Normalizing and annotating Slovenian dialect text with no gold standard AKA LLMs for dialectal corpus compilation & analysis

Varya Magomedova · UNG

Main goals of the project

- Collect scattered dialectal materials into one user-friendly corpus
- Annotate them syntactically
- Overcome the incorrect annotationы that the current model (CLASSLA: standard, non-standard, spoken) assigns to dialectal variations using **human-corrected annotations and LLMs**

In preparation to that:

- Perform a user needs study to avoid building an inefficient resource

The work includes: questionnaires and data analysis, data collection from various sources, data cleaning and standardization, developing an annotation pipeline, developing a searchable data-base with a web-interface.

SMASH
machine learning for science and humanities postdoctoral program

User needs study

- Multiple choice questions survey (Q=8,N=94):
  - Q5: annoyance rate question: problem severity (14 problems, 1–5 scale),
  - then abandonment rates
  - desired improvements.
- Semi-structured interviews (N=13)

To summarize in one slide:

57% * of respondents have abandoned their idea because of corpus problems:

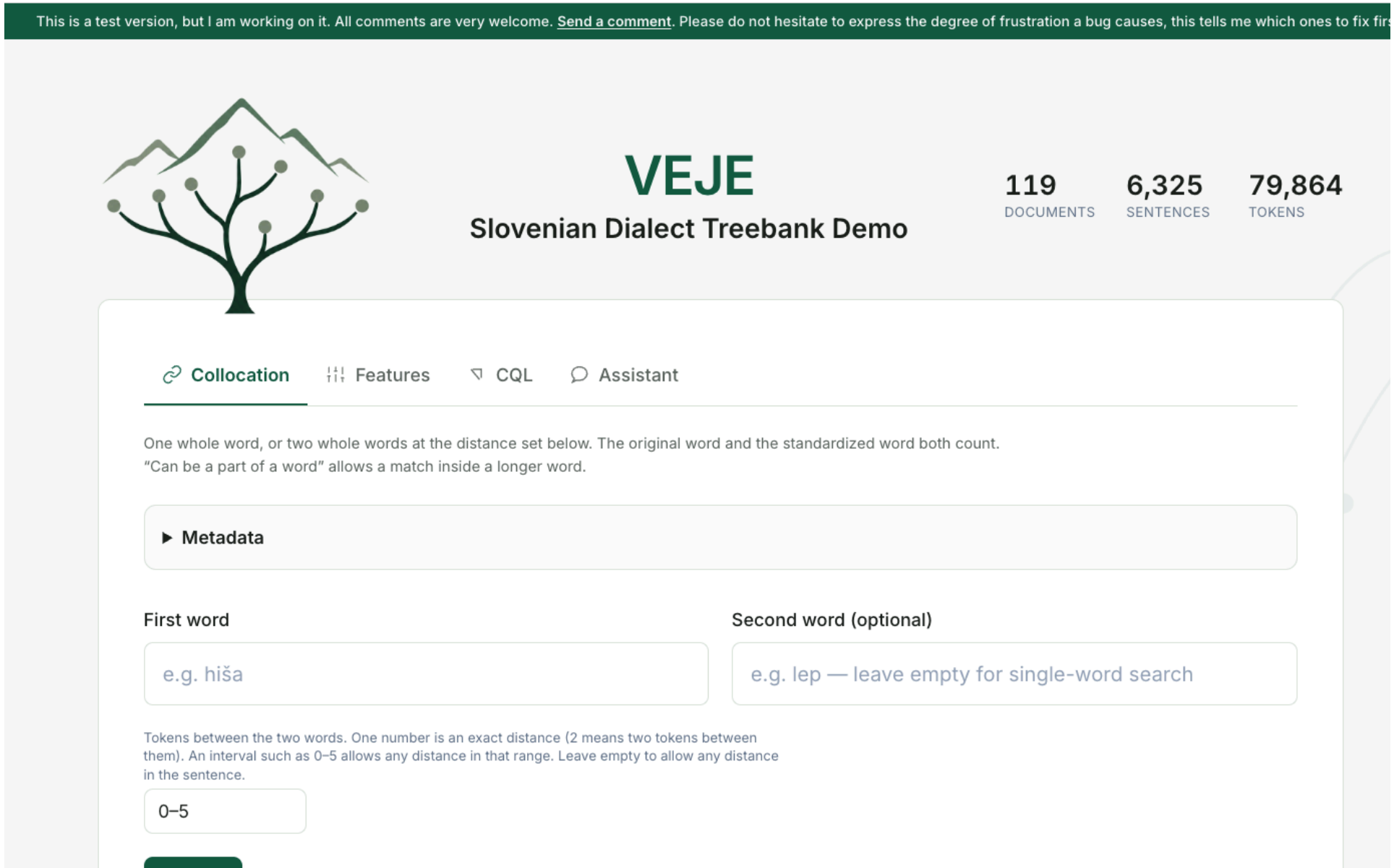
cleaning irrelevant examples, no access to full result set, no corpus exists for the research question, annotation/ tokenization errors, general distrust of corpus annotation and frequencies, missing metadata, interface-related problems,

