Periphrastic Verb Forms in Universal Dependencies

Lenka Krippnerová and Daniel Zeman
Charles University, Faculty of Mathematics and Physics
Institute of Formal and Applied Linguistics (ÚFAL)
Prague, Czechia
lenka.krippnerova@volny.cz, zeman@ufal.mff.cuni.cz

Abstract

We propose a generalization of the morphological annotation in Universal Dependencies (UD) to phrases spanning multiple words, possibly discontinuous. Our focus area is that of periphrastic tenses, voices and other forms, typically consisting of a non-finite content verb combined with one or more auxiliaries; however, the same approach can be applied to other morphosyntactic constructions. We present a software tool that can detect periphrastic verb forms, extract the relevant morphological features from member words and combine them into new, phrase-level annotation. The tool currently detects periphrastic verb forms in 15 Slavic languages that are represented in UD and it is easily adaptable to other constructions and languages. Both the tool and the processed Slavic data are freely available.

1 Introduction

Since the basic annotation units in Universal Dependencies (de Marneffe et al., 2021) are morphosyntactic words, UPOS tags and morphological features always relate to a single word. This can be viewed as a limitation. Many languages have multiword expressions[1] that could be described by feature-value pairs from an inventory similar to morphological features, but the features would apply to the whole expression, and not to any of its member words alone (Zeman, 2023). Periphrastic verb forms are a prime example of this. For example, the English present perfect *I have left* can be described as Mood=Ind, Tense=Pres, Aspect=Perf, Voice=Act, Number=Sing, Person=1; however, in UD annotation some of these features are scattered on individual words and others, such as the aspect, are not annotated at all because they do not characterize any of the words in isolation.

Another issue is that words are not always easy to delimit (Evang and Zeman, 2024). For example, the Japanese writing system does not insert spaces between words, and several approaches have been proposed to break the text up to word-like units (Murawaki, 2019). Consider (1) below:

(1) 行ってきました
ittekimashita
'went'

This could be treated as one verb in the polite form of the past tense. It contains two lexical roots, so it could be also considered a compound predicate *itte kimashita*, consisting of a converb of *iku* 'go' and a polite past form of *kuru* 'come'. But in fact, Japanese UD[2] decomposes it into two verbs, two auxiliaries, and one subordinator: *it te ki mashi ta*. Naturally, the selected segmentation directly affects which features can be annotated and where to find them.

In some cases, a word participating in a periphrastic form may even bear a feature that conflicts with the feature of the whole expression. For example, in Czech (2), *řekl jsem* 'I told' is a periphrastic past tense composed of past participle and present auxiliary; in *by přišel* 'he would come', the "past" participle is used in a present conditional construction.

(2)

Řekl	*jsem*	*mu,*	*a=*	*=by*	*přišel*
told	I.have	him	that	would	he.come
V.F.=Part	M.=Ind			M.=Cnd	V.F.=Part
T.=Past	T.=Pres				T.=Past

'I told him to come.'

We present a rule-based software tool that takes the existing UD annotation as input and enriches it with features for periphrastic verb forms. The tool currently covers all 15 Slavic languages in

[1] By multiword expression we now mean just an expression of multiple words; no idiosyncrasy is required.

[2] Japanese GSD in UD 2.15.

Proceedings of the Eighth International Conference on Dependency Linguistics (Depling, SyntaxFest 2025), pages 140–149

ID	FORM	UPOS	MISC
1	Řekl	VERB	Phrase=[1, 2]|PhraseAspect=Perf|PhraseForm=Fin|PhraseGender=Masc|PhraseMood=Ind|PhraseNumber=Sing|PhrasePerson=1|PhraseTense=Past|PhraseVoice=Act
2	jsem	AUX	_
3	mu	PRON	SpaceAfter=No
4	,	PUNCT	_
5-6	aby	_	_
5	aby	SCONJ	_
6	by	AUX	_
7	přišel	VERB	Phrase=[6, 7]|PhraseAspect=Perf|PhraseForm=Fin|PhraseGender=Masc|PhraseMood=Cnd|PhraseNumber=Sing|PhrasePerson=3|PhraseVoice=Act|SpaceAfter=No
8	.	PUNCT	_

Table 1: Sample output; for glosses and translation, see (2) in the text. The new annotations are placed in the MISC column at the head node of the verb form. The Phrase attribute identifies the nodes that belong to the periphrastic form, the other Phrase* attributes correspond to morphological features as defined for the FEATS column.

