Developing a Universal Dependencies Treebank for Ukrainian Parliamentary Speech

Maria Shvedova, Arsenii Lukashevskyi, Andry Rysin

UNLP
31.07–01.08.2025

Why Universal Dependencies for Ukrainian?

- A universal framework for syntactic and morphological annotation
- Enables integration of Ukrainian data into projects such as ParlaMint, InterCorp, ParaRook, and many others
- De facto standard and the only available solution for syntactic annotation of the Ukrainian language

Why creating new treebank?

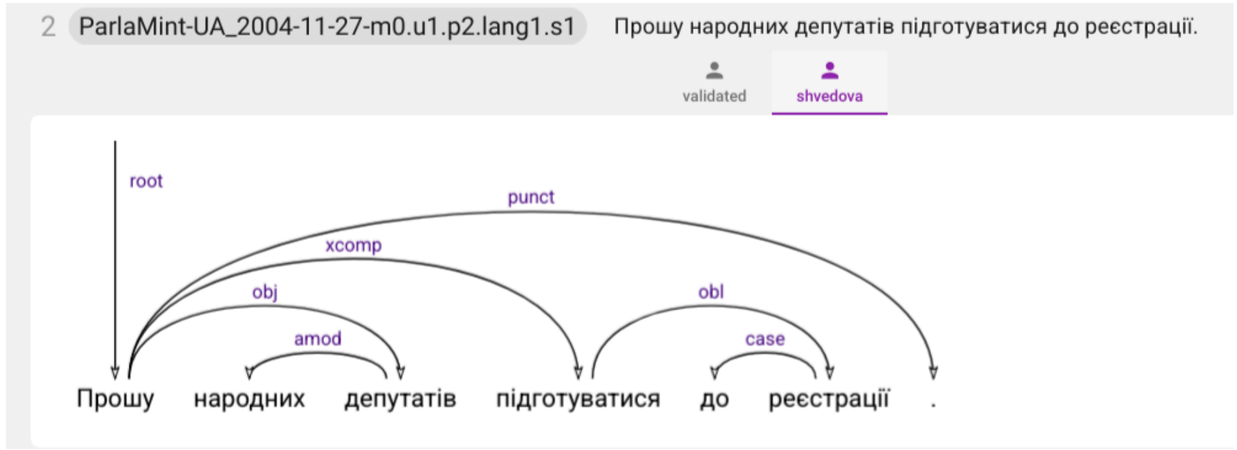
- There are currently two treebanks: IU and ParlaMint
- Despite having more tokens, IU is not entirely suitable for annotating spoken texts because it consists of written texts
- Recent changes to the UD standard
- Complex spoken syntax, conversational syntactic patterns, live speech errors, and self-corrections...
- Growing international interest in parliamentary treebanks

Which data do we use?

- Parliamentary transcripts of sessions of the Verkhovna Rada
- Transcripts from the National Security and Defense Council
- Additional audio verification of new transcripts due to potential Whisper over-normalization

The first step

- Syntactic annotation using ArboratorGrew
- Initially using ParlaMint files as bootstrap (morphosyntactic annotation using IU model), transitioning to a model based on our treebank
- Significant reduction in annotation time



Dual-parser approach

- **Dual-parser morphological annotation**: rule-based **TagText** + neural **UDPipe2**
- **Complementary error patterns** enable quality control through disagreement detection
- **UDPipe2**: better case disambiguation and pronoun recognition
- **TagText**: superior lemma accuracy and morphological features
- **Agreement = confidence; disagreement = manual review**

https://github.com/brown-uk/nlp_uk

TagText

https://github.com/ufal/udpipe/tree/udpipe-2

ÚFAL

Tagset conversion

- Need for VESUM-to-UD tag conversion
- VESUM tagset: 100 POS, morphological, and additional tags
- Most tags have direct UD equivalents (number, gender, case, person, tense, aspect, mood, etc.)
- 16 VESUM tags with no UD correspondence: style, spelling standards, dates, numbers, hashtags
- New UD feature created: BadStyle=Yes for non-standard but common forms

