PhD Thesis Proposal

Formalisation of the word-formation meaning in language data resources

Lukáš Kyjánek

March 17, 2023

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

Outline

Introduction of word-formation meaning

State-of-the-art in the formalisation of word-formation meaning

Experiments

- Delimiting word-formation meanings
- Labelling word-formation meanings
- Selecting different affixes within one word-formation meaning
- Conveying the same word-formation meaning across languages

Conclusion & Future perspectives

Word-formation meaning? Example of affixation.

- The sum of the input words and the undergone changes is denoted as WORD-FORMATION or STRUCTURAL MEANING (Dokulil, 1962, Štekauer, 2005)

(x)		**(y: a person who does [x])**
učit ‘to teach’	→	*učitel ‘teacher’*
hrát ‘to act’	→	*herec ‘actor’*
bojovat ‘to fight’	→	*bojovník ‘fighter’*
kouřit ‘to smoke’	→	*kuřák ‘smoker’*

- Many-to-many relationship of form and meaning

(x)			
skříň ‘cupboard’	→	*skříňka ‘small cupboard’*	**(y: a small [x])**
učitel ‘teacher’	→	*učitelka ‘female teacher’*	**(y: a female counterpart of [x])**
obalit ‘to wrap’	→	*obálka ‘envelope’*	**(y: an instrument of [x])**

Formalisation of word-formation meaning in affixation

COMPARATIVE SEMANTIC CONCEPTS

= fundamental concepts of cognition rooted in cognitive linguistics; relevant for cross-linguistic research (Haspelmath, 2010)

- Bagasheva (2018) elaborates on the idea for affixation (52 labels applicable across pos)

ability
abstraction
action
agent
anticausative
augmentative[iii]
causative
collectivity
comitative
composition
cumulative
desiderative
diminutive[iv]
directional
distributive
durative
dweller
entity
experiencer
female
hyperonymy
hyponymy
inceptive
instrument
iterative
location
manner[i]
ornative
patient
pejorative
perceptive
pluriactionality
possessive
privative
process
purposive
quality
reciprocal
reflexive
relational
resultative
reversative
saturative[ii]
semelfactive
similative
singulative
state
subitive
terminative
temporal
undergoer

[i] *viewpoint* [ii] *total* [iii] *ameliorative/intensive* [iv] *attenuative*

In Bagasheva (2018, pp.53–56):

agent	*killer*
dweller	*villager*
patient	*amputee*
undergoer	*čuhppojuvvot* [Saami, Uralic] (*to be cut (of somebody)*)

Formalisation in language resources

- Bagasheva's labels implemented in the research into derivational networks for 40 languages (Körtvélyessy et al., 2020)

Labels in CroDeriv (Croatian).

- action
- **agent, female**
- anatomical part
- animal, female
- deprivation
- diminutive
- disease
- drink
- event
- instrument
- linguistic term
- literary type
- location
- number of men involved
- **person, both sexes**
- plant
- possibility
- quantity
- result
- temporal mark
- thing

Labels in Morpho-semantic database (English).

- **agent**
- body-part
- by-means-of
- destination
- event
- instrument
- location
- material
- property
- result
- state
- **undergoer**
- uses
- vehicle

Labels in Derivancze (Czech).

Label	Example
k1verb	*bití* ← *bít*
k2pas	*bit* ← *bít*
k2rpas	*bitý* ← *bít*
k2proc	*bijící* ← *bít*
k2rakt	*zabivší* ← *zabít*
k2ucel	*bicí* ← *bít*
k1ag	*badatel* ← *bádat*
k1prop	*hluchota* ← *hluchý*
k6a	*dobře* ← *dobrý*
k2pos	*otcův* ← *otec*
k2rel	*mrazový* ← *mráz*
k1f	*doktorka* ← *doktor*
k1jmf	*Novotná* ← *Novotný*
k1jmr	*Novotní* ← *Novotný*
k1obyv	*Kanaďan* ← *Kanada*
k1dem	*domek* ← *dům*
k5freq	*bádávat* ← *bádat*
var	*komunismus* ↔ *komunizmus*

Labels in DeriNet (Czech).

