Developing a Universal Dependencies Treebank for Ukrainian Parliamentary Speech

Maria Shvedova[1,2], Arsenii Lukashevskyi[1], Andriy Rysin[3]
[1]National Technical University "Kharkiv Polytechnic Institute", Ukraine
[2]University of Jena, Germany
[3]Independent researcher
Mariia.Shvedova@khpi.edu.ua
Arsenii.Lukashevskyi@sgt.khpi.edu.ua
arysin@gmail.com

Abstract

This paper presents a new Universal Dependencies (UD) treebank based on Ukrainian parliamentary transcripts, complementing the existing UD resources for Ukrainian. The corpus includes manually annotated texts from key historical sessions of the Verkhovna Rada, capturing not only official rhetoric but also features of colloquial spoken language. The annotation combines UDPipe2 and TagText parsers, with subsequent manual correction to ensure syntactic and morphological accuracy. A detailed comparison of tagsets and the disambiguation strategy employed by TagText is provided. To demonstrate the applicability of the resource, the study examines vocative and nominative case variation in direct address using a large-scale UD-annotated corpus of parliamentary texts.

1 Introduction

Universal Dependencies (UD) is a framework that aims to create a consistent, multilingual annotation scheme for syntactic structures across languages (Nivre et al., 2020), and it has become an important tool for Ukrainian language processing. On the one hand, it enables deeper integration into international multilingual projects that rely on a unified annotation scheme across languages. On the other hand, it provides a valuable resource for studying the Ukrainian language itself, as UD currently offers the only publicly available system for syntactic annotation of Ukrainian texts. UD annotation has already been used in multilingual projects involving Ukrainian, such as the ParlaMint parliamentary transcript corpora (Erjavec et al., 2024) (Kopp et al., 2023), and the parallel corpora collections, namely InterCorp (Čermák and Rosen, 2012) and ParaRook (Shvedova and Lukashevskyi, 2024). As the list of such multilingual projects tends to expand (CLARIN, 2023), the importance of having universal tools like UD becomes even more critical. This ensures that Ukrainian data is compatible with existing and future multilingual projects, allowing us to actively participate in their development.

2 UD Treebanks for Ukrainian

Currently, there are two UD treebanks for Ukrainians. The first is Ukrainian IU[1] by Natalia Kotsyba, Bohdan Moskalevskyi, and Mykhailo Romanenko, published in 2018 (Kotsyba and Moskalevskyi, 2018). The treebank consists of 122,000 tokens in 7,000 sentences drawn from various sources, including fiction, news, opinion articles, Wikipedia, legal documents, letters, posts, and comments. The texts span the last 15 years and the first half of the 20th century, offering a diverse corpus of Ukrainian written speech. The second is Ukrainian ParlaMint Treebank of 52,000 tokens in 3,400 sentences, which was published in the UD repository in 2024 and is a corpus of Ukrainian parliamentary transcripts.[2] The transcripts published on the official website of Verkhovna Rada provide a fairly accurate record of real speech, preserving elements of colloquial syntax, grammatical inconsistencies, lexical errors, and Ukrainian-Russian code switching (Kanishcheva et al., 2023). As such, they serve as valuable material for studying spoken Ukrainian and complement the corpora of written texts. For example, although the Ukrainian IU treebank is larger in volume, it includes only about a hundred instances of direct address, whereas Ukrainian ParlaMint treebank features more than 500. The UDPipe2 model[3] (Straka, 2018) trained on UD_Ukrainian-ParlaMint makes fewer errors in detecting vocative dependency relations, in particular in less regular positions of direct address in the middle and at the end of the sentence (90%

[1] https://universaldependencies.org/treebanks/uk_iu/index.html

[2] https://universaldependencies.org/treebanks/uk_parlamint/index.html

[3] https://lindat.mff.cuni.cz/services/udpipe/

Proceedings of the Fourth Ukrainian Natural Language Processing Workshop (UNLP 2025), pages 55–63
July 31 - August 1, 2025 ©2025 Association for Computational Linguistics

precision for the vocative dependency relation; see Appendix A). Thus, the treebank of parliamentary transcripts complements the existing treebank of written texts by providing grammatical patterns that are more typical of spoken language and less frequent in written sources.

3 The Construction and Annotation of UD Ukrainian ParlaMint Treebank

3.1 Text Selection

Parliamentary transcripts officially released as open data are both a valuable and accessible resource for corpus creation, and the ParlaMint project is the most prominent example of such kind of corpora (Erjavec et al., 2024). In 2024, the Universal Dependencies collection was expanded with three treebanks based on parliamentary transcripts: UD_ParlaMint-It for the Italian Parliament (developed specifically as part of the ParlaMint initiative) (Alzetta et al., 2024), UD_Hebrew-IAHLTknesset for the Knesset of Israel (Goldin et al., 2024), and the third Ukrainian one described in this article.

