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.
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

<table>
<thead>
<tr>
<th>agent noun</th>
<th>verb.IPVF</th>
<th>PFV</th>
<th>noun</th>
<th>adjective</th>
</tr>
</thead>
<tbody>
<tr>
<td>sjednot-i-tel</td>
<td>-</td>
<td>sjednot-i-t</td>
<td>‘unify’</td>
<td></td>
</tr>
<tr>
<td>sjednoc-ova-tel</td>
<td>sjednoc-ova-t</td>
<td>- ‘unify’</td>
<td></td>
<td></td>
</tr>
<tr>
<td>model-ář</td>
<td>model-ova-t</td>
<td>- ‘model’</td>
<td>model ‘model’</td>
<td></td>
</tr>
<tr>
<td>zvon-ík</td>
<td>zvon-i-t</td>
<td>- ‘ring’</td>
<td>zvon ‘bell’</td>
<td></td>
</tr>
<tr>
<td>závod/n/ik</td>
<td>závod-i-t</td>
<td>- ‘race’</td>
<td>závod ‘race’</td>
<td>závod-n-í ‘racing’</td>
</tr>
<tr>
<td>boj-ov/n/ik</td>
<td>boj-ova-t</td>
<td>- ‘fight’</td>
<td></td>
<td>boj-ov-n-ý ‘fighting’</td>
</tr>
<tr>
<td>střel-ec</td>
<td>stříl-e-t</td>
<td>střel-i-t ‘shoot’</td>
<td>střel-a ‘shot’</td>
<td></td>
</tr>
<tr>
<td>kup-ec</td>
<td>kup-ova-t</td>
<td>koup-i-t ‘purchase’</td>
<td>koup-če ‘purchase’</td>
<td></td>
</tr>
</tbody>
</table>
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

<table>
<thead>
<tr>
<th>Suffix</th>
<th>-tel</th>
<th>-č</th>
<th>-ník/-ík</th>
<th>-ář/-ař</th>
<th>-ce</th>
<th>-ák</th>
<th>-ec</th>
<th>-čí</th>
<th>Σ</th>
</tr>
</thead>
<tbody>
<tr>
<td>Count</td>
<td>426</td>
<td>388</td>
<td>106</td>
<td>96</td>
<td>66</td>
<td>50</td>
<td>32</td>
<td>14</td>
<td>1,178</td>
</tr>
</tbody>
</table>

- 20 features assumed as potentially relevant for modeling the rivalry
  (Strnadoňová 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

<table>
<thead>
<tr>
<th>Feature</th>
<th>Value</th>
</tr>
</thead>
<tbody>
<tr>
<td>válečník válčit válka válečný</td>
<td>warrior make war war.n war.adj</td>
</tr>
<tr>
<td>target_noun_suffix</td>
<td>-ník/-ík</td>
</tr>
<tr>
<td>root_final</td>
<td>č</td>
</tr>
<tr>
<td>root_final_cvs</td>
<td>consonant</td>
</tr>
<tr>
<td>root_final_vertical</td>
<td>africate</td>
</tr>
<tr>
<td>root_final_horizontal</td>
<td>postalveolar</td>
</tr>
<tr>
<td>number_prefixes</td>
<td>0</td>
</tr>
<tr>
<td>v1_theme</td>
<td>i</td>
</tr>
<tr>
<td>v1_aspect</td>
<td>imp</td>
</tr>
<tr>
<td>v1_conjug</td>
<td>4</td>
</tr>
<tr>
<td>v1_suf_asp_counterpart</td>
<td>no</td>
</tr>
<tr>
<td>v2_theme</td>
<td>–</td>
</tr>
<tr>
<td>v2_aspect</td>
<td>–</td>
</tr>
<tr>
<td>v2_conjug</td>
<td>–</td>
</tr>
<tr>
<td>paradigm_type</td>
<td>NNA-V-</td>
</tr>
<tr>
<td>inanim_noun</td>
<td>no</td>
</tr>
<tr>
<td>freq_parent_noun</td>
<td>25,895</td>
</tr>
<tr>
<td>freq_parent_adj</td>
<td>4,953</td>
</tr>
<tr>
<td>freq_parent_v1</td>
<td>499</td>
</tr>
<tr>
<td>freq_parent_v2</td>
<td>–</td>
</tr>
<tr>
<td>freq_slots</td>
<td>VAN</td>
</tr>
</tbody>
</table>

