

A Free/Open-Source Morphological Analyser and Generator for Sakha

Sardana
Ivanova
Helsingin yliopisto

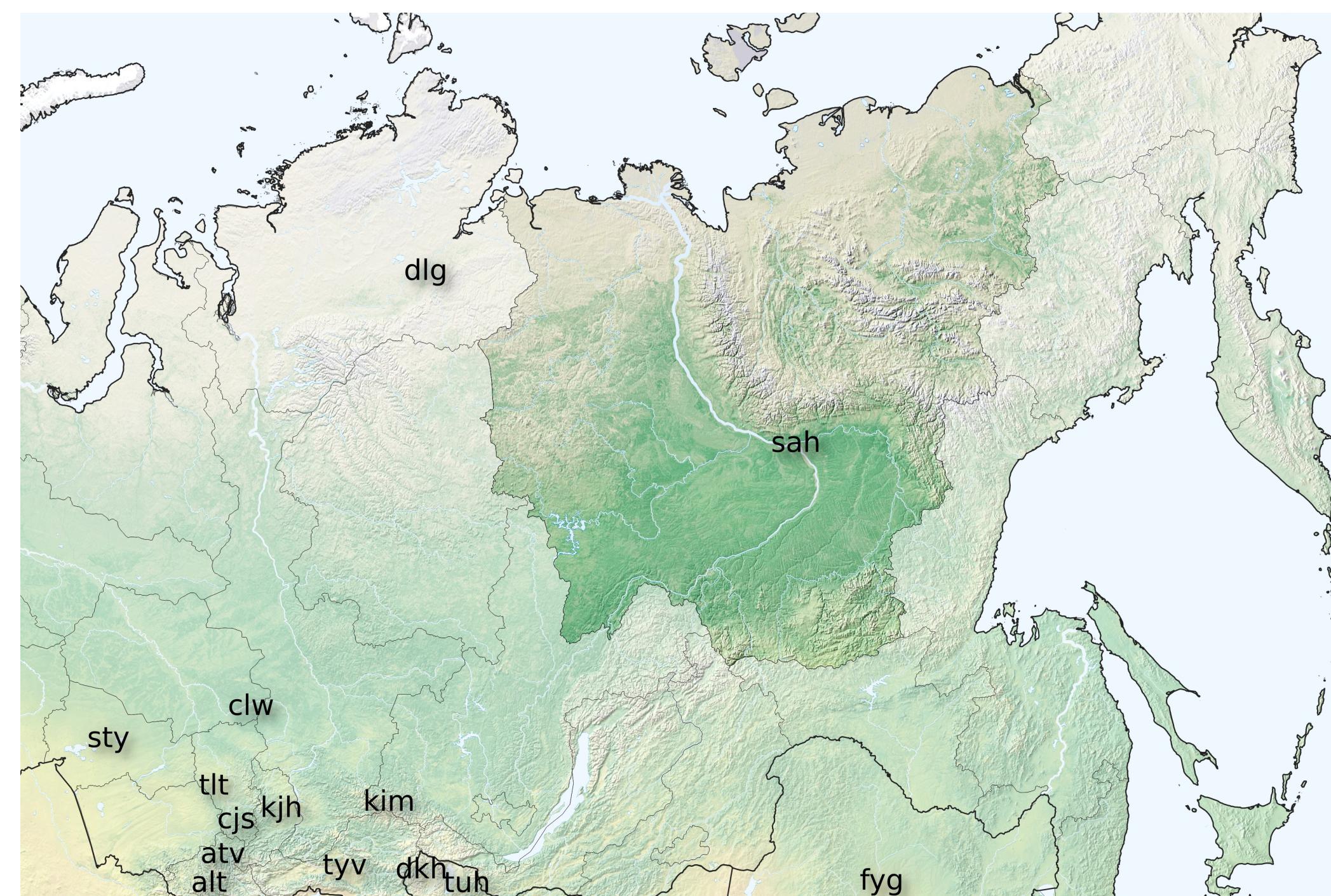
Jonathan N.
Washington
Swarthmore College

Francis M.
Tyers
Indiana University

Background

Sakha (sah)

