Merging Data Resources for Inflectional and Derivational Morphology in Czech

Zdeněk Žabokrtský, Magda Ševčíková, Milan Straka,
Jonáš Vidra, Adéla Limburská

Charles University in Prague
Institute of Formal and Applied Linguistics

LREC, 25th May 2016, Portorož

Outline

- Motivation for processing inflection and derivation together
- Inflectional and derivation resources for Czech
- The resulting (merged) data resource
- User interfaces to the data
- Conclusions

Basic notions

- morphological inflection: *to derive* $\rightarrow$ *derives*, *derived*, *deriving*
- morphological derivation: *to derive* $\rightarrow$ *derivative*, *derivation*, *derivator*

Motivation

- an omnipresent problem of NLP: zillions of different words
- one of the reasons: morphological variation
- standards ways to reduce the lexical space:
  - lemmatization – replacing inflectionally related words by a selected representative
  - stemming – replacing related words by a common stem (usually approximated very roughly)

Motivation, cont.

- in morphologically complex languages:
  - possibly several tens (or more) inflected word forms per lemma
  - but possibly several tens (or more) derived lemmas too!
- a common-sense expectation: extending lemmatization (as *anti-inflection*) with nesting (as *anti-derivation*) might help NLP apps
- in Czech, derivation is the most productive word formation method (hundreds of suffixes)
- surprisingly few data resources for derivation (e.g., Derivancze for Czech, DerivBase for German, Démonette for French)

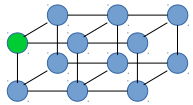
Derivation vs. inflection: similarities

For both it holds that

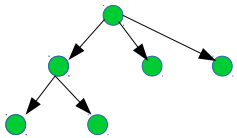
- there is a strong **form-function** asymetry, e.g.
  - there are several suffixes that express the same meaning (e.g. an actor)
  - one specific suffix can express several roles
- the way how forms are combined is **far from simple catenation**
  - consonant and vowel changes (not limited to morpheme boundaries, can appear inside roots too)
  - sometimes similar changes for inflection and derivation: *sníh* - *sn***ě***hu* (inflection: snow gen.sg.), *sníh* - *sn***ě***žný* (derivation: snowy adj.)
- **fuzzy boundaries** of parole
  - exhaustive enumeration of all potentially inflected/derived forms often reaches language periphery

Derivation vs. inflection: differences

- different data structure
  - a set of words connected by **inflection**:
    - ★ typically a full **Cartesian product** of morphological categories



  - a set of lemmas connected by **derivation**:
    - ★ rather an **oriented graph** (a nest), a rooted tree is often enough



- in inflection, the paradigm representative is chosen by a convention, while in derivation, the tree root seems more tangible
- semantic relatedness gradully weakens for more distant words in a derivation nest
- in NLP, lemmatization is widely used while nesting is not

MorfFlex CZ

- Czech morphological dictionary
- developed originally by Jan Hajič as a spelling checker and lemmatizer
- more than two decades of improvements
- 985 thousand unique lemmas with their inflectional paradigms
- associated with a positional tagset
- capable of analyzing/generating 120 million word forms (form-lemma-tag tripples)
- used *inter alia* in the Prague Dependency Treebank and Czech National Corpus

