Evaluation of taggers for 19th-century fiction

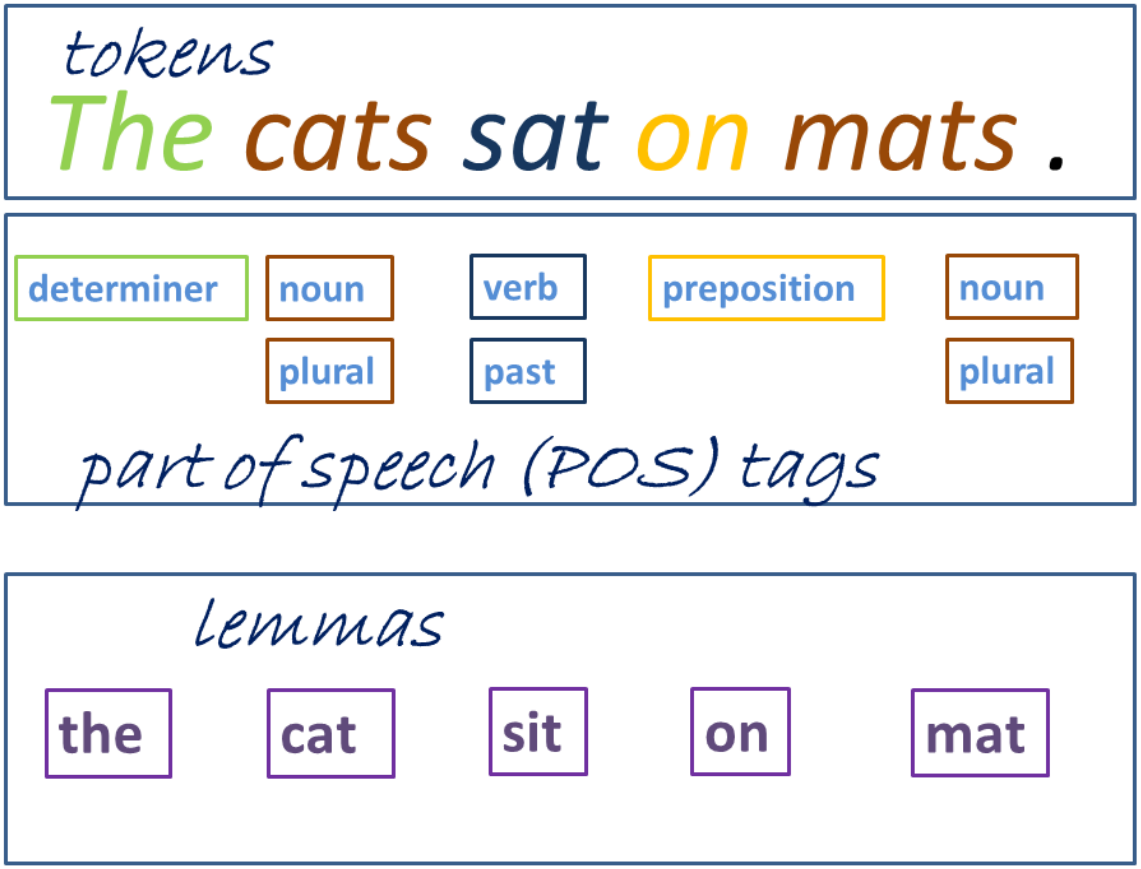
Silvie Cinková – Tomaž Erjavec –
Cláudia Freitas – Ioana Galleron –
Péter Horváth – Christian-Emil Ore –
Pavel Smrž – Balázs Indig

INTER-EXCELLENCE
LINDAT CLARIAH-CZ
DARIAH-CZ
CLARIN
cost EUROPEAN COOPERATION IN SCIENCE & TECHNOLOGY
CA16204 - Distant Reading for European Literary History

