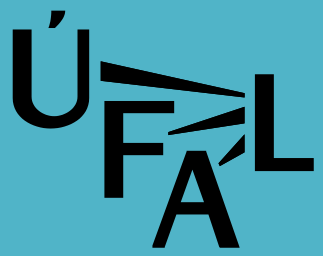
A data-based approach to competition in word-formation: diminutives and gender marking across seven languages

Morphological meeting at Laboratoire de linguistique formelle
Paris 2022

Lukáš Kyjánek
KYJANEK@UFAL.MFF.CUNI.CZ
https://lukyjanek.github.io/

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



This work was supported by the Grant No. START/HUM/010 of Grant schemes at Charles University (reg. No. CZ.02.2.69/0.0/0.0/19_073/0016935).

START Grant

- 03/2021 – 03/2022

- Morphological research into competition in Germanic, Romance, and Slavic langs.
  - **G**: Dutch English German | **R**: French Spanish | **S**: Czech Russian

- Mgr. Magda Ševčíková, PhD. (mentor)
- Mgr. Lukáš Kyjánek (PI) : semantics in derivational morphology, language resources
- Mgr. Jan Bodnár : morphological segmentation
- Mgr. Emil Svoboda : compounding
- Mgr. Jonáš Vidra : linguistic transfer methods, language resources

Outline

1. Basic notions
2. Data Resources
   - DeriNet
   - Universal Derivations
   - DeriNet.RU
   - Universal Segmentations
3. Methodology
   - Searching for spelling variants (in Czech)
   - Labelling derivational meanings (in Czech)
   - Analysing agent noun formation (in Czech)
   - Transfering word-formation networks (from Czech)
4. Ongoing work
   - Analysis of gender marking formation (in Czech)
   - Comparison of diminutiveness and gender marking across languages

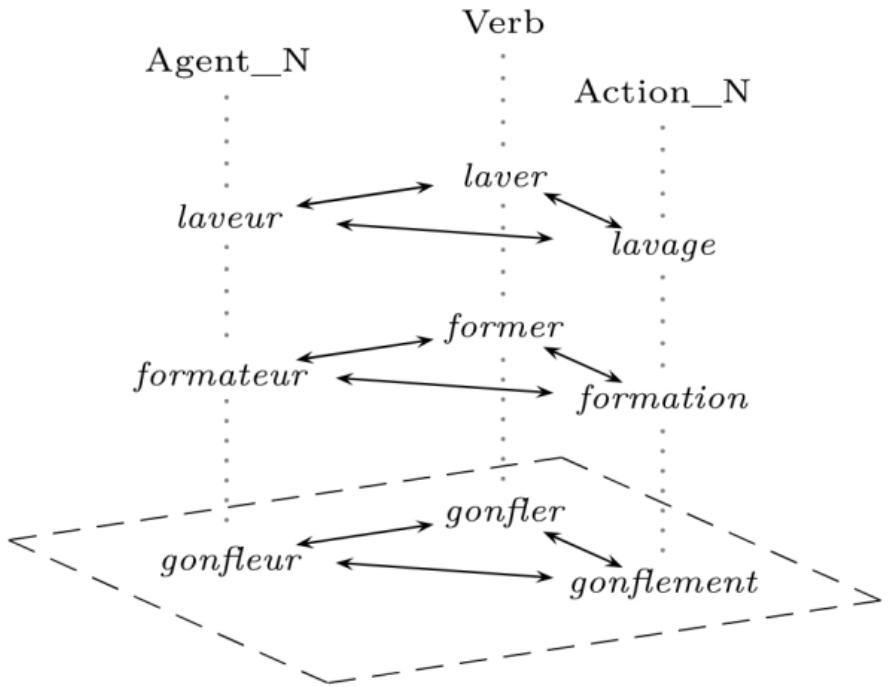
Basic notions

Approaches to derivational morphology

Körtvélyessy et al. (2020:10-11)

1. Direct derivatives (paradigm)
   dom → *dom-ov*
   → *dom-ček*
   → *dom-ík*
   → *dom-isko*
2. Subsequent derivatives (series)
   dom → *dom-ov* → *dom-ov-ina* → *dom-ov-in-ový*
   dom → *dom-ček* → *dom-ček-ový*
   dom → *dom-ík* → *dom-ík-ový*
   dom → *dom-isko* → *dom-isk-ový*
3. Semantic categories of each derivational step
   agent, female, location, quality, agmentative, etc.
4. Derivational network
   = derivatives derived from a simple underived word
   (combination of (1) and (2) and (3))

Bonami and Strnadová (2019:172)



Derivational meaning

- *odesílat* $\xrightarrow[\text{agent}]{}$ *odesíla-tel* *(to send > sender)*
  - *odesílat* = activity
  - *odesílatel* = someone who does the activity

- One affix can convey many meanings
  - *úředník* $\xrightarrow[\text{female}]{}$ *úředn-ice* *(officer > female officer)*
  - *věznit* $\xrightarrow[\text{location}]{}$ *vězn-ice* *(to imprison > jail)*
  - *kytka* $\xrightarrow[\text{augmentative}]{}$ *kyt-ice* *(flower > bouquet)*

- One meaning can be conveyed by many affixes
  - *úředník* $\xrightarrow[\text{female}]{}$ *úředn-ice* *(officer > female officer)*
  - *šéf* $\xrightarrow[\text{female}]{}$ *šéf-ová* *(boss > female boss)*
  - *učitel* $\xrightarrow[\text{female}]{}$ *učitel-ka* *(teacher > female teacher)*
  - *ministr* $\xrightarrow[\text{female}]{}$ *ministr-yně* *(minister > female minister)*