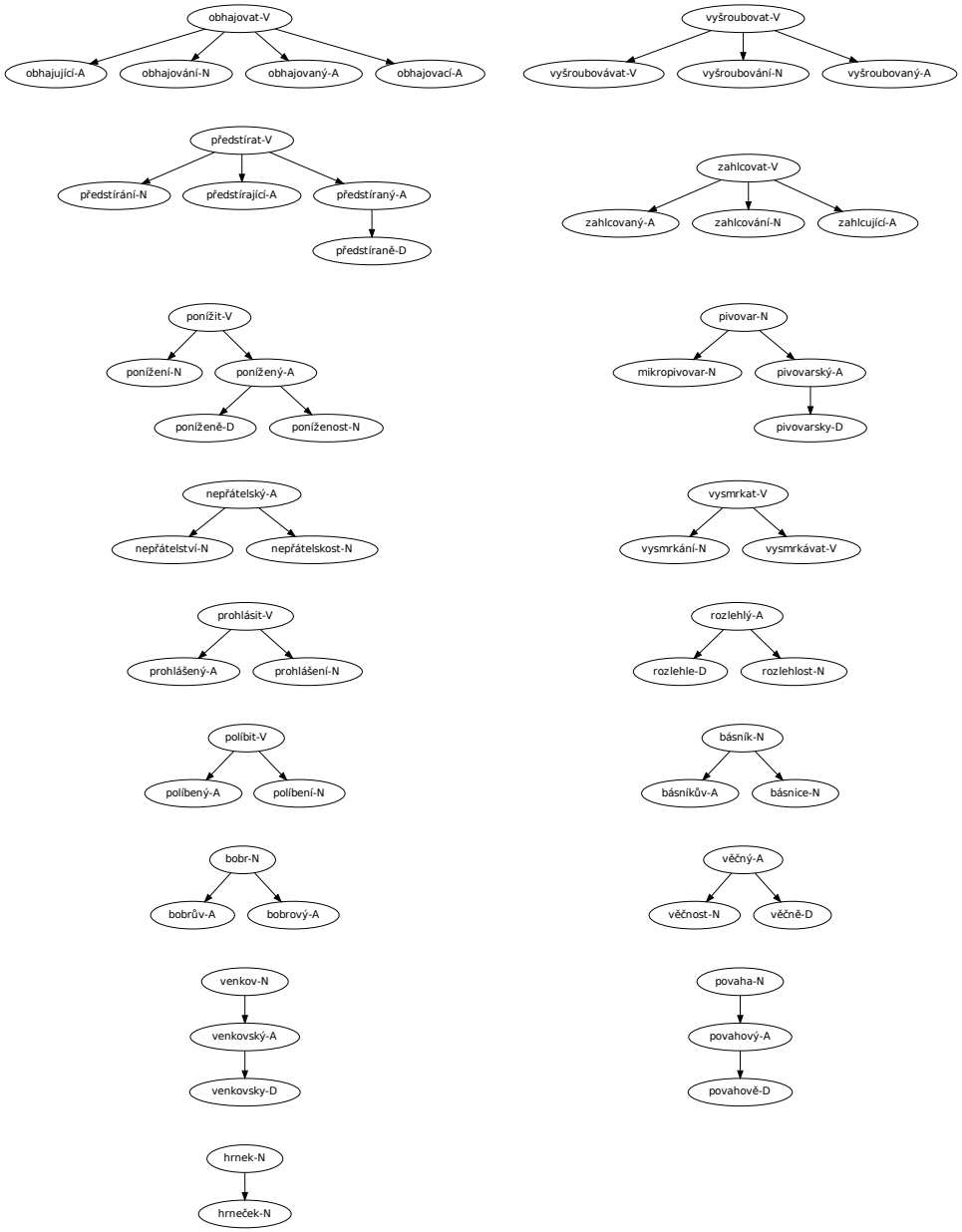
A glimpse at the MorfFlex CZ data

```
podle-1_^(*3ý-1)  Dg-------3N---6  nejnepodlejc
podle-1_^(*3ý-1)  Dg-------3N----  nejnepodleji
podle-1_^(*3ý-1)  Dg-------3A---6  nejpodlejc
podle-1_^(*3ý-1)  Dg-------3A----  nejpodleji
podle-1_^(*3ý-1)  Dg-------1N----  nepodle
podle-1_^(*3ý-1)  Dg-------2N---6  nepodlejc
podle-1_^(*3ý-1)  Dg-------2N----  nepodleji
podle-1_^(*3ý-1)  Dg-------1A----  podle
podle-1_^(*3ý-1)  Dg-------2A---6  podlejc
podle-1_^(*3ý-1)  Dg-------2A----  podleji
podle-2           RR--2----------  podle
```

