Overabundance as hybrid inflection
Quantitative evidence from Czech

Matías Guzmán Naranjo and Olivier Bonami

09-11.11.2016, Mannheim
1. Overabundance

2. The Czech system

3. Materials

4. Methodology

5. Results
   - Singular locative
   - Overabundance as hybrid inflection
   - Instrumental plural as sociolinguistic variation
Defining overabundance

Overabundance: two different words in free variation fill the same cell in an inflectional paradigm.

- **Example:** Spanish \texttt{SBJV.IMP.3SG} \texttt{canta-ra} vs. \texttt{canta-se}

Not to be confused with:

1. **Extended (multiple) exponence:** two separate exponents realizing the same features within the same word.
   - **Example:** French \texttt{FUT.3PL} \texttt{chant-er-ont}

2. **Heteroclisis:** one lexeme uses a paradigm that is a mix of two inflection classes
   - **Example:** Czech neuter nouns

\begin{center}
\begin{tabular}{lccc}
 & \textquote{town} & \textquote{chicken} & \textquote{sea} \\
\texttt{NOM.SG} & měst-o & kuř-e & moř-e \\
\texttt{NOM.PL} & měst-a & kuřat-a & moř-e \\
\end{tabular}
\end{center}
Overabundance and morphological theory

- The phenomenon was mostly ignored by morphologists until the pioneering work of Thornton, (2011, 2012).
- Few efforts to date to accommodate overabundance within morphological theory (see Bonami and Stump, in press for a sketchy proposal).
- The conceptual characterization of overabundance is still unclear. In particular:
  1. Do overabundant lexemes belong to discrete classes, contrasting with nonoverabundant inflection classes? Or is morphological realization inherently variable (Aronoff and Lindsay, 2016)?
  2. How are competing inflection strategies distributed?
     ▪ Given that a lexeme is overabundant, are there linguistic/extralinguistic factors governing the distribution of its alternate forms?
     ▪ Do overabundant lexemes differ in their preference for one or the other realization?
     ▪ If so, are a lexeme’s preferences predictable from its form and/or meaning?
Our project

Our goals:

1. Show that the answers to these questions are not uniform: there are different kinds of overabundance, calling for different kinds of analyses.
2. Show that, in some cases, overabundance amounts to hybridization of inflection classes: a group of lexemes forms a class that is a hybrid between two other inflection classes in that it simultaneously allows inflection strategies from both.

The method:

- We use statistical modeling to explore the distribution of inflection strategies in a large corpus.
- We focus on Czech declension for opportunistic reasons:
  1. High prevalence of overabundance
  2. Good documentation of the phenomenon (Bermel and Knittl, 2012a,b; Bermel, Knittl, and Russell, 2015; Cvrček et al., 2010)
  3. Availability of large corpora with high quality annotation through the Czech National Corpus
The data set

- We examine all nouns from the SYN2015 corpus (Křen et al., 2015), a 120M token balanced corpus of written, edited Czech documenting usage between 2010 and 2014.
- We estimate whether a lexeme is overabundant over the larger (2200M token) SYN v3 collection of corpora (Hnátková et al., 2014).
  - This diminishes the proportion of incorrect classification as non-overabundant due to data sparsity.
- Lemmatization and tagging provided with the corpus.
- Semi-automatic identification of case-number exponents.

<table>
<thead>
<tr>
<th></th>
<th>NOM.SG</th>
<th>LOC.SG</th>
</tr>
</thead>
<tbody>
<tr>
<td>‘oak tree’</td>
<td>dub</td>
<td>dubu</td>
</tr>
<tr>
<td>‘zebu’</td>
<td>zebu</td>
<td>zebu</td>
</tr>
<tr>
<td>‘cold’</td>
<td>zima</td>
<td>zimě</td>
</tr>
<tr>
<td>‘sister’</td>
<td>sestra</td>
<td>sestře</td>
</tr>
<tr>
<td>‘book’</td>
<td>kniha</td>
<td>knize</td>
</tr>
</tbody>
</table>
Almost all paradigm cells give rise to some amount of overabundance in the corpus. Some nonsystematic instances involve:

- Spelling variation, e.g. *analýza* INS.SG: *analýzou* vs. *analyzou*
- Semi-undeclinables, e.g. *whisky* INS.SG: *whisky* vs. *whiskou*

<table>
<thead>
<tr>
<th></th>
<th>NOM</th>
<th>GEN</th>
<th>DAT</th>
<th>ACC</th>
<th>VOC</th>
<th>LOC</th>
<th>INS</th>
</tr>
</thead>
<tbody>
<tr>
<td>SG</td>
<td>0.0179</td>
<td>0.0135</td>
<td>0.0219</td>
<td>0.0127</td>
<td>0.0045</td>
<td>0.0111</td>
<td>0.0097</td>
</tr>
<tr>
<td>PL</td>
<td>0.0313</td>
<td>0.0129</td>
<td>0.0046</td>
<td>0.0104</td>
<td>0.0</td>
<td>0.0088</td>
<td>0.0206</td>
</tr>
</tbody>
</table>
Example 1: the GEN.SG of masculine animate nouns

- Masculine animate nouns ending with a consonant-final NOM.SG have two possibilities in the GEN.SG:
  1. ‘hard nouns’: -a, cf. PÁN ‘sir’: pána
  2. ‘soft nouns’: -e, cf. MUŽ ‘man’: muže

- ‘Hard’ or ‘soft’ status is predictable from the phonological and morphological makeup of the stem.

- However, our corpus shows a handful of overabundant nouns (8 out of 1400), all proper names ending in /s/.

<table>
<thead>
<tr>
<th>Lexeme</th>
<th>Prop. -a</th>
<th>Lexeme</th>
<th>Prop. -a</th>
</tr>
</thead>
<tbody>
<tr>
<td>COLUMBUS</td>
<td>0.25</td>
<td>PARIS</td>
<td>0.25</td>
</tr>
<tr>
<td>SMITH</td>
<td>0.21</td>
<td>KEITH</td>
<td>0.38</td>
</tr>
<tr>
<td>JULIUS</td>
<td>0.98</td>
<td>LOS</td>
<td>0.76</td>
</tr>
<tr>
<td>JOHANNES</td>
<td>0.58</td>
<td>JACQUES</td>
<td>0.31</td>
</tr>
</tbody>
</table>

- This we call ERRATIC OVERABUNDANCE
Example 2: locative singular of hard inanimate nouns

- Masculine inanimate nouns ending in a so-called hard consonant may use two different endings in the LOC.SG: -u or -ě.
  - DUB ‘oak tree’, GEN.SG dubu
  - DŮM ‘house’, GEN.SG domě
- Many of these are overabundant. In our corpus:
  - -u only 7146
  - both 1820
  - -ě only 363

- Overabundant nouns tend to have strong preferences, but some nouns exhibit a balanced distribution.
- This is a good candidate for HYBRIDIZATION: overabundant nouns form a class of their own.
Example 3: the instrumental plural

- All Czech nouns may occur in two forms in the instrumental plural, one of which involves the sequence \(-ma\).
- Sociolinguistic conditioning: the \(-ma\) form is informal.
  - In particular, it is unexpected in writing.
- The distribution of overabundant forms in our corpus is as expected, given its stylistic makeup.

<table>
<thead>
<tr>
<th>MÚŽ ‘man’: muži~mužema</th>
<th>ŽENA ‘woman’: ženami~ženama</th>
<th>MĚSTO ‘town’: městy~městama</th>
</tr>
</thead>
<tbody>
<tr>
<td>Only non-(ma)</td>
<td>439</td>
<td></td>
</tr>
<tr>
<td>Both</td>
<td>551</td>
<td></td>
</tr>
<tr>
<td>Only (ma)</td>
<td>0</td>
<td></td>
</tr>
</tbody>
</table>
Our goal is twofold:

1. modelling the general Czech inflectional system as a proof of concept, and
2. modelling the last two particular cases (-u vs. ě in the LOC.SG, -ma vs. other forms in the INS.PL) to confirm how they contrast.