	UPOS	VerbForm	Mood	Aspect	Tense	Voice	Number	Person	Gender	Animacy	Clitic	Variant	Phrase
Několik	DET												
jsem	AUX	Fin	Ind	Imp	Pres	Act	Sing	1					*
jich	PRON												
našel	VERB	Part		Perf	Past	Act	Sing		Masc				*
jsem našel	VERB	Fin	Ind	Perf	Past	Act	Sing	1	Masc				[2, 4]
Znalazł-	VERB	Fin	Ind	Perf	Past	Act	Sing		Masc	Hum			*
-em	AUX			Imp			Sing	1			Yes	Long	*
ich	PRON												
kilka	DET												
Znalazłem	VERB	Fin	Ind	Perf	Past	Act	Sing	1	Masc	Hum			[1, 2]

Table 2: Propagation of word features to phrase features shown on a Czech and a Polish sentence with the same meaning: [cs] *Několik jsem jich našel.* / [pl] *Znalazłem ich kilka.* 'I found several of them.' Blue color indicates the periphrastic form (phrase) and its features. Orange are the contributing features of the member words. Word features shown in black are not copied to the phrase annotation. The two languages differ in word order. In Czech, the periphrastic form is discontinuous, while in Polish the auxiliary is a clitic on the main verb. The feature profiles are very similar except that the Polish participle expresses Animacy and the Polish auxiliary lacks the Tense=Pres annotation.

UD[3] and it is easily extensible to other languages and other grammatical constructions. The tool has been used to prepare Czech data for the shared task on "Morpho-Syntactic Parsing" that is being organized[4] as part of SyntaxFest 2025.

2 The Tool

The tool relies on the Udapi[5] Python framework (Popel et al., 2017). Udapi works as a processing pipeline that reads data in the CoNLL-U format, applies selected processing *blocks* to the data and saves the modified data in CoNLL-U again. We created a number of blocks that take care of verb forms found in Slavic languages. When a periphrastic form is found, the features that describe it are encoded as MISC attributes of the word that heads the periphrastic expression in the

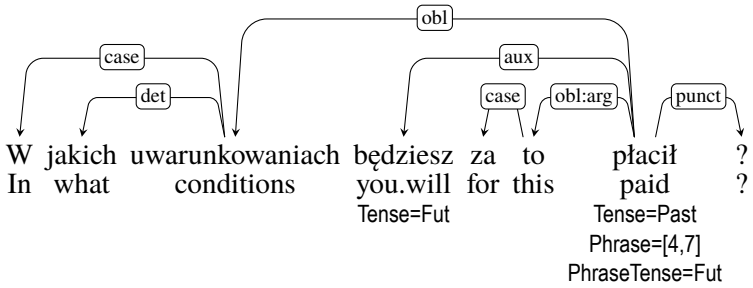
[3] Belarusian, Bulgarian, Croatian, Czech, Macedonian, Old Church Slavonic, Old East Slavic, Polish, Pomak, Russian, Serbian, Slovak, Slovenian, Ukrainian, and Upper Sorbian.

[4] https://unidive.lisn.upsaclay.fr/doku.php?id=other-events:msp

[5] https://udapi.github.io/

future indicative		*budu dělat*	*budu se dělat*	*budu dělán*
present indicative		*dělám*	*dělám se*	*jsem dělán*
past indicative		*dělal jsem*	*dělal jsem se*	*byl jsem dělán*
conditional		*dělal bych*	*dělal bych se*	*byl bych dělán*
imperative	active	*dělej*	*dělej se*	*buď dělán*
	× middle / reflexive passive			
participle	passive	*dělající*	*dělající se*	*dělaný*
converb		*dělaje*	*dělaje se*	*jsa dělán*
infinitive		*dělat*	*dělat se*	*být dělán*
		‘to do’	‘to be done’	‘to be done’

Table 3: Overview of verb forms with examples from Czech. A few other forms that are not attested in Modern Czech exist in other Slavic languages: supine (Old Church Slavonic, Slovenian); aorist and imperfect past (Old Church Slavonic, Bulgarian, Macedonian, Upper Sorbian); impersonal non-passive form (Polish, Ukrainian).



‘Under what conditions will you pay for this?’