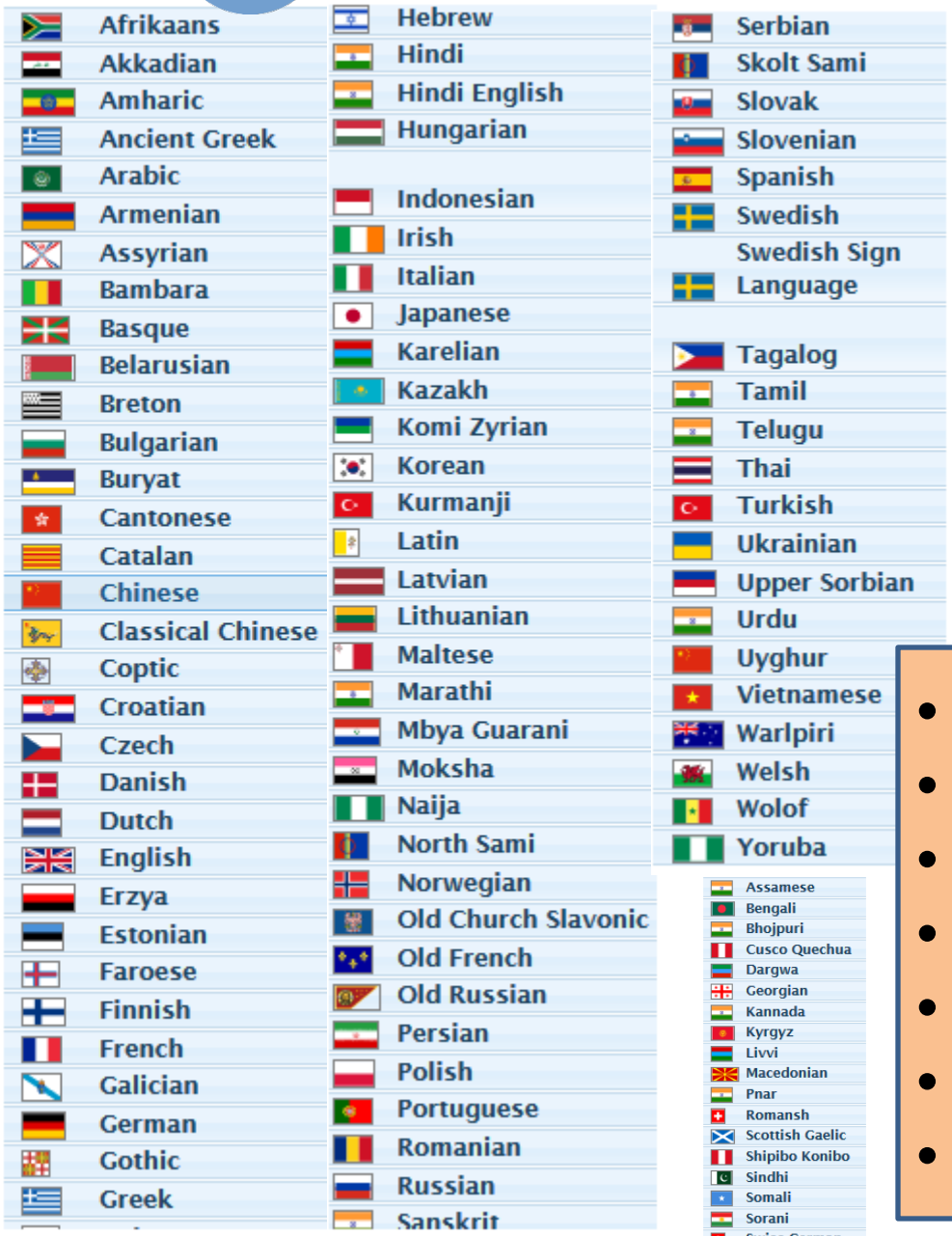
Distant Reading

19th-century fiction corpora

Language	Texts	Words
cze	23	692936
deu	44	8422194
eng	91	11990064
fra	72	6035694
hun	101	7563930
ita	34	3328244
nor	27	1114092
por	53	3513274
rom	5	200037
slv	72	3894267
spa	17	1981499
srp	24	886794



universaldependencies.org



form	lemma	upos	features
Her	she	PRON	Gender=Fem|Number=Sing|Person=3|Poss=Yes|PronType=Prs
diamonds	diamond	VERB	Number=Pl
blazed	blaze	VERB	Tense=Past|VerbForm=Part
out	out	ADP	_

- universal tagset + dependencies
- over 100 treebanks
- over 70 languages
- regular version releases
- since 2015
- new languages upcoming
- range of tools

universaldependencies.org

Universal POS (parts of speech)		
ADJ	adjective	*good*
ADP	adposition	*before, ago, in*
ADV	adverb	*fortunately*
AUX	auxiliary verb	*be, have*
CCONJ	coord. conjunction	*and, but*
DET	determiner	*this, a, both*
INTJ	interjection	*oh, thanks*
NOUN	noun, cf. PROPN	*woman*
NUM	numeral	*two, 12, VII*
PRON	pronoun	*you*
PART	particle	*indeed, just*
PUNCT	punctuation	*,;!.*
SCONJ	subord. conjunction	*because*
SYM	symbol	*$,%*
VERB	verb	*speak*

Feature	Values
Number	Plur, Sing, ...
Animacy	Anim, Inan
Gender	Masc, Fem, ...
VerbForm	Inf, Fin, Part
Definite	Def, Indef
Voice	Act, Pass,...

form	lemma	upos	features
Her	she	PRON	Gender=Fem|Number=Sing|Person=3|Poss=Yes|PronType=Prs
diamonds	diamond	VERB	Number=Pl
blazed	blaze	VERB	Tense=Past|VerbForm=Part
out	out	ADP	_

UDPipe

About | Run | REST API Documentation

UDPipe is an trainable pipeline for tokenization, tagging, lemmatization and dependency parsing of CoNLL-U files. UDPipe is language-agnostic and can be trained given only annotated data in CoNLL-U

http://lindat.mff.cuni.cz/services/udpipe/run.php

Description of the available methods is available in the API Documentation and the models are described in the UDPipe User's Manual.

