Agent noun formation in Czech: An empirical study on suffix rivalry

Magda Ševčíková, Lukáš Kyjánek, Barbora Vidová Hladká
{sevcikova|kyjanek|hladka}@ufal.mff.cuni.cz

Charles University, Prague, Czech Republic
Faculty of Mathematics and Physics
Institute of Formal and Applied Linguistics

ParadigMo 2, June 4, 2021

Introduction: Agent nouns across languages

- one of the most frequent categories attested cross-linguistically (Bauer 2002, Štekauer et al. 2012)
- derived from verbs (nomina agentis)
  - *writer* < *write*
- agentive meaning ascribed also to denominal nouns (Rainer 2015; nomina actoris)
  - *paintballer* < *paintball*
- often both a directly related noun and verb attested (oed.com):
  - *fisher* < *fish.v* (*fish.v* < *fish.n*)
  - *footballer* < *football.n* or *footballer* < *football.v* (*football.v* < *football.n*)

Agent noun formation in Czech

- 35 different agent suffixes to combine with verbs (Daneš et al. 1967, Dokulil et al. 1986, Štícha et al. 2018)

  - 8 most frequent of them covered by the paper:

a. *uč-i-**tel*** 'teacher' < *uč-i-t* 'to teach'
b. *řid-i-**č*** 'driver' < *říd-i-t* 'to drive'
c. *řez-**ník*** 'butcher' < *řez-a-t* 'to cut'
d. *kov-**ář*** 'blacksmith' < *kov-a-t* 'to forge'
e. *soud-**ce*** 'judge' < *soud-i-t* 'to judge'
f. *kuř-**ák*** 'smoker' < *kouř-i-t* 'to smoke'
g. *kup-**ec*** 'buyer' < *koup-i-t* 'to buy'
h. *mluv-**čí*** 'speaker' < *mluv-i-t* 'to speak'

- *-tel* only in agents, but most of the suffixes convey more than one semantic category:

  e.g. the suffix *-ec* in
  1. agents (*letec* 'pilot' < *létat* 'to fly'), 2. inhabitants (*Nepálec* 'Nepali' < *Nepál* 'Nepal'), 3. bearers of social roles (*vdovec* 'widower' < *vdova* 'widow'), 4. bearers of qualities (*stařec* 'old man' < *starý* 'old'), 5. animal names (*dravec* 'predator' < *dravý* 'predatory'), 6. instruments (*bodec* 'spike' < *bodat* 'to stab'), 7. toponyms (*Hradec* < *hrad* 'castle'), etc.

Outline

1. Design of the data
    - A data-based approach to the agent suffix rivalry
    - Extraction of the agent nouns from the corpus
    - Features to assign
2. Baseline solution
3. Machine learning experiments: logistic regression vs. decision trees
    - Experiments on all features
    - Experimenting with feature sets
4. Discussion & conclusions
    - Comparison of the methods
    - Incorrect predictions
    - Final remarks

A data-based approach to the agent suffix rivalry

- paradigmatic approach (Bonami & Strnadová 2019)
    - agent nouns as members of morphological families
    - all potential predecessors considered

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'	

Extraction of the agent nouns from the corpus

- all masculine animate nouns ending in one of the suffix strings extracted from the SYN2015 corpus (Křen et al. 2015)
- non-agents, nouns where the string is not a suffix, compounds, typos, etc. excluded
- potential predecessors listed: verb (imperfective | perfective), noun, adjective
- nouns without a verbal predecessor removed

>>> 1,178 nouns in the final set

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

- 20 features assumed as potentially relevant for modeling the rivalry (Strnadová 2015, Santana-Lario & Valera 2017, Bonami & Thuilier 2019, Wauquier et al. 2020)

Features to assign

- 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*
v1_suf_asp_counterpart	*no*
v2_theme	–
v2_aspect	–
v2_conjug	–
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*

Baseline solution

- data set divided into a training set, an evaluation set, and a hold-out set (60:20:20)
- random baseline predicting one of the eight suffixes in a uniform distribution
  - weighted average of **F-score=0.16** calculated on the hold-out data set

