Constructing a Lexical Resource of Russian Derivational Morphology

Lukáš Kyjánek[1] & Olga Lyashevskaya[2] & Anna Nedoluzhko[1]
& Daniil Vodolazsky[3] & Zdeněk Žabokrtský[1]

kyjanek@ufal.mff.cuni.cz

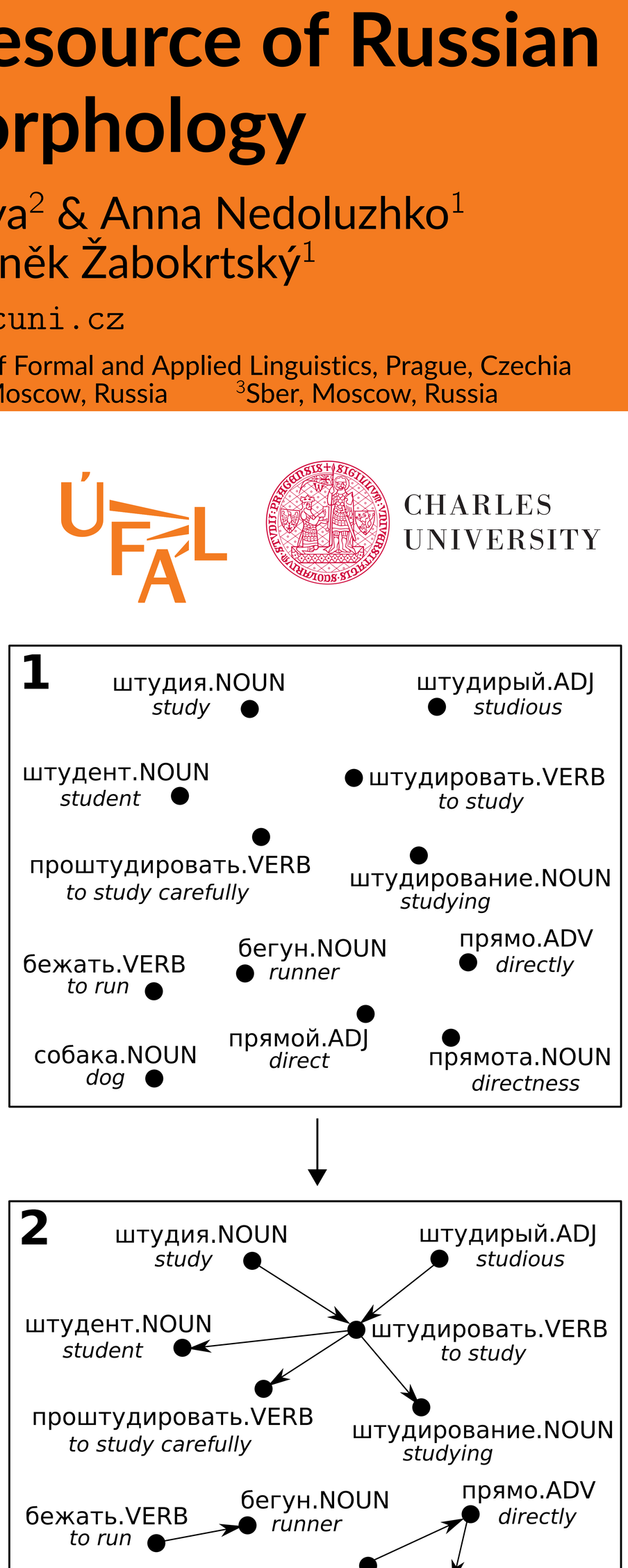
[1]Charles University, Faculty of Mathematics and Physics, Institute of Formal and Applied Linguistics, Prague, Czechia
[2]National Research University Higher School of Economics, Moscow, Russia [3]Sber, Moscow, Russia

Summary

- The verb *exempl-ify* and the noun *example-s* are both based on the word *example*, but the verb is derived, while the noun is inflected. Russian has both of these morphological processes very rich, but there are only a few language resources that capture derivations.
- We improve the existing techniques and create the **largest resource of Russian derivational morphology** dubbed *DeriNet.RU*.
- The resource represents derivations as binary relations, and thus models derivationally-related words in rooted trees like other *DeriNets* (CS, ES, FA) and *Universal Derivations*. It is **freely available**.

How did we create *DeriNet.RU*? In three steps.