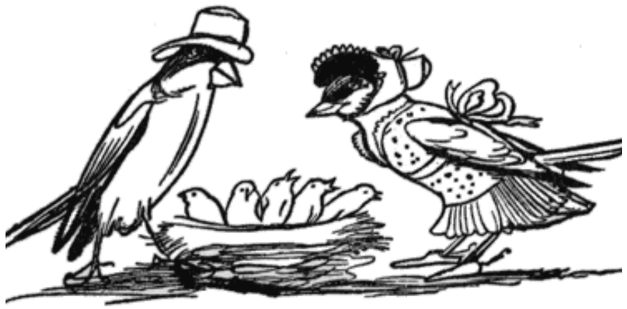
VESUM	UD	VESUM	UD
noun	NOUN	ns	Number=Ptan
anim	Animacy=Anim	p	Number=Plur
fname	NameType=Giv	s	Number=Sing
lname	NameType=Sur	m	Gender=Masc
pname	NameType=Pat	f	Gender=Fem
inanim	Animacy=Inan	n	Gender=Neut
unanim	Animacy=Anim,Inan	abbr	Abbr=Yes
prop	PROPN	bad	BadStyle=Yes
geo	NameType=Geo	subst	-
verb	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		

User-friendly review interface

	A	B	D	E	G	J	K	L	N	P	Q
1	# sent_id = 2										
2	# text = Шановні колеги, Рада національної безпеки має	Секретар	а Ради національної безпеки, найближчим часом будуть сформовані фактично всі підрозділи РНБО, призначені заступники Секретаря Ради національної безпеки та оборони								
3	ID	FORM_orig	LEMMA_orig	LEMMA_new	UPOS_ori	XPOS_orig	XPOS_new	HE/	DEPREL_o	FEATS_orig	FEATS_new
4	1	Шановні	шановний	шановний	ADJ	ADJ	adj:p:v_kly:compb	2	amod	Case=Voc|Degree=Pos|Number=Plur	Case=Voc|Degree=Pos|Number=Plur
5	2	колеги	колега	колега	NOUN	NOUN	noun:anim:p:v_kly	7	vocative	Animacy=Anim|Case=Voc|Gender=Fem,Masc|Number=Plur	Animacy=Anim|Case=Voc|Gender=Fem|Number=Plur
6	3	,	,	,	PUNCT	PUNCT	punct	2	punct	_	_
7	4	Рада	рада	рада	NOUN	NOUN	noun:inanim:f:v_naz	7	nsubj	Animacy=Inan|Case=Nom|Gender=Fem|Number=Sing	Animacy=Inan|Case=Nom|Gender=Fem|Number=Sing
8	5	національної	національний	національний	ADJ	ADJ	adj:f:v_rod	6	amod	Case=Gen|Gender=Fem|Number=Sing	Case=Gen|Gender=Fem|Number=Sing
9	6	безпеки	безпека	безпека	NOUN	NOUN	noun:inanim:f:v_rod	4	nmod	Animacy=Inan|Case=Gen|Gender=Fem|Number=Sing	Animacy=Inan|Case=Gen|Gender=Fem|Number=Sing
10	7	має	мати	мати	VERB	VERB	verb:imperf:pres:s:3	0	root	Aspect=Imp|Mood=Ind|Number=Sing|Person=3|Tense=Pres|Ve	Aspect=Imp|Mood=Ind|Number=Sing|Person=3|Tense=Pres|
11	8	Секретаря	секретар	секретар	NOUN	NOUN	noun:anim:m:v_zna	7	obj	Animacy=Anim|Case=Acc|Gender=Masc|Number=Sing	Animacy=Anim|Case=Acc|Gender=Masc|Number=Sing
12	9	Ради	рада	рада	NOUN	NOUN	noun:inanim:f:v_rod	8	nmod	Animacy=Inan|Case=Gen|Gender=Fem|Number=Sing	Animacy=Inan|Case=Gen|Gender=Fem|Number=Sing
13	10	національної	національний	національний	ADJ	ADJ	adj:f:v_rod	11	amod	Case=Gen|Gender=Fem|Number=Sing	Case=Gen|Gender=Fem|Number=Sing
14	11	безпеки	безпека	безпека	NOUN	NOUN	noun:inanim:f:v_rod	9	nmod	Animacy=Inan|Case=Gen|Gender=Fem|Number=Sing	Animacy=Inan|Case=Gen|Gender=Fem|Number=Sing
15	12	,	,	,	PUNCT	PUNCT	punct	16	punct	_	_
16	13	найближчим	найближчий	найближчий	ADJ	ADJ	adj:m:v_oru:comps	14	amod	Case=Ins|Degree=Sup|Gender=Masc|Number=Sing	Case=Ins|Degree=Sup|Gender=Masc|Number=Sing