Suffix	all	*-tel*	*-č*	*-ník|-ík*	*-ář|-ař*	*-ce*	*-ák*	*-ec*	*-čí*
Instances	233	85	77	21	19	13	10	6	2
Precision	0.28	0.43	0.32	0.10	0.07	0.08	0.04	0.04	0.04
Recall	0.13	0.14	0.10	0.14	0.11	0.23	0.10	0.17	0.50
F-score	**0.16**	0.21	0.16	0.12	0.09	0.12	0.05	0.06	0.07

Machine learning experiments

- which agent suffix is chosen by a particular verb?
  - the agent suffix used as the target class in the experiments
  - the other features as predictors
- two different machine learning methods applied
  - hyper-parameter settings tuned in the first experiment on all features
  - results compared to experiments on four different feature subsets

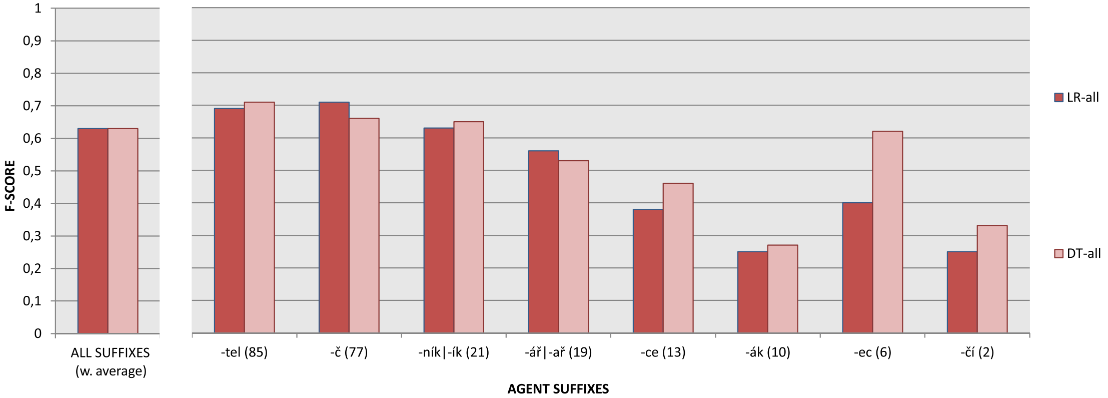
- Logistic regression

```
classifier_LR = LogisticRegression(
    multi_class='multinomial',
    class_weight='balanced',
    solver='newton-cg',
    penalty='l2',
    C=1e30)
```

- Decision trees

```
classifier = DecisionTreeClassifier(
    criterion='entropy',
    class_weight='balanced',
    splitter='best',
    max_depth=10)
```