   Grammatical vs. sociolinguistic conditioning

Our model was fitted using the \texttt{nnet} (Venables and Ripley, 2002) package in R, with a softmax link function, and 10 hidden nodes. We performed ten-fold cross-validation on all of our models. The set of predictors that best fitted the data was:

\begin{verbatim}
final_segment + penultimate_segment + antepenultimate_segment + length_in_letters + number_vowels + frequency
\end{verbatim}

We did not find any improvements from adding additional factors, interactions, or hidden nodes.
Confusion matrices and accuracy measures

We make use of two basic tools for evaluating the analogical systems: Confusion matrices and accuracy measures.

Suppose we have two groups A, and B, and the following words:

A: lama, lara, lado, laso, lerr, liz
B: pama, ra, dal, kar, olor, gin, grip, wek.

We can postulate two models:
Model 1: all words starting with an ‘l’ belong to group A, all others to group B
Model 2: all words with an ‘a’ as first vowel belong to group A, all others to group B
Model 1, a perfectly predictive model, produces the following results:

A: lama, lara, lado, laso, lerr, liz
B: pama, ra, dal, kar,olor, gin, grip, wek.

<table>
<thead>
<tr>
<th>Prediction</th>
<th>A</th>
<th>B</th>
</tr>
</thead>
<tbody>
<tr>
<td>A</td>
<td>6</td>
<td>0</td>
</tr>
<tr>
<td>B</td>
<td>0</td>
<td>8</td>
</tr>
</tbody>
</table>

Accuracy: 1
95% CI: (0.7684, 1)
No Information Rate: 0.5714
Model 2, a completely unpredictive model, produces the following results:

A: lama, lara, lado, laso, pama, ra, dal, kar
B: lerr, liz, olor, gin, grip, wek.

<table>
<thead>
<tr>
<th>Prediction</th>
<th>Reference</th>
<th>A</th>
<th>B</th>
</tr>
</thead>
<tbody>
<tr>
<td>A</td>
<td>4</td>
<td>4</td>
<td></td>
</tr>
<tr>
<td>B</td>
<td>2</td>
<td>4</td>
<td></td>
</tr>
</tbody>
</table>

Accuracy : 0.5714
95% CI : (0.2886, 0.8234)
No Information Rate : 0.5714
Results

1. We first present the results of our model in the complete system for each individual cell of the paradigm.

2. The point of this initial step is to provide some evidence that inflectional class in Czech nouns is strongly correlated with the phonological shape of nouns.