17	14	часом	час	час	NOUN	NOUN	noun:inanim:m:v_oru	16	obl	Animacy=Inan|Case=Ins|Gender=Masc|Number=Sing	Animacy=Inan|Case=Ins|Gender=Masc|Number=Sing
18	15	будуть	бути	бути	AUX	AUX	verb:imperf:futr:p:3	16	cop	Aspect=Imp|Mood=Ind|Number=Plur|Person=3|Tense=Fut|Verb	Aspect=Imp|Mood=Ind|Number=Plur|Person=3|Tense=Fut|V
19	16	сформовані	сформований	сформований	ADJ	ADJ	adj:p:v_naz:&adjp:pasv:perf	7	conj	Aspect=Perf|Case=Nom|Number=Plur|VerbForm=Part|Voice=Pa	Aspect=Perf|Case=Nom|Number=Plur|VerbForm=Part|Voice
20	17	фактично	фактично	фактично	ADV	ADV	adv:compb	16	advmod	_	Degree=Pos
21	18	всі	весь	весь	DET	DET	adj:p:v_zna:rinanim:&pron:ger	19	det	Case=Nom|Number=Plur|PronType=Tot	Animacy=Inan|Case=Acc|Number=Plur|PronType=Tot
22	19	підрозділи	підрозділ	підрозділ	NOUN	NOUN	noun:inanim:p:v_zna	16	nsubj:pass	Animacy=Inan|Case=Nom|Gender=Masc|Number=Plur	Animacy=Inan|Case=Acc|Gender=Masc|Number=Plur
23	20	РНБО	РНБО	РНБО	PROPN	PROPN	noun:inanim:f:v_rod:nv:abbr:p	19	nmod	Abbr=Yes|Animacy=Inan|Case=Gen|Gender=Fem|InflClass=Ind|	Abbr=Yes|Animacy=Inan|Case=Gen|Gender=Fem|InflClass=In
24	21	,	,	,	PUNCT	PUNCT	punct	22	punct	_	_
25	22	призначені	призначений	призначений	ADJ	ADJ	adj:p:v_naz:&adjp:pasv:perf	16	conj	Aspect=Perf|Case=Nom|Number=Plur|VerbForm=Part|Voice=Pa	Aspect=Perf|Case=Nom|Number=Plur|VerbForm=Part|Voice
26	23	заступники	заступник	заступник	NOUN	NOUN	noun:anim:p:v_naz	22	nsubj:pass	Animacy=Anim|Case=Nom|Gender=Masc|Number=Plur	Animacy=Anim|Case=Nom|Gender=Masc|Number=Plur
27	24	Секретаря	секретар	секретар	NOUN	NOUN	noun:anim:m:v_rod	23	nmod	Animacy=Anim|Case=Gen|Gender=Masc|Number=Sing	Animacy=Anim|Case=Gen|Gender=Masc|Number=Sing
28	25	Ради	рада	рада	NOUN	NOUN	noun:inanim:f:v_rod	24	nmod	Animacy=Inan|Case=Gen|Gender=Fem|Number=Sing	Animacy=Inan|Case=Gen|Gender=Fem|Number=Sing
29	26	національної	національний	національний	ADJ	ADJ	adj:f:v_rod	27	amod	Case=Gen|Gender=Fem|Number=Sing	Case=Gen|Gender=Fem|Number=Sing
30	27	безпеки	безпека	безпека	NOUN	NOUN	noun:inanim:f:v_rod	25	nmod	Animacy=Inan|Case=Gen|Gender=Fem|Number=Sing	Animacy=Inan|Case=Gen|Gender=Fem|Number=Sing

.nest for annotations

```
<nest>
  <sentence sent_id="1" text="Гніздо">
    <token id="1">
      <form>Гніздо</form>
      <lemma>гніздо</lemma>
      <upos>NOUN</upos>
      <feats Animacy="Inan" Case="Nom"
Gender="Neut" Number="Sing" />
      <head>0</head>
      <deprel>root</deprel>
    </token>
  </sentence>
</nest>
```



