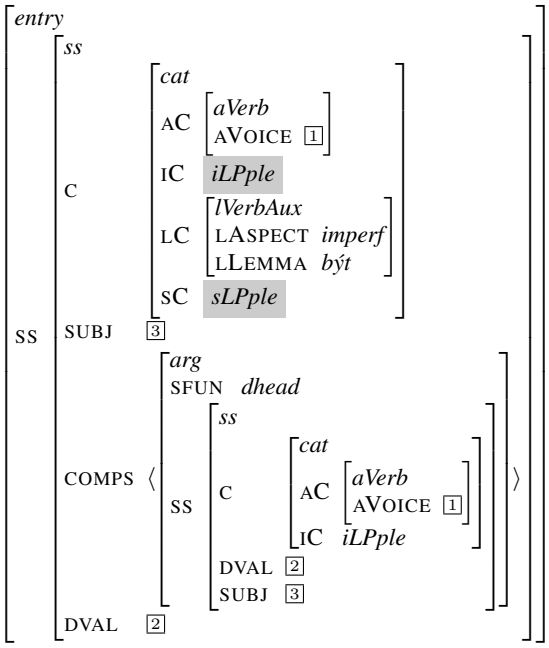


Figure 9: Lexical entry for the past auxiliary *l*-participle (e.g., *nebyl*).

ber of the *l*-participles (11), polarity of the entire AVF is positive unless any of its constituents is negated (12).

Using the same mechanism with different attributes, prepositions and conjuctions as surface heads can model the AC of prepositional phrases and subordinate clauses.

(8) $sdheaded \rightarrow \begin{bmatrix} \text{SH|..ALEMMA} & \boxed{1} \\ \text{DH|..ALEMMA} & \boxed{1} \end{bmatrix}$

(9) $\begin{bmatrix} sdheaded \\ \text{SH|..LC} \;\; lVerbAux \end{bmatrix} \rightarrow \begin{bmatrix} \text{SH|..AVOICE} & \boxed{1} \\ \text{DH|..AVOICE} & \boxed{1} \end{bmatrix}$

(10) $\begin{bmatrix} sdheaded \\ \text{SH|..LC} \;\; lVerbAux \\ \text{SH|..SC} \;\; sLPple \\ \text{DH|..SC} \;\; sLPple \end{bmatrix} \rightarrow \begin{bmatrix} \text{SH|..APPLE} \;\; more \end{bmatrix}$

(11) $\begin{bmatrix} sdheaded \\ \text{SH|..LC} \;\; lVerbAux \\ \text{SH|..IC} \;\; iFin \\ \text{DH|..SC} \;\; sLPple \end{bmatrix} \rightarrow \begin{bmatrix} \text{SH|..AMOOD} & \boxed{1} \\ \text{SH|..ATENSE} & \boxed{2} \\ \text{DH|..APPLE} & \boxed{3} \end{bmatrix}$

$$\begin{array}{ll} & mood\_tense(\ \boxed{1},\ \boxed{2},\ \boxed{3}\ ) \\ & mood\_tense(\ ind,\ past,\ one\ ) \\ \wedge & mood\_tense(\ ind,\ plusq,\ more\ ) \\ & mood\_tense(\ cond,\ pres,\ one\ ) \\ & mood\_tense(\ cond,\ past,\ more\ ) \end{array}$$

(12) $\begin{bmatrix} sdheaded \\ \text{SH|..LC} \;\; lVerbAux \\ \text{DH|..SC} \;\; sLPple \end{bmatrix} \rightarrow \begin{bmatrix} \text{DH|..APOL} & \boxed{1} \\ \text{SH|..LPOL} & \boxed{2} \\ \text{SH|..APOL} & \boxed{3} \end{bmatrix}$

$$\begin{array}{ll} & polarity(\ \boxed{1},\ \boxed{2},\ \boxed{3}\ ) \\ & polarity(\ bool,\ minus,\ minus\ ) \\ \wedge & polarity(\ minus,\ plus,\ minus\ ) \\ & polarity(\ plus,\ plus,\ plus\ ) \end{array}$$

## 4 Discussion

We presented a uniform and compact approach to the annotation of AVFs, supporting effective search options in a treebank. Information about an AVF as a whole is contained in its analytic category (AC, Fig. 7) in the phrasal node representing this form (Fig. 5). E.g., content verbs in past conditional can be retrieved by a straightforward query quoting appropriate values of the ACAT attributes. The entire AVF, including auxiliaries, is retrieved when the selection is extended by all surface and deep heads along the analytic projection of the content verb.

This is an advantage over an approach adopted, e.g., in the Prague Dependency Treebank (PDT).[14] On its *analytic level*,[15] auxiliaries are immediate dependents of a content verb (unless coordination is involved) and sisters of dependents of other types. Thus it is not easy to identify AVFs and their type or to infer their properties, e.g., as a response to a query. On the PDT's *tectogrammatical level* the auxiliaries are absent: an AVF is represented as a single complex node, but components of the complex node on the analytic level can be recovered since the representations on the two levels are interlinked. However, the corpus annotated on the tectogrammatical level is too small for many research tasks.

AVFs can have a complex internal syntax. If there is a single auxiliary for two or more coordinated content verbs (e.g., in *Já jsem přišel a viděl.* 'I came and saw'), the two content verbs as well as the predicate are identified as active past indicative forms. On the other hand, such structures are very difficult to identify on the PDT analytic layer. Searching, e.g., for all present conditionals, requires a complex query, based on detailed knowledge of the PDT representation.