Figure 1: Example of future tense from the Polish PDB treebank. The value of the PhraseTense is copied from the auxiliary.

dependency tree (Tables 1 and 2).[6]

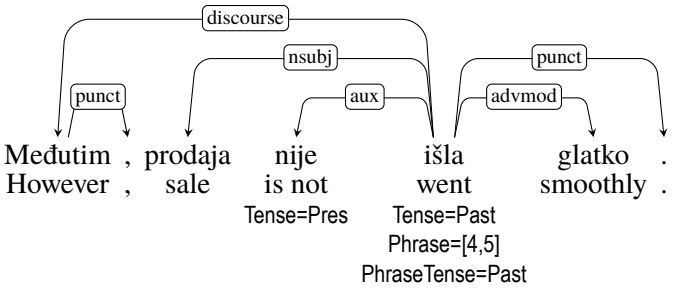
Each of the blocks specializes in finding one of the verb forms expressed in Slavic languages. There are blocks for present tense, future tense, past tense, conditional, imperative, transgressive,[7] and infinitive. The blocks must also handle negation and passivization, as each of these forms can be passivized using either a reflexive marker or a passive participle with an auxiliary. Some verb forms are not periphrastic (they are expressed by one word) but for completeness we capture them, too. This is the case with present tense, transgressive, imperative, and infinitive in the active voice. In some Slavic languages, the past tense and some forms of the future tense are also simple (Zeman, 2016). Periphrastic verb forms include the future and past tenses, the conditional, and all passive forms. Besides standard verbal predicates, the forms are also marked for non-verbal predicates with copula. Table 3 gives an overview of the forms with examples.

The detection of a periphrastic form normally involves identification of the head word (verb, other word with the VerbForm feature, or a non-verbal predicate) and collection of its children with relations `aux`, `cop`, or `expl`. Each block has its own set of conditions over these nodes to verify whether they represent an instance of the construction the block focuses on. When the conditions are met, it is necessary to determine which features should be propagated to the entire periphrastic phrase. For example, in the periphrastic future tense, the value of the attribute PhraseTense is copied from the Tense feature of the auxiliary (Figure 1). In contrast, in the periphrastic past tense, the value of PhraseTense is copied from the Tense feature of the content verb (Figure 2).

In some cases, a value of a Phrase* attribute must be added even though it is not present on the head word or any of its dependents. This occurs,

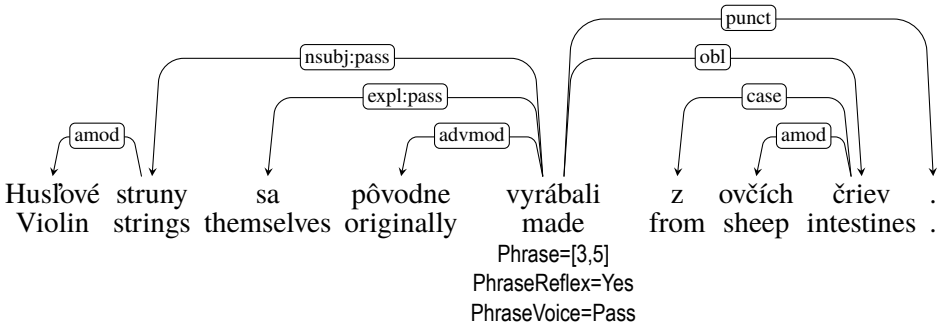
[6]Our blocks are now available directly in the Udapi GitHub (`https://github.com/udapi/udapi-python`). They have their own namespace `msf.slavic`, e.g., `msf.slavic.Future` is the block for future tense.

[7]Also known as gerund or converb.



'However, sales did not go smoothly.'

Figure 2: Example of past tense from the Serbian SET treebank. The value of the PhraseTense is copied from the content verb.



'Violin strings were originally made from sheep intestines.'

Figure 3: Example of reflexive passive from the Slovak SNK treebank (abridged).

for example, in the case of the reflexive passive (Figure 3). The need to assign PhraseVoice=Pass is inferred solely from the presence of a reflexive marker in an `expl:pass` relation. Voice=Pass is annotated neither on the content verb nor the reflexive marker; in fact, the content verb is marked with Voice=Act.