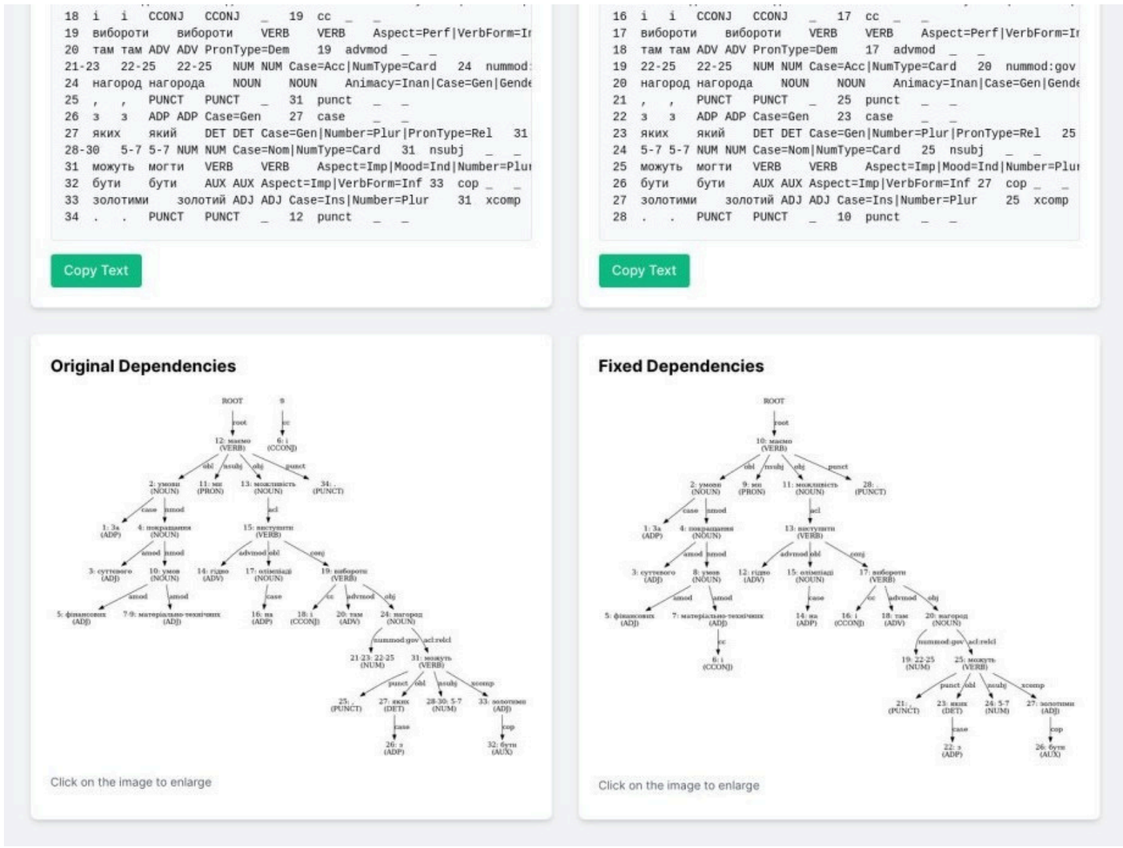
Case=Gen->Nom;Number=Sing->Plur
-Case=Gen;-Gender=Neut;-NumType=Ord;-Number=Sing
-Animacy=Inan;Gender=Neut->Masc
Number=Sing->Plur
Case=Gen->Nom;Number=Sing->Plur

Fixing over-splitting

- Cross-parser tokenization inconsistency
- Sometimes UD over-splitting causes alignment problems (e.g., [’Po-tretje’] <=> [’Po-’, ’tretje’] ’thirdly’; [’Prem’jer-ministr’] <=> [’Prem’jer’, ’-’, ’ministr’] ’prime minister’)
- Direct matching → component checking → difflib analysis

