Overabundance as hybrid inflection. Quantitative evidence from Czech

Matías Guzmán Naranjo (Universität Leipzig) and Olivier Bonami (Université Paris Diderot)

Conventional wisdom holds that there is a unique form filling each cell of a lexeme's paradigm. When this does not hold, the lexeme is said to be in that paradigm cell (Thornton, 2012). Table 1 provides some relevant examples from Czech declension. Although descriptive grammars have always acknowledged situations of overabundance, theoretical work in morphology has only started addressing the issue very recently.

On the face of it, overabundance may correspond to a wide variety of situations. We focus on cases where overabundance can be seen as competition between two inflection strategies, and identify four situations leading to overabundance:

1. Unconditional free variation: all lexemes compatible with one of the two strategies are compatible with both, with no lexical conditioning of the distribution of alternants.
2. Lexically gradient overabundance: lexemes have random individual preferences as to the use of the two strategies, filling up the space of possible distributions of alternants.
3. Overabundance as noise: all lexemes normally use a single strategy, but random errors lead to small amounts of use of the alternate strategy, leading to overabundance.
4. Overabundance as hybridization: some, but not all, lexemes are definitely compatible with both strategies, and form an identifiable hybrid class.

We believe that all four situations are likely to be attested in some systems. In this paper we show that situation 4 is attested in the Czech locative singular. We focus on the alternation between -ě and -u in hard inanimate nouns, whose syntactic and sociolinguistic distribution have previously been studied by Bermel and Knittl (2012); Bermel et al. (2015). Figure 1 already suggests that we are dealing with situation 4: different lexemes clearly have different preferences, but there are high frequency lexemes with no strong preference; hence we are not in situation 3.

Recent work has focused on the related issue of computationally modelling affix rivalry in derivational morphology (Arndt-Lappe, 2014; Gouskova et al., 2015) from an analogical perspective. These approaches assume that analogy occurs parallel to the grammar with little or no interaction between the two systems and hence make no prediction as to the influence of particular grammatical features. We also pursue an analogical approach to the problem, but explore the influence of specific linguistic features on classification through the use of a neural network.

We use data from the SYN2015 portion of the Czech National Corpus (Křen et al., 2015). We extracted all nouns in the locative singular from the SYN2015 corpus, but estimated whether lexemes are overabundant from the larger SYN corpus (Hnátková et al., 2014). After this we ended up with 9336 nouns with the following suffix distribution: -ě : 363, -ě/u: 1821, -u: 7152.

We fitted a neural network using the nnet package in R, with the softmax link function, one hidden layer and 10 neurons. For the outcomes we set three classes: -ě, -u and -ě/u. The predictors we chose were the last three segments of the nominative singular, its length (in letters), the number of vowels, and the lexeme frequency.

Table 2 presents the results of the model. The first thing we can see is that predicting inflection class of Czech nouns is possible and relatively easy from an analogical perspective. If we consider that our predictors are fairly rough, the high accuracy means that there is a very clear systematicity to the way different endings are chosen. But what is more interesting is where most of the errors lie. There is little confusion between words that exclusively take -ě and words that exclusively take -u, but high confusion between those that take -ě and -u, and those that exclusively take -ě or -u. In other words, there is little similarity between the non-overabundant words, but the overabundant words are similar to the non-overabundant. This strongly suggests that the overabundant class is a true hybrid of two other classes rather than just another ordinary inflection class.

In the talk we will put these results in perspective by looking at other cells in the Czech nominal paradigm and showing how they instantiate the different types of overabundance.

Lexeme	*-ě* form	*-u* form	Prop. *u*	Freq. *-ě* + *-u*
severovýchod 'north-east'	severovýchodě	—	0	211
dům 'house'	domě	domu	0.01	10279
úřad 'bureau'	úřadě	úřadu	0.46	1504
důchod 'pension'	důchodě	důchodu	0.99	425
článek 'article'	—	článku	1	1625

Table 1: Overabundance in the LOC.SG in hard masculine inanimate nouns

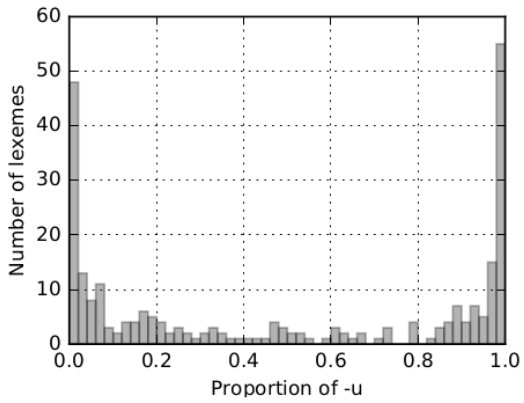

Figure 1: Distribution of overabundant LOC.SG hard masculine animate nouns by proportion of *-u* forms (SYN2015)

predicted distribution	actual distribution *u*	*ě-u*	*ě*
u	6726	451	35
ě-u	393	1233	148
ě	27	136	180
Accuracy : 0.87		Kappa : 0.65	

Table 2: Confusion matrix of analogical model

References

Arndt-Lappe, S. (2014). 'Analogy in suffix rivalry: the case of English -ity and-ness'. 18:497.

Bermel, N. and Knittl, L. (2012). 'Morphosyntactic variation and syntactic constructions in Czech nominal declension: corpus frequency and native-speaker judgments'. Russian Linguistics, 36:91–119.

Bermel, N., Knittl, L., and Russell, J. (2015). 'Morphological variation and sensitivity to frequency of forms among native speakers of Czech'. Russian Linguistics, 39:283–308.

Gouskova, M., Newlin-Łukowicz, L., and Kasyanenko, S. (2015). 'Selectional restrictions as phonotactics over sublexicons'. 167:41–81.

Hnátková, M., Křen, M., Procházka, P., and Skoumalová, H. (2014). 'The SYN-series corpora of written Czech'. In Proceedings of the Ninth International Conference on Language Resources and Evaluation. 160–164.

Křen, M., Cvrček, V., Čapka, T., Čermáková, A., Hnátková, M., Chlumská, L., Jelínek, T.,

Kováříková, D., Petkevič, V., Procházka, P., and Skoumalová, M. T. P. V. P. Z. A., H. Škrabal (2015). SYN2015: reprezentativní korpus psané češtiny.

Thornton, A. M. (2012). 'Reduction and maintenance of overabundance. A case study on Italian verb paradigms'. Word Structure, 5:183–207.