For the treebank, we selected full transcripts of Verkhovna Rada plenary sessions for several days from the official website [4]. In order to have the most authentic material, we did not use texts from before 2003, where we noticed partial grammatical corrections, and texts from after 2023, where there are signs of speech-to-text recognition that in many cases overly normalizes the text, up to replacing colloquial words with literary ones (e.g., change *ščas* to *zaraz*, 'now'). We did not include texts with Ukrainian-Russian code switching in the corpus; the sentences in Russian were previously removed. When selecting the texts, we chose transcripts of meetings related related to key events in modern events important for modern Ukrainian history, where there is a larger share of spontaneous speech. The corpus includes transcripts of the sessions on 10.10.2003 (Ukrainian state border violated by Russia, building a dam towards Tuzla), 4.04.2014 (first session after the annexation of Crimea), 25.01 and 24.02.2022 (political tension before the full-scale invasion and declaration of martial law), and the transcript of the National Security Council meeting on 28.02.2014 after the annexation of Crimea. The corpus also features samples of the routine work of the Ukrainian parliament during which regular laws are considered.

3.2 Corpus Annotation

Ukrainian ParlaMint treebank has both syntactic and morphological annotation, manually checked by a single annotator. Syntactic dependencies were revised in files initially annotated by the UDPipe2 ukrainian-iu-ud-2.15 model, using the Arborator-Grew graphical annotation interface (Guibon et al., 2020). The part-of-speech and morphological features were annotated on the basis of a comparison of tagging provided by two parsers: UDPipe2 ukrainian-iu-ud-2.15 model with precision for lemmas – 98%, pos – 98%, morphological features – 95%[5] and TagText, which is based on a Ukrainian morphological dictionary, rules and statistical algorithms with precision for lemmas – 99.3%, pos – 98.7%, full morphological tags (including pos and lemmas) – 94.5%[6].

Disambiguation in TagText is performed on three levels. The first two are coming from the Ukrainian module of LanguageTool that the TagText is based on. These two layers are used in grammar and style checking so they are needed to be more precise. The third one is based on statistics from BRUK corpus (Starko and Rysin, 2023) and used only for tagging texts.

1. Discarding extremely rarely used word forms. The VESUM dictionary (Starko and Rysin, 2022) on which the tagger is based provides a full set of possible standard forms no matter how frequently they are used in text, and many such forms could be easily discarded to decrease the noise in the result; e.g. *rozpalenij* 'fired up' can in theory be an imperative form of the verb *rozpalenity* 'flame up,' but in texts it is almost always an adjective. Currently, there are about 600 words in this module.

2. Disambiguation based on rules. These range from simple ones, applied to particular words, for example, discarding the verb *derty* 'to scratch' in compounds like *van der Vala*, or the plural form of *kyj* 'pole' in *Kyiv*, to more complex rules, such as keeping only the locative case in phrases like *v/u/na Ukrajini* 'in Ukraine', or selecting the genitive case in *Petra Poroshenka*, derived from *Petro Poroshenko*, while discarding the feminine name *Petra*. The system also applies more general rules, such as discarding vocative

[4] https://static.rada.gov.ua/zakon/new/STENOGR/index.htm

[5] https://ufal.mff.cuni.cz/udpipe/2/models

[6] https://github.com/brown-uk/nlp_uk

forms after prepositions, etc. The layer includes around 470 rules.

For most complicated disambiguation rules, the logic is implemented in Java. For example, *ledi Čerčil'* where we leave only feminine forms of the surname, or removing locative case if there are no prepositions which requires it. We also discard a vocative case for inanimate nouns which overlaps with other cases (excluding some common uses like *misjačen'ku* 'moon' etc). Total about 10 rules.

3. The statistical module is based on statistics collected from BRUK. Statistics of the forms, morphological tags, and previous context (currently with depth=1) and, for some cases, the following context (currently with depth=1) are collected from the corpus and then used to rate the probability of each lemma and morphological tag for a word in the context. The lemma and tag with the highest probability are kept and the others are discarded[7].

The present approach to disambiguation was developed independently of previous contributions to this problem in Ukrainian linguistics, including traditional rule-based methods described in works (Gryaznukhina et al., 1989) (Shypnivska, 2007), as well as the interesting experience of using a valency dictionary to improve the performance of a syntactic parser (Kotsyba and Moskalevskyi, 2019).

Although both parsers (UDPipe2 and TagText) make mistakes, their errors are mostly different. Comparison of annotation choices is therefore useful for detecting errors in cases of disagreement. UDPipe2 is much better than TagText in the disambiguation of noun forms, including the challenging homonymy of the nominative and accusative cases. It also accurately detects relative and interrogative pronouns, for which TagText has just one double tag. On the other hand, TagText is better than UDPipe2 in identifying known lemmas without distorting them, since it is dictionary-based, and the morphological features attributed to a lemma by the dictionary, such as verbal aspect, nominal gender.

