Annotating Verbal Periphrasis: Leveraging Universal Dependencies for Phrase-Level Enrichment

Lenka Krippnerová[1], **Adriana S. Pagano**[2], **Patricia Chiril**[3], **Daniel Zeman**[1]
[1]Charles University, Prague
[2]Federal University of Minas Gerais, Brazil
[3]Télécom Paris, Institut Polytechnique de Paris, France

Relevant UniDive working groups: WG1, WG2, WG3, WG4

1 Introduction

Verbal periphrasis is challenging for corpus annotation because it lies at the interface of morphosyntax, the lexicon, and semantics. In many languages, meanings such as MODALITY, CONATION, and PHASE (Halliday and Matthiessen, 2013) are often expressed through multiword verbal constructions rather than through inflection alone. These constructions are important for linguistic description and computational analysis, but they are not always straightforwardly represented in syntactic annotation.

For example, in Spanish, modal meanings may be expressed by verbs such as *poder* 'be able to/may' and *deber* 'must/should' followed by an infinitive, as in *puede salir* 'may leave' and *debe responder* 'must answer'. Conative meanings are illustrated by constructions such as *intentó resolver el problema* 'tried to solve the problem'. Phase meanings are likewise expressed periphrastically, for instance in *empezó a llover* 'it began to rain', *siguió trabajando* 'continued working', and *acabó de llegar* 'has just arrived'. Similar constructions are common in Brazilian Portuguese, where some verbal periphrases express overlapping semantic values. For instance, *começar a tentar resolver o problema* 'begin to try to solve the problem' combines phase and conation, while *pode começar a tentar resolver o problema* 'may begin to try to solve the problem' brings together modality, phase, and conation within a single verbal complex.

The Universal Dependencies (UD) framework (Nivre et al., 2020) provides a useful basis for studying verbal periphrasis, since part-of-speech labels, dependency relations, and morphological features make many candidate patterns searchable. However, because UD annotation is primarily token-based, periphrastic meaning is usually distributed across several tokens rather than encoded as a single unit. Its identification therefore depends not only on structural cues such as `aux`, `xcomp`, and finite/non-finite contrasts, but also on lexical information, especially the lemmas of verbs that recurrently participate in these constructions. Against this background, and drawing on Krippnerová and Zeman (2025), this paper examines how verbal periphrasis can be automatically identified in UD treebanks and how `CoNLL-U` files can be enriched with tags that make these constructions more explicitly retrievable.

2 Aim

The study has two aims: first, to assess how far UD annotations can support the identification of verbal periphrastic constructions expressing MODALITY, CONATION, and PHASE; and second, to propose and test a phrase-level enrichment procedure for `CoNLL-U` files that automatically detects such constructions and adds functional tags to make them explicitly retrievable. To test this proposal, the study applies the enrichment procedure to two UD treebanks from closely related Romance languages: `Porttinari`, for Brazilian Portuguese, and `AnCora`, for Spanish.

3 Methodology

The first step was to identify the verbal periphrastic constructions targeted in the study. We focused on patterns expressing MODALITY, CONATION, and PHASE, considering both structural configurations such as `AUX + VERB`, `VERB + infinitive`, `VERB + preposition + infinitive`, and `VERB + gerund`, and the verb lemmas that typically occur in them. Examples from our data illustrate constructions of phase, modality and conation. (1) exemplifies PHASE, since the finite verb *começa* (begins) combines with the infinitive *investir* (to invest) to construe the beginning of a process.

(1) Quem começa a investir aos 20
who begin.PRS.3SG to invest.INF at.the 20
anos de idade terá, aos 65
year.PL of age have.FUT.3SG at.the 65
anos, quase R$ 1 milhão.
year.PL almost BRL 1 million

'Whoever starts investing at the age of 20 will have almost BRL 1 million at the age of 65.'

(2) exemplifies MODALITY, since *precisa* (needs) functions as a modal predicate expressing necessity in relation to the infinitive *pagar* (to pay).

(2) O investidor precisa pagar para
the investor need.PRS.3SG pay.INF for
comprar esses papéis.
buy.INF these paper.PL
'The investor needs to pay in order to buy these securities.'

(3) exemplifies CONATION, since the finite verb *tentam* (try) combines with the infinitive *recuperar* (to recover) to construe an attempted but not necessarily completed process.

(3) As empresas europeias tentam
the.PL company.PL European.PL try.PRS.3PL
recuperar terreno.
recover.INF ground
'European companies try to regain ground.'