- *diminutive*
- *female*
- *iterative*
- *aspect*
- *possessive*

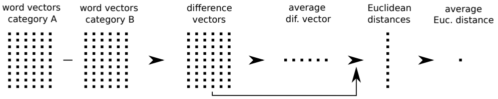
Labels in Démonette (French).

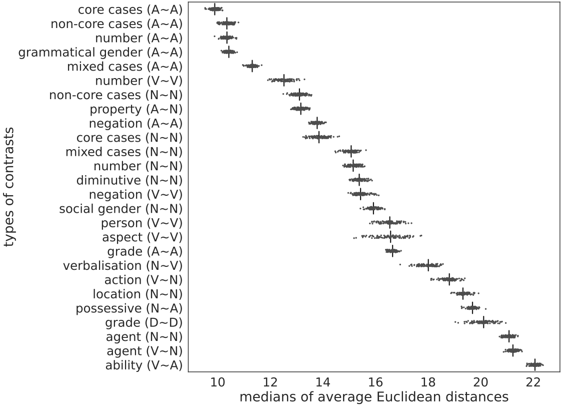
- *action*
- ***agent***
- *property*

(1) How to delimit word-formation meanings?

An experiment to observe the inflexion-derivation scale and diversity of meanings in Czech.

- Word pairs from DeriNet (Vidra et al., 2021) / MorfFlex (Hajič et al., 2020); represented as vectors of distributional semantics (Mikolov et al., 2013)
- Bootstrapped samples of 200 word pairs (token freq $>$ 50); cf. Mickus et al. (2019)





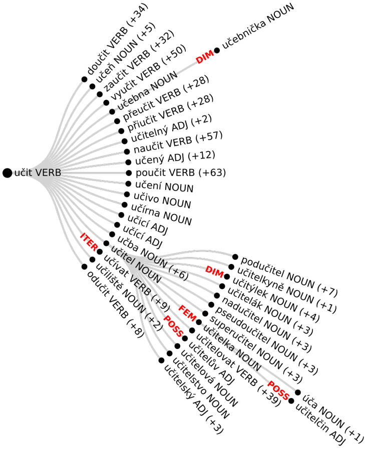
L. Kyjánek and O. Bonami. A Distributional Approach to Inflection vs. Derivation in Czech.

In *Word-Formation Theories VI & Typology and Universals in Word-Formation V*, pages 21–22, Košice, 2022

(2) How to label word-formation meanings?

An experiment on labelling several selected word-formation meanings in DeriNet for Czech.

- Selected meanings: FEMALE, DIMINUTIVE, ASPECT, ITERATIVE, and POSSESSIVE.
- Multinomial Logistic Regression model developed on the existing annotations from the digitised language data resource for Czech
- Features: [pos, gender, aspect, final character n-grams ($n = 2, ..., 6$), possesivity tag of derivatives and base words]
- Evaluated on testing data set; F-score = 98 %



M. Ševčíková and L. Kyjánek. Introducing semantic labels into the DeriNet network. *Journal of Linguistics/Jazykovedný časopis*, 70(2):412–423, 2019

(3) What selects the affix within word-formation meanings?

An experiment to observe formal-linguistic features in the agent nouns formations in Czech.

- Agent nouns formed by 8 most frequent suffixes (*-tel, -č, -ík|-ník, -ář|-ař, -ce, -ák, -ec, -čí*)