However, there are still cases where both parsers make the same mistake, so focusing only on instances of disagreement is not sufficient for comprehensive error correction. This can occur in cases containing irregular syntactic structure, e.g. *u serpni misjaci* 'in August' (literally, 'in the month of August'): a rare construction with the month names; both parsers misinterpreted the second noun as a plural. Similar parser errors occur in some cases with homonymous case forms. For example, in the following sentence, where the subject is dropped, and the sentence opens, irregularly, with the object in the accusative case, formally identical to the nominative: *Rankove zasidannja ogološuju vidkrytym* 'I call the morning meeting to order'. In rare cases, the distinction between object and subject is challenging even for a human expert, e.g.: *Bezperervnist' roboty Verchovnoji Rady obumovljuje takož bezperervnist' roboty komitetiv* 'The continuity of the Verkhovna Rada's work also determines the continuity of the committees' work' (or vice versa). The complexity of annotating words like *ïx* 'their', *joho* 'his', and *ïï* 'her', homonymous forms that can function either as possessive pronouns or as genitive forms of personal pronouns, and which are sometimes difficult to disambiguate even for an expert, is discussed in (Kotsyba and Moskalevskyi, 2019).

Thus, although the combination of parsers facilitates the task of annotation correction, human control is necessary on the entire corpus.

3.3 Converting and Comparing Morphological Tags from UDPipe2 and TagText Parsers

To automatically compare the annotations from the two parsers, we converted the VESUM dictionary tags[8] into the Universal Dependencies format (Appendix B). The VESUM tagset contains 100 part-of-speech, morphological, and additional tags, mostly with a direct equivalent in the UD tagset; they define POS and morphological features, such as number, gender, grammatical case, person, tense, aspect, mood, degrees of comparison. 16 tags from VESUM have no correspondence in the UD tagset. These tags are related to style, spelling standards (1992 and 2019), date, time, number, and hashtag that we did not preserve during conversion. We created a new UD tag for the VESUM 'bad' tag, which marks non-standard but still common words and grammatical forms, as well as stylistically unrecommended variants: BadStyle=Yes.[9]

The UD system requires the annotation of some

[7] Disambiguation in TagText https://github.com/brown-uk/nlp_uk/blob/master/doc/disambig.md

[8] https://github.com/brown-uk/dict_uk/blob/master/doc/tags.txt

[9] https://universaldependencies.org/uk/feat/BadStyle.html

phenomena that are not represented in the traditional Ukrainian grammar or in the VESUM tagset. This was partially harmonized during the conversion as follows.

- **AUX: auxiliary verb.** The auxiliary verb in Ukrainian is *buty*, *buvaty* ‘to be’, as well as *by (b)*, which forms the conditional mood and is considered a particle in Ukrainian grammar (historically it is a form of the same verb *buty*). However, *buty*, *buvaty* also have lexical meanings (‘to exist’), and in VESUM it is tagged as a regular verb. Therefore, we can automatically assign the AUX tag only to particle *by (b)*, which has no homonyms.

- **Cnd: conditional mood.** The Ukrainian conditional is formed analytically and therefore has no tag in either VESUM or UD.

- **Ind: indicative mood.** This attribute is not present in the VESUM tagset but can be added automatically to all verb forms that already have tense or impersonal form tags.

- **Fin: finite verb.** Attribute indicating a finite verb form as opposed to the infinitive, participle, or converb is not present in the VESUM but can be added automatically to the verb forms that already have tags of personal and impersonal verb forms.

- **DET: determiner.** In traditional Ukrainian grammar and in VESUM, determiners are not defined as a separate class of words. In the UD system, “determiners are words that modify nouns or noun phrases and express the reference of the noun phrase in context. That is, a determiner may indicate whether the noun is referring to a definite or indefinite element of a class, to a closer or more distant element, to an element belonging to a specified person or thing, to a particular number or quantity, etc.”[10] Since Ukrainian has no articles, most determiners are attributive pronouns (but they do not cover all possible determiners). In the VESUM system, all pronouns are tagged with the corresponding parts of speech (noun/adv/numr/adj) and the :&pron tag. We convert attributive and numeral pronouns (adj.*pron; numr.*pron;) to DET, and nominative and adverbial pronouns (noun.*pron; adv.*pron) to PRON. The determiner category also definitely includes the words *odyn* ‘one’ and *druhyj* ‘second’ in the pronoun sense of ‘one’ and ‘another’. However, it is impossible to tag them unambiguously as DET, because they can also be numerals. It is also not possible to unambiguously tag adverbs with the meaning of quantity or degree (*bahato, čymalo, bil’še, najbil’še, dosyt’, malo, nebahato, menše, najmenše*), which may be close to determiners in certain contexts; this difficulty for Slavic languages is described on the UD website.[11]

In cases difficult for full automatic conversion (such as DET or AUX), ambiguity was resolved manually after partial automatic processing.