3. This is not just a property of overabundant classes.
# Singular locative

<table>
<thead>
<tr>
<th>Prediction</th>
<th>Reference</th>
<th>i</th>
<th>é</th>
<th>é</th>
<th>0</th>
<th>0-u</th>
<th>u</th>
<th>é-u</th>
<th>i-u</th>
<th>ovi-u</th>
<th>ovi</th>
<th>m</th>
<th>i-ovi</th>
<th>ém</th>
<th>ti</th>
<th>tu</th>
<th>ý</th>
<th>é-ý</th>
</tr>
</thead>
<tbody>
<tr>
<td>i</td>
<td>6692</td>
<td>27</td>
<td>0</td>
<td>92</td>
<td>1</td>
<td>13</td>
<td>2</td>
<td>0</td>
<td>1</td>
<td>20</td>
<td>4</td>
<td>19</td>
<td>2</td>
<td>5</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td></td>
</tr>
<tr>
<td>é</td>
<td>41</td>
<td>6834</td>
<td>2</td>
<td>20</td>
<td>0</td>
<td>23</td>
<td>23</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>1</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td></td>
</tr>
<tr>
<td>éé</td>
<td>0</td>
<td>4</td>
<td>471</td>
<td>6</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>1</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>26</td>
</tr>
<tr>
<td>0</td>
<td>71</td>
<td>6</td>
<td>1</td>
<td>9353</td>
<td>14</td>
<td>31</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>3</td>
<td>9</td>
<td>0</td>
<td>0</td>
<td>14</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td></td>
</tr>
<tr>
<td>0-u</td>
<td>3</td>
<td>8</td>
<td>0</td>
<td>335</td>
<td>29</td>
<td>287</td>
<td>12</td>
<td>0</td>
<td>1</td>
<td>1</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td></td>
</tr>
<tr>
<td>u</td>
<td>23</td>
<td>59</td>
<td>0</td>
<td>79</td>
<td>11</td>
<td>7707</td>
<td>170</td>
<td>12</td>
<td>4</td>
<td>10</td>
<td>0</td>
<td>1</td>
<td>0</td>
<td>10</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td></td>
</tr>
<tr>
<td>ééú</td>
<td>6</td>
<td>768</td>
<td>0</td>
<td>30</td>
<td>7</td>
<td>1438</td>
<td>421</td>
<td>5</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>2</td>
<td>0</td>
</tr>
<tr>
<td>iú</td>
<td>25</td>
<td>0</td>
<td>0</td>
<td>2</td>
<td>1</td>
<td>44</td>
<td>1</td>
<td>5</td>
<td>1</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>1</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td></td>
</tr>
<tr>
<td>oviú</td>
<td>14</td>
<td>0</td>
<td>0</td>
<td>9</td>
<td>0</td>
<td>886</td>
<td>0</td>
<td>1</td>
<td>602</td>
<td>2276</td>
<td>6</td>
<td>14</td>
<td>0</td>
<td>1</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td></td>
</tr>
<tr>
<td>oviú</td>
<td>23</td>
<td>2</td>
<td>0</td>
<td>19</td>
<td>0</td>
<td>25</td>
<td>0</td>
<td>0</td>
<td>14</td>
<td>830</td>
<td>17</td>
<td>7</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td></td>
</tr>
<tr>
<td>m</td>
<td>10</td>
<td>0</td>
<td>0</td>
<td>25</td>
<td>0</td>
<td>3</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>30</td>
<td>282</td>
<td>1</td>
<td>2</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td></td>
</tr>
<tr>
<td>i-ovi</td>
<td>340</td>
<td>0</td>
<td>0</td>
<td>1</td>
<td>0</td>
<td>5</td>
<td>0</td>
<td>0</td>
<td>4</td>
<td>221</td>
<td>4</td>
<td>50</td>
<td>2</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td></td>
</tr>
<tr>
<td>éém</td>
<td>1</td>
<td>0</td>
<td>1</td>
<td>2</td>
<td>0</td>
<td>3</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>3</td>
<td>6</td>
<td>0</td>
<td>180</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td></td>
</tr>
<tr>
<td>tiú</td>
<td>4</td>
<td>0</td>
<td>0</td>
<td>7</td>
<td>0</td>
<td>2</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>16</td>
<td>0</td>
<td>0</td>
</tr>
<tr>
<td>tuú</td>
<td>0</td>
<td>1</td>
<td>0</td>
<td>4</td>
<td>1</td>
<td>5</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>29</td>
<td>0</td>
<td>0</td>
</tr>
<tr>
<td>ý</td>
<td>0</td>
<td>0</td>
<td>3</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
</tr>
<tr>
<td>ééý</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
</tr>
</tbody>
</table>
Statistics for the singular locative

Overall Statistics

Accuracy : 0.8105
95% CI : (0.8066, 0.8142)
No Information Rate : 0.2533
P-Value [Acc > NIR] : < 2.2e-16
Kappa : 0.7716
Clustering singular locative

Dendrogram with negative correlation distance
Interim summary

1. For all three cases the accuracy of the models was well above random chance.
2. Most of the errors were due to overabundance
Modelling overabundance

- Here now we focus on the ě-u alternation specifically and try to distinguish those nouns that only take -ě, nouns that only take -u, and overabundant nouns.