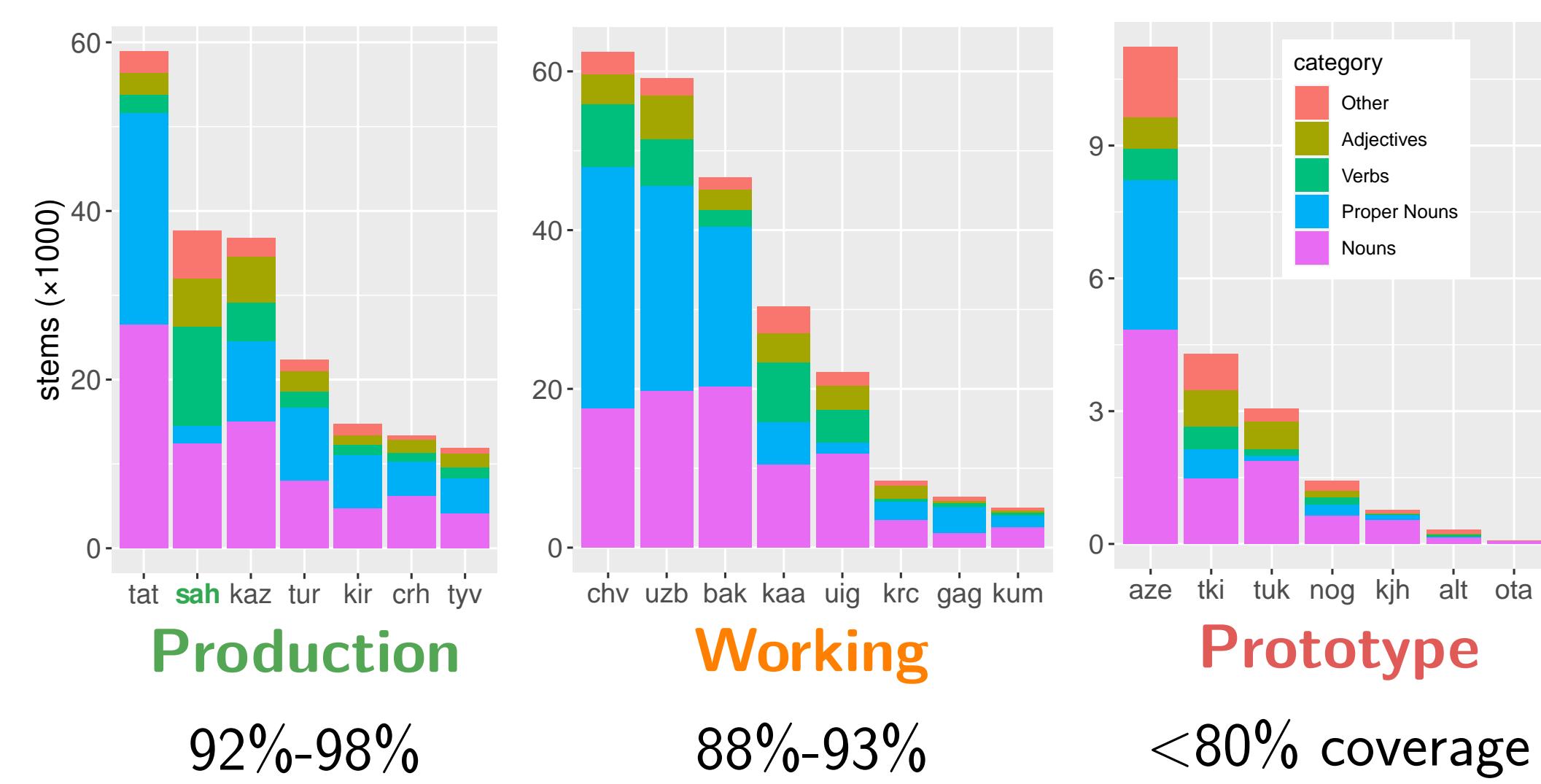
- Indigenous language of **Siberia**, official in Sakha Republic
- **Turkic** language: agglutinating, vowel harmony
- ~0.5M speakers, under pressure from Russian



Morphological transducers

- provide morphological **analysis** and **generation**:
атын ↔ ат<n><px3sg><acc>/атын<adj>/атын<post>
- Useful in **language technology** and **downstream tasks**:
spell checking, MT, CALL, text processing, etc.
- Only one development cycle!