Due to its efficiency in parsing with Pandas and the ability to edit it manually in the Microsoft Excel interface, it was decided to use XLSX as the format for outputting the difference between the results. The main difficulties in processing data in this way were conversion between non-standard formats, comparison of annotations, design of user output, and subsequent comparison of annotation results, including handling of different tokenizations (e.g., *1,5* for uk_iu is three tokens, while for TagText it is one token, similarly with the hyphenated compound words, which uk_iu also tends to split into three separate tokens).

The solution to such problems was to create an intermediate XML-like .nest format to store CONLL-U tokens in an easily parsable form and convert them without making a separate converter for each pair of formats. Difflib (Python Foundation, 2025) is used to align different tokenizations. The tokenization alignment establishes a partition-to-partition mapping $\phi : \{O_1, O_2, ...\} \rightarrow \{A_1, A_2, ...\}$ between contiguous subsequences of original and annotated tokens, where form$(O_i) \approx$ form(A_j) while preserving the lexical integrity of aligned subsequences. In other words, during the alignment process, we combine consecutive tokens from the source and target annotations into pairs or groups, and then process them as a single lexical unit (e.g., [’Po-tretje’] <=> [’Po-’, ’tretje’] ’thirdly’; [’Prem’jer-ministr’] <=> [’Prem’jer’, ’-’, ’ministr’] ’prime minister’).

Manual processing of treebank files in Excel

[10] https://universaldependencies.org/u/pos/DET.html

[11] https://universaldependencies.org/sla/pos/PRON.html

can lead to inconsistent numbering of sentence tokens, resulting in validation failures and other parsing complications, since the CONLL-U format assumes consistent numbering within a sentence. To solve this problem, we created an algorithm for the normalization of numbering. The renumbering algorithm implements a surjective mapping function $\phi : O \rightarrow N$ from the original ID space O to a normalized sequential space $N = \{1, 2, ..., n\}$, preserving the directed graph structure of dependency trees under transformation $e(i, j) \rightarrow e(\phi(i), \phi(j))$. In essence, we rebuild the same dependency graph, but with the numeration corrected.

The resulting output of the algorithm is standard CONLL-U[12]. The programs can be applied to future projects involving semi-automatic annotation of syntactic relations and morphology.

4 Vocative vs. Nominative in Direct Address: Study on a Large Corpus Annotated with UDPipe2

Although the modern norm of the Ukrainian language recommends using only the vocative case in addresses (ukr, 2019), in practice there is a variation between the vocative and nominative cases. The study of this variation in a corpus with only morphological annotation, without syntactic one, like GRAC[13], is practically impossible due to the difficulty of distinguishing between the different syntactic functions of the nominative case (address, subject, predicate, appositional modifier, list element) and homonymy with the accusative case forms. The UD annotation makes it possible to analyze the use of vocative and nominative cases within the vocative dependency relation, and thus to assess trends in a large textual material.

Using the UDPipe2 ukrainian-parlamint-ud-2.15-241121 model, we annotated the corpus of Ukrainian parliament transcripts from 1990 to 2024, totaling 88 million tokens[14], from which we obtained more than 128 thousand contexts with the vocative relation. The precision of the data was manually verified. We included only singular masculine and feminine nouns, except for indeclinable nouns (e.g., *pani* 'madam', *Jerry*, *Geo*), and nouns that decline according to the adjectival paradigm (e.g., *včenyj* 'scholar'). We also excluded examples consisting of a single surname, as the model often fails to distinguish between masculine and homonymous feminine surnames that do not decline.

The corpus shows significant variation between vocative and nominative in addresses, except for the data before 1995 and for 1997–2001, which show 100% use of the vocative and were likely edited. The proportion of nominative or vocative varies considerably for different lemmas, thus the material requires a deeper linguistic study to find the reasons for the variation (Appendix C).

The resource appears to be highly promising both for corpus-based studies of Ukrainian grammar, in particular, the grammar of spoken language, and for providing annotation of Ukrainian corpora.

In future work, we plan to expand the size of the corpus and explore new annotation possibilities within the UD framework. One such direction is the annotation of ExtPos (external part of speech), which has already been added to the Ukrainian ParlaMint corpus in its second release, completed shortly after the main work on this paper[15]. We also plan to explore the possibility of annotating morphosyntactic features of multiword expressions, so that analytical grammatical forms in Ukrainian, such as the conditional mood or the analytical future, can be represented as annotation features. This would significantly enhance the resource's potential for advanced grammatical research and facilitate more fine-grained linguistic analysis.

Limitations

The corpus contains transcripts of selected plenary sessions of the Verkhovna Rada and is not representative of the entire parliamentary discourse of Ukraine's period of independence. In particular, transcripts featuring Ukrainian-Russian code switching have been excluded, which limits the applicability of the resource for the study of bilingualism and language contact.

Although all annotations were reviewed manually, the process was performed by a single annotator. This may introduce subjectivity, particularly in cases where multiple annotation solutions are possible. Double annotation in future work may improve consistency and reliability.