The blocks also handle negation. In Slavic languages, negation can be expressed in two ways: either with a negative prefix or with a negative particle. In addition to searching for `aux`, `cop`, and `expl` relations among the descendants of the head word, we also look for the presence of a negative particle to determine whether the attribute PhrasePolarity=Neg should be generated. If no negative particle is found among the descendants, we then check whether the negation is expressed via a prefix. This can be challenging, as different verb forms may realize the negative prefix in different parts of the verb phrase. For example, in the Czech active past tense, the negative prefix appears on the content verb, whereas in the passive, it can be expressed on the auxiliary, on the content verb, or both.

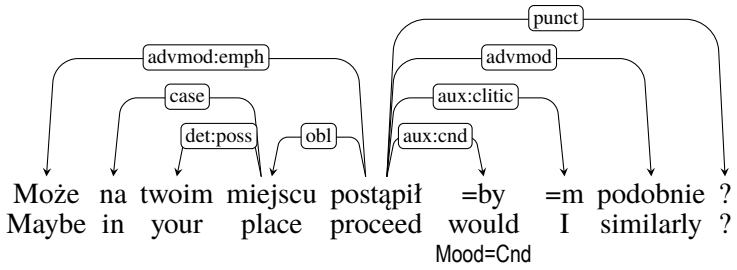
3 Harmonization of Annotations

Even though the annotations in Universal Dependencies are supposed to be consistent, there are still cases across different languages where the annotations are not unified sufficiently. Whenever such discrepancies directly affect the retrieval of periphrastic verb forms, we harmonize them, meaning that even word-level features in our output may differ from the input data. The benefit is twofold: Besides making the identification of verb forms easier, the resulting data is also more suitable for cross-linguistic studies, very much in the UD spirit.

The conditional mood may serve as an example. In Polish, the conditional auxiliary is not tagged with Mood=Cnd, but its incoming relation is subtyped as `aux:cnd`. However, in other Slavic languages, the conditional auxiliary is marked with Mood=Cnd, therefore we assign this feature to the corresponding auxiliaries in Polish as well (Figure 4).

4 Participles

We decided to harmonize the UPOS annotation of participles. Since participles express both verbal



'Maybe I would do something similar in your place?'

Figure 4: Example of conditional from the Polish PDB treebank. The Mood=Cnd feature was added by the preprocessing block.

past active (l-participle)	VERB	*читал*	*čital*	used mainly with auxiliaries
present active	ADJ	*читающий*	*čitajuščij*	used mainly as attribute
past active (adjectival)	ADJ	*читавший*	*čitavšij*	used mainly as attribute
present passive[8]	ADJ	*читаемый*	*čitaemyj*	used mainly as attribute
past passive (long variant)	ADJ	*прочитанный*	*pročitannyj*	used as attribute or predicate
past passive (short variant)	ADJ	*прочитан*	*pročitan*	used as attribute or predicate

Table 4: Overview of participles with examples from Russian *читать (čitat')* 'to read'. The l-participle is used predicatively, with or without auxiliaries, to form the past tense (or resultative / perfect in old languages), conditional and future tense (in Polish and Slovenian; other languages use the infinitive instead). The past passive participle with an auxiliary forms the passive voice. In addition, Polish and Ukrainian have a special impersonal verb form, which is not considered participle, but it bears some similarities to passive participles.

features (such as aspect) and adjectival features (such as case), they occupy an intermediate position between verbs and adjectives. This leads to inconsistencies in the annotations. In some treebanks, these forms are tagged with UPOS ADJ, while in others they receive UPOS VERB. To resolve this, we apply a simple rule of thumb: participle types that can express case (i.e., all types except so called l-participles[9]) are now annotated as adjectives; it is still easy to recognize them thanks to the VerbForm=Part feature. Despite the UPOS tag, we continue to treat participles as potential members of periphrastic verb forms. The fact that some participles are used attributively rather than predicatively will be visible in syntactic annotation (which we carry over unmodified to the output); in such cases, our Phrase* features will only reflect the features of the participle itself.

Table 4 exemplifies the various participle types that can be found in Slavic languages (Sussex and Cubberley, 2006).

[8]The present passive participle is found only in Russian, Old Church Slavonic, and Old East Slavic.

[9]Also excluded are converbs, which developed from participles but their forms are frozen w.r.t. Case.

5 Reflexive / Middle Voice

One of the Phrase* attributes placed in the MISC column at the head node of a verb phrase is PhraseReflex. This is a Boolean feature that appears only with the value Yes; when absent, it is interpreted as No. We mark as reflexive only those verb phrases that contain the reflexive marker in an expletive relation. Reflexive pronouns that function as objects or obliques are not considered part of the verb phrase and therefore do not justify reflexive marking.