Frustration patterns varied by specialization.

* My published CLARIN poster says 43% - this is miscalculation. Other than that - see the poster

SMASH
machine learning for science and humanities postdoctoral program

veje.ung.si - working prototype



IN A GLANCE

Eight types of problems

01

Sources

Dialect texts, papers on dialects, spoken corpora

02

“Gold” resources

Few annotated datasets and Example picking from papers on dialects

03

Notation

Every transcription mapped to one alphabet - a random choice, links to the sources attached

04

Normalization

Dictionary, ByT5, dictionary
But I tried many other ways

05

Training the normalizer

Training data, regimes, sweeps

06

Annotation

Classla on the normalized text

07

Evaluation

Held-out dialects and documents

08

Heavy Calculations

Vega, which is unpredictably down when I need it most

Design principle: token-preserving normalization; the dialect form stays linked to every standard token (D46, D70 - these little numbers are decisions in my decision log)

SMASH

THE PROBLEM OF THE PROBLEMS

Tools for (non)standard Slovenian do not read dialect text well (enough)

Ziljsko, Lipalja book

pole, k je priš tis Xitler, ne

authors' own standard text

Potem, ko je prišel tisti Hitler, ne…

Dialect word	Classla on the raw word	After normalization	On the authors' text
pole	NOUN · pol	potem · ADV	ADV · potem
k	ADP · k	ko · SCONJ	SCONJ · ko
je	AUX · biti	je · AUX	AUX · biti
priš	PROPN · priš	prišel · VERB	VERB · priti
tis	NOUN · tis	tisto · DET	DET · tisti
Xitler	PROPN · Xitler	Hitler · PROPN	PROPN · Hitler
ne	PART · ne	ne · PART	PART · ne

POS accuracy	Lemma accuracy
70 - 74 %	47 - 60 %

Classla (non-standard) on the raw unseen or seen dialect (not unseen text)

83 - 96 %	75 - 90 %

after normalization - which tells us that normalization is still far from perfect, but I have another year to fix it

Unseen dialects: 2,444 words of Ziljsko and Rož texts (Lipalja book, ZRC papers): two dialects the normalizer never saw in training. These dialects have no gold data

D90, D91, D102 ·

SMASH

Five mines of gold

Resource	Material	Size	Held out for tests	Standard side
GOKO	Kopriva na Krasu, aligned words	112 sentences	23 sentences	Full standard vocabulary
Narecja.si	111 texts, 32 dialect labels	81 usable documents	19 documents	Poknjižitev; dialect words kept
Repositories: UNG · JZ · Lipalja book	Theses, a book, papers	3,016 pairs	Ziljsko, Rož, Planina	Mixed
RUL	Ljubljana repository	1,722 pairs	Carinthian	Mixed
DKUM	Maribor repository	9,132 pairs	3 reviewed papers	Poknjižena iztočnica

Two conventions are mixed in training: full standard vocabulary, and poknjižitev, which standardizes spelling and morphology but keeps dialect words. This is because I don’t have enough data, and competence to manually revise it myself.
On the bright side, once I have manually revised trees with dialectal words, we can tune classla maybe?