The currently used utilities solve narrow problems within the project and have not yet been adapted to be used seamlessly and automatically with other tools for UD. In addition, using the utili-

[12] https://universaldependencies.org/format.html

[13] https://uacorpus.org/

[14] Available for download at https://huggingface.co/datasets/uacorpus/Rada_Trees

[15] https://universaldependencies.org/uk/feat/ExtPos.html

ties still involves manual steps to validate the result, which is also worth automating.

Acknowledgments

We are grateful to the reviewers for their helpful feedback, to Daniel Zeman for his expert assistance, and to Kyrylo Zakharov for collecting and sharing the parliamentary transcripts.

This work was supported by the CA21167 COST Action UniDive and the Humboldt Foundation.

References

2019. *Ukrajins'kyj pravopys [Ukrainian Orthography]*. Naukova dumka, Kyiv.

Chiara Alzetta, Simonetta Montemagni, Marta Sartor, and Giulia Venturi. 2024. Parlamint-it: an 18-karat UD treebank of Italian parliamentary speeches. *Language Resources and Evaluation*.

CLARIN. 2023. CLARIN K-Centre for Ukrainian NLP and Corpora. University of Jena, Institute of Slavic and Caucasus Studies. Available at: https://k-centre.uacorpus.org/, accessed: June 9, 2025.

Tomaž Erjavec, Matyáš Kopp, Nikola Ljubešić, Taja Kuzman, Paul Rayson, Petya Osenova, Maciej Ogrodniczuk, Çağrı Çöltekin, Danijel Koržinek, Katja Meden, Jure Skubic, Peter Rupnik, Tommaso Agnoloni, José Aires, Starkaður Barkarson, Roberto Bartolini, Núria Bel, María Calzada Pérez, Roberts Darģis, and 18 others. 2024. ParlaMint II: advancing comparable parliamentary corpora across Europe. *Language Resources and Evaluation*.

Gili Goldin, Nick Howell, Noam Ordan, Ella Rabinovich, and Shuly Wintner. 2024. The Knesset corpus: An annotated corpus of Hebrew parliamentary proceedings. *Preprint*, arXiv:2405.18115.

T. O. Gryaznukhina, L. H. Bratyshchenko, N. P. Darchuk, V. I. Krytska, T. K. Puzdyryeva, and L. V. Orlova. 1989. Šljaxy unyknennja omonimiï v systemi avtomatyčnoho morfolohičnoho analizu [Ways of avoiding homonymy in an automatic morphological analysis system]. *Movoznavstvo*, (5):3–12.

Gaël Guibon, Marine Courtin, Kim Gerdes, and Bruno Guillaume. 2020. When collaborative treebank curation meets graph grammars. In *Proceedings of The 12th Language Resources and Evaluation Conference*, pages 5293–5302, Marseille, France. European Language Resources Association.

Olha Kanishcheva, Tetiana Kovalova, Maria Shvedova, and Ruprecht von Waldenfels. 2023. The parliamentary code-switching corpus: Bilingualism in the Ukrainian parliament in the 1990s-2020s. In *Proceedings of the Second Ukrainian Natural Language Processing Workshop (UNLP)*, pages 79–90, Dubrovnik, Croatia. ACL.

Matyáš Kopp, Anna Kryvenko, and Andriana Rii. 2023. Ukrainian parliamentary corpus ParlaMint-UA 4.0.1. https://www.clarin.si/repository/xmlui/handle/11356/1900.

Natalia Kotsyba and Bohdan Moskalevskyi. 2018. An essential infrastructure of Ukrainian language resources and its possible applications. In *SlaviCorp 2018. Book of Abstracts*, pages 94–95, Prague, Czech Republic. Charles University.

Natalia Kotsyba and Bohdan Moskalevskyi. 2019. Using transitivity information for morphological and syntactic disambiguation of pronouns in Ukrainian. *Journal of Lviv Polytechnic National University "Information Systems and Networks"*, 5:101–115.

Joakim Nivre, Marie-Catherine de Marneffe, Filip Ginter, Jan Hajič, Christopher D. Manning, Sampo Pyysalo, Sebastian Schuster, Francis Tyers, and Daniel Zeman. 2020. Universal dependencies v2: An evergrowing multilingual treebank collection. In *Proceedings of the Twelfth Language Resources and Evaluation Conference*, pages 4034–4043, Marseille, France. European Language Resources Association.

Python Foundation. 2025. difflib — Helpers for computing deltas. Python 3.13.2 documentation. Accessed: June 9, 2025.

Maria Shvedova and Arsenii Lukashevskyi. 2024. Creating parallel corpora for Ukrainian: A German-Ukrainian parallel corpus (ParaRook||DE-UK). In *Proceedings of the Third Ukrainian Natural Language Processing Workshop (UNLP) @ LREC-COLING 2024*, pages 14–22, Torino, Italia. ELRA and ICCL.