Experimenting with all features: F-score on hold-out data

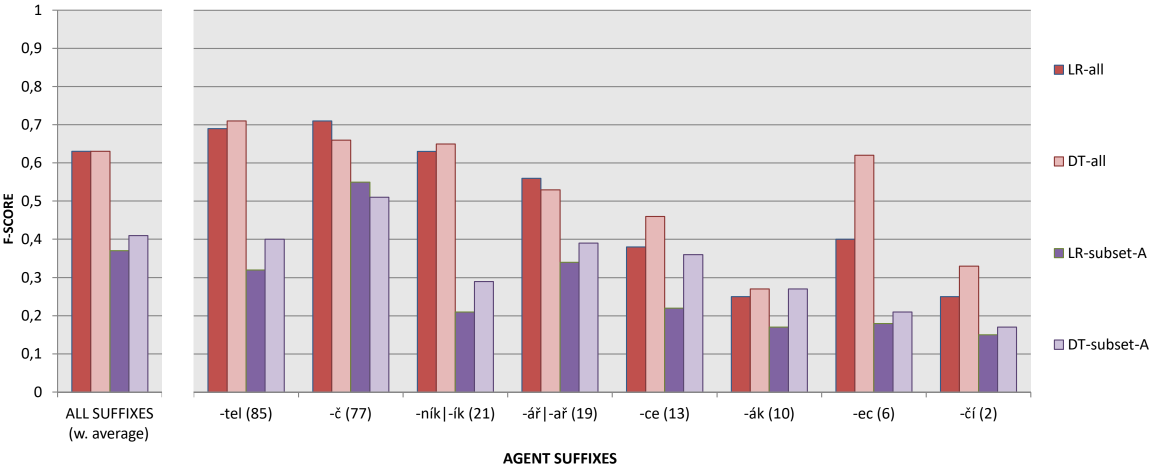


Experimenting with feature subsets: Subsets A to D

- A: the motivating verb(s): root's final character and theme
  [root_final, root_final_cvs, root_final_vertical, root_final_horizontal, v1_theme, v2_theme]
- B: the motivating verb(s): number of prefixes, theme, aspect, conjugation class
  [number_prefixes, v1_theme, v1_aspect, v1_conjug, v2_theme, v2_aspect, v2_conjug]
- C: the derivational paradigm: which motivating items available?, does the verb have a suffixed aspectual counterpart?, does an inanimate homonym exist?
  [paradigm_type, v1_suf_asp_counterpart, inanim_noun]
- D: corpus frequency of the motivating items
  [freq_parent_noun, freq_parent_adj, freq_parent_v1, freq_parent_v2, freq_slots]

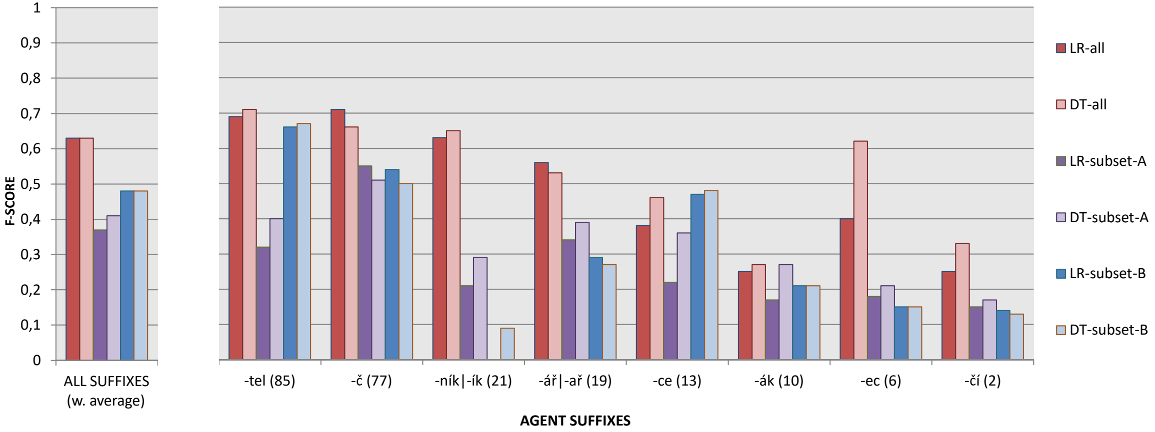
Experiments with the subset A: F-score on hold-out data

subset A: root_final, root_final_cvs, root_final_vertical, root_final_horizontal, v1_theme, v2_theme



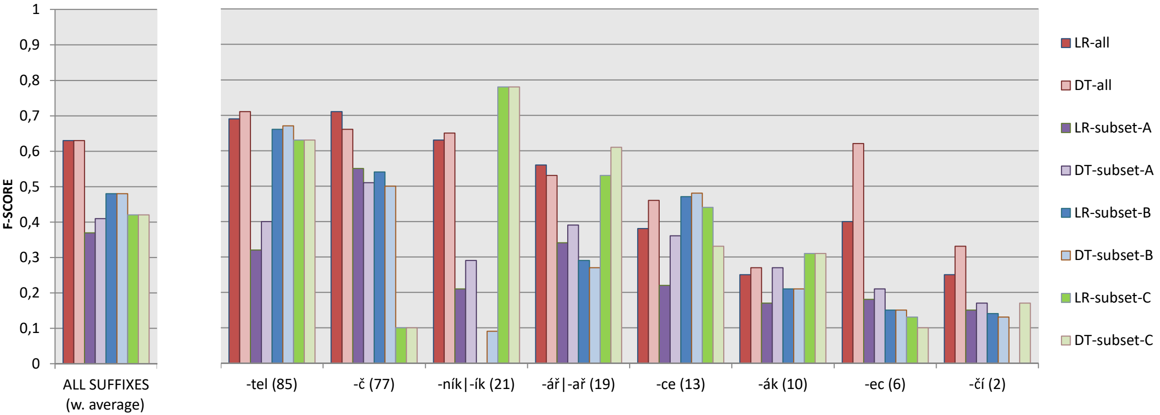
Experiments with the subset B: F-score on hold-out data

subset B: number_prefixes, v1_theme, v1_aspect, v1_conjug, v2_theme, v2_aspect, v2_conjug