D51, D77, D83, D87, D94, D95 ·

PROBLEM 3 OF 8 · NOTATION

Every transcription is mapped to a simplified corpus alphabet

GOKO
ZRCola glyphs, IPA

→

MojJezik
own symbols, colon for length

→

Theses and papers
notation of each author

→

Corpus alphabet
Standard letters, ə and w: the ByT5 input

Mapping	Source	Why	Evidence
ɛ → e ɔ → o ɪ → i	GOKO	The corpus alphabet has no open or lax vowel symbols; Slovenian spelling writes e, o, i	about 1,250 glyph-only differences disappeared; GOKO's narrow phonetic layer, mapped this way, loses only 2 to 3 points
ɐ → ə	GOKO	A reduced central vowel; the corpus has no ɐ and writes reduced vowels with ə	decision D52; not tested separately
i: → ie u: → uo	MojJezik	Its README names these as the long vowels that Narecja.si and GOKO write ie and uo, which the model knows	valid Slovenian output 57 % → 78 % (proxy, no gold); tri:ba → treba
ə → ə	MojJezik	Cyrillic ə (U+04D9) looks like the schwa ə (U+0259) but is another character, unseen in training	an encoding fix

Schwa words (sə, kə, nə) cannot be mapped by rule: unanimous cases go to a lexicon, the rest are decided per occurrence.

D52, D72, D84, D86, D89 · also see notation notes, `NOTATION.md` §5.

SMASH

STAGE 4 OF 8 · NORMALIZATION

Dictionary, then ByT5, then the dictionary again, then rules (clitic split)

1 Dictionary
exact match on the dialect form; a hit wins

→

2 ByT5
base model, three words of context each side, geo tag

→

3 Dictionary again
on ByT5's output, only if it is not a standard form

→

4 Clitic split
This is all what is left from rule-based corrections :)

→

text_norm
the same number of tokens as the input; a split keeps its link

štejnhe → štenge → stopnice

dialect form

after ByT5: second dictionary pass

po doməče se rieče bərzuota
po domače se reče brzota

A real sentence from a thesis on Velike Žablje.

Do the rules help? Not really

Dialect gold, 10,311 scored words: 21 changed; 7 fixed, 4 broken, 10 still wrong

Spoken test Krohot: word error 16.74 % without rules, 16.89 % with

Rules before the second dictionary pass: the same scores on all ten gold sets

On 22,521 real dialect words: 120 changes, of which 23 are clitic splits (k-je → ko je)

D43–D46, D70, D73–D76, D80 · guards: Sloleks 3.1 standard forms, gender and number flip check

SMASH

PROBLEM 5 OF 8 · SOURCES FOR TRAINING AND TESTING THE NORMALIZER

“Gold” pieces used for ByT5 training

Used

Material	What it is	Used for
GOKO	Karst village; 112 sentences, aligned word by word	gold: training and test
Narecja.si	81 usable texts, 32 dialects; the authors' standardization	gold: training and test
UNG theses, Lipalja book, ZRC papers	village dictionaries, Ziljsko and Rož texts	3,016 gold pairs; 2,860 dialect sentences
RUL repository	theses and papers, Univ. of Ljubljana; 24 of 86 gave material	1,722 gold pairs; 2,792 sentences
DKUM repository	theses, Univ. of Maribor	9,132 gold pairs; 10,642 sentences
GOS, GOS-VL, Artur	spoken, colloquial; region labels only	ByT5 training; GOS also the dictionary
Krohot	spoken, Maribor and Savinjska	held-out test on spoken text

Could not be used

MojJezik, Idrija and Ribnica
no standard version, so no training pairs

30 of 111 Narecja.si texts
empty, phonetic text in the standard column, or too freely rephrased to align

62 of 86 RUL documents
no dialect and standard pairs: font damage, phraseology, manuscripts, reflex tables

DKUM theses in legacy fonts
glyph readings undecided and no PDFs to check against

UNG syntax thesis, interviews
no dialect and standard pairs

D67, D77, D83, D87, D94, D95 · also used as resources: Sloleks 3.1 (standard forms), Classla (annotation)

SMASH

PROBLEM 6 OF 8 · ANNOTATION

Token aligned - so classla parses the normalized version and it applies

phonetic source, verbatim	A sej ˈmi smo ˈnaː-pašo sˈkup ˈxoːdəɬi.
dialect, corpus alphabet	A sej mi smo na-pašo skup xodəli .
orthographic, parsed	A saj mi smo na pašo skupaj hodili .

