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 oveabundance

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

- Example: Spanish SBJV.IMP.3SG *canta-ra* vs. *canta-se*

Not to be confused with:

1. Extended (multiple) exponence: two separate exponents realizing the same features within the same word.
   - Example: French FUT.3PL *chant-er-ont*
2. Heteroclisis: one lexeme uses a paradigm that is a mix of two inflection classes
   - Example: Czech neuter nouns

	'town'	'chicken'	'sea'
NOM.SG	měst-o	kuř-e	moř-e
NOM.PL	měst-a	kuřat-a	moř-e

Overabundance and morphological theory

- The phenomenon was mostly ignored by morphologists until the pioneering work of Thornton, (2011, 2012).
- Few efforts to date to accomodate 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, is 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

	NOM.SG	LOC.SG
'oak tree'	dub	dubu
'zebu'	zebu	zebu

	NOM.SG	LOC.SG
'cold'	zima	zimě
'sister'	sestra	sestře
'book'	kniha	knize

Overall distribution of overabundance

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*

	NOM	GEN	DAT	ACC	VOC	LOC	INS
SG	0.0179	0.0135	0.0219	0.0127	0.0045	0.0111	0.0097
PL	0.0313	0.0129	0.0046	0.0104	0	0.0088	0.0206

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/*.

Lexeme	Prop. *-a*	Lexeme	Prop. *-a*
COLUMBUS	0.25	PARIS	0.25
SMITH	0.21	KEITH	0.38
JULIUS	0.98	LOS	0.76
JOHANNES	0.58	JACQUES	0.31

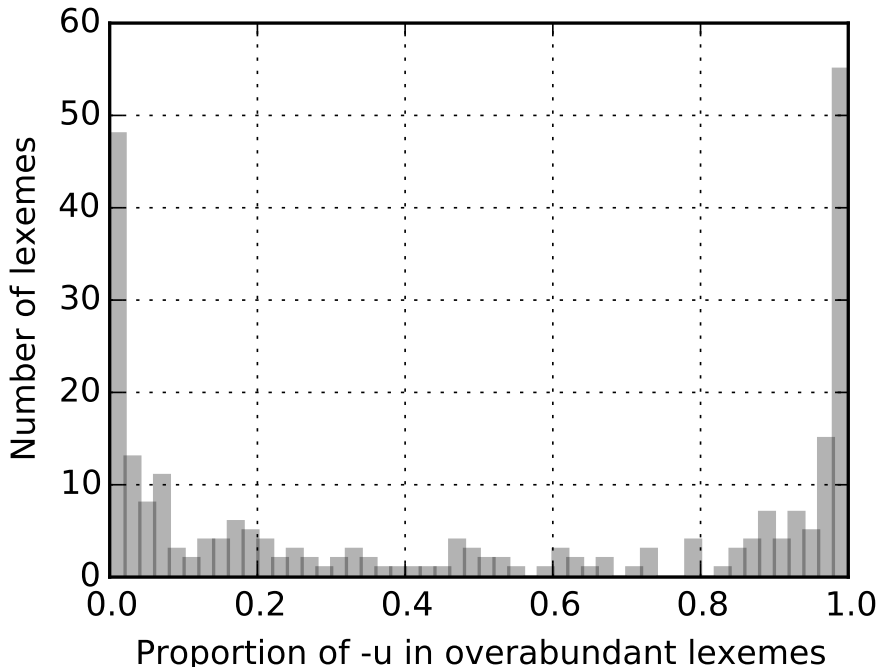
- This we call ERRATIC OVERABUNDANCE

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.