These examples show that identification depends both on the structural relation between the verbs and on the lexical contribution of the governing verb.

The second step consisted in defining structural detection criteria based on UD annotation. These criteria drew on part-of-speech labels such as `AUX` and `VERB`, on dependency relations such as `aux`, `cop`, `xcomp`, `ccomp`, and `mark`, and on relevant morphological features, including `VerbForm`, `Mood`, `Tense`, and related information. On the basis of these criteria, we developed a script implemented using the `Udapi`[1] Python framework (Popel et al., 2017). `Udapi` operates as a processing pipeline that reads data in the `CoNLL-U` format, applies specific processing *blocks*, and saves the resulting output back to the `CoNLL-U` format. When a verbal periphrasis is detected, the features that describe it are encoded as MISC attributes of the word that heads the periphrastic expression in the dependency tree. Table 1 shows a `ConLL-U` file with phrase enrichment.

The script was then tested on the treebanks `Porttinari` (Brazilian Portuguese) and `AnCora` (Spanish).

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

4 Results

Table 2 displays an overall count of periphrases annotated by our script on the complete `Porttinari` dataset (8,418 sentences) and an equivalent subset comprising the first 8,418 sentences of the `AnCora` dataset.

The results show that MODALITY is the most frequent type of verbal periphrasis in both treebanks, followed by PHASE, while CONATION is the least frequent category. This distribution suggests that modal and phasal constructions are especially robust targets for automatic retrieval in UD treebanks. Overall, the results indicate that the script is effective in identifying the targeted categories across both datasets and that the combination of structural and lexical information provides a productive basis for phrase-level enrichment.

5 Conclusion

Our results show that verbal periphrasis is more effectively identified when structural and lexical information are combined. They also support the view that phrase-level enrichment is a productive complement to token-level UD annotation, particularly for categories whose meaning is distributed across multiple tokens and cannot be recovered from morphology alone. More broadly, the approach proposed here has the potential to be extended to other languages and treebanks, provided that the structural patterns and lexical inventories relevant to each language are adequately modeled.

Acknowledgements

This work received support from the CA21167 COST action UniDive, funded by COST (European Cooperation in Science and Technology).

References

Michael Alexander Kirkwood Halliday and Christian MIM Matthiessen. 2013. *Halliday's introduction to functional grammar*. Routledge.

Lenka Krippnerová and Daniel Zeman. 2025. Periphrastic verb forms in universal dependencies. In *Proceedings of the Eighth International Conference on Dependency Linguistics (Depling, SyntaxFest 2025)*, pages 140–149.

Joakim Nivre, Marie-Catherine De Marneffe, Filip Ginter, Jan Hajic, Christopher D Manning, Sampo Pyysalo, Sebastian Schuster, Francis Tyers, and Daniel Zeman. 2020. Universal dependencies v2:

ID	FORM	LEMMA	UPOS	XPOS	FEATS	HEAD	DEPREL	DEPS	MISC
1	As	o	DET	_	Definite=Def|Gender=Fem|Number=Plur|PronType=Art	2	det	_	_
2	empresas	empresa	NOUN	_	Gender=Fem|Number=Plur	4	nsubj	_	_
3	europeias	europeu	ADJ	_	Gender=Fem|Number=Plur	2	amod	_	_
4	tentam	tentar	VERB	_	Mood=Ind|Number=Plur|Person=3|Tense=Pres|VerbForm=Fin	0	root	_	PeriphrasisType=Conation|Phrase=[4,5]|PhraseForm=Fin|PhraseMood=Ind|PhraseNumber=Plur|PhrasePerson=3|PhraseTense=Pres
5	recuperar	recuperar	VERB	_	VerbForm=Inf	4	xcomp	_	_
6	terreno	terreno	NOUN	_	Gender=Masc|Number=Sing	5	obj	_	SpaceAfter=No
7	.	.	PUNCT	_	_	4	punct	_	SpaceAfter=No

Table 1: Annotated CoNLL-U with phrase enrichment for (3) *As empresas europeias tentam recuperar terreno*.

Periphrasis type	Portuguese	Spanish
MODALITY	952	1,334
PHASE	536	510
CONATION	100	194

Table 2: Overall counts of periphrasis types in the Porttinari (Portuguese) and AnCora (Spanish) treebanks.

An evergrowing multilingual treebank collection. In *Proceedings of the twelfth language resources and evaluation conference*, pages 4034–4043.

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.