- To control for the possibility of false negatives (failing to see a noun appear with -u does not mean it only appears with -ě we make use of two corpora, the SYN2015 and the larger SYN data-set.
Results for the -ě/-u classes

<table>
<thead>
<tr>
<th>Prediction</th>
<th>Reference</th>
<th>-ě/-u</th>
<th>u</th>
<th>-ě</th>
</tr>
</thead>
<tbody>
<tr>
<td>-ě/-u</td>
<td></td>
<td>678</td>
<td>86</td>
<td>176</td>
</tr>
<tr>
<td>u</td>
<td></td>
<td>137</td>
<td>507</td>
<td>5</td>
</tr>
<tr>
<td>-ě</td>
<td></td>
<td>174</td>
<td>2</td>
<td>181</td>
</tr>
</tbody>
</table>
Results

Overall Statistics

Accuracy : 0.702
95% CI : (0.6811, 0.7222)
No Information Rate : 0.5082
P-Value [Acc > NIR] : < 2.2e-16
Kappa : 0.518
This is what we expect to see if the grammatical system treats overabundant nouns to be hybridization between -ě and -u nouns. Our system classifies nouns on the basic idea of nouns like look alike behave alike. The overabundant cases inherit from both types, and thus look like either of both types, leading to higher confusability.
Overabundance as hybridization

- This situation is readily accounted for within a view of inflection class systems as semi-lattices of subclasses and superclasses.

```
classes
  /     \
 /       /
A       B
  \
...

C

  \
...
```

- Can readily be modeled in frameworks that rely on multiple inheritance hierarchies (Boas and Sag, 2012; Brown and Hippisley, 2012; Pollard and Sag, 1994).
- Convergence with abstractive modeling of inflection class systems using formal concept analysis (Beniamine and Bonami, 2016).
A different kind of overabundance

- The plural instrumental presents systematic and lexically unrestricted overabundance between the forms: -ama and -y, -ama and -ami, -ema and -emi, -ma and -mi, and -ema and -i

- Overabundance seems to be sociolinguistically and stylistically conditioned

- If this is a fundamentally different kind of overabundance, we expect our models to perform in radically different ways (ie. not very well)
## Plural instrumental

<table>
<thead>
<tr>
<th>Prediction</th>
<th>ami</th>
<th>mi</th>
<th>emi</th>
<th>ama</th>
<th>ema</th>
<th>ma</th>
<th>y</th>
<th>i</th>
<th>y-ama</th>
<th>ami-emi-mi-ama ema ma</th>
<th>i-ema</th>
</tr>
</thead>
<tbody>
<tr>
<td>ami</td>
<td>172</td>
<td>0</td>
<td>3</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>5</td>
<td>1</td>
<td>0</td>
<td>3</td>
<td>0</td>
</tr>
<tr>
<td>mi</td>
<td>0</td>
<td>33</td>
<td>2</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>3</td>
<td>1</td>
<td>0</td>
<td>0</td>
<td>0</td>
</tr>
<tr>
<td>emi</td>
<td>0</td>
<td>3</td>
<td>67</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>3</td>
</tr>
<tr>
<td>ama</td>
<td>1</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>4</td>
<td>1</td>
<td>326</td>
<td>210</td>
<td>3</td>
</tr>
<tr>
<td>ema</td>
<td>0</td>
<td>0</td>
<td>1</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>5</td>
<td>1</td>
<td>78</td>
</tr>
<tr>
<td>ma</td>
<td>0</td>
<td>0</td>
<td>2</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>1</td>
<td>3</td>
<td>5</td>
<td>22</td>
</tr>
<tr>
<td>y</td>
<td>0</td>
<td>1</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>458</td>
<td>9</td>
<td>11</td>
<td>0</td>
<td>0</td>
</tr>
<tr>
<td>i</td>
<td>0</td>
<td>1</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>1</td>
<td>96</td>
<td>0</td>
<td>0</td>
<td>0</td>
</tr>
<tr>
<td>y-ama</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>7</td>
<td>1</td>
<td>0</td>
<td>0</td>
<td>0</td>
</tr>
<tr>
<td>ami-ama</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>1</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
</tr>
<tr>
<td>emi-ema</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
</tr>
<tr>
<td>mi-ma</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
</tr>
<tr>
<td>i-ema</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
</tr>
</tbody>
</table>
Overall Statistics