agent noun	**verb.IPVF|PFV**	**noun**	**adjective**
*sjednot-i-**tel*** ‘unifier’	- | *sjednot-i-t* ‘unify’		
*sjednoc-ova-**tel*** ‘unifier’	*sjednoc-ova-t* | - ‘unify’		
*model-**ář*** ‘modeler’	*model-ova-t* | - ‘model’	*model* ‘model’	
*zvon-**ík*** ‘bell-ringer’	*zvon-i-t* | - ‘ring’	*zvon* ‘bell’	
*závod/**n/ík*** ‘racer’	*závod-i-t* | - ‘race’	*závod* ‘race’	*závod-n-í* ‘racing’
*boj-ov/**n/ík*** ‘fighter’	*boj-ova-t* | - ‘fight’		*boj-ov-n-ý* ‘fighting’
*střel-**ec*** ‘shooter’	*stříl-e-t* | *střel-i-t* ‘shoot’	*střel-a* ‘shot’	
*kup-**ec*** ‘purchaser’	*kup-ova-t* | *koup-i-t* ‘purchase’	*koup-ě* ‘purchase’	

- Assigned with formal-linguistic features divided into four subsets
- Decision Tree and Logistic Regression models were trained for each subset of features to predict an agent affix for an input verb

(3) What selects the affix within word-formation meanings?

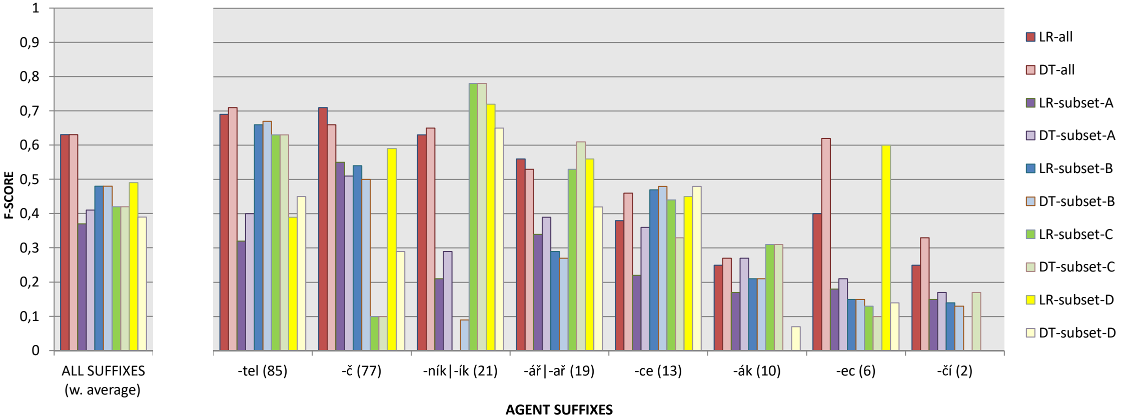
- related to the motivating verb(s)
  - final consonant of the root
  - number of prefixes
  - theme
  - aspect
  - conjugation class

- related to the derivational paradigm
  - which motivating items available?
  - does the verb have a suffixed aspectual counterpart?
  - does an inanimate homonym exist?
  - absolute corpus frequency of all items
  - motivating items ordered by frequency

válečník	*válčit*	–	*válka*	–	*válečný*
warrior	make war	–	war.n	–	war.adj

target_noun_suffix	***-ník*|*-ík***
root_final	*č*
root_final_cvs	*consonant*
root_final_vertical	*africate*
root_final_horizontal	*postalveolar*
number_prefixes	*0*
v1_theme	*i*
v1_aspect	*imp*
v1_conjug	*4*
v2_theme	–
v2_aspect	–
v2_conjug	–
v1_suf_asp_counterpart	*no*
paradigm_type	*NNA-V-*
inanim_noun	*no*
freq_parent_noun	*25,895*
freq_parent_adj	*4,953*
freq_parent_v1	*499*
freq_parent_v2	–
freq_slots	*VAN*

(3) What selects the affix within word-formation meanings?



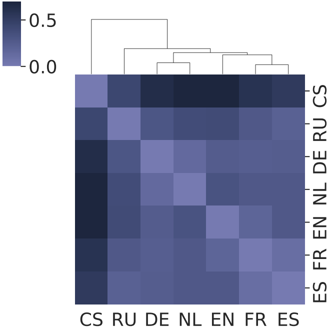
M. Ševčíková, L. Kyjánek, and B. Vidová Hladká. Agent noun formation in Czech: An empirical study on suffix rivalry. In *Second Workshop on Paradigmatic Word Formation Modelling (ParadigMo II)*, page 65, Bordeaux, 2021

(4) How are word-formation meanings conveyed across languages?

