# The future of derivational resources

Zdeněk Žabokrtský

📅 October 6, 2023, panel discussion

Charles University
Faculty of Mathematics and Physics
Institute of Formal and Applied Linguistics

# Developments in the last decade

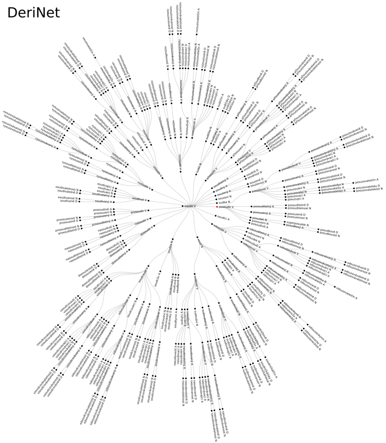
- positive:
  - **many new derivational resources** came into existence, or resulted from digitization of older dictionaries
  - efforts towards **“universal”** multilingual solutions too
    - UniMorph (Kirov et al., 2016), 22 languages with derivational information (out of 182 languages) in 2022
    - Universal Derivations (Kyjánek, 2019), 21 languages
- less positive:
  - sometimes too many (unjustified) **differences** among the data resources
  - most studies still **trapped in a single language** or a small group of languages, even if data for many more languages exist
  - **no more motivation** coming **from** the **NLP** industry
    - end-to-end deep-learning solutions dominate nowadays
    - their own notion of “morphology”: mechanically induced subwords

# Diversity as a limiting factor

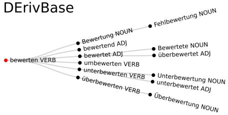
- sure, there are phenomena that are really language specific
- **but most issues** (and perhaps the hardest ones) are **shared**:
  - homography, allomorphy
  - **fuzzy boundaries** between derivation/inflection/compounding
  - fuzzy boundary between morphology and syntax
  - problems induced by diachrony (etymology vs. analogy) ...
  - **different "data structures"** too: tree-based vs. paradigm-based representation of derivational relations
  - and also technical questions, such as how to choose the inventory of lexical units for a resource ...
- even after technical harmonization, many **differences remain**...

# A glimpse at Universal Derivations, v. 1.0

DeriNet

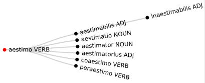


DErivBase



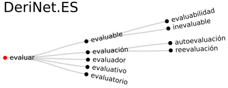
bewerten VERB
Bewertung NOUN
bewertend ADJ
bewertet ADJ
umbewerten VERB
unterbewerten VERB
überbewerten VERB
Fehlbewertung NOUN
Bewertete NOUN
überbewertet ADJ
Unterbewertung NOUN
unterbewertet ADJ
Überbewertung NOUN

Word Formation Latin



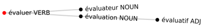
aestimo VERB
aestimabilis ADJ
aestimatio NOUN
aestimator NOUN
aestimatorius ADJ
coaestimo VERB
peraestimo VERB
inaestimabilis ADJ

DeriNet.ES



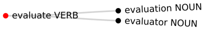
evaluar
evaluable
evaluación
evaluador
evaluativo
evaluatorio
evaluabilidad
inevaluable
autoevaluación
reevaluación

Démonette



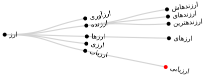
évaluer VERB
évaluateur NOUN
évaluation NOUN
évaluatif ADJ

English WordNet



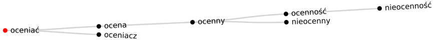
evaluate VERB
evaluation NOUN
evaluator NOUN

DeriNet.FA



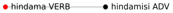
ارز
ارزآوری
ارزنده
ارزها
ارزی
ارزیاب
ارزندهاش
ارزندهای
ارزندهترین
ارزهای
ارزیابی

The Polish Word-Formation Network



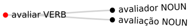
oceniać
ocena
oceniacz
ocenny
ocenność
nieocenny
nieocenność

EstWordNet



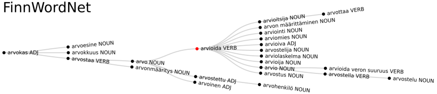
hindama VERB
hindamisi ADV

NomLex-PT



avaliar VERB
avaliador NOUN
avaliação NOUN

FinnWordNet



arvokas ADJ
arvoesine NOUN
arvokkuus NOUN
arvostaa VERB
arvo NOUN
arvonmääritys NOUN
arvioida VERB
arvostettu ADJ
arvoinen ADJ
arvioitsija NOUN
arvon määrittäminen NOUN
arviointi NOUN
arviomies NOUN
arvioiva ADJ
arvostelija NOUN
arviolaskelma NOUN
arvioija NOUN
arvio NOUN
arvostus NOUN
arvohenkilö NOUN
arvottaa VERB
arvioida veron suuruus VERB
arvostella VERB
arvostelu NOUN

# A turn to morphological segmentation?

- opinion: maybe some of the hard questions would become less problematic if our **primary representation** of a word form is a **sequence of morphs**
- segmentation vs. derivation+compounding+inflection – two sides of the same coin (almost)
    - some notions such as lemmatization are entrenched in our **linguistic traditions**, but may seem arbitrary from other languages' perspectives; e.g. dictionary organization by root morphs could be less problematic than by lemmas
- speculating wildly: morphs/subwords as a possible **meeting point** between NLP and linguistics?
- UFAL+UniCatt: an attempt at **Universal Segmentations guidelines** planned soon

# DeriMo and Sigmorphon events

- Almost **disjoint communities**. Why?
- A lesson from Sigmorphon: a strong **shared task** tradition

# Conclusion

My stand: it would be great to

- try to apply **multilingual perspectives** whenever possible
- at the representation level, try to **suppress traditional “fundamental oppositions”** (such as the one between inflection and word formation), esp. for the benefit of easier applicability to more languages
- put **more focus on morphological segmentation**
- organize a (multilingual) **shared task** at some further DeriMo edition