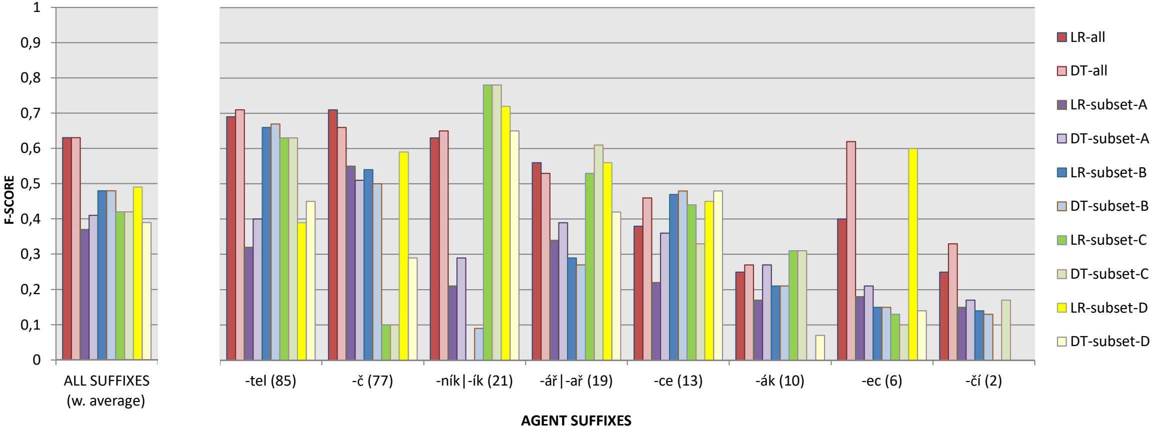
Experiments with the subset C: F-score on hold-out data

subset C: paradigm_type, v1_suf_asp_counterpart, inanim_noun



Experiments with the subset D: F-score on hold-out data

subset D: freq_parent_noun, freq_parent_adj, freq_parent_v1, freq_parent_v2, freq_slots



Discussion: predicting all suffixes by logistic regression *vs.* decision trees

- the methods model the impact of the features differently
  - logistic regression estimates dependencies among the given features
  - decision trees propose a set of decisions over the features such that their disorder (entropy) is minimized
- all suffixes best predicted based on all features
  - logistic regression with all features: F-score=0.63
  - decision trees with all features: F-score=0.63 (vs. baseline F-score=0.16)

- features seem to be relevant
- there must be more relevant features not yet covered by the data

Results on individual suffixes

- *-tel*, *-č*, *-ec*, *-čí*: best results with all features
- *-ce* the same results on the subset B (detailed features of the verb) and D (frequency)
- *-ník|-ík*, *-ář|-ař*, *-ák* best predicted from the derivational paradigm (subset C)
  - *-ník|-ík* motivated by a verb/verbs and by an adjective (*pracovník* 'worker')
  - *-ář|-ař* motivated by a noun and a verb/verbs, never has an inanimate homonym (*záchranář* 'rescuer', *tiskař* 'printer')
  - *-ák* based on a verb/verbs, can have an inanimate homonym (*piják* 'drunkard x blotter')
- subset A (root & themes) not sufficient