The automatically determined analytic categories can be projected to a different annotation format, including PDT or CoNLL-U.[16] At the very least, the annotation of content verbs can be extended by analytically determined specification of mood, tense and voice. In addition to theoretical interest, some NLP applications may profit from the identification of AVFs as a distinctive unit with specific properties. While a certain lan-

[14] http://ufal.mff.cuni.cz/pdt3.0

[15] Note that *analytic level* denotes a *a level of surface syntax* rather than anything related to AVFs.

[16] http://universaldependencies.github.io/docs/format.html

guage tends to express morphological meanings analytically, using auxiliaries and other function words, a different (synthetic) language may avoid AVFs. Identification of such equivalent units may improve the quality of parallel texts alignment and machine translation. Similarly, a parser trained on texts where such units are identified can produce better results.

The first release of a part of the Czech National Corpus annotated in the style of the PDT analytic level is due soon. A pilot treebank including the proposed annotation of analytic categories will follow, supplemented by the formal grammar and lexicon. The planned size is in the order of tens of millions of words. The annotation will include analytic categories and other information added by the grammar and the lexicon, or a flag identifying a failure in the application of the grammar and its possible reason, while the annotation will retain only information from the parser. At present, the grammar and the lexicon are developed and tested on a sample of 1000 sentences from the PDT annotation manual,[17] covering a wide range of linguistic phenomena. A proper evaluation is previewed on a larger sample extracted from real corpus texts.

## Acknowledgments

This research was supported by the Grant Agency of the Czech Republic, grant no. 13-27184S.

## References

Tania Avgustinova, Alla Bémová, Eva Hajičová, Karel Oliva, Jarmila Panevová, Vladimír Petkevič, Petr Sgall, and Hana Skoumalová. 1995. Linguistic problems of Czech. Project Peco 2924. Technical report, Charles University, Prague.

Robert D Borsley. 1999. Auxiliaries, verbs and complementizers in Polish. *Slavic in Head-Driven Phrase Structure Grammar*, pages 29–59.

Václav Cvrček, Vilém Kodýtek, Marie Kopřivová, Dominika Kováříková, Petr Sgall, Michal Šulc, Jan Táborský, Jan Volín, and Martina Waclawičová. 2010. *Mluvnice současné češtiny*. Karolinum, Praha.

Mary Dalrymple. 1999. Lexical-functional grammar. In Rob Wilson and Frank Keil, editors, *MIT Encyclopedia of the Cognitive Sciences*. The MIT Press.

Jan Hajič, Jarmila Panevová, Eva Buráňová, Zdeňka Urešová, and Alla Bémová. 1999. A Manual for Analytic Layer Annotation of the Prague Dependency Treebank. Technical report, ÚFAL MFF UK, Prague.

Tomáš Jelínek, Vladimír Petkevič, Alexandr Rosen, Hana Skoumalová, Přemysl Vítovec, and Jiří Znamenáček. 2014. A grammar-licensed treebank of Czech. In Verena Henrich, Erhard Hinrichs, Daniël de Kok, Petya Osenova, and Adam Przepiórkowski, editors, *Proceedings of TLT13*, pages 218–229, Tübingen.

Petr Karlík, Marek Nekula, and Zdenka Rusínová. 1995. *Příruční mluvnice češtiny*. Lidové noviny, Praha.

Miroslav Komárek, Jan Kořenský, Jan Petr, and Jarmila Veselková, editors. 1986. *Mluvnice češtiny 2 – Tvarosloví*. Academia, Praha.

Anna Kupść and Jesse Tseng. 2005. A New HPSG Approach to Polish Auxiliary Constructions. In Stefan Müller, editor, *Proceedings of HPSG12*, CSLI Publications, pages 253–273. University of Lisbon.

Anna Kupść. 2000. *A HPSG Grammar of Polish Clitics*. Ph.D. thesis, Atelier national de Reproduction des Thèses.

Carl J. Pollard and Ivan A. Sag. 1994. *Head-Driven Phrase Structure Grammar*. University of Chicago Press, Chicago.

Adam Przepiórkowski. 2007. On heads and coordination in valence acquisition. In Alexander Gelbukh, editor, *CICLing 2007*, pages 50–61, Berlin. Springer.

Alexandr Rosen. 2014. A 3D taxonomy of word classes at work. In Ludmila Veselovská and Markéta Janebová, editors, *Complex Visibles Out There. Proceedings of OLINCO 2014*, volume 4, pages 575–590, Olomouc. Palacký University.

Jesse Tseng and Anna Kupść. 2006. A cross-linguistic approach to Slavic past tense and conditional constructions. *Proceedings of FDSL6*.

Jesse Tseng. 2009. A formal model of grammaticalization in Slavic past tense constructions. *Current Issues in Unity and Diversity of Languages*, pages 749–762.

Ludmila Veselovská and Petr Karlík. 2004. Analytic passives in Czech. *Zeitschrift für Slawistik*, pages 163–235.

Ludmila Veselovská. 2003. Analytické préteritum a opisné pasivum v češtině: dvojí způsob saturace silného rysu <+v> hlavy v*. *Sborník filozofické fakulty Masarykovy Univerzity*, pages 161–177.

Gert Webelhuth, editor. 1995. *Government and Binding Theory and the Minimalist Program*. Blackwell Publishers, Oxford, UK.

[17] Hajič et al. (1999).