{sevcikova|kyjanek}@ufal.mff.cuni.cz

Agent suffix rivalry in Czech
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

<table>
<thead>
<tr>
<th>Suffix</th>
<th>all</th>
<th>-tel</th>
<th>-č</th>
<th>-ník/-ík</th>
<th>-ář/-ař</th>
<th>-ce</th>
<th>-ák</th>
<th>-ec</th>
<th>-čí</th>
</tr>
</thead>
<tbody>
<tr>
<td>Instances</td>
<td>233</td>
<td>85</td>
<td>77</td>
<td>21</td>
<td>19</td>
<td>13</td>
<td>10</td>
<td>6</td>
<td>2</td>
</tr>
<tr>
<td>Precision</td>
<td>0.28</td>
<td>0.43</td>
<td>0.32</td>
<td>0.10</td>
<td>0.07</td>
<td>0.08</td>
<td>0.04</td>
<td>0.04</td>
<td>0.04</td>
</tr>
<tr>
<td>Recall</td>
<td>0.13</td>
<td>0.14</td>
<td>0.10</td>
<td>0.14</td>
<td>0.11</td>
<td>0.23</td>
<td>0.10</td>
<td>0.17</td>
<td>0.50</td>
</tr>
<tr>
<td>F-score</td>
<td><strong>0.16</strong></td>
<td>0.21</td>
<td>0.16</td>
<td>0.12</td>
<td>0.09</td>
<td>0.12</td>
<td>0.05</td>
<td>0.06</td>
<td>0.07</td>
</tr>
</tbody>
</table>
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
  ```python
classifier_LR = LogisticRegression(
    multi_class='multinomial',
    class_weight='balanced',
    solver='newton-cg',
    penalty='l2',
    C=1e30)
```

- Decision trees
  ```python
classifier = DecisionTreeClassifier(
    criterion='entropy',
    class_weight='balanced',
    splitter='best',
    max_depth=10)
```
Experimenting with all features: F-score on hold-out data

![Graph showing F-scores for different agent suffixes.]

- ALL SUFFIXES (w. average)
- -tel (85)
- -č (77)
- -nik/-ik (21)
- -ár/-ař (19)
- -ce (13)
- -ák (10)
- -ec (6)
- -cl (2)

F-score results for different agent suffixes using logistic regression (LR) and decision tree (DT) models. The highest F-score is observed for -č (77) followed by -nik/-ik (21) and -tel (85).
Experimenting with feature subsets: Subsets A to D

- **A:** the motivating verb(s): root’s final character and theme
  
  \[
  \text{root\_final, root\_final\_cvs, root\_final\_vertical, root\_final\_horizontal, v1\_theme, v2\_theme} \]

- **B:** the motivating verb(s): number of prefixes, theme, aspect, conjugation class
  
  \[
  \text{number\_prefixes, v1\_theme, v1\_aspect, v1\_conjug, v2\_theme, v2\_aspect, v2\_conjug} \]

- **C:** the derivational paradigm: which motivating items available?, does the verb have a suffixed aspectual counterpart?, does an inanimate homonym exist?
  
  \[
  \text{paradigm\_type, v1\_suf\_asp\_counterpart, inanim\_noun} \]

- **D:** corpus frequency of the motivating items
  
  \[
  \text{freq\_parent\_noun, freq\_parent\_adj, freq\_parent\_v1, freq\_parent\_v2, freq\_slots} \]
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

---

{sevcikova|kyjanek}@ufal.mff.cuni.cz
Agent suffix rivalry in Czech
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 (píják ‘drunkard x blotter’)
- **subset A (root & themes)** not sufficient

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

- Daneš, F. et al. 1967. Tvoření slov v češtině 2: Odvozování podstatných jmen. ČSAV.
- Štekauer, P. et al. (eds.). 2012. Word-Formation in the World’s Languages. CUP.

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).

{sevcikova|kyjanek}@ufal.mff.cuni.cz

Agent suffix rivalry in Czech