Accuracy : 0.5127
95% CI : (0.488, 0.5374)
No Information Rate : 0.2973
P-Value [Acc > NIR] : < 2.2e-16
Kappa : 0.4532
Assessing the role of overabundance

- The preceding model suggests that it is quite hard to predict the behavior in the instrumental plural from properties of the lemma.
- Possible causes:
  1. Predicting overabundance is hard.
  2. Predicting possible exponents (irrespective of whether they are overabundant or not) is hard.
  3. Both are hard.
- To tell these hypotheses apart, we construct a new dataset where overabundant lexemes are grouped together with lexemes exhibiting only one of the two forms.
- Thus the effect of overabundance is neutralized in this dataset.
### Statistics plural instrumental after collapsing classes

<table>
<thead>
<tr>
<th>Prediction</th>
<th>ami:ama</th>
<th>mi:ma</th>
<th>emi:ema</th>
<th>y:ama</th>
<th>i:ema</th>
</tr>
</thead>
<tbody>
<tr>
<td>ami:ama</td>
<td>388</td>
<td>1</td>
<td>3</td>
<td>0</td>
<td>0</td>
</tr>
<tr>
<td>mi:ma</td>
<td>1</td>
<td>59</td>
<td>5</td>
<td>0</td>
<td>1</td>
</tr>
<tr>
<td>emi:ema</td>
<td>1</td>
<td>3</td>
<td>155</td>
<td>0</td>
<td>0</td>
</tr>
<tr>
<td>y:ama</td>
<td>0</td>
<td>1</td>
<td>0</td>
<td>814</td>
<td>11</td>
</tr>
<tr>
<td>i:ema</td>
<td>0</td>
<td>3</td>
<td>1</td>
<td>8</td>
<td>156</td>
</tr>
</tbody>
</table>
Overall Statistics

Accuracy : 0.9758
95% CI : (0.9671, 0.9827)
No Information Rate : 0.5102
P-Value [Acc > NIR] : < 2.2e-16
Kappa : 0.9631
We compare with \textit{non\_xma vs xma vs overabundant}

<table>
<thead>
<tr>
<th>Prediction</th>
<th>non_xma</th>
<th>xma</th>
<th>overabundant</th>
</tr>
</thead>
<tbody>
<tr>
<td>non_xma</td>
<td>805</td>
<td>0</td>
<td>669</td>
</tr>
<tr>
<td>xma</td>
<td>69</td>
<td>0</td>
<td>68</td>
</tr>
<tr>
<td>overabundant</td>
<td>0</td>
<td>0</td>
<td>0</td>
</tr>
</tbody>
</table>

We can see that the model distinguishes the cases without -\textit{ma}, but otherwise predicts that the rest of the cases should also be overabundant. That is, all cases seen with -\textit{ma} are predicted to also be possible with the alternative form.
The new model based on inflectional classes performs extremely well. This suggests that

- Phonological and morphosyntactic properties of the lemma do not allow to predict whether a lexeme will use a -ma form, a non-ma form, or both, in the INS.PL.
- However, they are very good predictors if which -ma form (resp. which non-ma form) is used.
- Thus overabundance is not predictable, but inflection class is highly predictable.

This is in stark contrast with the situation in the LOC.SG, where we saw that overabundance was indeed predictable on the basis of grammatical information.
We have shown that:

- Overabundance is not a single homogeneous phenomenon, but there are multiple different types.
- One of these types of overabundance can be analyzed as hybridization of two different inflectional classes.
- We find statistical evidence for this analysis in the form of models that predict inflectional class on the basis of phonological shape.
- Quantitatively, sociolinguistic overabundance behaves very differently from hybrid-class overabundance.
Děkuji, gracias, merci...