Service

The service is freely available for testing. Respect the CC BY-NC-SA licence of the models – **explicit written permission of the authors is required for any commercial exploitation of the system**. If you use the service, you agree that data obtained by us during such use can be used for further improvements of the systems at UFAL. All comments and reactions are welcome.

Model: (•) Universal Dependencies 1.2

czech-ud-1.2-160523

Input: (•) Plain text ( ) CoNLL-U ( ) Horizontal ( ) Vertical

Actions: [x] Tag and Lemmatize [x] Parse

A Input Text | Input File

Podle agentury TASS, která se odvolala na očité svědky, neznámý ozbrojenec zasáhl velvyslance do zad.

Process Input

A Output Text | Show Table | Show Trees

Save Tree as SVG

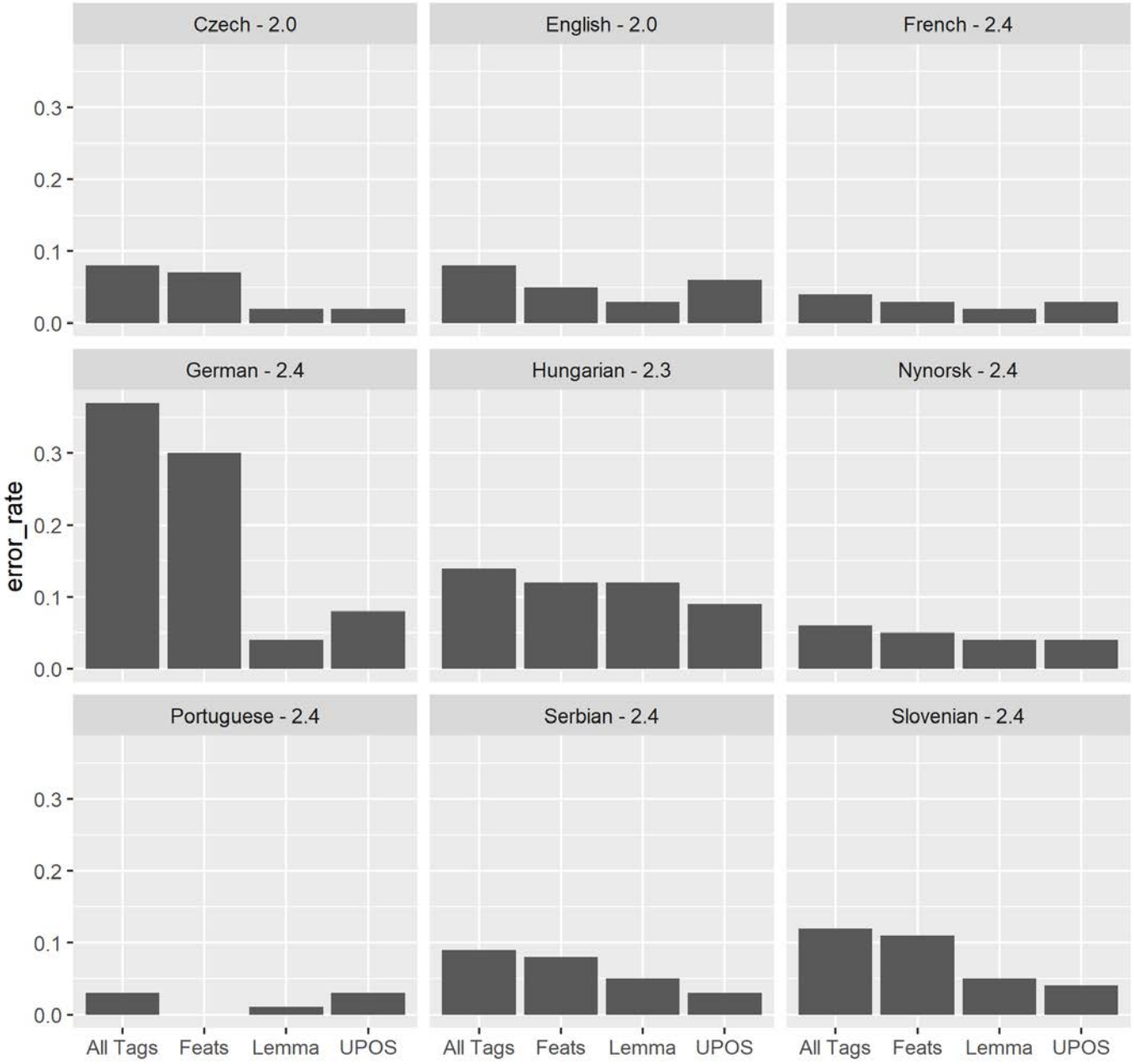
Previous 1 2 Next

Podle agentury TASS, která se odvolala na očité svědky, neznámý ozbrojenec zasáhl velvyslance do zad.