The expl relation of reflexive pronouns can include subtypes such as expl:pv (pronominal verb), expl:pass (reflexive passive), and expl:impers (impersonal construction). Among these, expl:pass is essential for identifying reflexive passives (Figure 3). However, because this subtype is not distinguished in many treebanks, we are often unable to recognize reflexive passives and must instead annotate such verb phrases with PhraseVoice=Act.

In East Slavic languages (Belarusian, Russian, and Ukrainian, partly also in Old East Slavic), reflexive markers are suffixed on the verb. In such cases, neither the Reflex=Yes feature nor the expl relation is present in the data.[10] Instead, the feature

[10]Old East Slavic contains both suffixed and separate re-

Feature	Precision	Recall	F_1-score
Phrase	1	0.99	0.99
PhraseAspect	1	0.63	0.78
PhraseForm	1	0.99	0.99
PhraseMood	1	0.99	0.99
PhraseNumber	0.94	0.92	0.93
PhrasePerson	1	0.99	0.99
PhraseTense	0.96	0.95	0.95
PhraseVoice	1	0.99	0.99

Table 5: Evaluation of Czech.

Feature	Precision	Recall	F_1-score
Phrase	0.99	0.99	0.99
PhraseAspect	1	1	1
PhraseForm	1	1	1
PhraseMood	0.98	0.96	0.97
PhraseNumber	1	1	1
PhrasePerson	1	1	1
PhraseTense	0.98	1	0.99
PhraseVoice	1	1	1

Table 6: Evaluation of Ukrainian.

Voice=Mid (middle voice) indicates that the verb is reflexive (Figure 5).[11]

When a reflexive verb phrase is identified, we assign PhraseReflex=Yes. Additionally, if the head verb form contains the feature Voice=Mid, we also assign PhraseVoice=Mid.

6 Data Release

While we believe that installation and usage of Udapi with our blocks is easy, we are simplifying it even more by releasing the processed Slavic treebanks from UD 2.16 at `http://hdl.handle.net/11234/1-5936`. The blocks are still useful when one wants to process other versions of UD, or even one's own data processed by an automatic parser.

7 Evaluation

We have manually verified the output and calculated the precision, recall, and F_1-score on 100 Czech and 100 Ukrainian sentences.

7.1 Evaluation of Czech

For the evaluation, we used the first 100 sentences of the Czech PUD treebank v2.15 with 275 periphrastic verb forms (Table 5). The recall of the feature PhraseAspect is low due to the fact that this feature is often missing in the input data for individual verbs. 90 verb tokens out of 286 lack the Aspect feature.[12] Because of this, we have decided to simplify the detection of the future tense. In Czech, perfective verbs have a simple future tense, which looks morphologically like the present and is labeled Tense=Pres in the input features; due to the absence of the Aspect feature, it is not possible to reliably discriminate present from future in these cases. Therefore, we mark all simple present-like forms as PhraseTense=Pres, which decreases precision of PhraseTense in Table 5. There are five perfective verbs in the test data that we marked as PhraseTense=Pres even though they express the future and the aspect is specified. For 22 verb tokens in the present-like form, the aspect is not specified and we marked all of them as PhraseTense=Pres, while three of them should actually be PhraseTense=Fut. The precision of the PhraseNumber feature is lower because some verb forms have Number=Plur,Sing and it is not always easy to decide which number to choose.

7.2 Evaluation of Ukrainian

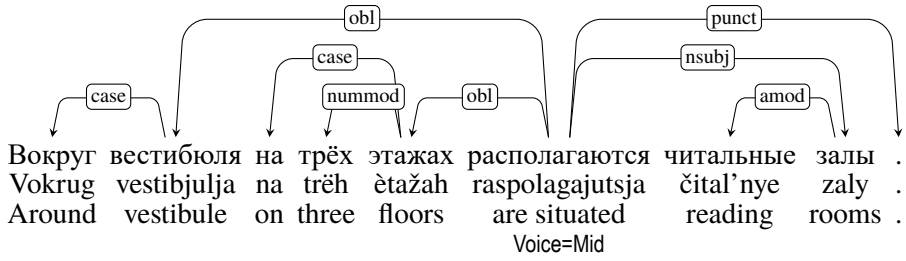
Ukrainian was evaluated on the first 100 sentences of the IU test treebank v2.15 with 250 periphrastic verb forms (Table 6). Unlike the Czech test data, there are no issues with missing Aspect. The present-like form of perfective verbs is already tagged Tense=Fut in the input data; copying it to PhraseTense is all we need to do.