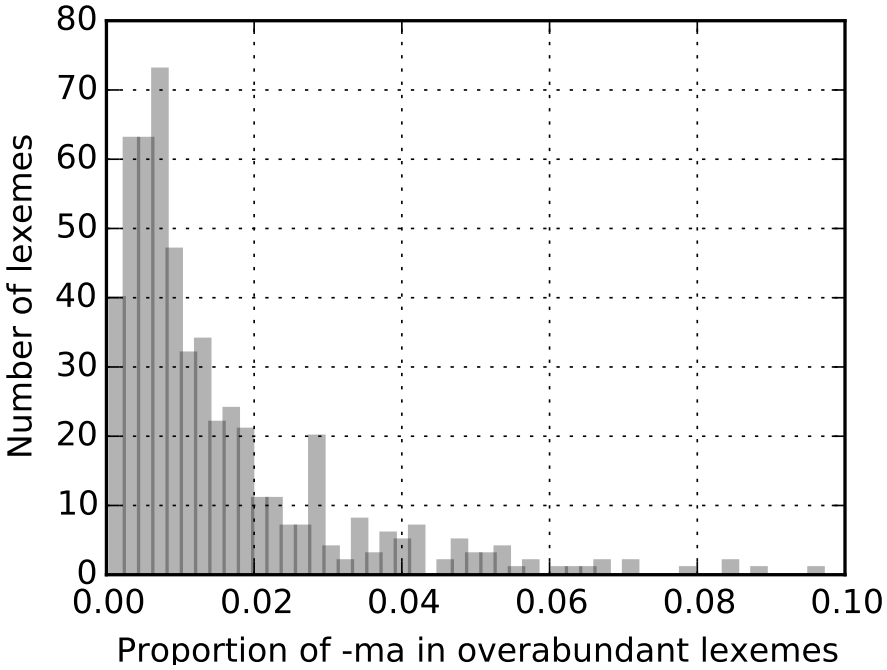


Example 3: the instrumental plural

- 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.

MUŽ 'man': *muži*∼*mužema*
ŽENA 'woman': *ženami*∼*ženama*
MĚSTO 'town': *městy*∼*městama*

Only non-*ma*	439
Both	551
Only *ma*	0



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 `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:

```
final_segment + penultimate_segment +
antepenultimate_segment + length_in_letters + number_vowels
+ frequency
```

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.

	Reference	
Prediction	A	B
A	6	0
B	0	8

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.

	Reference	
Prediction	A	B
A	4	4
B	2	4

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

Prediction \ Reference	i	ě	é	0	0-u	u	ě-u	i-u	ovi-u	ovi	m	i-ovi	ém	ti	tu	ý	é-ý
i	6692	27	0	92	1	13	2	0	1	20	4	19	2	5	0	0	0
ě	41	6834	2	20	0	23	23	0	0	0	0	0	1	0	0	0	0
é	0	4	471	6	0	0	0	0	0	1	0	0	0	0	0	0	26
0	71	6	1	9353	14	31	0	0	0	3	9	0	0	14	0	0	0
0-u	3	8	0	335	29	287	12	0	1	1	0	0	0	0	0	0	0
u	23	59	0	79	11	7707	170	12	4	10	0	1	0	0	1	0	0
ě-u	6	768	0	30	7	1438	421	5	0	0	0	0	0	0	2	0	0
i-u	25	0	0	2	1	44	1	5	1	0	0	0	0	1	0	0	0
ovi-u	14	0	0	9	0	886	0	1	602	2276	6	14	0	1	0	0	0
ovi	23	2	0	19	0	25	0	0	14	830	17	7	0	0	0	0	0
m	10	0	0	25	0	3	0	0	0	30	282	1	2	0	0	0	0
i-ovi	340	0	0	1	0	5	0	0	4	221	4	50	2	0	0	0	0
ém	1	0	1	2	0	3	0	0	0	3	6	0	180	0	0	0	0
ti	4	0	0	7	0	2	0	0	0	0	0	0	0	16	0	0	0
tu	0	1	0	4	1	5	0	0	0	0	0	0	0	0	29	0	0
ý	0	0	3	0	0	0	0	0	0	0	0	0	0	0	0	0	0
é-ý	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0