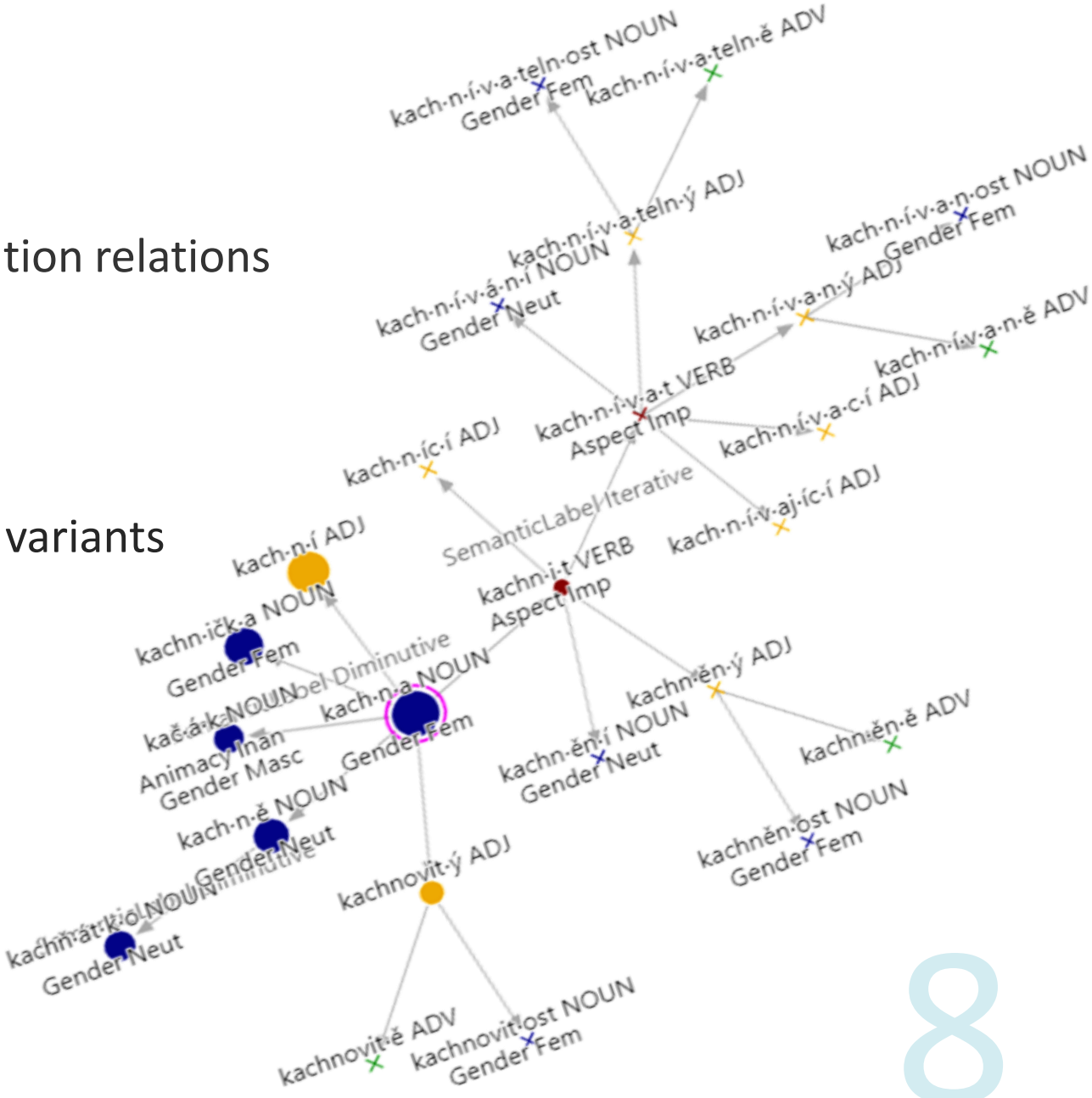
Data Resources

DeriNet

- Lexical network which models word-formation relations in the lexicon of Czech

- Over 1 milion lexemes
  - 782 thousand derivational relations
  - 50 thousand links for orthographic variants
  - 1,952 links for compounding
  - 144 relations of conversion
  - 295 relations of univerbisation

- http://www.ufal.cz/derinet



Universal Derivations

- Collection of harmonized lexical networks capturing word-formation, especially derivation, in a cross-linguistically consistent annotation scheme for many languages (UDer 1.1 contains 31 harmonized resources covering 21 languages)
- http://www.ufal.cz/universal-derivations
- Harmonisation process:
  - Assembling the existing resources
  - Scoring derivational relations
  - Finding maximum spanning tree

