Article:

Bermel, N., Knittl, L. and Russell, J. (2015) Morphological variation and sensitivity to frequency of forms among native speakers of Czech. Russian Linguistics, 39 (3). pp. 283-308. ISSN 0304-3487

https://doi.org/10.1007/s11185-015-9149-2

Morphological variation and sensitivity to frequency of forms among native speakers of Czech[1]

Neil Bermel, Luděk Knittl and Jean Russell (University of Sheffield)

Abstract: This article looks at inter-speaker variation in two environments: the genitive and locative singular cases of masculine “hard inanimate” nouns in Czech, using a large-scale survey of native speakers that tested their preferences for certain forms and their choices. Our hypothesis that such variation exists was upheld, but only within limited parameters. Most biographical data (age, gender, education) played no role in respondents’ choices or preferences. Their region of origin played a small but significant role, although not the one expected. Relating the two types of tasks to each other, we found that respondents’ use of the ratings scale did not correlate to their choice of forms, but their overall strength of preference for one form over another did correlate with their choices. Inter-speaker variation does thus go some way to explaining the persistent diversity in this paradigm and arguably may contribute to its maintenance.

Резюме: Настоящая статья рассматривает вариацию между говорящими в двух средах: в родительном и предложном падежах ед. ч. «твердых» мужских неодушевленных существительных в чешском языке. Материалом исследования стал широкий опрос носителей чешского языка, с целью проверки предпочтений и выбора используемых форм. Наша гипотеза, состоящая в том, что такая вариация существует, была до некоторой степени подтверждена. С одним исключением биографические данные носителей (возраст, пол, образование) не играли роли в предпочтениях и выборе наших респондентов. Место происхождения, однако, играло небольшую, но существенную роль, хотя результаты оказались иными, чем мы ожидали. Стараясь соотнести эти два типа задачи между собой, мы пришли к выводу, что способ использования шкалы предпочтений не соответствовал выбору форм, но общая тенденция в предпочтениях той или иной формы соответствовала выбору окончаний, который сделали участники анкеты. Таким образом, причины устойчивого разнообразия в этой парадигме частично объясняет вариативность в языке носителей, и есть основания полагать, что она является условием сохранения этого разнообразия

Keywords: Czech, morphology, variation, experimental linguistics, questionnaires,

1. Introduction

The observation that language users tolerate a certain amount of variation across speakers and instances of usage is uncontroversial. By *variation*, we mean that for a given proposition, there are multiple ways of realizing it, such that equivalent messages can be conveyed using different formal signs. In the terms set out by Baayen et al. (2013:255), these largely constitute the class of “identical rival forms”, where meaning and environment are held constant but the form nonetheless differs from item to item.

Different levels of linguistic analysis have proven to be more or less fruitful areas for such variation. For example, in the lexicon and phraseology examples abound (*enter* vs. *go in*); syntax (*bored with* vs. *bored of*, ни один студент там не был vs. ни одного студента там не было 'not one student (nom. sg. vs. gen. sg.) was (masc. vs. neut.) there' (Borschev & Partee 2002).

Frequently, the variation found is conditioned by social factors, such that the realization of material in a slot differs depending on the background of the producer or the social situation he finds himself in. In particular, the geographic and social isoglosses demarcating such variation in phonology and phonetics have largely formed the material explored in dialectology and sociolinguistics over recent decades.

Pierrehumbert describes this situation as effectively two realizations of variation in our data: *statistical variation in usage*, which, as she demonstrates, we are capable of forming mental representations of; and the *variation amongst individuals* in how they deal with the usage they encounter, which she, in agreement with Labov et al. (1991), says contributes to many of the results found in experiments (1994, 233-234, 241).

Variation is especially interesting when looked at from an emergentist perspective. As Bybee (2006, 714) puts it:

> Viewed in this way, language is a complex dynamic system similar to complex systems that have been identified, for instance, in biology (Lindblom et al. 1984, Larsen-Freeman 1997). It does not have structure a priori, but rather the apparent structure emerges from the repetition of many local events (in this case speech events).

In recent years, scholars working from this perspective have suggested that it is possible for groups and individuals to acquire subtly different grammars as a result of divergent input and processing strategies, and have shown that in some cases we can identify differences in the way groups respond to grammatical prompts. Dąbrowska (2010) demonstrated that linguists and non-linguists tend to react differently to judgement tasks of long-distance dependencies in English, and Dąbrowska (2008) showed that while all Polish speakers performed more or less at ceiling in an inflection task using real words, their level of education had an effect on their ability to inflect nonce words according to the expected pattern. Rácz et al. (2014), in a study of the English past tense observed that vocabulary size and gender had an effect on the application of grammatical rules. Larger vocabulary size was reflected in a greater tolerance for irregularity due to the existence of more robust models in smaller inflectional classes, and men showed greater sensitivity to levels of analogical support for generalization of rules than women did.