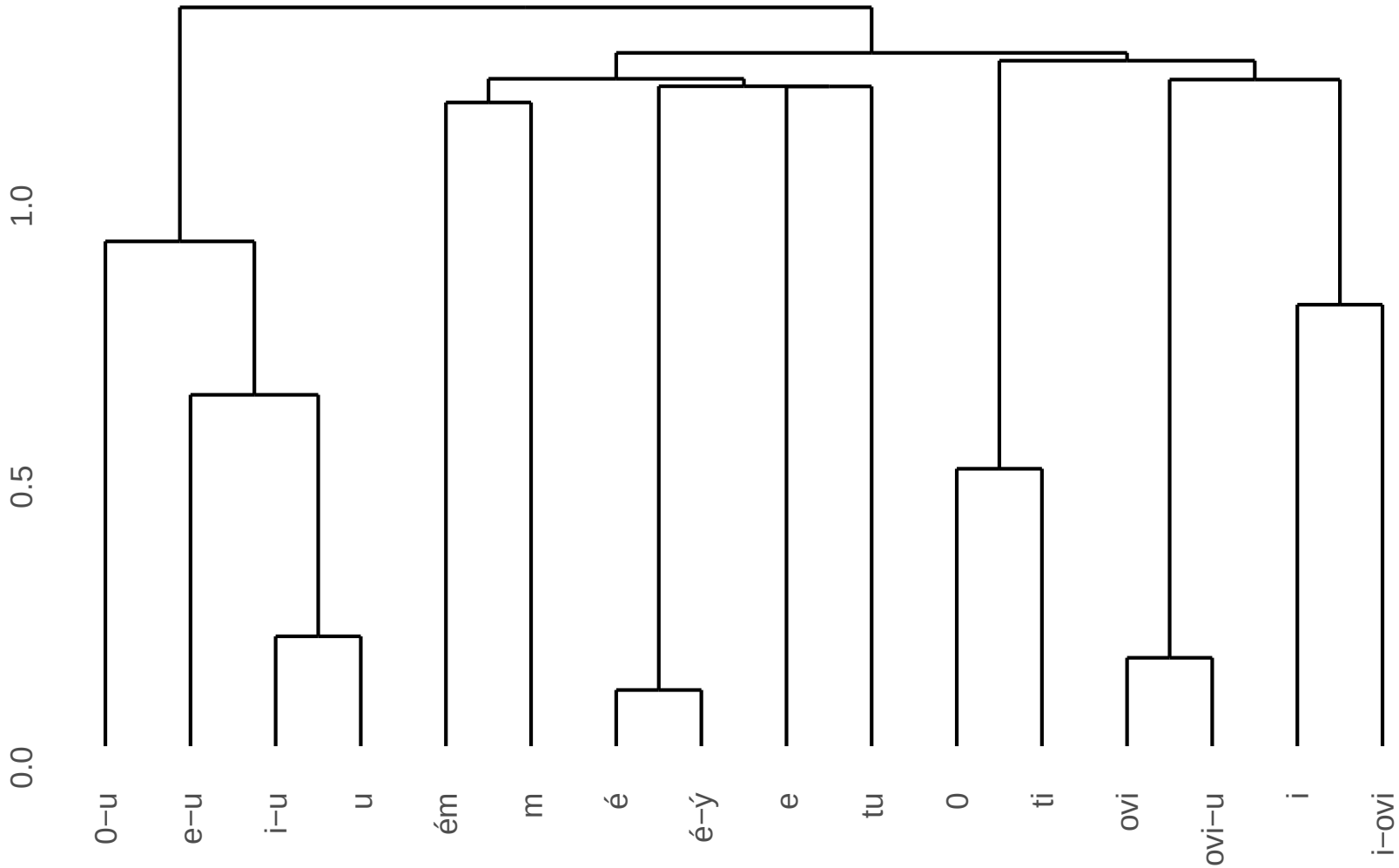
Statistics for the singular locative

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

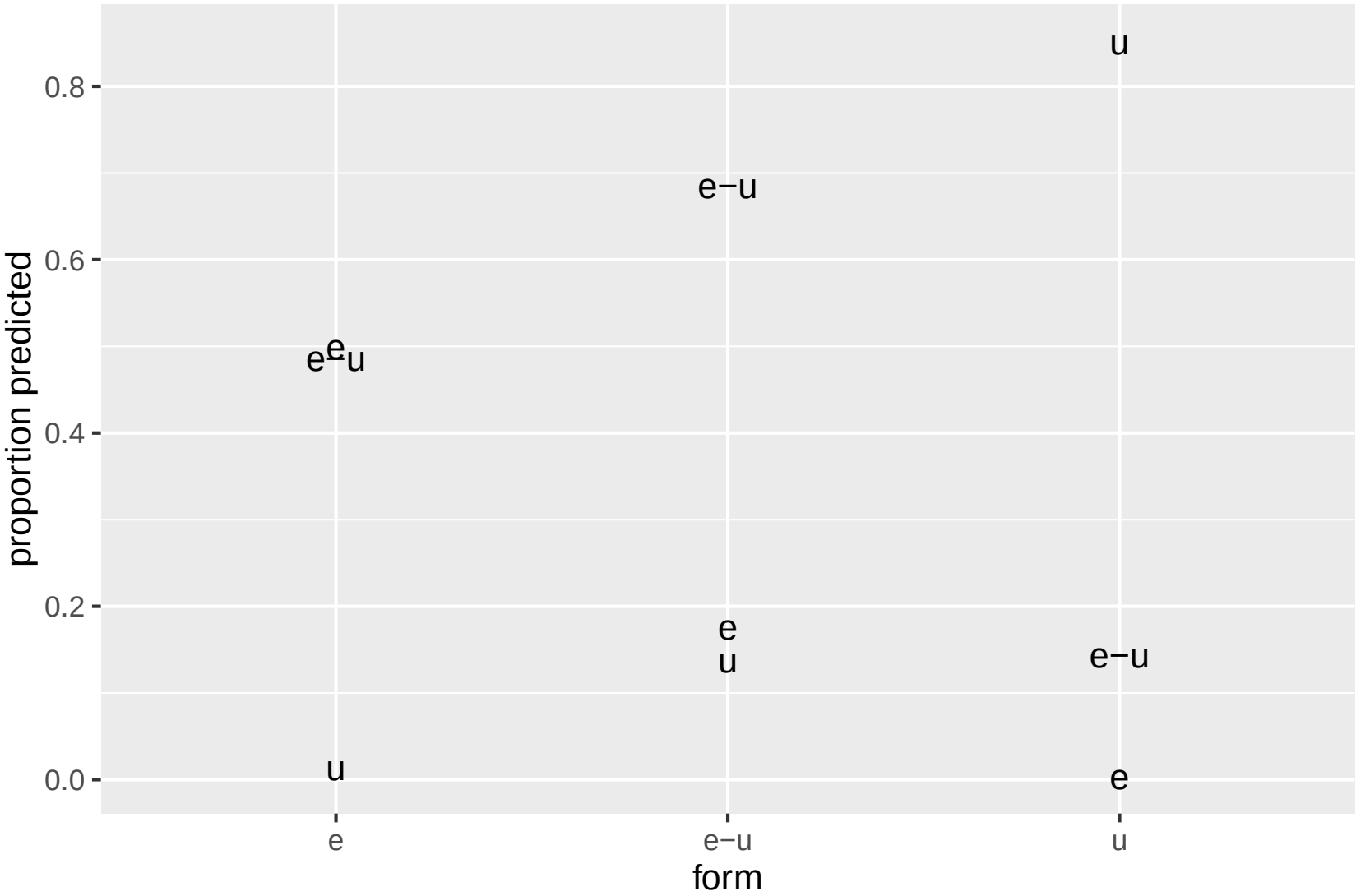
- Here now we focus on the ě-u alternation specifically and try to distinguish those nouns that only take -*ě*, nouns that only take -*u*, and overabundant nouns.
- To control for the possibility of false negatives (failing to see a noun appear with -*u* does not mean it only appears with -*ě* we make use of two corpora, the SYN2015 and the larger SYN data-set.