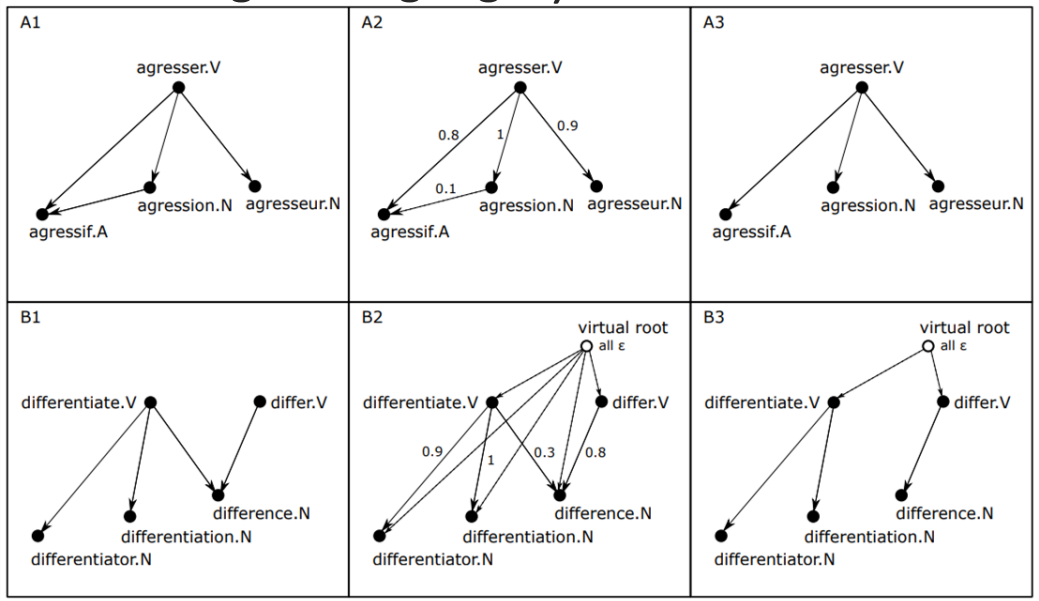
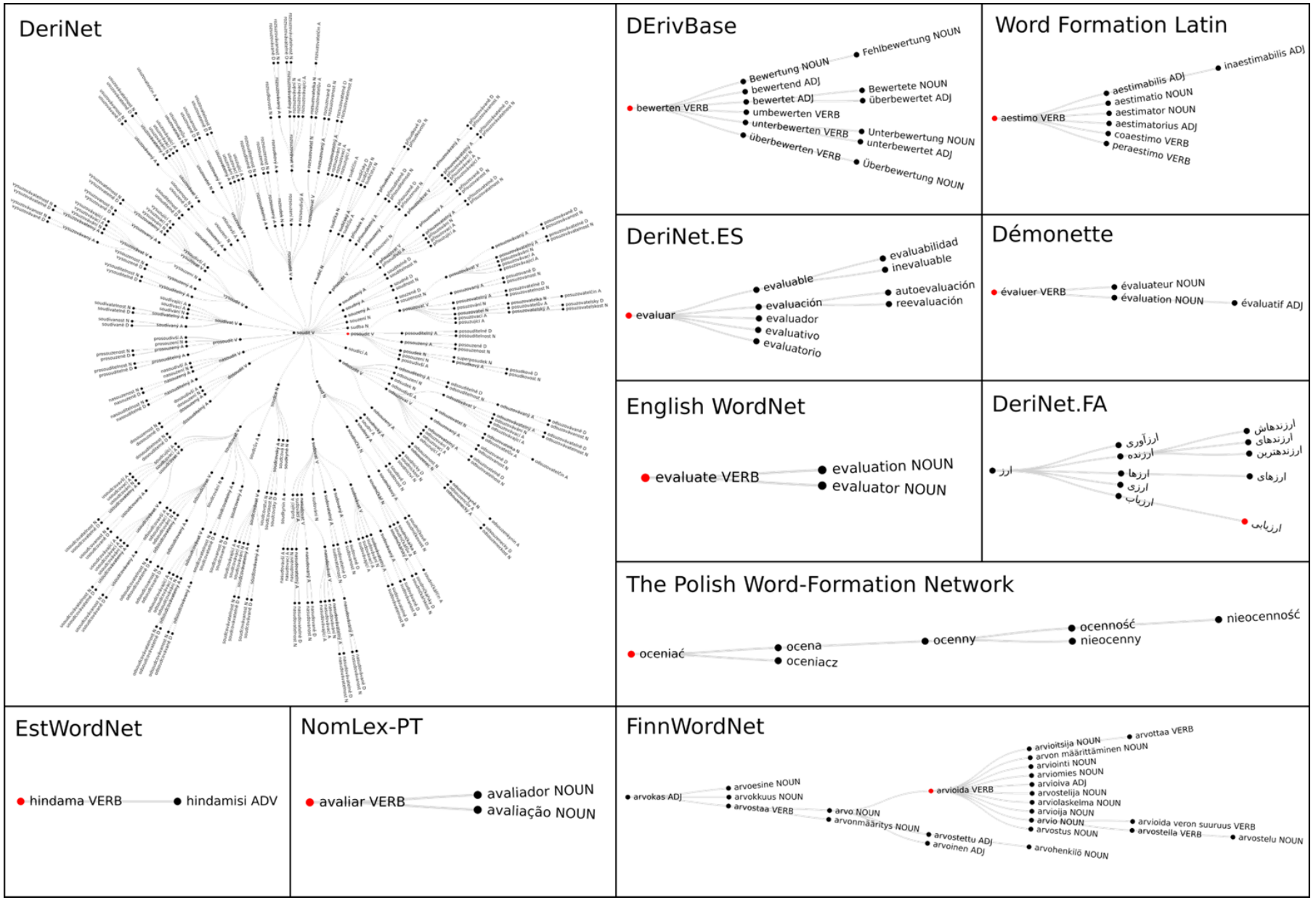


Figure 3.6: Illustration of identifying rooted trees by maximising a sum of scores. While just one tree is obtainable from family A (The Morpho-Semantic Database), family B (Démonette) has to be divided. The virtual root prevents failing Maximum Spanning Tree algorithm, and provides smoothing based on the value of ε.