Olha Shypnivska. 2007. *Strukturno-semantyčni ta funkcional'ni xarakterystyky mižčastynomovnoï morfolohičnoï omonimiï sučasnoï ukraïns'koï movy [The structural-semantic and functional characteristics of the morphological homonyms belonging to different part-of-speech in the contemporary Ukrainian language]*. Candidate of philological sciences dissertation, NAS of Ukraine; ULIF, Kyiv.

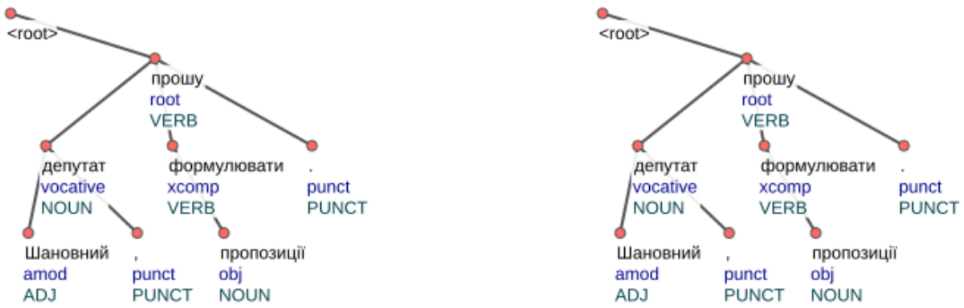
Vasyl Starko and Andriy Rysin. 2022. VESUM: A large morphological dictionary of Ukrainian as a dynamic tool. In *Computational Linguistics and Intelligent Systems*, volume 6th Int. Conf, pages 71–80, Gliwice. COLINS.

Vasyl Starko and Andriy Rysin. 2023. Creating a POS gold standard corpus of modern Ukrainian. In *Proceedings of the Second Ukrainian Natural Language Processing Workshop (UNLP)*, pages 91–95.

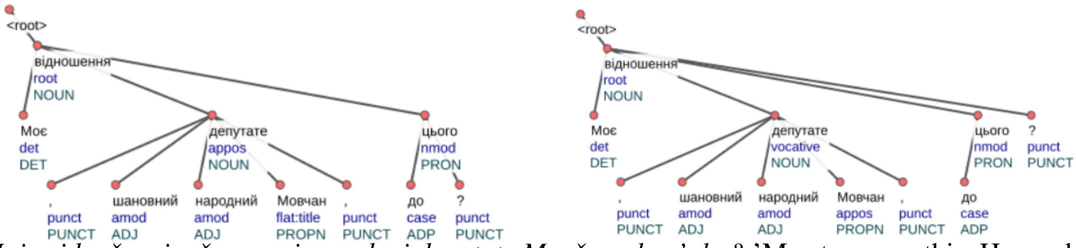
Milan Straka. 2018. UDPipe 2.0 prototype at CoNLL 2018 UD shared task. In *Proceedings of the CoNLL 2018 Shared Task: Multilingual Parsing from Raw Text to Universal Dependencies*, pages 197–207, Brussels, Belgium. ACL.

F. Čermák and A. Rosen. 2012. The case of InterCorp, a multilingual parallel corpus. *International Journal of Corpus Linguistics*, 17(3):411–427.

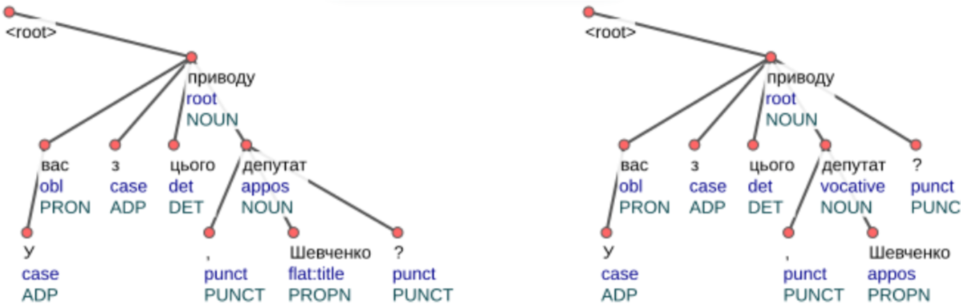
A Vocative Sentence Graphs: ukrainian-iu-ud-2.15-241121 (Left) vs. ukrainian-parlamint-ud-2.15-241121



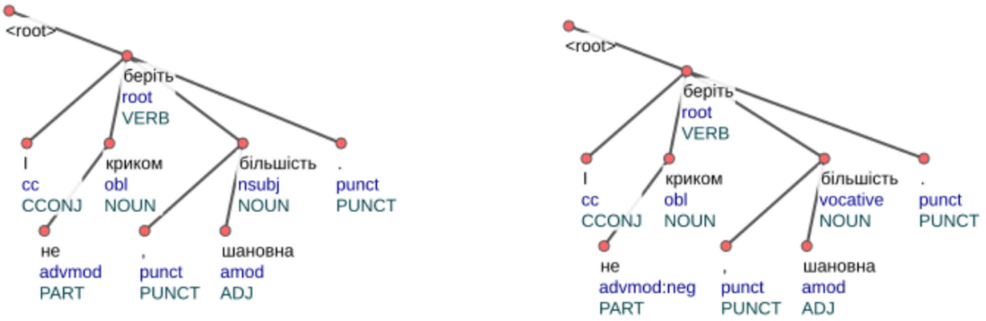
(a) *Šanovnyj deputat, prošu formuljuvaty propozyciї.* ’Honorable Member, please formulate your proposals.’