DeriNet

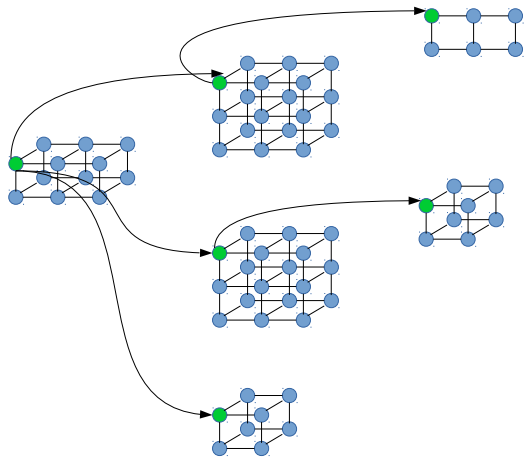
- a network capturing derivation in Czech, developed since 2013
- oriented graph (forest, each rooted tree = one derivational nest)
  - nodes = lemmas
  - edges = derivation relations (from base to derived lemmas)
- size before merging with MorfFlex CZ
  - 306 thousand nodes (chosen according to frequency in the Czech National Corpus)
  - 117 thousand edges
- compiled using semi-automatic procedure, based especially on
  - suffix substitution rules (extracted both from grammar books and from data)
  - manually assembled lists of exceptions
  - patterns for vowel and consonants changes

A glimpse at the DeriNet data



Merging process

- set of lemmas of the previous DeriNet version extended to that of MorfFlex CZ
- the pipeline for building DeriNet re-executed on the new lemma set
- only minor modifications of substitution rules and exception lists needed
- resulting data: 970 thousand lemmas connected with 715 thousand derivational relations