An experiment on conveying the same word-formation meaning across seven languages.

CS	RU	NL	DE	EN	FR	ES
uklízečka	**[ubórshchitsa]**	**schoonmaakster**	**Putzfrau**	**cleaning lady**	**femme de ménage**	**mujer de limpieza**
derivative	derivative	derivative	compound	synt. phrase	synt. phrase	synt. phrase

- 3,746 female counterparts translated from Czech to six above-listed languages
- Naming strategies annotated automatically (Svoboda and Ševčíková, 2022): *derivation*, *compounding*, *syntactic phrase*, *unmotivated word*, *unmarked for female social gender*
- Similarity between probability distributions measured by information radius (Jensen-Shannon divergence)



- Distributions for conveying female social gender demonstrate some resemblances with the genetic classification of languages

Conclusion & Future perspectives

- Promising approaches:
  - **Distributional semantics**: Bonami and Naranjo (2023) exemplify modelling and predicting word-formation meanings in affixation
  - **Machine translation**: Gast (2022) and Gast and Borges (2022) exemplify cross-lingual transfer of word-formation meanings by using back translation
- ToDo: Formalisation of word-formation meanings in conversion and compounding

Future: Formalisation of Conversion

to butcher	↔	*a butcher*
to kennel	↔	*a kennel*

- The same set of word-formation meanings as for affixation seems viable
- Identification of particular meanings is difficult (no formal changes)
- Clark and Clark (1979) exploit paraphrases
  - *agent and experiencer verbs*, e.g.,
    to butcher in ‘*John butchered the cow.*’ vs. *a butcher* in ‘*John did to the cow the act that one would normally expect [a butcher to do to a cow].*’
  - *location and duration verbs*, e.g.,
    to kennel in ‘*John kenneled the dog.*’ vs. *a kennel* in ‘*John did something to cause it to come about that [the dog was in a kennel].*’
- Models of distributional semantics might be promising here, not only in affixation

Future: Formalisation of Compounding

strong	+	*man*	→	*strongman*
ice	+	*man*	→	*iceman*
fire	+	*fighter*	→	*firefighter*
pick	+	*pocket*	→	*pickpocket*
guitar	+	*player*	→	*guitar_player*

- Scalise and Bisetto (2009) propose rather syntactic classification
  - Tectogramatical functors (PDT)?
    - ACT for *dřevorubec* ‘*lumberjack*’
    - PAT for *senoseč* ‘*haymaking*’
    - RSTR for *modrooký* ‘*blue eyed*’
- Štekauer (2016) applies the onomas. theory to compounds with regular components

References I

R. Harald Baayen, Yu-Ying Chuang, Elnaz Shafaei-Bajestan, and J. Blevins. The discriminative lexicon: A unified computational model for the lexicon and lexical processing in comprehension and production grounded not in (de)composition but in linear discriminative learning. *Complex.*, 2019: 4895891:1–4895891:39, 2019.

Alexandra Bagasheva. Comparative semantic concepts in affixation. In Juan Santana-Lario and Salvador Valera-Hernández, editors, *Competing Patterns in English Affixation*, pages 33–65. Peter Lang Verlag, Lausanne, 2018. URL https://www.peterlang.com/document/1055047.

Olivier Bonami and Gilles Boyé. Paradigm uniformity and the French gender system. In Matthew Baerman, Oliver Bond, and Andrew Hippisley, editors, *Perspectives on morphology: Papers in honour of Greville G. Corbett*, pages 171–192. Edinburgh University Press, 2019. ISBN 978-1474446006. URL http://www.llf.cnrs.fr/sites/llf.cnrs.fr/files/biblio//BonamiBoye19.pdf.