suffix	noun	all features (log.regr./dec.trees)	A	B	C	D
-tel	85	0.69/0.71	0.32/0.40	0.66/0.67	0.63/0.63	0.39/0.45
-č	77	0.71/0.66	0.55/0.51	0.54/0.50	0.10/0.10	0.59/0.29
-ník|-ík	21	0.63/0.65	0.21/0.29	0.00/0.09	0.78/0.78	0.72/0.65
-ář|-ař	19	0.56/0.53	0.34/0.39	0.29/0.27	0.53/0.61	0.56/0.42
-ce	13	0.38/0.46	0.22/0.36	0.47/0.48	0.44/0.33	0.45/0.48
-ák	10	0.25/0.27	0.17/0.27	0.21/0.21	0.31/0.31	0.00/0.07
-ec	6	0.40/0.62	0.18/0.21	0.15/0.15	0.13/0.10	0.60/0.14
-čí	2	0.25/0.33	0.15/0.17	0.14/0.13	0.00/0.17	0.00/0.00
all	233	0.63/0.63	0.37/0.41	0.48/0.48	0.42/0.42	0.49/0.39

Incorrect predictions

- *-ník/ík* predicted in **signatník* (expected *signatář* 'signatory')
  - the native suffix incompatible with the foreign base (cf. German *Signatar*)

- *-č* predicted in **oblehač* (vs. *oblehatel* 'besieger'), **budič* (vs. *buditel* 'revivalist')
  - differences in registers (formal register of the base vs. informal suffix)
  - *budič* attested as an inanimate noun

- *-ce* predicted in **ulejvce* (vs. *ulejvák* 'loafer'), **výčepce* (vs. *výčepák* 'bartender')
  - different registers (informal base vs. formal suffix)

Conclusions

- study on rivalry among eight suffixes used in Czech agent nouns
- 1,178 agent nouns with verbal predecessors
  - provided with 20 features (phonology, morphology, paradigmatic info)
- random baseline model's F-score 0.16
- two machine-learning methods applied
  - experiments with all features vs. with feature subsets
  - best prediction of all suffixes based on all features
    - F-score 0.63 both with logistic regression and decision trees
  - derivational paradigms relevant for predicting individual suffixes

- not considered:
  - diachronic features (date of attestation), registers, origin (foreign vs. native)
  - speakers's preferences, lexicalization

References

- Bauer, L. 2002. What you can do with derivational morphology. In S. Bendjaballah et al. (eds.), *Morphology 2000. Selected Papers from the 9th Morphology Meeting*, 37–48. John Benjamins.
- Bonami, O. & J. Strnadová. 2019. Paradigm structure and predictability in derivational morphology. *Morphology* 29. 167–197.
- Bonami, O. & J. Thuilier. 2019. A statistical approach to rivalry in lexeme formation: French *-iser* and *-ifier*. *Word Structure* 12: 4–41.
- Daneš, F. et al. 1967. *Tvoření slov v češtině 2: Odvozování podstatných jmen.* ČSAV.
- Dokulil, M. et al. 1986. *Mluvnice češtiny 1.* Academia.
- Křen, M. et al. 2015: *SYN2015*. http://www.korpus.cz
- Pedrogosa, F. et al. 2011. Scikit-learn: Machine Learning in Python. *Journal of Machine Learning Research* 12. 2825-2830.
- Rainer, F. 2015. Agent and instrument nouns. In Peter O. Müller et al. (eds.), *Word-Formation. An International Handbook of the Languages of Europe*, vol. 2, 1304–1316. De Gruyter.
- Santana-Lario, J. & S. Valera. 2017. *Competing patterns in English affixation.* Peter Lang.
- Strnadová, J. 2015. Multiple Derivation in French Denominal Adjectives. In *Carnets de Grammaire* 22, 327–346. CLLE-ERSS.
- Štekauer, P. et al. (eds.). 2012. *Word-Formation in the World's Languages*. CUP.
- Štícha, F. et al. 2018. *Velká akademická gramatika spisovné češtiny 1.* Academia.
- Wauquier, M. et al. 2020. Contributions of distributional semantics to the semantic study of French morphologically derived agent nouns. In J. Audring et al. (eds.), *Online Proceedings of the MMM12* 2, 111–122. Pasithee.

The research has been supported by the Grant No. GA19-14534S of the Czech Science Foundation,
by the LINDAT/CLARIAH-CZ project of the Ministry of Education, Youth and Sports of the Czech Republic (project LM2018101),
and 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).