Conll-u format

Id	Form	Lemma	UPosTag	XPosTag	Feats
# newdoc					
# newpar					
# sent_id = 1					
# text = This is one of the finest of our castles.					
1	This	this	PRON	DT	Number=Sing|PronType=Dem
2	is	be	AUX	VBZ	Mood=Ind|Number=Sing|Person=3|Tense=Pres|VerbForm=Fin
3	one	one	NUM	CD	NumType=Card
4	of	of	ADP	IN	_
5	the	the	DET	DT	Definite=Def|PronType=Art
6	finest	finest	NOUN	NN	Number=Sing
7	of	of	ADP	IN	_
8	our	we	PRON	PRP$	Number=Plur|Person=1|Poss=Yes|PronType=Prs
9	castles	castle	NOUN	NNS	Number=Plur
10	.	.	PUNCT	.	_

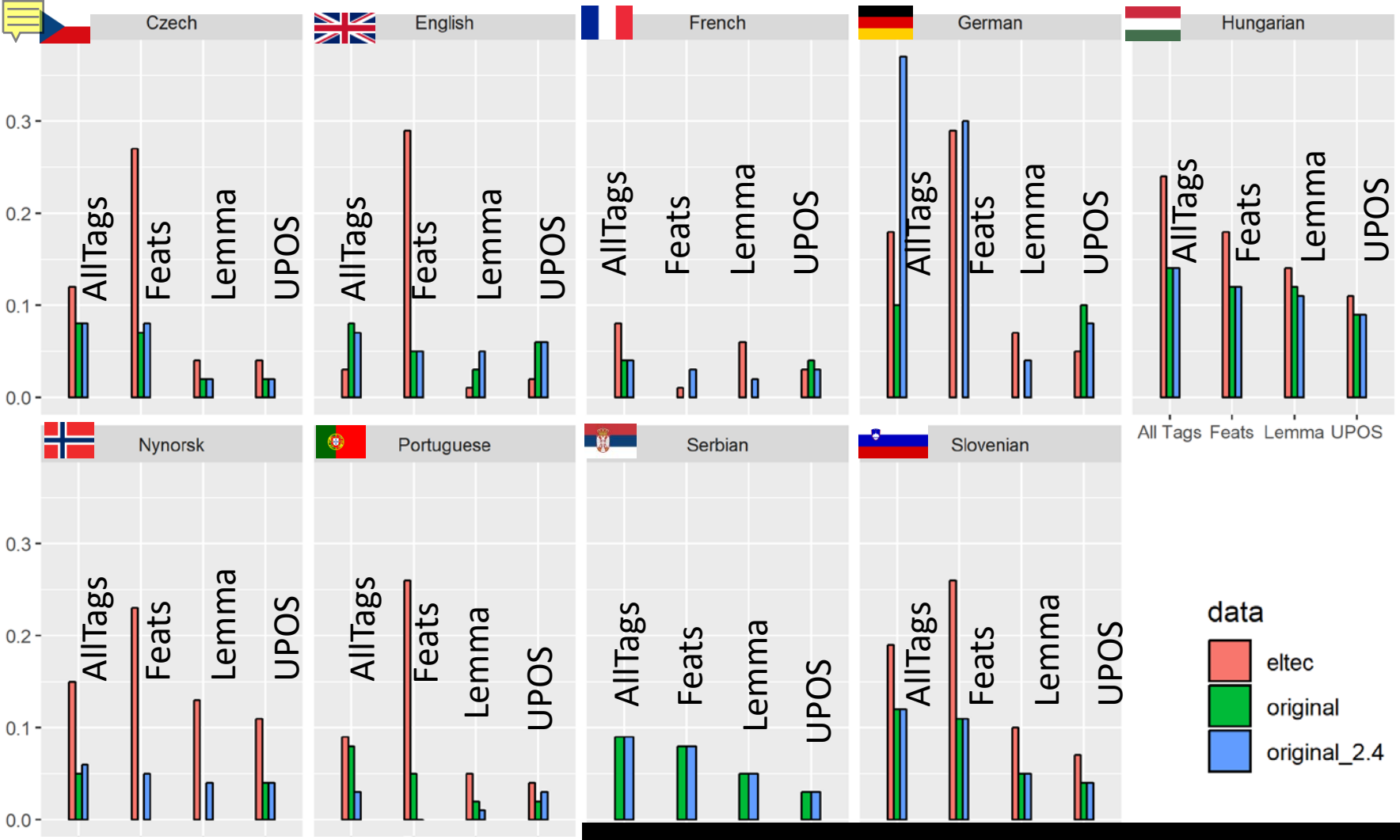
UDPipe error rate in documentation



Manual annotation

Form	Lemma	upos	xpos	feats	EV_lemma	EV_tok	EV_upos	EV_feats_errors
said	say	VERB	VBD	Mood=Ind|Tense=Past|VerbForm=Fin				
Egremo	Egremo	PROPN	NNP	Number=Sing		F		
nt	not	PART	RB	_	F	F	F	1

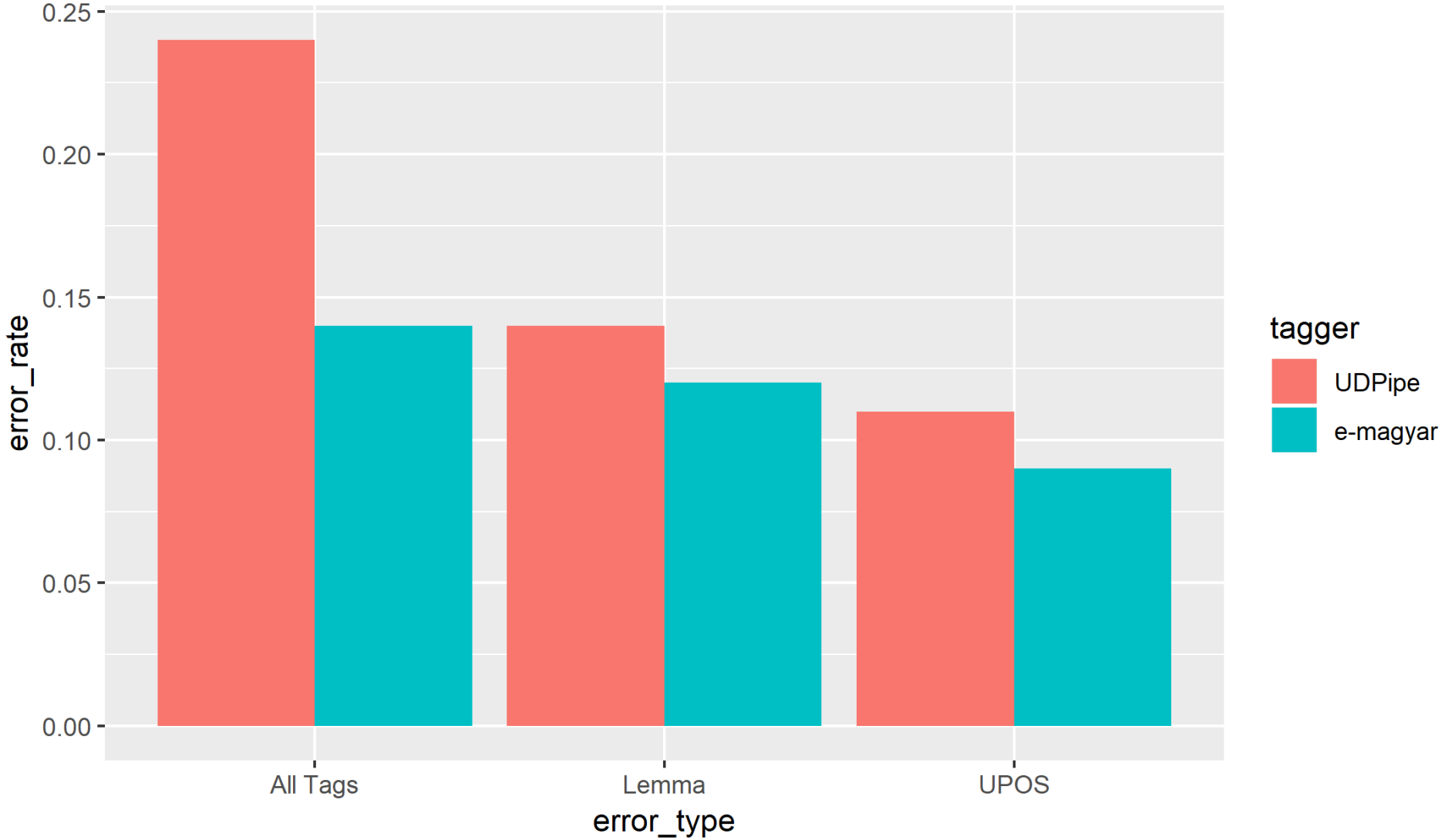
- Random sample of 5,000 tokens for each language + tagger
  - across all documents
  - length 3 tokens < complete sentences < length 30 tokens
- Annotation in a spreadsheet editor (MS Excel)
- True/False, number of errors in tokens