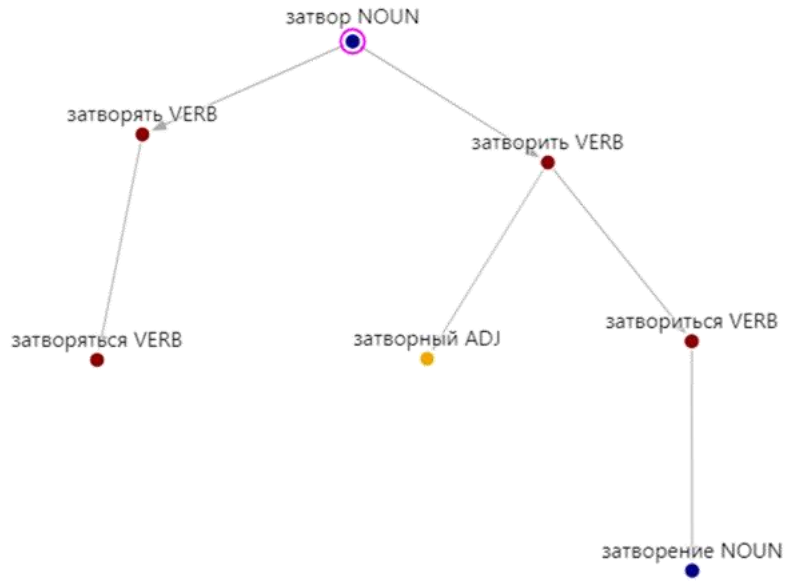


DeriSearch v2: http://www.ufal.cz/derisearch

Lukáš Kyjánek & Olga Lyashevskaya & Anna Nedoluzhko & Daniil Vodolazsky & Zdeněk Žabokrtský 2022
(submitted to LREC 2022, in January); data included to Universal Derivations

DeriNet.RU

- Lexical network which models word-formation relations in the lexicon of Russian
- Over 337 thousand lexemes connected by more than 164 thousand derivational relations into 172 thousand derivational families
- Created on the basis of:
  - Grammar-based model of derivational rules from Russian grammar books, e.g.,
    rule343(noun + ucm > noun)
    анархия [anarchy] > анархист [anarchist]
  - Harmonisation procedure (improved)



Zdeněk Žabokrtský & Niyati Bafna & Jan Bodnár & Lukáš Kyjánek & Emil Svoboda & Magda Ševčíková & Jonáš Vidra 2022
(submitted to LREC 2022, in January; data will be published in LINDAT/CLARIAH-CZ, in January)

Universal Segmentations

- Collection of lexical resources capturing morphological segmentations harmonised into a cross-linguistically consistent annotation scheme for many languages
(17 harmonized resources providing 48 data sets covering 37 languages)
- http://www.ufal. cz/universal-segmentations

	Resource	Original format	→	UniSegments format
Ex. 1	Démonette	"abaissement","tlfnome","abaisser","tlfnome","Ncms","tlfnome", "Vmn----","tlfnome", "simple","derif","suf","ment","derif",,,, "\@RES","demonette","\@","demonette", "résultat de abaisser", "derif","résultat de \@","demonette","descendant", "demonette", "abaiss","derif",,,"derif"	→	abaiss + e + ment (*lowering*)
Ex. 2	DerIvaTario	3951;ABBATTIMENTO;BATTERE:vrb_th; ACons:ad:mt2:ms2b;MENTO:mento:mt4:ms1;;;;	→	ab + batt + i + mento (*breakdown*)
Ex. 3	DerivBase.Ru	вымор noun повыморить verb rule887(по + noun + и1(ть) -> verb) PFX,SFX	→	по + вымори + ть (*become extinct*)
Ex. 4	MorphoLex	rafraîchissant [VB]>>sant>	→	r + a + fraîchis + sant (*refreshing*)
Ex. 5	Word Formation Latin	(23891,'malaxo','V1','','VmF','m0158','malaxo', 'VERB',NULL,'B') (23890,'malaxatio','N3B','f','NcC','m0157', 'malaxatio','NOUN',NULL,'B') (23891,1,23890,'86','a','2016-03-29 12:45:48') ('V-To-N','Derivation_Suffix','86','','n6p1; n2np; Regular PP: v1*; v2*; v3*; v4*; v5*; v6*','','(t)io(n)', 'n31','abiurat-io, -ion-is; abstrus-io, -ion-is')	→	malax + a + tio (*comminution*)