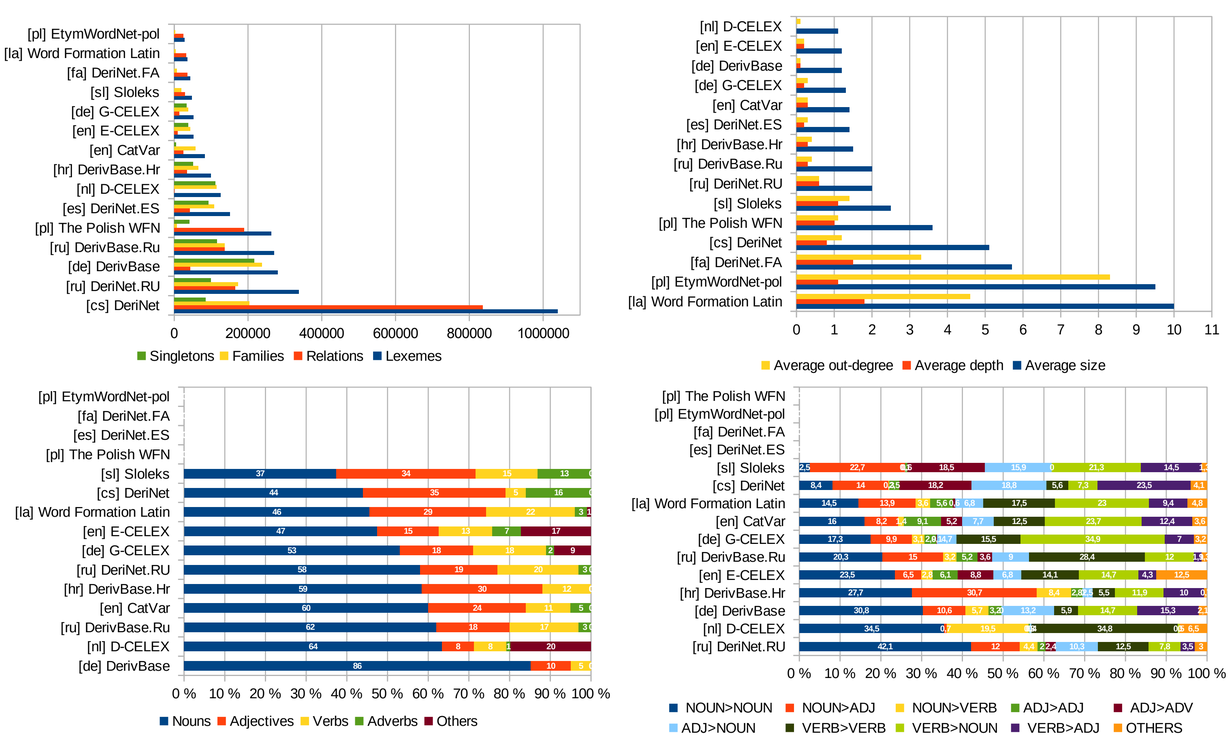
1. Corpus-attested nouns, adjectives, verbs, and adverbs were extracted from Araneum Russicum Corpus Maius.
2. Derivational rules extracted from grammar books were applied to lexemes, organising related lexemes into families. For example:
   - rule887(y + adj + ить → verb), e.g.,
     простой '*prostoj*' (*simple*) → упростить '*uprostit*' (*to simplify*).
3. A machine-learning scorer of relations and Maximum Spanning Tree algorithm restructured families into rooted trees.



Results & Future Work

- Our construction achieved **62.9% accuracy** (baseline 52.6%, maximum oracle score 87.3%); measured on a new data set of manual, parallel annotations by two Russian native speakers.
- *DeriNet.RU* includes more than **300 thousand lexemes** connected with more than **164 thousand derivational relations**.
- It is freely available under **CC BY-NC-SA 4.0**; it was **incorporated into *Universal Derivations*** as one of the largest resources.

Figure: Quantitative comparison between *DeriNet.RU* and the 15 largest resources from the *Universal Derivations* collection.



ARTICLE

Presented at International Conference on Language Resources and Evaluation (LREC), 21–23 Jun 2022. Supported by grants No. GA19-14534S of the Czech Science Foundation and No. START/HUM/010 of Grant schemes at Charles University (reg. No. CZ.02.2.69/0.0/0.0/19_073/0016935), LINDAT/CLARIAH-CZ project of the Ministry of Education (LM2015071, LM2018101), and SVV project No. 260 575. It was using language resources developed, stored, and distributed by the LINDAT/CLARIAH-CZ project. We thank Novosibirsk State University for the computational time on HCI NSU graphic cluster.