Olivier Bonami and Matías Guzmán Naranjo. Distributional evidence for derivational paradigms. In Sven Kotowski and Ingo Plag, editors, *The semantics of derivational morphology: theory, methods, evidence*. De Gruyter, Berlin, 2023.

Eve V. Clark and Herbert H. Clark. When nouns surface as verbs. *Language*, 55(4):767–811, 1979. doi: 10.2307/412745.

Miloš Dokulil. *Tvoření slov v češtině 1: Teorie odvozování slov*. Academia, Prague, 1962.

Volker Gast. Compounds in translation and interpreting. A study of English and German based on a multimodal parallel corpus, 6 2022. URL http://kaa.ff.upjs.sk/en/event/43/word-formation-theories-vi-typology-and-universals-in-word-formation-v#toc-plenary-speakers-2. Plenary speech at the conference Word-Formation Theories VI & Typology and Universals in Word-Formation V.

Volker Gast and Robert Borges. A triangular translation corpus, and a case study of nominal compounds in English and German. In *Translation in Transition 6*, pages 45–57, Prague, 2022. URL https://tt2022.ff.cuni.cz/wp-content/uploads/sites/69/2022/09/tt2022_book_of_abstracts.pdf.

Jan Hajič, Jaroslava Hlaváčová, Marie Mikulová, Milan Straka, and Barbora Štěpánková. MorfFlex CZ 2.0, 2020. URL http://hdl.handle.net/11234/1-3186. LINDAT/CLARIAH-CZ digital library at the Institute of Formal and Applied Linguistics (ÚFAL), Faculty of Mathematics and Physics, Charles University.

Martin Haspelmath. Comparative concepts and descriptive categories in crosslinguistic studies. *Language*, 86(3):663–687, 2010. doi: https://doi.org/10.1353/lan.2010.0021.

Dominika Kovarikova, Lucie Chlumska, and Vaclav Cvrcek. What belongs in a dictionary? The Example of Negation in Czech. In *Proceedings of the 15th Euralex international congress*, pages 822–827, Oslo, 2012.

References II

L. Kyjánek and O. Bonami. A Distributional Approach to Inflection vs. Derivation in Czech. In *Word-Formation Theories VI & Typology and Universals in Word-Formation V*, pages 21–22, Košice, 2022.

Lívia Körtvélyessy, Alexandra Bagasheva, and Pavol Štekauer, editors. *Derivational Networks Across Languages*. De Gruyter Mouton, Berlin, 2020. doi: https://doi.org/10.1515/9783110686630. URL `https://www.degruyter.com/document/doi/10.1515/9783110686630/html?lang=en`.

Timothee Mickus, Olivier Bonami, and Denis Paperno. Distributional Effects of Gender Contrasts Across Categories. In *Proceedings of the Society for Computation in Linguistics (SCiL) 2019*, volume 2, pages 174–184, 2019. doi: https://doi.org/10.7275/g11b-3s25. URL `https://aclanthology.org/W19-0118`.

Tomas Mikolov, Kai Chen, Greg Corrado, and Jeffrey Dean. Efficient estimation of word representations in vector space. *arXiv preprint arXiv:1301.3781*, 2013.

Tore Nesset, Alexander Piperski, and Svetlana Sokolova. Russian feminitives: what can corpus data tell us? *Russian Linguistics*, 46:95–113, 2022. ISSN 1572-8714. doi: 10.1007/s11185-022-09253-w. URL `https://doi.org/10.1007/s11185-022-09253-w`.

Sergio Scalise and Antonietta Bisetto. The classification of compounds. In Rochelle Lieber and Pavol Štekauer, editors, *The Oxford handbook of compounding*, pages 34–53. Oxford University Press, New York, 2009. ISBN 9780191743566. doi: https://doi.org/10.1093/oxfordhb/9780199695720.013.0003.

Emil Svoboda and Magda Ševčíková. Word Formation Analyzer for Czech: Automatic Parent Retrieval and Classification of Word Formation Processes. *The Prague Bulletin of Mathematical Linguistics*, 118:55–73, 2022. ISSN 0032-6585. doi: 10.14712/00326585.019. URL `https://ufal.mff.cuni.cz/pbml/118/art-svoboda-sevcikova.pdf`.