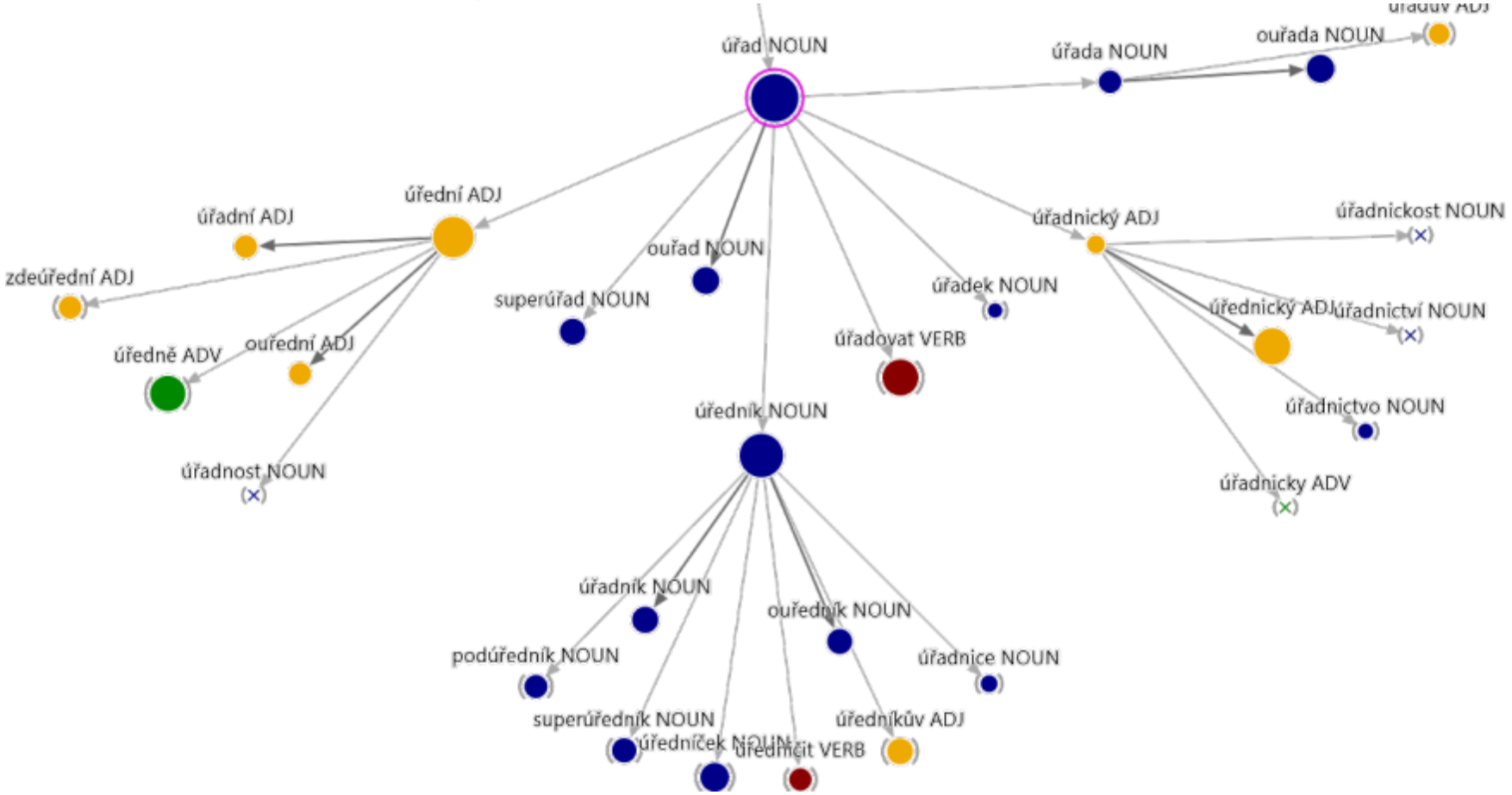
Methodology

Lukáš Kyjánek & Magda Ševčíková 2021
(included in DeriNet)

Searching for spelling variants

Examples:

- ***úř**ad, **ou**řad*
- *předhřát, pře**de**hřát*
- *ohra**ž**ování, ohra**z**ování*
- *jakkoli**v**, jakkoli*
- *doping**o**vý, dopin**k**ový*
- *b**ý**k, b**ej**k*
- *Tcha**iv**an, Tcha**jw**an*
- *odbydl**et**, odbydl**it***
- ***troj**nožka, **tří**nožka*
- ***ž**ebřina, **ř**ebřina*
- *berl**a**, berl**e***
- (?) *bezkolej**ný**, bezkolej**ový***
- (?) *bezhlesý, bezhles**ný***
- (?) *bleďou**č**ký, bleďou**n**ký*
- (?) *bočn**ý**, bočn**í***
- (?) *drobí**n**ek, drobí**t**ek*
- (?) *rozechvě**n**ý, rozechvě**l**ý*



Magda Ševčíková & Lukáš Kyjánek 2021 (Journal of Linguistics; also included in DeriNet)
https://www.juls.savba.sk/ediela/jc/2019/2/jc19-02.pdf#page=283