FORM	LEMMA	UPOS	DEPREL	MISC
A	a	CCONJ	cc	Dia=A
saj	saj	CCONJ	advmod	Dia=sej
mi	jaz	PRON	iobj	Dia=mi
smo	biti	AUX	aux	Dia=smo
na	na	ADP	case	**Dia=na-pašo**
pašo	paša	NOUN	obl	**Dia=na-pašo**
skupaj	skupaj	ADV	advmod	Dia=skup
hodili	hoditi	VERB	root	Dia=xodəlil SpaceAfter=No
.	.	PUNCT	punct	SpaceAfter=No

Classla 2.2.1, ***standard model***, pretokenized. The sentence header keeps source, citation, licence, dialect and location.

D62, D64, D85, D91 · a real sentence from a thesis on Velike Žablje (columns HEAD, FEATS and XPOS not shown)

SMASH

PROBLEM 5 OF 8 · STILL TRAINING THE NORMALIZER

One word plus context in, one standard form out

d:krasko narecje l

l3 l2 l1 <t> word </t> r1 r2 r3

→ standard form of word

Two regimes

gold: test and dev documents held out, so scores are clean (225,356 windows)

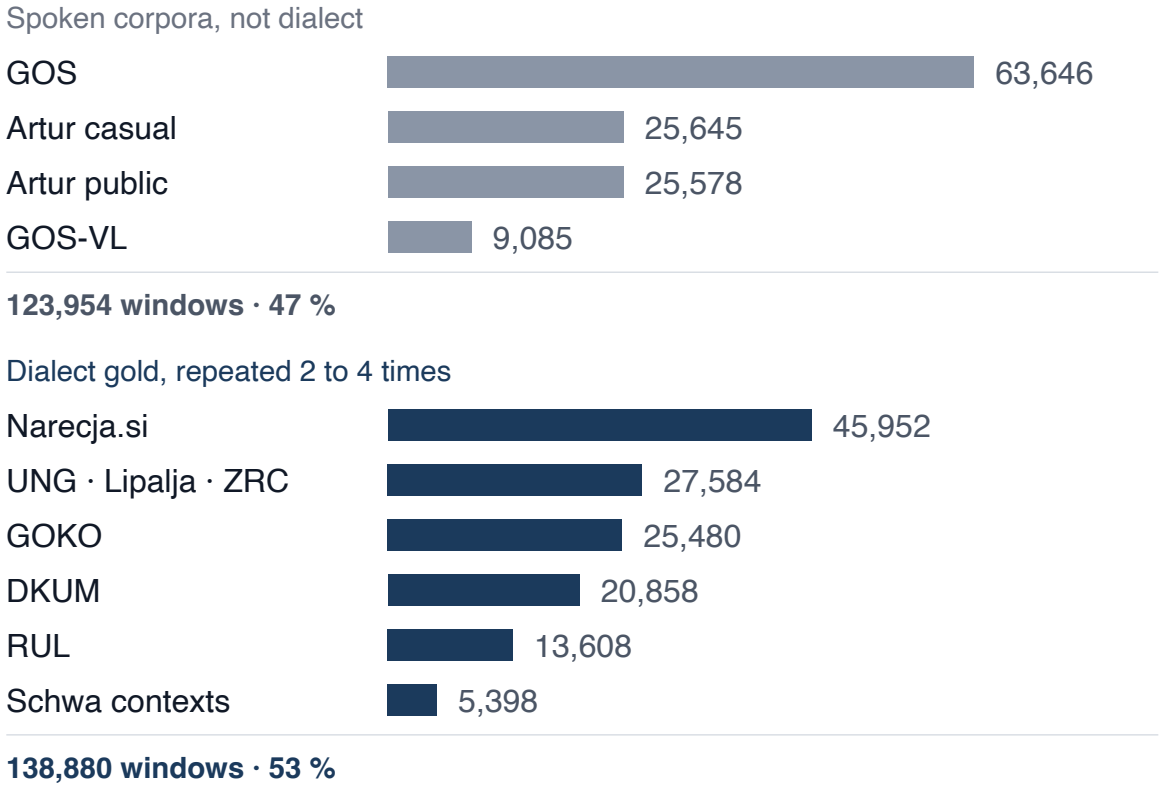
final: everything in, the model to ship (262,834 windows) - so I cannot evaluate this one honestly