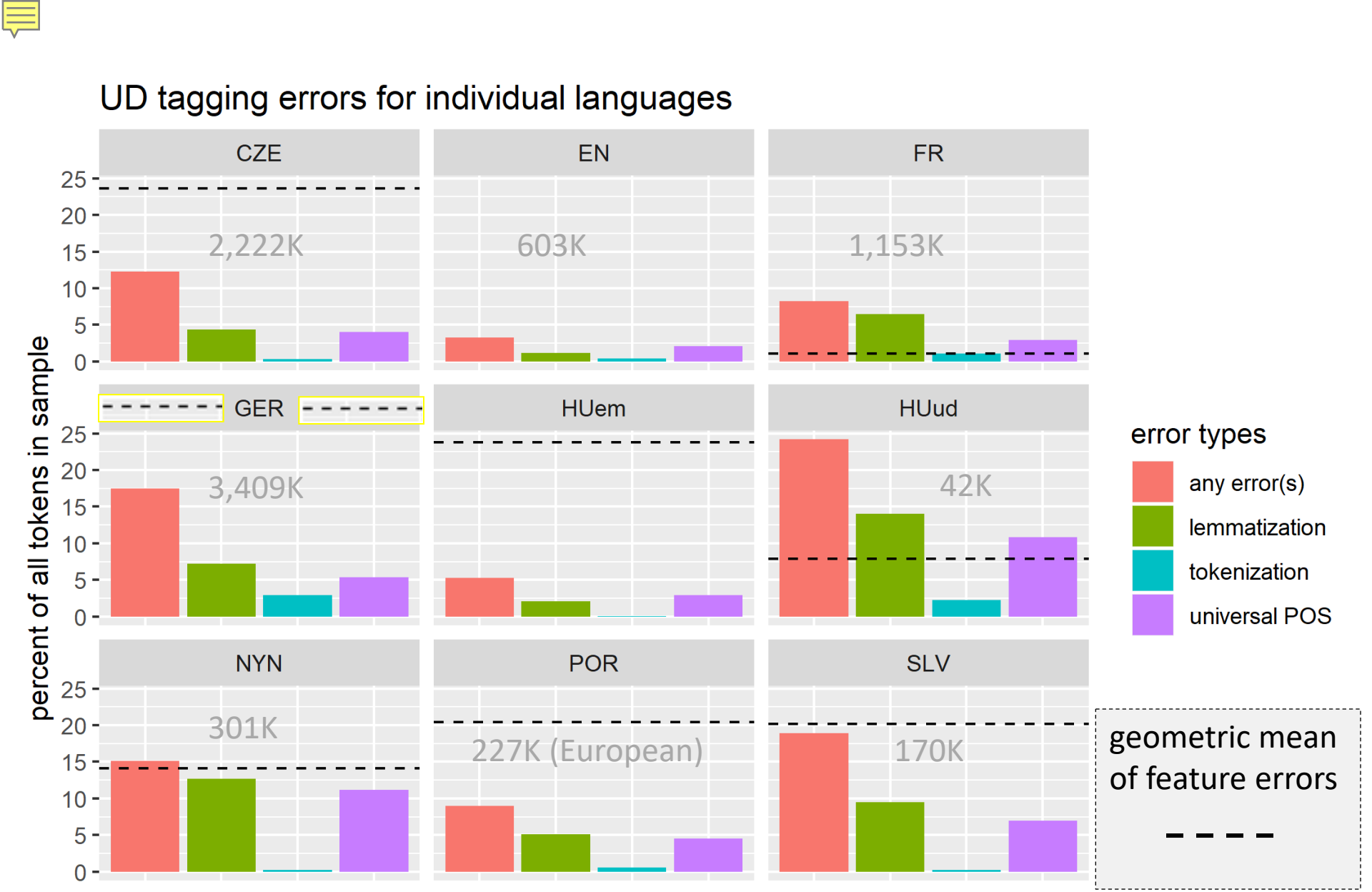


UDPipe reference vs. ELTeC

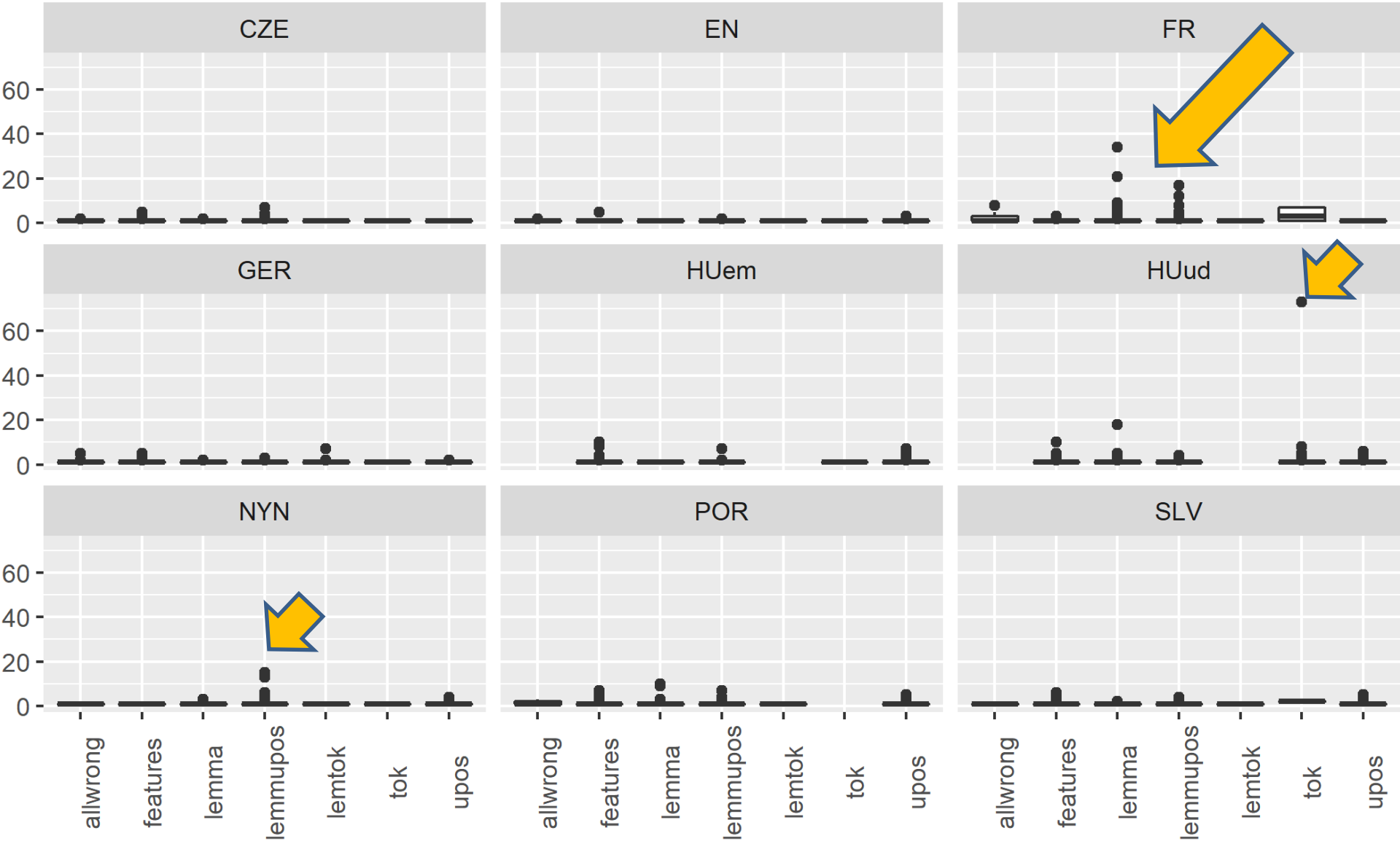
	reference	ELTeC
AllTags	F-measure all	errors/all except features
Feats	F-measure	log100k(geom.mean errors)
Lemma	F-measure	errors/all
UPOS	F-measure	errors/all