(b) *Moje vidnošennja, šanovnyj narodnyj deputate Movčan, do c’oho?* ’My stance on this, Honorable MP Movchan?’



(c) *U vas z c’oho pryvodu, deputat Ševčenko?* ’Do you have a comment on this, MP Shevchenko?’

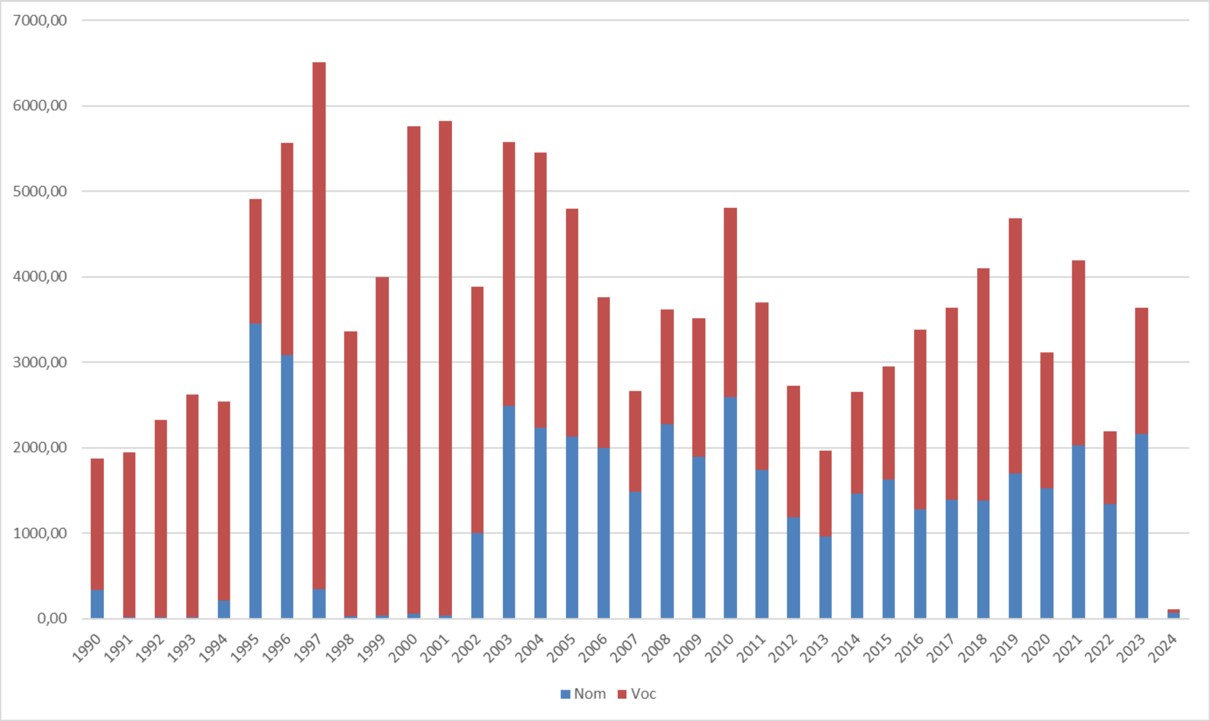


(d) *I ne krykom berit’, šanovna bil’šist’.* ’Don’t try to win by shouting, dear majority.’