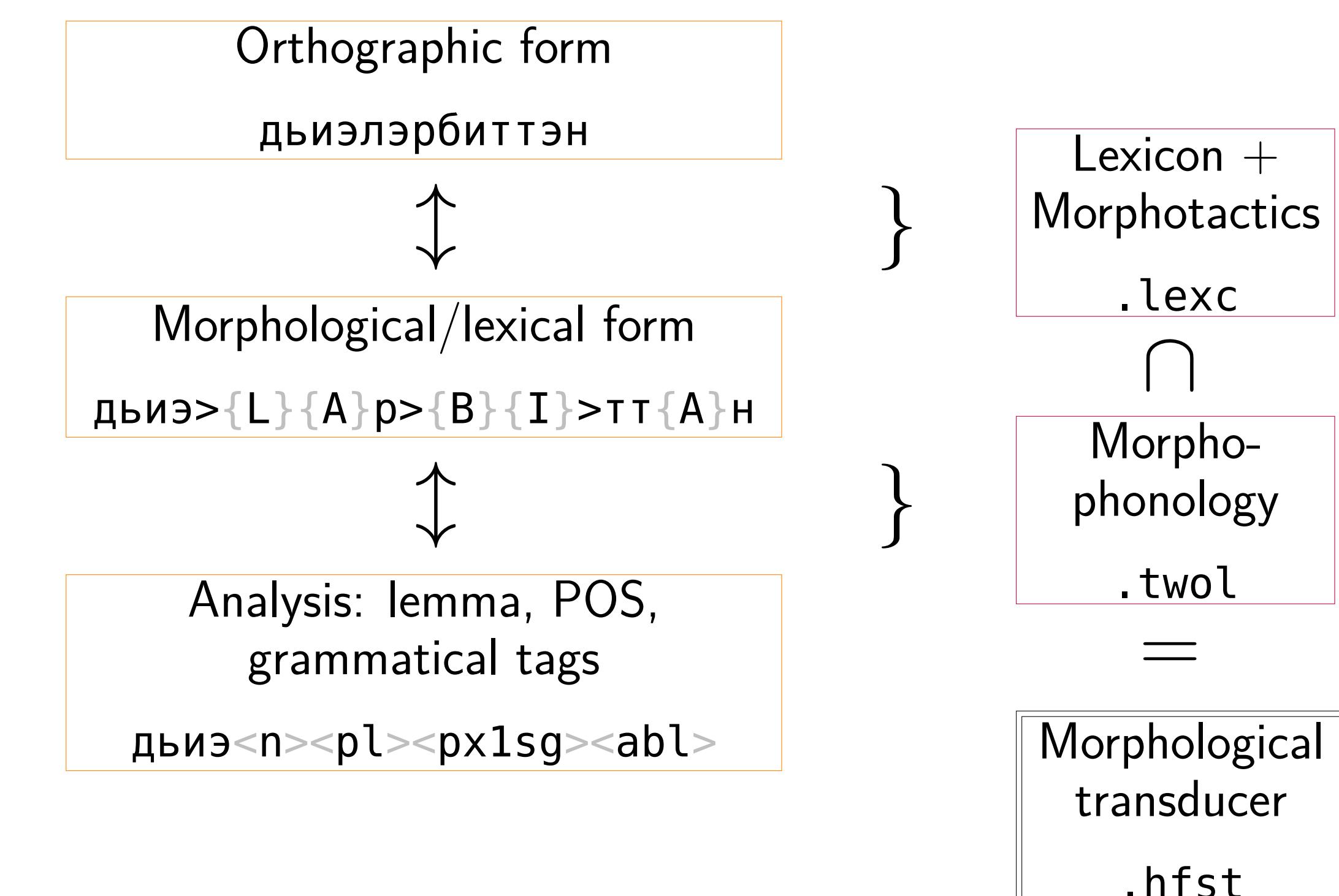
Existing Turkic transducers



The first ever published morphological analyser & generator for Sakha

Implementation

Two-level approach using HFST, entirely hand-coded



Evaluation: naïve coverage

Number of stems that receive an analysis, correct or not

Corpus Tokens Coverage

Corpus	Tokens	Coverage
Newspapers	16M	91.04%
Wikipedia	2.4M	91.30%
New Testament	188K	94.53%

Evaluation: quality

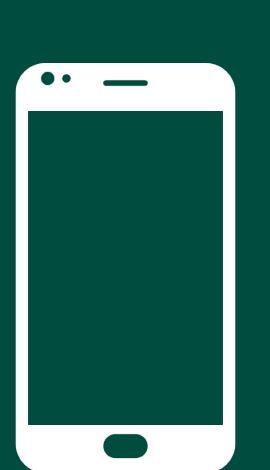
- Selected 1000 random valid wordforms from Wikipedia corpus
- Manually annotated output of transducer over forms
- **Precision:** 98.52% (of provided analyses are correct)
- **Recall:** 75.42% (of correct analyses are provided)

Future work

- Refine morphology and morphophonology
- Morphological and syntactic disambiguation

Conclusion

- Robust transducer, >30k stems
 - high coverage
 - high precision (moderate recall)
- Usable in language technology applications, downstream tasks
- Creation contributed to documentation of Sakha grammar
- Free/Open Source Software (GNU GPL v3)



Scan with mobile camera
to access the code

<https://github.com/apertium/apertium-sah>