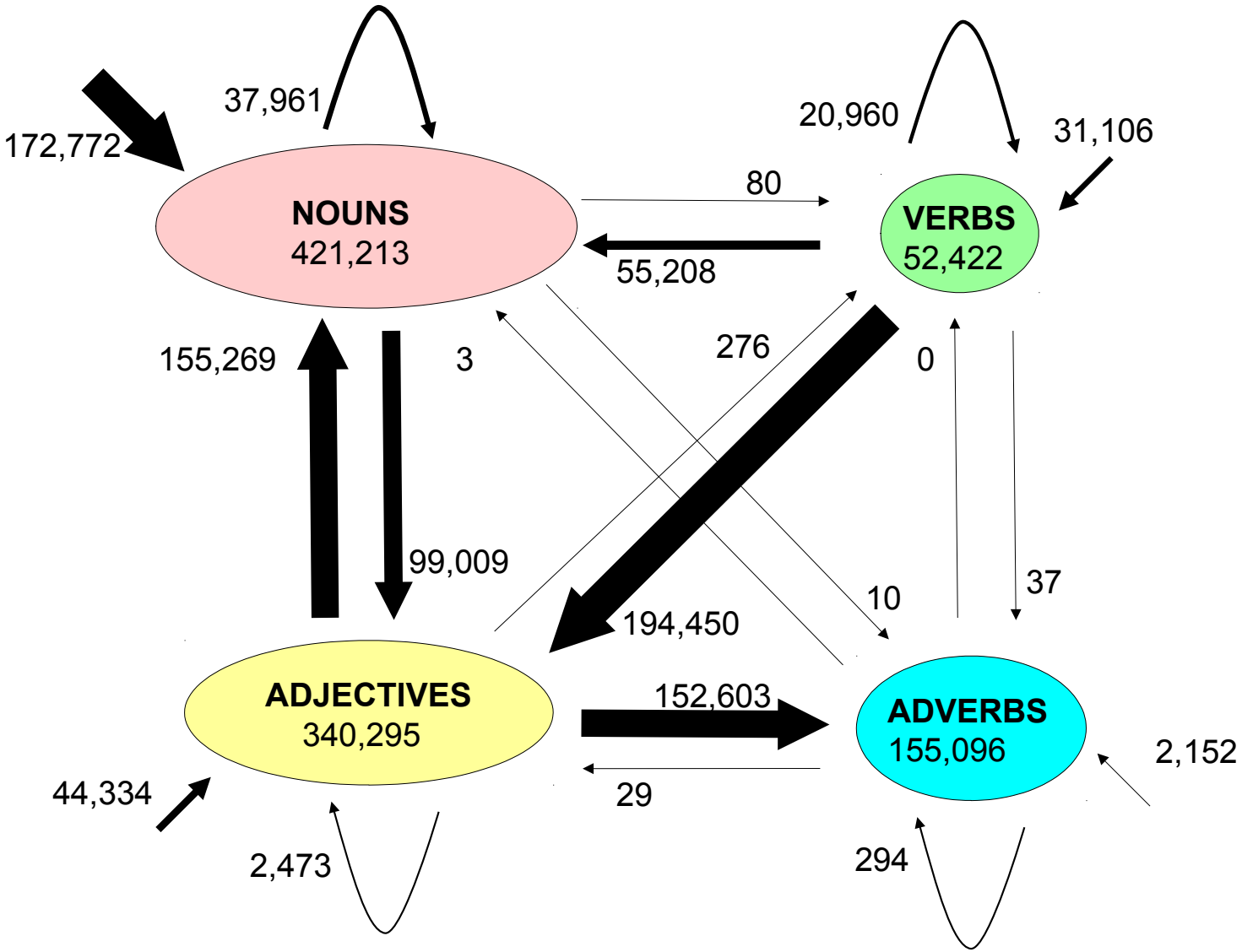


Extension of the derivation forest

after merging DeriNet with MorfFlex CZ

- in the derivational forest
  - #nodes increased 3.2 times
  - #edges increased 6.1 times
- evaluation (based on a manually annotated sample) shows that
  - precision of derivations stayed at 99 %
  - recall increased from 75 % to 85 %
- we attribute both observations to language economy:
  - lower-frequency words tend to be derived more frequently...
  - ...and they tend to be derived in a more regular way

POS and POS→POS counts in the merged data

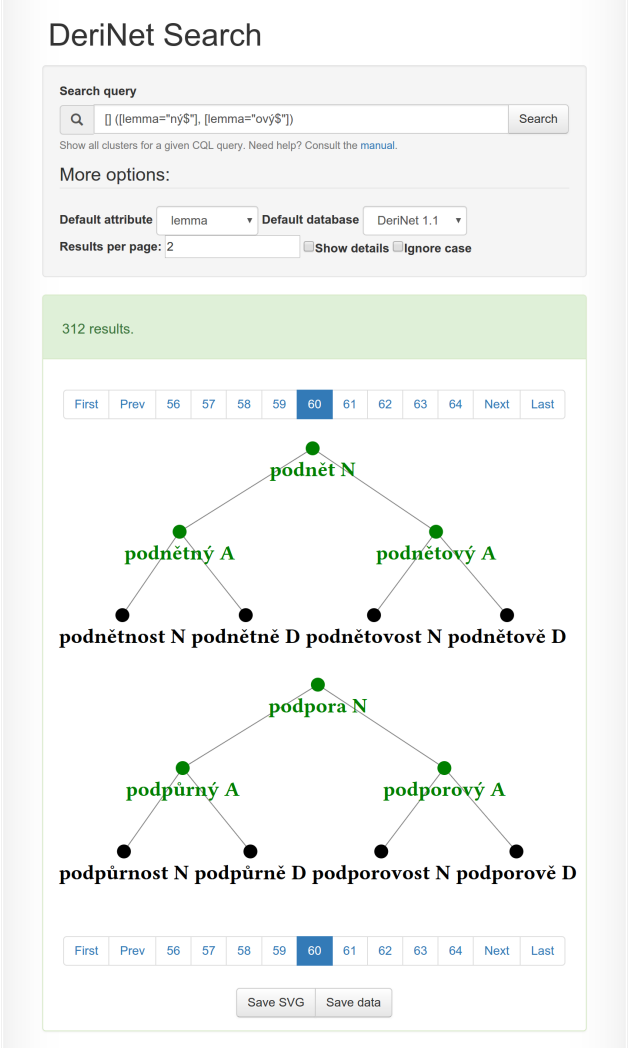


Access to the data

- Application Programming Interfaces
  - derivations integrated in the MorphoDiTa tool since version 2.0
  - REST API
- Graphical User Interfaces (in web browsers)
  - MorphoDiTa online demo - shows both derivations and inflections
  - DeriNet Viewer - for browsing derivation trees
  - DeriNet Search - query language allowing quite complex search queries