Although there are no errors in future tense, the precision of Phrase, PhraseMood and PhraseTense is less than 1. This is because the conditional mood is not detected correctly. Ukrainian conditional is formed using the past participle of the content verb and a special form of the auxiliary verb *б (b)*. However, this auxiliary can be encliticized to a subordinator, forming *щоб (ščob)* 'so that', but these cases are not labeled conditional in the treebank. As a result, the periphrastic verb form is not fully detected and only the content verb in the past tense is recognized. Consequently, the feature PhraseTense=Past is generated, but the conditional mood does not express tense, so the feature PhraseTense should not be

flexive markers. When they are separate words, they have Reflex=Yes and `expl(:pv)`.

[11]The Voice=Mid feature is currently not used in Ukrainian treebanks. To maintain consistency, we add it in our harmonization step (Section 3), based on verb suffixes.

[12]In the rare case of biaspectual verbs, omitting the Aspect feature would be legitimate. The verbs in our test sample are not biaspectual.



'There are reading rooms on three floors around the vestibule.'

Figure 5: Example of middle voice from the Russian GSD treebank.

generated at all. Figure 6 gives an example of such conditional clause with the correct annotation that our tool failed to deliver.

7.3 Old Church Slavonic

We do not have the same kind of manual evaluation for Old Church Slavonic as we do for Czech and Ukrainian. Nevertheless, this language's data is an outlier in many respects and we believe that some observations are worth sharing here. Some of them are related to OCS being different than the other languages; quite a few, however, reflect divergent approaches to annotation of phenomena that are not so different in nature.

Infinitives do not express tense, nevertheless, in Old Church Slavonic they are annotated with Tense=Pres. We remove this feature in our preharmonization step.

The future tense seems to be the youngest grammaticalized tense (Vepřek, 2015) and in Old Church Slavonic it is often expressed using several pseudo-auxiliary verbs that may still keep a shade of their original lexical meaning. The UD treebank does not distinguish the original present tense meaning of the auxiliaries from periphrastic future. We cannot reliably make this distinction on the fly, so we annotate all such forms as PhraseTense=Pres, although it is probably not always correct.

The Aspect in modern Slavic languages is lexical: If an imperfective verb has a perfective counterpart, they will have different lemmas and will be considered different lexemes. This is how the Aspect feature is handled in languages where its annotation is present.[13] However, in OCS the lexical aspect is not annotated and the feature is used to distinguish the two simple past tenses: imperfect (Aspect=Imp) and aorist (Aspect=Perf). This generates inconsistency because in the other languages where these tenses have been preserved (most notably Bulgarian), the tenses are distinguished by the Tense feature (Tense=Imp for imperfect, Tense=Past for aorist).

Moving from tense to mood, we observe a terminological mismatch: Some authors (Huntley, 2002, p. 156) use the term 'subjunctive' for the form that is usually called conditional (Mood=Cnd) in Slavic languages including Old Church Slavonic (Vepřek, 2015, 5.17.1). Unfortunately, the authors of the OCS treebank preferred the former term and used Mood=Sub instead of Mood=Cnd. We eliminate this inconsistency in the preprocessing step.

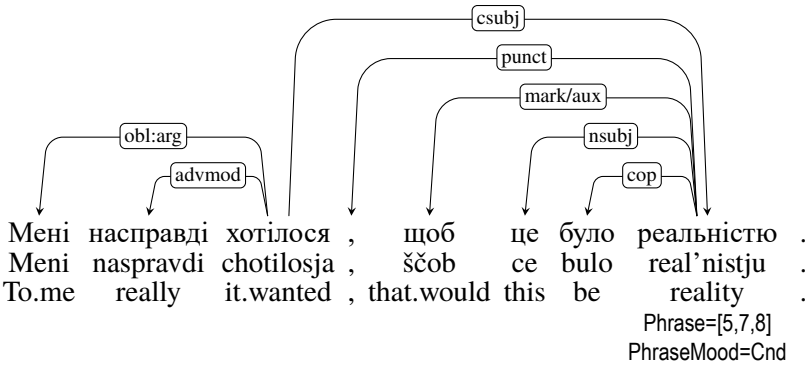
Passive participles have present and past forms, unlike all the other Slavic languages except Russian. The periphrastic passive (the auxiliary *byti* 'be' + passive participle) is difficult to distinguish from a similar deverbative adjective used as a non-verbal predicate with a copula; in the data, most such cases are annotated as `cop` rather than `aux:pass`. There was also the reflexive (medio)passive but again it is not recognizable in the data. The reflexive clitic is always attached as `expl:pv`, although some occurrences should probably receive `expl:pass`.