ByT5-base, 8 epochs, bf16, three seeds per comparison

Still based more on spoken corpora 😢

47 % of the windows as trained; 70 % of the distinct windows, because the dialect gold is repeated 2 to 4 times.

Training windows by source, final regime

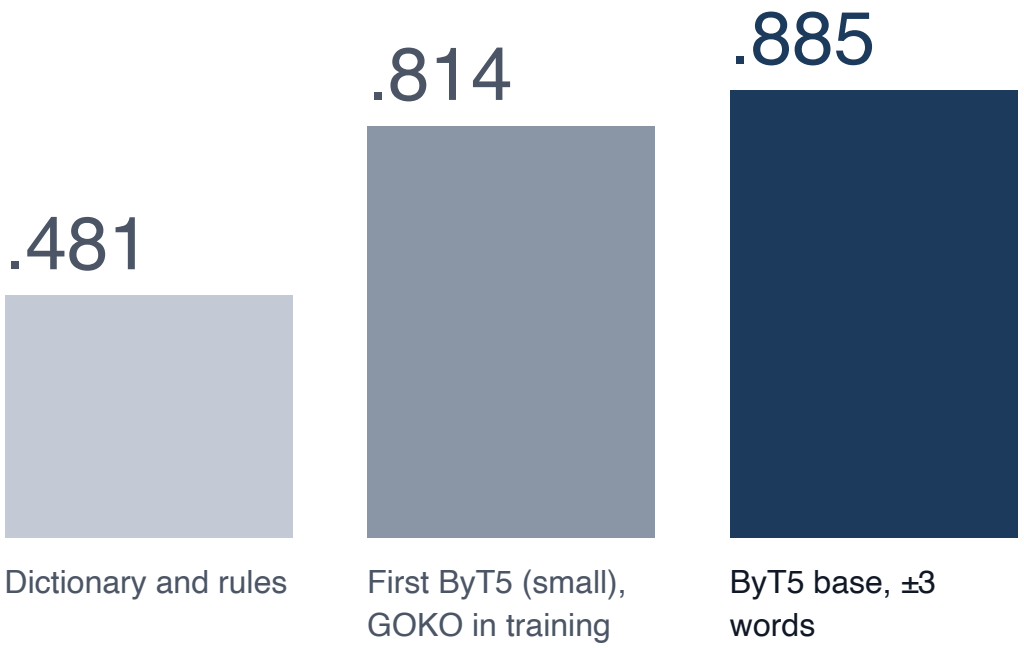


D66–D68, D82, D90, D98 · counts from the final-regime training set rebuilt on 2026-10-12

SMASH

DECISION 1 OF 6

ByT5 replaced the dictionary as the main instrument



Why

The dictionary pipeline recalled 19 % of the words that must change on GOKO, at 95 % precision.

A byte-level model with context handles spelling variation that no word list covers.

The dictionary stays for lexical swaps such as pole → potem.

D51, D66, D69, D73, D99 · token accuracy on GOKO test, 23 sentences, 1,180 scored tokens

SMASH

DECISION 2 OF 6

Size, window and geo tag were chosen by sweeps, and some gaps are noise

Choice	Compared	Finding	Decision
Model size	small, base, large	base beats small in 9 of 9 comparisons; base ties large at ±3	Base
Context window	±2, ±3, wider	±3 is never worse, but the gap is inside the noise	±3 words
Geo tag	with, without	helps only when the dialect has no in-domain training data	Keep
GOKO in training	with, without	without it GOKO accuracy falls by about 11 points	Keep
More GOS data	1×, 3×	no gain (GOKO .839 → .831)	Not added

Chose on cost: large ByT5 needs twice the compute and buys nothing over base.