M. Ševčíková and L. Kyjánek. Introducing semantic labels into the DeriNet network. *Journal of Linguistics/Jazykovedný časopis*, 70(2):412–423, 2019.

M. Ševčíková, L. Kyjánek, and B. Vidová Hladká. Agent noun formation in Czech: An empirical study on suffix rivalry. In *Second Workshop on Paradigmatic Word Formation Modelling (ParadigMo II)*, page 65, Bordeaux, 2021.

Pavol Štekauer. Onomasiological Approach to Word-Formation. In Pavol Štekauer and Rochelle Lieber, editors, *Handbook of Word-Formation*, pages 207–232. Springer, Dordrecht, 2005. ISBN 978-1-4020-3596-8.

Pavol Štekauer. Compounding from an onomasiological perspective. In Pius ten Hacken, editor, *The Semantics of Compounding*, page 54–68. Cambridge University Press, 2016. doi: 10.1017/CBO9781316163122.004.

Fabian Tomaschek, Ingo Plag, Mirjam Ernestus, and R. Harald Baayen. Phonetic effects of morphology and context: Modeling the duration of word-final S in English with naïve discriminative learning. *Journal of Linguistics*, 57(1):123–161, 2021. doi: 10.1017/S0022226719000203.

Jonáš Vidra, Zdeněk Žabokrtský, Lukáš Kyjánek, Magda Ševčíková, Šárka Dohnalová, Emil Svoboda, and Jan Bodnár. DeriNet 2.1, 2021. URL `http://hdl.handle.net/11234/1-3765`. LINDAT/CLARIAH-CZ digital library at the Institute of Formal and Applied Linguistics (ÚFAL), Faculty of Mathematics and Physics, Charles University.

Q: Discriminative approach for the purpose of annotation

- Tomaschek et al. (2021): Phonetic effects of morphology and context
  - discriminative learning to observe differences in the duration of word-final S in English inflexion
  - structures of words ending with S (for inflection) learnt and the influence of phonetics effects measured based on the learnt models
- Baayen et al. (2019): The Discriminative Lexicon
  - incorporates the insight from machine learning that end-to-end modeling instead of a cascade of models targeting individual subtasks
  - simple linear networks are used for mapping form onto meaning and meaning onto form
  - their model: recognises words and produces words correctly, understands and produces novel complex words, and correctly predicts a wide array of experimental phenomena in lexical processing

Q: Criteria of harmonization for picking (or not picking) one nomenclature over the other?

- The existing approaches have a similar goal but labelling, e.g.,
  - *female* in Bagasheva (2018) expects only affixation;
  - *social gender* in Bonami and Boyé (2019) includes derivatives and compounds;
  - *feminitives* in Nesset et al. (2022) refers to female professionals but not animals.

- The task is not to harmonise the existing approaches.
- To label word-formation meanings consistently, we finding ”optimal”:
  - **granularity of meanings**
  - **formal-linguistic features**
  - **method for labelling**
- Proceeding ***with respect to the data***

Q: The sampling method in the inflexion–derivation experiment

- MorfFlexCZ to find word pairs more precisely
- The bootstrapping samples were created from the corpus (text)
- The samples have the ability to represent the distribution in the text
  (with frequent pairs more likely to be chosen)

Q: Negation for adjectives *vs.* negation for verbs

- Kovarikova et al. (2012):
    - **frequency criterion**: negation compared to ADJ.num
      — vebr. neg. is closer to ADJ.num than the adj. neg. is;
    - **gramatical criteria**: the same meaning, number of constituents, and full coverage within a word group; the first two are met in all cases, and the last one is more complicated with adjectives and adverbs
      — verb. neg. closer to other inflectional categories.