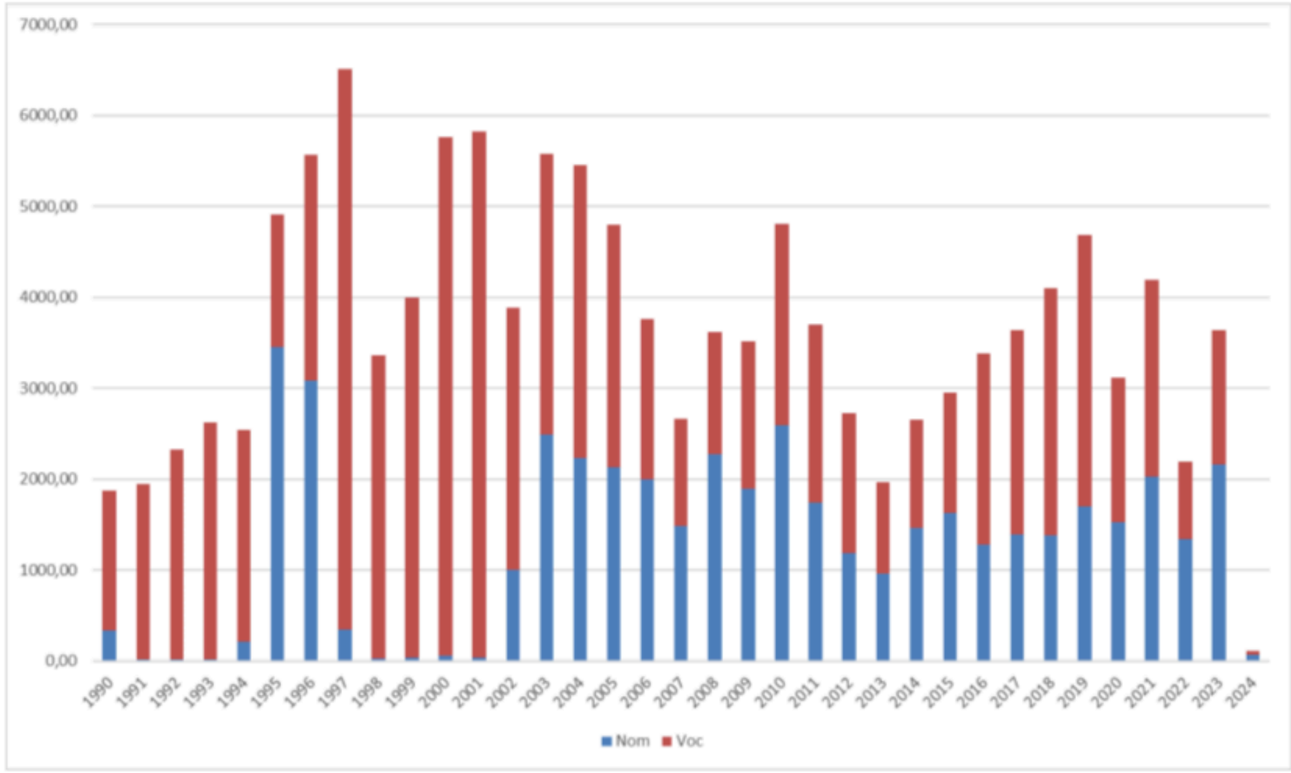
ID normalization

- Incorrect CONLL-U IDs despite correct syntactic dependencies
- Caused by manual Excel editing or excessive token splitting
- Syntactic relations point to distorted tokens/the IDs' order is not subsequent
- Track all syntactic relations + reconstruct dependency graph
- "Renumbering" effectively addresses ID inconsistencies