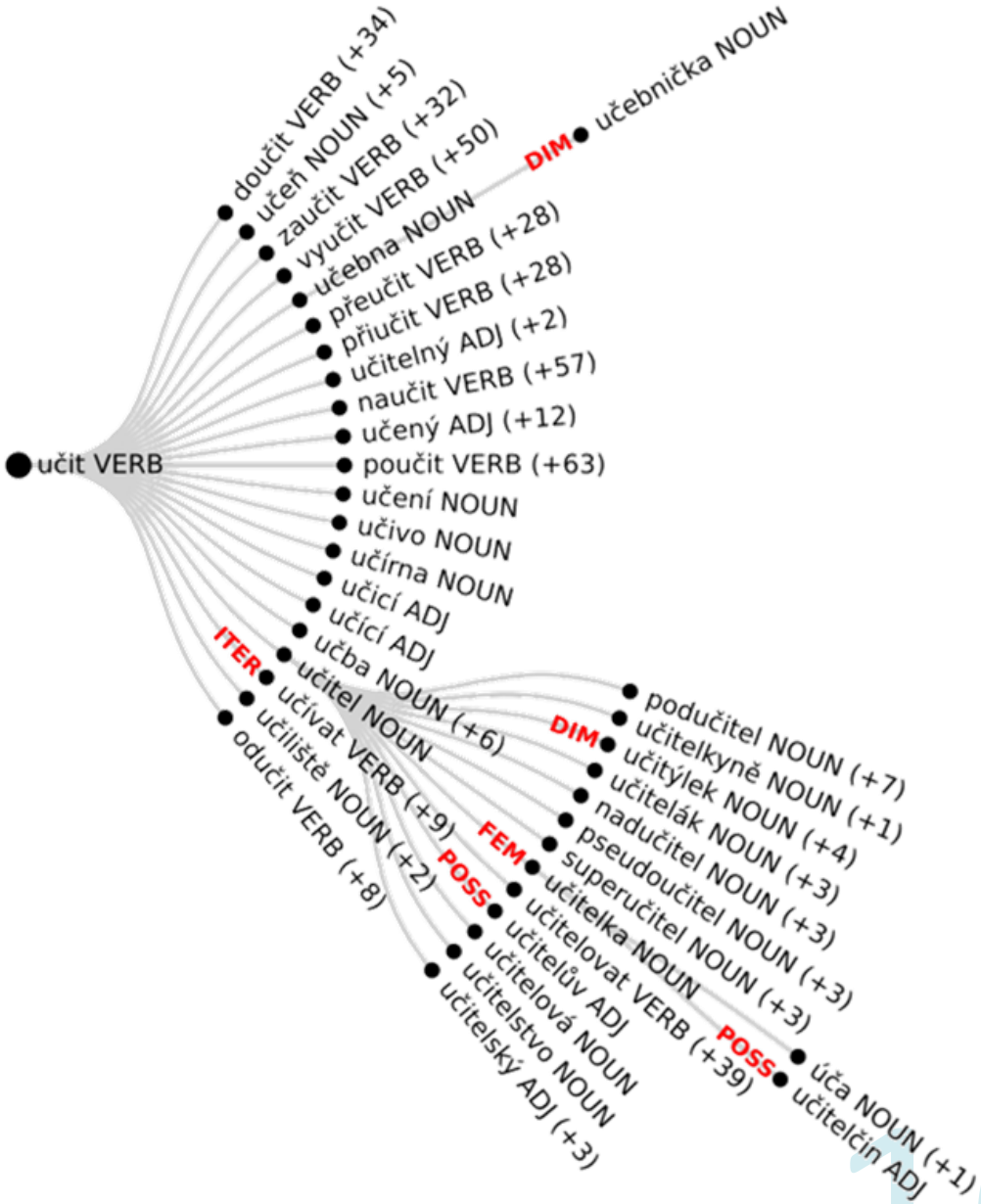
Labelling derivational meanings

- Pilot experiment: to add 5 labels limited to suffixation into DeriNet for Czech
  - *pes* $\xrightarrow[\text{diminutive}]{}$ *psík* *(dog > small dog)*
  - *učitel* $\xrightarrow[\text{female}]{}$ *učitelka* *(teacher > female teacher)*
  - *učitel* $\xrightarrow[\text{possessive}]{}$ *učitelův* *(teacher > teacher's)*
  - *chodit* $\xrightarrow[\text{iterative}]{}$ *chodívat* *(to walk (IPFV) > to walk repeatedly (IPFV))*
  - *obalit* $\xrightarrow[\text{aspect}]{}$ *obalovat* *(to wrap (PFV) > to wrap (IPFV))*
- Input data: 14,752 semantically labelled base-derivative pairs from SSJČ (Havránek 1960-1971), MorfFlexCZ (Hajič and Hlaváčová 2013), VALLEX 3.0 (Lopatková et al. 2016), and PMČ (Nekula et al. 2012); each label around 2.5 thousand pairs
- Features: part-of-speech categories, genders, aspects, possessivity tags, final character n-grams (2-6)

- Task: to classify the most probable semantic label
- Method: Multinomial Logistic Regression with newton-cg solver
- F1-score = 98.4%

Label	Derivations
Diminutive	5,383
Female	28,623
Possessive	87,087
Iterative	11,778
Aspect	15,186

- Already available since DeriNet 2.0



Analysing agent noun formation

- 8 top-frequent suffixes forming agent nouns (SYN2015); manually created data
- Data set divided into training, evaluation, and hold-out subsets
- Settings of hyper-parameters of Logistic regression were obtained from the first experiment on dataset containig all features
- Other experiments used 5 different subsets of features, but the same settings

target_noun	***viník***	**target_noun_suffix**	***-ník/-ík***
base_number_syllables	*1*	paradigm_type	*NNA-V-*
base_number_prefixes	*0*	freq_target_noun	*1188*
base_shared_theme	*x*	freq_parent_noun	*6758*
base_ending	*n*	freq_parent_adj	*2274*
base_ending_cvs	*consonant*	freq_parent_oth	–
base_ending_vertical	*nasal*	freq_parent_v1	*689*
base_ending_horizontal	*alveolar*	freq_parent_v2	–
parent_noun	*vina*	freq_slots	*VxAN*
parent_adj	*vinný*	v1_theme	*i*
parent_oth	–	v1_aspect	*imp*
parent_v1	*vinit*	v1_conjug	*4*
parent_v2	–	v2_theme	–
inanim_noun	*no*	v2_aspect	–
v1_suf_asp_counterpart	*no*	v2_conjug	–

Table: Absolute numbers of individual agent suffixes in our data set.