Hungarian taggers on ELTeC texts

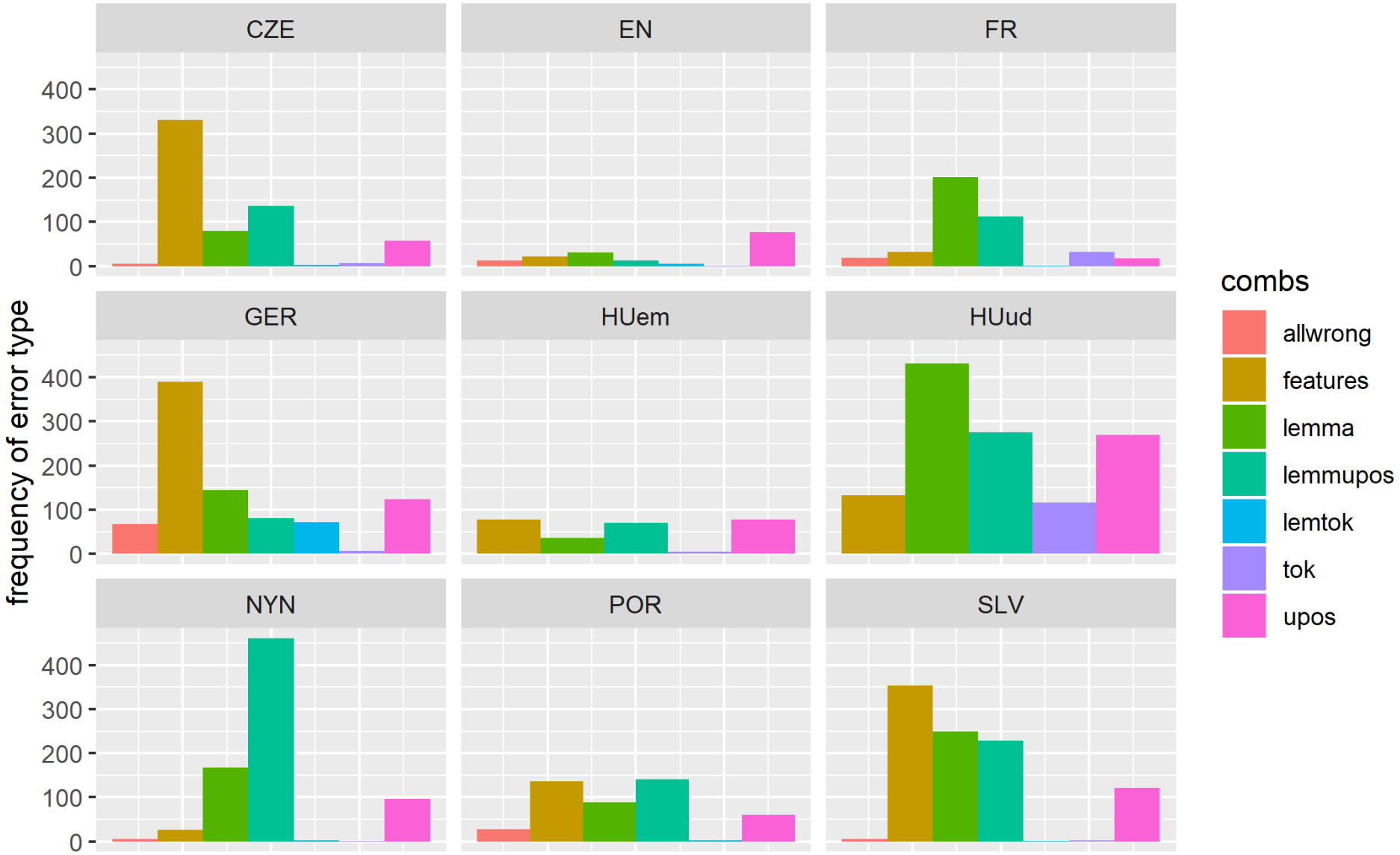



UD tagging errors for individual languages
CZE
EN
FR
GER
HUem
HUud
NYN
POR
SLV
2,222K
603K
1,153K
3,409K
42K
301K
227K (European)
170K
percent of all tokens in sample
0
5
10
15
20
25
error types
any error(s)
lemmatization
tokenization
universal POS
geometric mean of feature errors

Frequency distribution of erroneous tokens



Distribution of error combinations




allwrong
to
propose
anglo-
some
-
nt
an
morrow
hate—
features
when
what
terrified
flashed
ye
left
look
burned
mine
blazed
swelled
do
crept
that
further
her
waxed
where
lemma
wharves
scorned
parting
ornament
invigorating
mine
flouted
drawled
shot
bound
alarming
woven
throes
dared
's
's
'd
escaped
hers
're
bade
rushed
entreated
committed
hooted
hesitated
interferences
remembered
singing
slanting
worrying
lemmupos
kneeling
kindred
done
bed
yer
whom
'ere
're
yer
browsing
swathing
lemtok
i—
one
'i
?—
no—
tok
egremo
freq
a 1
a 2
a 3
a 4
a 5
combs
a allwrong
a features
a lemma
a lemmupos
a lemtok
a tok

EN: wrongly tagged as pronouns

Future work

- Serbian annotations
- manual error analysis - word clouds
- Training data for ELTeC domain necessary? **BAMTINOF** guessing error classification
  - count wrong/correct guesses on words absent in gold standard
  - classify guessing errors: BAMTINOF
    - diachr. word **B**order changes
    - syntactically **A**mbiguous word form
    - archaic or poetic **M**orphology
    - **T**ypo
    - token**I**zation error resulting in a non-existent token
    - **N**ame, proper noun
    - **O**ld spelling or archaic word typically replaced by a modern word
    - **F**oreign-language or heavy dialect