D69, D82, D99 · 13 configurations, one A100 each, 3 to 6 hours per run

SMASH

DECISION 4 OF 6

Discarded

SloBERTa gate

Flags non-standard tokens so that ByT5 touches only those. Token accuracy without and with the gate:

Test set	No gate	Gate	Gate recall
GOKO	89.6	87.3	92.8
Narecja.si, seen	83.8	83.3	93.6
Narecja.si, unseen	79.0	78.2	87.8
Lipalja	75.6	75.0	79.7
ZRC texts	61.9	61.5	95.7
UNG Planina	68.8	67.6	89.9

It costs more tokens than it saves on every set. Decision D104: no gate for dialect material.

GaMS-27B word oracle

Asked about words that every earlier pass left unchanged.

1,463 words queried, about 0.8 s each

Changed 298 of 238,814 tokens (0.12 %)

In ten sampled sentences: about 4 clear fixes, at least 3 damaged words

Share of X tags unchanged (0.37 %)

Off by default. To be reconsidered only after a hand judgement of 100 changes.

D71, D101, D103, D104 · gate applied after the fact to stored outputs, threshold 0.5; the gate was trained on GOS

DECISION 5 OF 6

The standard Classla model vs the non-standard one

(Tested on unseen dialects - the one that didn’t have “gold”)

Instrument, then input	UPOS agreement, %	Lemma, %
Standard model, raw text	61.0	48.4
Non-standard model, raw text	70.7	58.4
Dictionary, then standard	69.6	56.0
ByT5 chain, then standard	**80.4**	**70.9**
ByT5 chain, then non-standard	79.5	71.0

Normalization adds about 10 UPOS and 12 lemma points over the best plain instrument.

The non-standard model helps on raw text (+9.7 UPOS) but not on normalized text (–0.9).

Annotation is bounded by normalization: UPOS agrees for 97.5 % of tokens normalized correctly and 47.7 % of those normalized wrongly.

D91, D102 · 2,444 held-out Ziljsko and Rož tokens; agreement with Classla on the authors' own standard text, not gold annotation

SMASH

DECISION 6 OF 6

Phonetic notation is mapped and kept as a linked layer, not fed to the model

Words with symbols the model never saw in training

UNG theses

91 % → 2.3 %

Lipalja book

59 % → 1.2 %

before and after mapping each source notation to the training alphabet

Decision

One alphabet for the model: map each source notation to it, drop length marks, keep the schwa. Measured: dropping length marks gave +10 points on UNG words and +13 on Lipalja; mapping the schwa cost 2 to 4.

The treebank text is orthographic. The phonetic source stays in the sentence header and each token keeps its dialect form in MISC.

The schwa gets a lexicon, not a rule: the lexicon decides only what is unanimous, the rest is decided occurrence by occurrence from the sentence. 3,419 of 3,696 ambiguous cases were resolved by reading.

GOKO's private-use transcription is romanized before training or scoring.

D52, D84, D85, D86, D89 · measured with the first small model; mapping tables in NOTATION.md, some glyph identities still to be confirmed

LIMITS

More work to do, things will probably change

Known limits

Small test sets: GOKO has 23 sentences, so one point is 12 tokens. Seed spread is 0.2 to 1.4 points.

Accuracy depends on the source: .885 on GOKO, .79 on unseen Narecja.si, .75 on Lipalja, .62 on ZRC texts.

Isolated words of unseen dialects stay out of reach: .104 on 119 words.

Two target conventions, GOKO-style and the authors' standardization, are mixed in one model.

No gold syntax for dialects yet: UPOS and lemma are compared as agreement only.

To do (be do be do)

Comparison on the same tokens with Trankit, CLASSLA-spoken, the released Kuparinen et al. model and cSMTiser.

Moree data!

Get all the republish permissions

Get manual revisions for normalizations ond the UD annotations

More more more data. From local newspapers?

Figures from the literature, such as 87.3 % on GOS in Kuparinen et al., use other data and splits; they are indicative only.

D82, D90, D99, D101–D104 and the open items of the decision log

SMASH

Thank you for your attention!