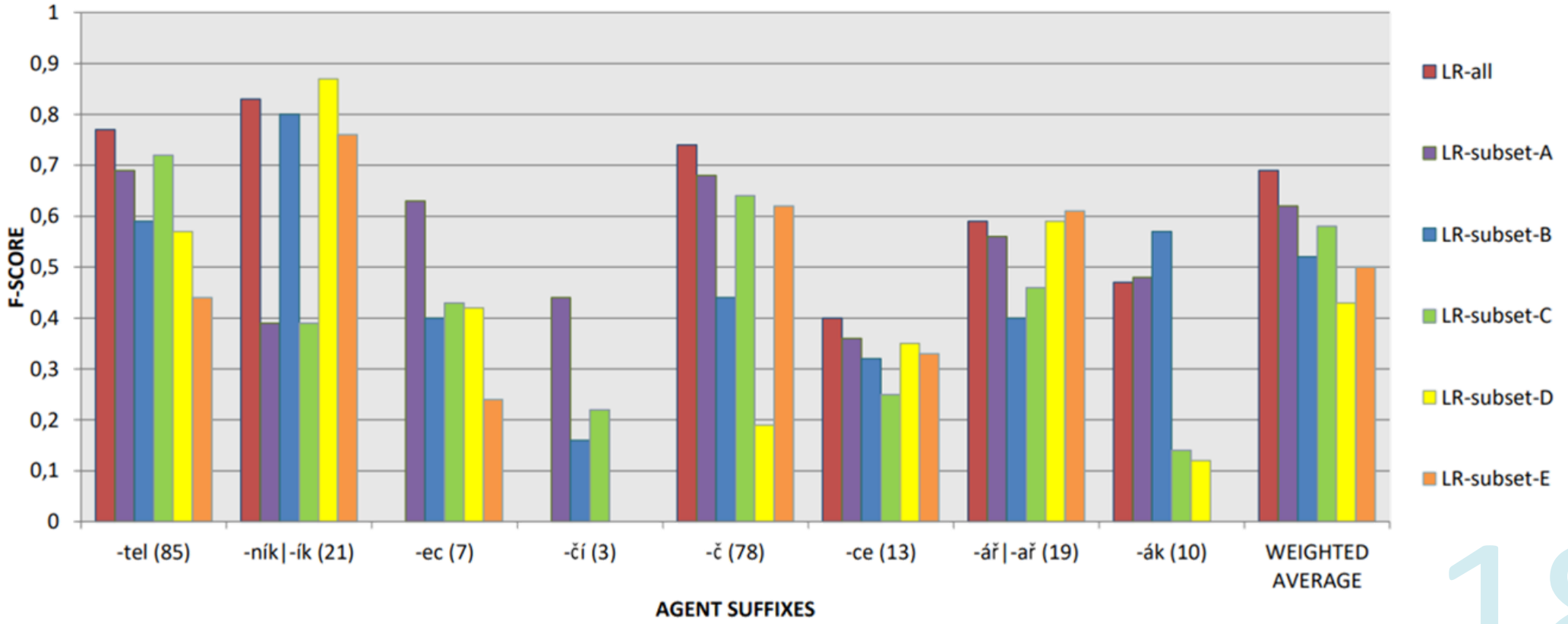
Suffix	*-tel*	*-č*	*-ník/-ík*	*-ář/-ař*	*-ce*	*-ák*	*-ec*	*-čí*	TOTAL
Count	426	388	106	96	66	50	32	14	**1,178**

Subsets

- Subset A: formal characteristics
- Subset B: phonological characteristics
- Subset C: morphological characteristics
- Subset D: morphological family characteristics
- Subset E: quantitative characteristics

Examples of results

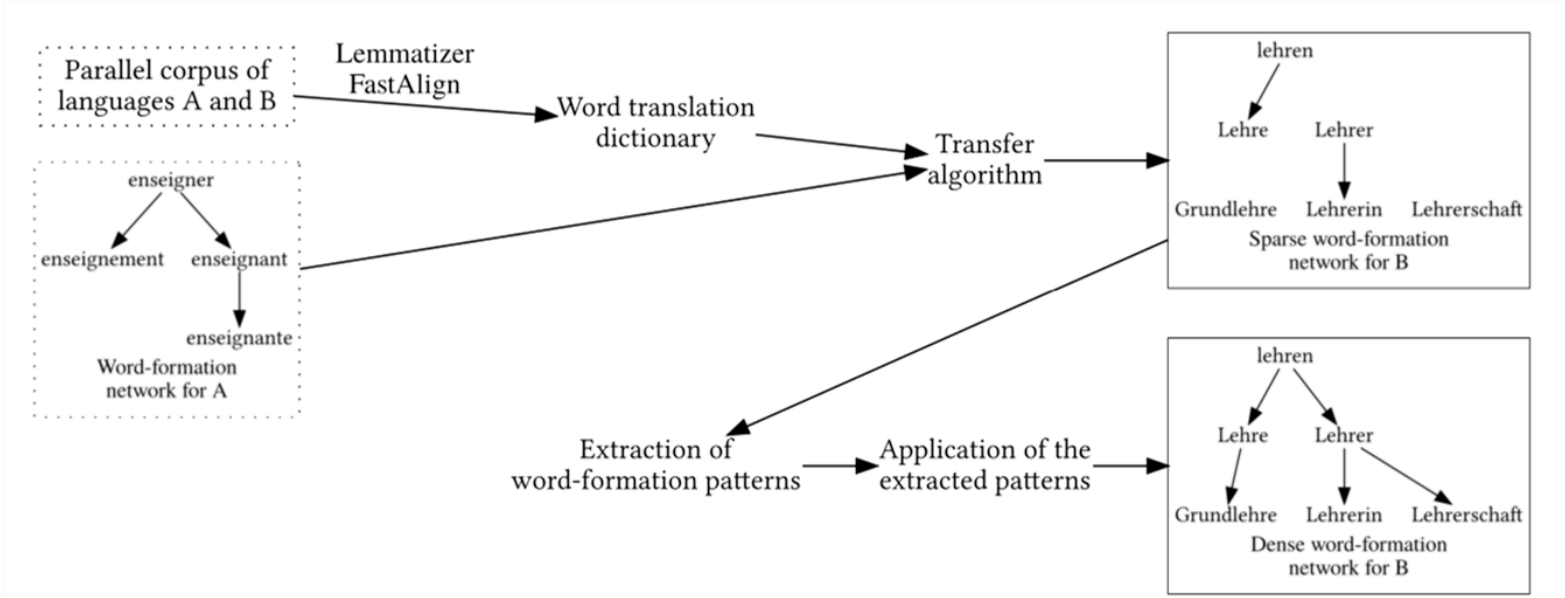
- There must be more relevant features not included
- The combination of features from different linguistic areas is necessary to model competition
- Results of *-ář|-ař* and *-ce* seems relatively balanced: instances are likely complex regarding competition



Jonáš Vidra & Zdeněk Žabokrtský 2021 (DeriMo 2021)
http://nabil.hathout.free.fr/DeriMo2021/pdf-files/DeriMo_2021_paper_16.pdf