Query example

- The query `[] ([lemma="ný$"], [lemma="ový$"])` searches for adjectives which were derived by the two different suffixes.



Future work and open questions

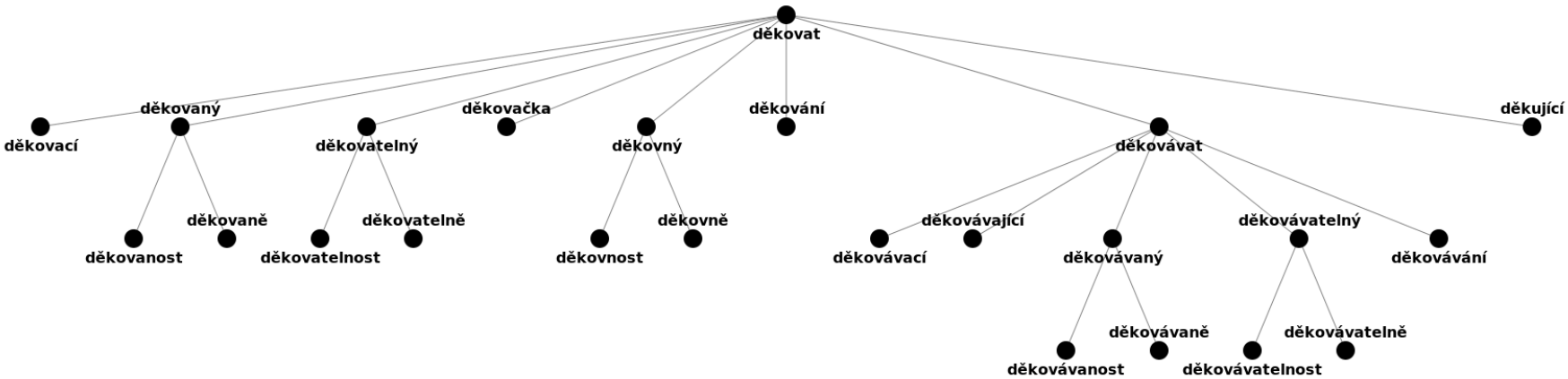
- add some **missing derivations** (e.g. verb prefixation, aspectual counterparts created by suffixation, etc.)
- **abandon the treeness** constraint to allow composition
- **semantic labelling** of derivation relations (diminutives, possessives. . .)
- resolve **homonymy** – inflection and derivation might pose different criteria on distingushing homonyms
- some problems **analogous to** that of **dependency trees**
  - clear presence of an edge, but unclear orientation
  - sometimes intermediate words are "predicted" that simply do not exist (phantom lexemes, similar to elipsis)
  - we know trees are actually not enough even for derivations, but are irresistibly attractive

Conclusions

There is a morphological resource for Czech that

- handles both morphological inflection and derivation
- covers roughly one million Czech lemmas
- is equiped with several user interfaces
- is available to your under CC-BY-NC-SA, see `http://ufal.mff.cuni.cz/derinet` or `http://ufal.mff.cuni.cz/morphodita`

Thank you!


děkovat
děkovací
děkovaný
děkovatelný
děkovačka
děkovný
děkování
děkující
děkovávat
děkovanost
děkovaně
děkovatelnost
děkovatelně
děkovnost
děkovně
děkovávající
děkovávací
děkovávaný
děkovávatelný
děkovávání
děkovávaně
děkovávatelně
děkovávanost
děkovávatelnost