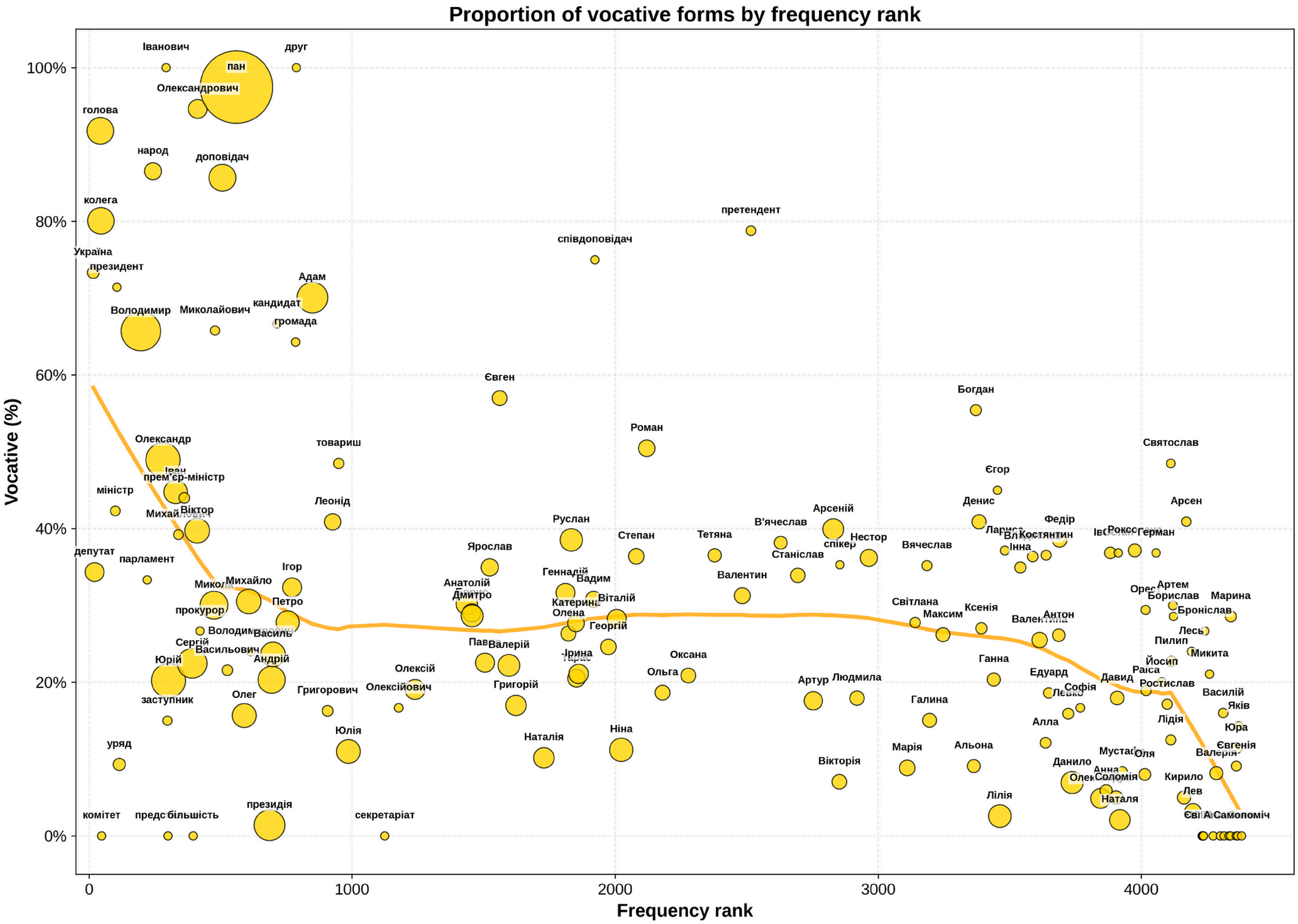


Rada_Trees corpus

- Ukrainian parliament transcripts (1990-2024)
- 88 million tokens
- ukrainian-parlamint-ud-2.15-241121 annotation
- >128 thousand contexts with the *vocative* relation
- Data from before 1995 and from 1997-2001 show 100% vocative use and are likely edited.



https://huggingface.co/datasets/uacorpus/Rada_Trees/tree/main


Proportion of vocative forms by frequency rank
Vocative (%)
Frequency rank
0%
20%
40%
60%
80%
100%
0
1000
2000
3000
4000
Іванович
пан
друг
Олександрович
голова
народ
доповідач
колега
Україна
президент
Адам
Володимир
Миколайович
кандидат
громада
претендент
співдоповідач
Євген
Богдан
Роман
Олександр
товариш
Святослав
прем'єр-міністр
міністр
Єгор
Леонід
Денис
Арсен
Руслан
Арсеній
В'ячеслав
Федір
Віктор
Степан
Тетяна
Нестор
Костянтин
Герман
депутат
спікер
Вячеслав
Інна
парламент
Ярослав
Станіслав
Геннадій
Вадим
Валентин
Ігор
Михайло
Анатолій
Дмитро
Артем
Марина
Борислав
Петро
Віталій
Світлана
Катерина
Ксенія
Броніслав
прокурор
Олена
Максим
Антон
Георгій
Лесь
Василь
Пилип
Сергій
Васильович
Павло
Валерій
Ірина
Оксана
Микита
Андрій
Йосип
Юрій
Ганна
Олексій
Ольга
Едуард
Давид
Раїса
Григорович
Олексійович
Людмила
Артур
Софія
Ростислав
Олег
заступник
Григорій
Галина
Василій
Яків
Лідія
Алла
Юлія
Ніна
Юра
Наталія
Євгенія
уряд
Марія
Альона
Валерія
Вікторія
Данило
Кирило
Соломія
Лев
Наталя
президія
Лілія
комітет
більшість
секретаріат
Самопоміч

Future plans

- Expand corpus size for broader coverage
- Add multiword expressions and analytical constructions
- Invite more expert annotators to ensure reliability
- Continue exploring spoken syntax phenomena

Thank you for your attention!