Transferring word-formation networks

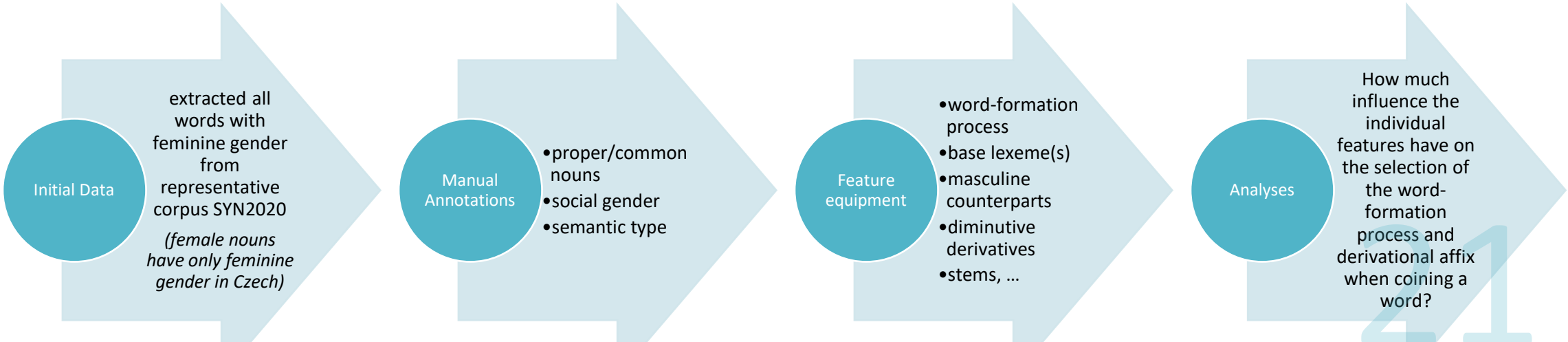
- proof-of-concept method for creating word-formation networks by transferring information from another language
- creates a low-precision and moderate-recall network in a language, for which no manual annotations need to be available



Ongoing work

Analysis of gender marking formation

- What are the base lexemes of the female representatives, and what is their distribution in terms of morphosyntactic categories, word-formation properties, and frequencies over time?
  - ***simplex:*** *matka*$_{N.fem}$ *(mother)* – *otec*$_{M.masc}$ *(father)*
    vdova$_{N.fem}$ *(widow)* > *vdovec*$_{M.masc}$ *(widower)*
  - ***derivatives:*** *učitelka*$_{N.fem}$ *(female teacher)* < *učitel*$_{N.masc}$ *(teacher)*
    kráska$_{N.fem}$ *(beautiful woman)* < *krása*$_{N.fem}$ *(beauty)*
    běhna$_{N.fem}$ *(floozy)* < *běhat*$_{V}$ *(to run)*
    světlovláska$_{N.fem}$ *(fair-haired woman)* < *světlovlasý*$_{A}$ *(fair-haired)*
  - ***conversion:*** *průvodčí*$_{N.fem}$ *(conductress)* <> *průvodčí*$_{N.masc}$ *(conductor)*
    hajná$_{N.fem}$ *(female ranger)* <> *hajný*$_{M.masc}$ *(ranger)*



Comparison of diminutiveness and gender marking across languages

- to quantify which strategies are used across the 7 languages to convey diminutiveness and gender marking => we need the same data across languages

- data: starts with a derivatives labelled as Diminutive/Female from DeriNet (cs) and translating them into other languages
  - several techniques of machine translation: *neural systems, bilingual dictionaries, custom dictionaries from parallel corpora, other resources*

- analyses: … soon ☺

¿ Distributional semantics ?

Thank you.

References

- Bonami, O., Strnadová, J. 2019. Paradigm Structure and Predictability in Derivational Morphology. Morphology, 29, 167-197. Springer. ISSN: 1871-5656.
- Körtvélyessy, L., Bagasheva, A., Štekauer, P. 2020. Derivational Networks Across Languages. De Gruyter Mouton. ISBN: 9783110686494.
- Kyjánek, L.; Žabokrtský, Z.; Vidra, J.; Ševčíková, M. 2021. Universal Derivations 1.1, LINDAT/CLARIAH-CZ digital library at the Institute of Formal and Applied Linguistics (ÚFAL), Faculty of Mathematics and Physics, Charles University, http://hdl.handle.net/11234/1-3247.
- Kyjánek, L.; Lyashevskaya, O.; Nedoluzhko, A.; Vodolazsky, D.; Žabokrtský, Z. 2021. DeriNet.RU 0.5, Institute of Formal and Applied Linguistics (ÚFAL), Faculty of Mathematics and Physics, Charles University, DeriNetRU-0.5.zip. Released also in the Universal Derivation collection v1.1.
- Ševčíková, M.; Kyjánek, L.; Vidová Hladká, B. 2021. Agent noun formation in Czech: An empirical study on suffix rivalry. *Second Workshop on Paradigmatic Word Formation Modelling*, 65-68. URL: ParadigMo-2-Booklet-of-abstracts.pdf.
- Ševčíková, M.; Kyjánek, L. 2019. Introducing Semantic Labels into the DeriNet Network. In *Journal of Linguistics*. Bratislava: Jazykovedný ústav Ľudovíta Štúra Slovenskej akadémie vied, pp. 412-423. ISSN: 0021-5597. URL: http://www.juls.savba.sk/ediela/jc/2019/2/jc19-02.pdf.
- Vidra, J.; Žabokrtský, Z. 2021. Transferring Word-Formation Networks Between Languages. In Proceeding of DeriMo 2021. ISBN: 978-2-9580006-0-8.
- Vidra, J.; Žabokrtský, Z.; Kyjánek, L.; Ševčíková, M.; Dohnalová, Š.; Svoboda, E.; Bodnár, J. 2021. DeriNet 2.1, LINDAT/CLARIAH-CZ digital library at the Institute of Formal and Applied Linguistics (ÚFAL), Faculty of Mathematics and Physics, Charles University, http://hdl.handle.net/11234/1-3765.