The periphrastic passive, combined with conditional, is illustrated in Figure 7.

8 Extensibility to Other Languages

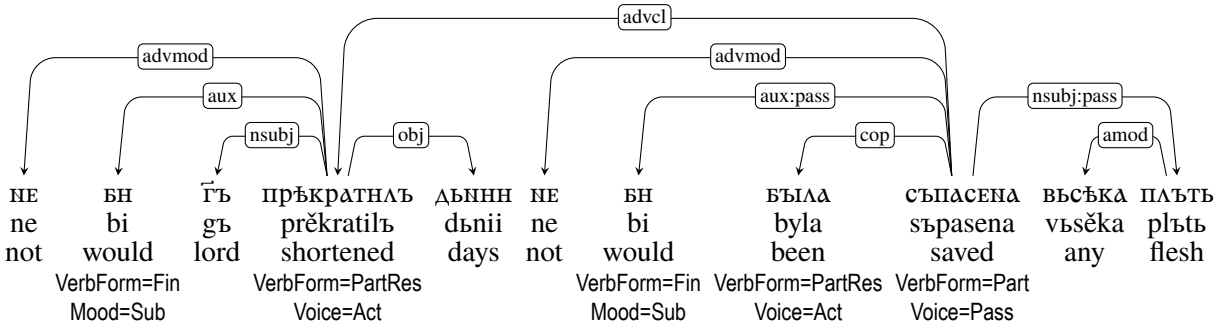
While the present version readily handles Slavic verb forms, the same approach can be used in other languages and for other phrase-level features. To facilitate such extensions, we have designed a generic Udapi block that reads a configuration file in YAML format. The YAML file defines rules for periphrastic forms in a particular language: how to identify nodes that belong to the form, and how to derive phrasal features from the features of the nodes. The rules have been designed to be simple enough that even a user without programming ex-

[13] Aspect annotation is not present in Upper Sorbian, Croatian and Serbian.



'I really wanted this to be a reality.'

Figure 6: Example of unrecognized conditional from the Ukrainian IU treebank.



'And except that the Lord had shortened those days, no flesh should be saved.'

Figure 7: Example of the source annotation in Old Church Slavonic PROIEL (sentence shortened). The l-participles have a language-specific feature VerbForm=PartRes. Conditional auxiliaries are tagged Mood=Sub instead of Mood=Cnd. The second l-participle is incorrectly attached to the passive participle as copula, while it should be the passive auxiliary. The conditional *bi* should be attached as auxiliary but not as passive auxiliary.

perience can create them.

We are currently working on a similar pipeline for Portuguese, Spanish, and Italian. We have created new rules for identifying periphrastic verb forms, based on the grammatical structures of these languages. A different approach to aspect is required: unlike Slavic languages, these languages do not express aspect lexically. Consequently, we introduced new values for the PhraseAspect attribute – ImpProg and PerfProg – to annotate completed and ongoing progressive actions, respectively. The example is illustrated in Figures 8 and 9.

Extending the approach to other languages is easy from the implementation perspective, as the logic of the existing Udapi blocks can be reused. However, when adapting the tool to a new language, it is necessary to develop specific rules for phrase identification. This process is relatively straightforward when rules already exist for a closely related language (for instance, once we developed rules for Portuguese, adapting them for Spanish was not difficult). In cases where no such rules are available, a careful analysis of the target language's grammar is necessary to formulate appropriate rules.

9 Conclusion

We have presented a software tool that reads UD treebanks and adds phrase-level features for periphrastic grammatical forms. The tool is freely available within the Udapi framework at https://github.com/udapi/udapi-python, and its output on UD v2.16 is available at http://hdl.handle.net/11234/1-5936.