Results for the -*ě*/-*u* classes

	Reference		
Prediction	-ě/-u	u	-ě
-ě/-u	678	86	176
u	137	507	5
-ě	174	2	181

Overall Statistics

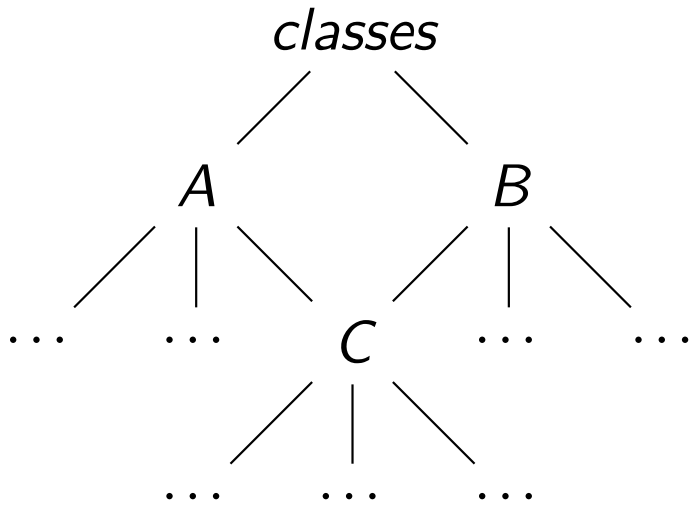
Accuracy : 0.702
95% CI : (0.6811, 0.7222)
No Information Rate : 0.5082
P-Value [Acc > NIR] : < 2.2e-16
Kappa : 0.518


0.8
0.6
0.4
0.2
0.0
proportion predicted
e
e–u
u
form

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.