- Kyjánek and Bonami (2022): **distributional semantics**
  — verb. neg. is closer to categories like diminution or social gender
    - *Fine-grained picture*: our results claim that the context of words conveys verb. neg. are more diverse than the context of those for adj. neg. (different view).
    - *Big picture*: even in the Czech linguistic tradition the diminution and social gender are treated as being on the borderline of inflexion-derivation. So our results do not change the view of the category of verb. neg. as being closer to inflectional categories.

Q: The results of automatic labelling and its evaluation

- The labelling experiment was designed to label only five word-formation meanings:

DIMINUTIVE	*psík ‘small dog’ ← pes ‘dog’*	5,072 rels. in DeriNet
FEMALE	*učitelka ‘female teacher’ ← učitel ‘teacher’*	27,938 rels. in DeriNet
POSSESSIVE	*učitelův ‘teacher’s’ ← učitel ‘teacher’*	85,327 rels. in DeriNet
ASPECT	*obalovat ‘to wrap’ ← obalit ‘to wrap’*	14,040 rels. in DeriNet
ITERATIVE	*chodívat ‘to walk repeat.’ ← chodit ‘to walk’*	10,969 rels. in DeriNet

- Evaluation on the **testing data set** (2,000 pairs separated from training data set):

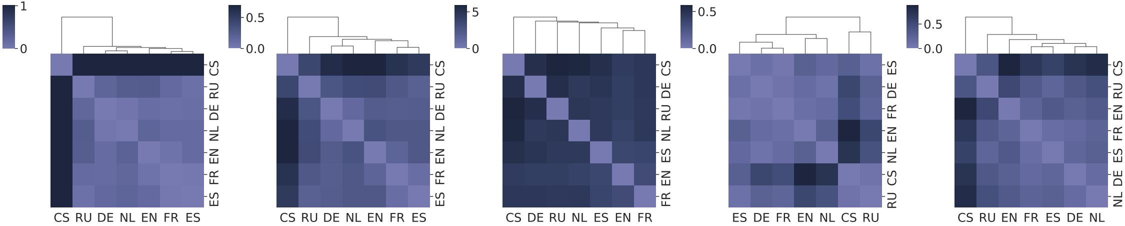
Model	**Accuracy**	**Precision**	**Recall**	**F1-score**
Baseline	0.827	0.813	0.827	0.792
MLR model	0.986	0.984	0.983	0.984

- Evaluation on the **predicted DeriNet data** (2,000 randomly selected pairs):

Gold/Pred.	**DIMINUTIVE**	**FEMALE**	**POSSESSIVE**	**ITERATIVE**	**ASPECT**	**NONE**
DIMINUTIVE	62	0	0	0	0	4
FEMALE	1	296	0	0	0	3
POSSESSIVE	0	0	905	0	0	1
ITERATIVE	0	0	0	135	4	0
ASPECT	0	0	0	3	170	1
NONE	1	39	1	0	0	374
PRECISION	0.969	0.982	0.999	0.985	0.987	0.948
RECALL	0.983	0.941	0.999	0.988	0.987	0.976

Q: Some genetic kinship and the Czech as a pivot language in the data-driven comparative research

- The Czech language served as a pivot language
  - we cannot claim: derivation is the most frequent naming strategy in Czech; but
  - we can claim: most of the seen instances which are formed as derivatives in Czech are also derivatives in Russian but gender-neural words in English
- German + Dutch: the usage of derivation and compounding
- French + Spanish: the usage of syntactic phrases.
- The genetic kindship is only a trend, but the situation seems more complicated (Distance measures: KL, JS, MI, Cos, Euc)



Acknowledgement

http://ufal.cz/node/2248

Individual parts of this work were supported by Grant No. START/HUM/010 of Grant schemes at Charles University (reg. No. CZ.02.2.69/0.0/0.0/19_073/0016935), Grant No. GA19-14534S of the Czech Science Foundation, and the Charles University Grant Agency (project No. 1176219). It has been using data, tools and services provided by the LINDAT/CLARIAH-CZ Research Infrastructure (https://lindat.cz), supported by the Ministry of Education, Youth and Sports of the Czech Republic (Project No. LM2018101).