The tool is ready to analyze verb forms in Slavic languages but it is easily extensible, both to other languages and to constructions other than verb forms. For example, it could be used to unify morphological and periphrastic comparatives (cf. English *smarter* vs. *more intelligent*). The tool can be used for cross-linguistic studies (e.g. the full verbal paradigms in two languages) but also in NLP appli-

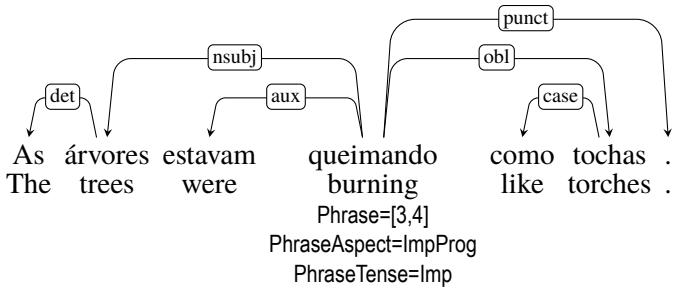


Figure 8: Example of a Portuguese verb phrase with PhraseAspect=ImpProg from the Porttinari treebank.

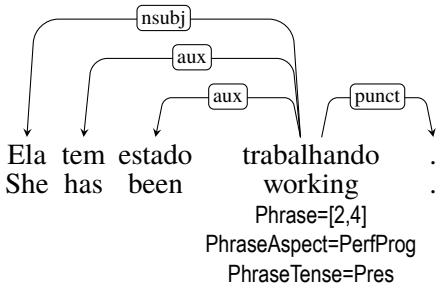


Figure 9: Example of a Portuguese verb phrase with PhraseAspect=PerfProg.

cations to overcome the difficulties of defining the word. The tool has been used to prepare Czech data for the UniDive Shared Task on Morphosyntactic Parsing, collocated with SyntaxFest 2025.

10 Limitations

For the most part, our tool just takes information from the input data and presents it in a restructured way. Whatever interpretation the tool adds is based on the knowledge of the grammatical rules of the given language as a whole, not on detailed understanding of individual words. Therefore, if some piece of the input annotation is missing or incorrect, it cannot be added or corrected in the output.

Acknowledgements

The authors are grateful to Adriana Pagano for useful insights when extending the work to Romance languages.

The work described herein has been supported by the grants *Language Understanding: from Syntax to Discourse* of the Czech Science Foundation (Project No. 20-16819X) and *LINDAT/CLARIAH-CZ* (Project No. LM2023062) of the Ministry of Education, Youth, and Sports of the Czech Republic. It has also received support from the CA21167 COST action UniDive, funded by COST (European Cooperation in Science and Technology).

References

Marie-Catherine de Marneffe, Christopher D. Manning, Joakim Nivre, and Daniel Zeman. 2021. Universal Dependencies. *Computational Linguistics*, 47(2):255–308.

Kilian Evang and Daniel Zeman. 2024. Word segmentation in universal dependencies. In *UniDive General Meeting in Naples posters*, Napoli, Italy.

David Huntley. 2002. Old Church Slavonic. In Bernard Comrie and Greville G. Corbett, editors, *The Slavonic Languages*, pages 125–187. Routledge, Oxon, UK.

Yugo Murawaki. 2019. On the definition of Japanese word. *Preprint*, arXiv:1906.09719.

Martin Popel, Zdeněk Žabokrtský, and Martin Vojtek. 2017. Udapi: Universal API for Universal Dependencies. In *Proceedings of the NoDaLiDa 2017 Workshop on Universal Dependencies (UDW 2017)*, pages 96–101, Gothenburg, Sweden. Association for Computational Linguistics.

Roland Sussex and Paul Cubberley. 2006. *The Slavic Languages*. Cambridge Language Surveys. Cambridge University Press.

Miroslav Vepřek. 2015. *Komparativní tvarosloví staroslověnštiny a staré češtiny (Comparative Morphology of Old Church Slavonic and Old Czech)*. Univerzita Palackého v Olomouci, Olomouc, Czechia.

Daniel Zeman. 2016. Universal annotation of Slavic verb forms. *The Prague Bulletin of Mathematical Linguistics*, (105):143–193.

Daniel Zeman. 2023. Subword relations, superword features. In *UniDive General Meeting at Paris-Saclay posters*, Orsay, France.