- 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 sociolinguisticaly 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

	Reference												
Prediction	ami	mi	emi	ama	ema	ma	y	i	y-ama	ami-ama	emi-ema	mi-ma	i-ema
ami	172	0	3	0	0	0	5	1	0	3	0	0	0
mi	0	33	2	0	0	0	3	1	0	0	0	0	0
emi	0	3	67	0	0	0	0	0	0	0	3	2	0
ama	1	0	0	0	0	0	4	1	326	210	3	2	2
ema	0	0	1	0	0	0	0	0	5	1	78	3	54
ma	0	0	2	0	0	0	0	0	1	3	5	22	2
y	0	1	0	0	0	0	458	9	11	0	0	0	1
i	0	1	0	0	0	0	1	96	0	0	0	0	0
y-ama	0	0	0	0	0	0	7	1	0	0	0	0	0
ami-ama	0	0	0	0	0	0	1	0	0	0	0	0	0
emi-ema	0	0	0	0	0	0	0	0	0	0	0	0	0
mi-ma	0	0	0	0	0	0	0	0	0	0	0	0	0
i-ema	0	0	0	0	0	0	0	0	0	0	0	0	0

Statistics plural instrumental

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

	Reference				
Prediction	ami:ama	mi:ma	emi:ema	y:ama	i:ema
ami:ama	388	1	3	0	0
mi:ma	1	59	5	0	1
emi:ema	1	3	155	0	0
y:ama	0	1	0	814	11
i:ema	0	3	1	8	156

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 *non_xma* vs *xma* vs overabundant

	Reference		
Prediction	non_xma	xma	overabundant
non_xma	805	0	669
xma	69	0	68
overabundant	0	0	0

We can see that the model distinguishes the cases without *-ma*, but otherwise predicts that the rest of the cases should also be overabundant. That is, all cases seen with *-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.

Concluding remarks

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...

Aronoff, Mark and Mark Lindsay (2016). "Partial organization in languages: la langue est un système où la plupart se tient". In: *Proceedings of the 8th Décembrettes*. Ed. by Sandra Augendre et al. CLLE-ERSS. Toulouse, pp. 1–14.

Beniamine, Sarah and Olivier Bonami (2016). "A comprehensive view on inflectional classification". Paper read at the LAGB Meeting, September 2016.

Bermel, Neil and Luděk Knittl (2012a). "Corpus frequency and acceptability judgments: A study of morphosyntactic variants in Czech". In: *Corpus Linguistics and Linguistic Theory* 8, pp. 241–275.

– (2012b). "Morphosyntactic variation and syntactic constructions in Czech nominal declension: corpus frequency and native-speaker judgments". In: *Russian Linguistics* 36, pp. 91–119.

Bermel, Neil, Luděk Knittl, and Jean Russell (2015). "Morphological variation and sensitivity to frequency of forms among native speakers of Czech". In: *Russian Linguistics* 39, pp. 283–308.

Boas, Hans and Ivan A. Sag, eds. (2012). *Sign-Based Construction Grammar*. Stanford: CSLI Publications.

Bonami, Olivier and Gregory T. Stump (in press). "Paradigm Function Morphology". In: *Cambridge Handbook of Morphology*. Ed. by Andrew Hippisley and Gregory T. Stump. Cambridge: Cambridge University Press.

Brown, Dunstan and Andrew Hippisley (2012). *Network Morphology: a defaults based theory of word structure*. Cambridge: Cambridge University Press.

Cvrček, Václav et al. (2010). *Mluvince soudobé češtiny*. Vol. 1. Prague: Karolinum.

Hnátková, M. et al. (2014). "The SYN-series corpora of written Czech". In: *Proceedings of the Ninth International Conference on Language Resources and Evaluation*, pp. 160–164.

Křen, Michal et al. (2015). *SYN2015: reprezentativní korpus psané češtiny*. Tech. rep. Ústav Českého národního korpusu FF UK, Praha.

Pollard, Carl and Ivan A. Sag (1994). *Head-driven Phrase Structure Grammar*. Stanford: CSLI Publications; The University of Chicago Press.

Thornton, Anna M. (2011). "Overabundance (Multiple Forms Realizing the Same Cell): A Non-Canonical Phenomenon in Italian Verb Morphology". In: *Morphological Autonomy: Perspectives from Romance Inflectional Morphology*. Ed. by Martin Maiden et al. Oxford: Oxford University Press, pp. 358–381.

– (2012). "Reduction and maintenance of overabundance. A case study on Italian verb paradigms". In: *Word Structure* 5, pp. 183–207.

Venables, William N. and Brian D. Ripley (2002). *Modern Applied Statistics with S*. Fourth. New York: Springer.