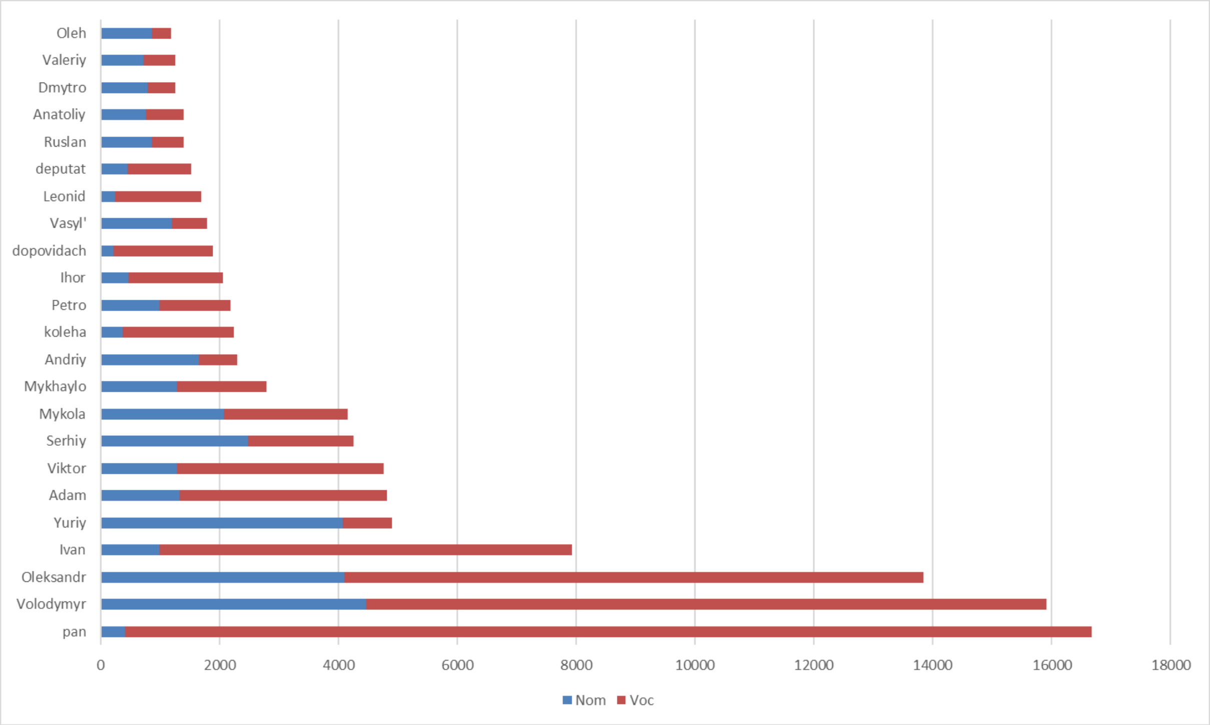
B Mapping between VESUM and UD Tags

VESUM	UD	VESUM	UD
noun	NOUN	ns	Number=Ptan
anim	Animacy=Anim	p	Number=Plur
fname	NameType=Giv	s	Number=Sing
lname	NameType=Sur	m	Gender=Masc
pname	NameType=Pat	f	Gender=Fem
inanim	Animacy=Inan	n	Gender=Neut
unanim	Animacy=Anim,Inan	abbr	Abbr=Yes
prop	PROPN	bad	BadStyle=Yes
geo	NameType=Geo	subst	-
verb	VERB	rare	Style=Rare
imperf	Aspect=Imp	coll	-
perf	Aspect=Perf	arch	Style=Arch
rev	Reflex=Yes	slang	-
inf	VerbForm=Inf	alt	Orth=Alt
futr	Tense=Fut; Mood=Ind	vulg	-
past	Tense=Past; Mood=Ind	ua_1992	-
pres	Tense=Pres; Mood=Ind	ua_2019	-
impr	Mood=Imp	var	Animacy[gram]=Anim
impers	VerbForm=Fin; Person=0; Mood=Ind	:xp[1-9]	-
1	VerbForm=Fin; Person=1	#	-
2	VerbForm=Fin; Person=2	v-u	-
3	VerbForm=Fin; Person=3	&pron	-
adj	ADJ	&numr	NumType=Ord
compb	Degree=Pos	&&numr	NumType=Card
compc	Degree=Cmp	&insert	-
comps	Degree=Sup	&predic	-
short	Variant=Short	pers	PronType=Prs
long	Variant=Uncontr	refl	Poss=Yes|PronType=Prs|Reflex=Yes
adjp	VerbForm=Part	pos	Poss=Yes|PronType=Prs
actv	Voice=Act	dem	PronType=Dem
pasv	Voice=Pass	def	PronType=Rel
v_zna:rinanim	Animacy=Inan	int	PronType=Int
v_zna:ranim	Animacy=Anim	rel	PronType=Rel
adv	ADV	neg	PronType=Neg
advp	VERB;VerbForm=Conv	ind	PronType=Ind
prep	ADP	gen	PronType=Tot
conj	-	emph	PronType=Emp
conj:subord	SCONJ	number	-
conj:coord	CCONJ	latin	-
part	PART	date	-
intj	INTJ	time	-
numr	NUM	hashtag	-
noninfl	Uninflect=Yes	punct	PUNCT
foreign	Foreign=Yes	symb	SYM
onomat	-	unknown	X
v_naz	Case=Nom	unclass	X
v_rod	Case=Gen	-	AUX
v_dav	Case=Dat	-	Mood=Cnd
v_zna	Case=Acc	noun.*pron	PRON
v_oru	Case=Ins	adv.*pron	ADV
v_mis	Case=Loc	numr.*pron	DET
v_kly	Case=Voc	adj.*pron	DET
nv	InflClass=Ind		

C Vocative and Nominative Usage Analysis



(a) Distribution of nouns in the vocative and nominative cases in direct address (1990–2024)



(b) Distribution of use in the vocative and nominative cases for the most frequent lemmas in direct address (after 2003)