These findings underscore the possibility that variation is not an exception to the general rule that all speakers of a language share its items, but is in fact an inbuilt factor accounted for in the way we learn language. As Blythe and Croft put it:

> In the basic evolutionary model of language change, speakers replicate linguistic structures in utterances while interacting with other speakers. Those tokens of linguistic structures are the replicators. The replication process generates variation (produces innovation), via mechanisms that will

> not be examined here. Once these variants are available to speakers, speakers choose—not necessarily consciously or intentionally—to produce certain variants. Mechanisms of linguistic selection lead to the differential replication, that is, propagation, of some variants at the expense of others. (2012, 271)

If Dąbrowska is thus correct, then inter-speaker variation not only provides evidence for an emergentist view of language acquisition, but also is convincing evidence for an emergentist view of language change.

2. Describing morphological variation

We resolved to look at three methods of describing such variation and contrast them against each other. Our project focused on morphology, a level of analysis that tends to display less variation for any given speaker than others do: most morphological slots are filled automatically with one and only one form, and in the places where variation exists, the number of variants is highly restricted. In particular, we were interested in how data from large-scale corpora could be used to approximate the sort of acceptability judgements that native speakers might give, and to predict the answers they might give in forced-choice tasks.

We began with a hypothesis that focused mainly on these three methods. It proposed that there is a relationship between the frequency with which a form occurs in a representative corpus and (1) its acceptability to users and (2) its frequency of use.

From previous research (Bermel & Knittl 2012a, 2012b) we identified proportional frequency – the percentage of time one variant occurs vis-à-vis other variants – as a type of frequency that has an effect on judgements. However, traditionally frequency is looked at in terms of absolute numbers (albeit often standardized to a corpus size of one million tokens, see e.g. Bybee 2002, 264), and so we included high vs. low absolute frequency – number of occurrences in a corpus, or in a “standardized” corpus of 1m tokens – as a further contributory factor.

Our findings suggest that proportional frequency continues to be the highest-ranking factor across the board in both production and offline judgements. Somewhat surprisingly, absolute frequency in a corpus seemed to be only a partial or occasional factor. (Headline results are given in Bermel et al. 2014.)

Particularly interesting in our results is the question of low-frequency endings and items. In some instances, we see forms being used at relatively low proportional frequencies, or rated highly despite their low frequency in a corpus. A frequent interpretation of this sort of event is that low-frequency items have fewer entrenched barriers to unconventional constructions, resulting in higher acceptability ratings (see Theakston 2004 for an examination of how non-canonical constructions are more acceptable with low-frequency verbs). It seems clear that in some instances, “recessive” items find support in clearly defined contexts (i.e. when we look at particular endings in particular contexts), and may achieve higher ratings than they would seem to merit from their overall frequency of usage in either corpora or experiments.

A second question we thus attempted to answer was: are there other factors at work that might explain the maintenance of these minority endings? We identified three potential types of inter-speaker variation (between-group effects) that we wished to test for significance:

- Personal data, i.e. age, gender, education, region
- Test-taking data, i.e. how the scale provided is used and manipulated
- Attitudinal data, i.e. categoricalness and permissiveness of responses.

In this contribution, we will examine the results of an experiment from this inter-speaker angle. Our operating hypothesis is that if there is significant between-group variation here, then this might help us to explain how certain minority endings are maintained over generations, as it may be that certain groups of people or certain types of people are more likely to maintain them than others. The reason for this particular emphasis on maintenance will become clear in the next section.

3. Background

The material for our study comes from two slots in the nominal morphology of Czech: the genitive and locative singular forms of the paradigm exemplified by the "hard masculine inanimate" noun *hrad* 'castle'.

Descriptions of Czech agree on six syntactic cases and a vocative form, and three genders. The number of nominal paradigms is a matter of taste, but grammar books list between 10 and 15.[2] In contrast to Russian, where paradigms are described in terms of one basic model for each gender with varying degrees of "hardness" and "softness" of the stem, we can observe two features with regard to Czech nominal morphology:

(1) relationships between paradigms, and thus any overall "shape" of the system, are more opaque due to the effects of sound change and subsequent analogical change, which have obscured original relationships;
(2) the system itself is more fluid due to numerous points at which variation is possible, making descriptions of the relationship between paradigms less helpful in any event.

The *hrad* paradigm in Czech arises as a result of the reorganization of the Proto-Slavonic o-stem and u-stem classes. The o-stem class was a large class comprising the bulk of masculine- and neuter-gender nouns; the u-stem class was very limited – not more than a dozen reliably attested nouns all told – and contained a small number of mostly high-frequency masculine animate and inanimate nouns; for a fuller discussion, see Janda (1996).[3]

In all Slavonic languages, following the loss of distinct nominative/accusative forms in these two classes in the Proto-Slavonic period, these two patterns saw increasing convergence across all their case forms. In Russian this resulted over time in the loss of the u-stem endings as nouns from this class were absorbed into the emerging "masculine inanimate" and "masculine animate" paradigms, which utilize the old o-stem endings. In two cases – the genitive and the locative – there eventually developed sub-cases in which the u-stem endings dominated, but these have been subject to ongoing attrition, with the current marginal status of the partitive genitive and the locative prepositional in Russian being relics of this (Brown 2007).

In Czech, by contrast, what happens is a more thorough reorganization and redistribution of the morphological material inherited from Proto-Slavonic. The old u-stem endings become the default endings for the new animate pattern *pán* ‘lord’ in the dative and locative singular, and for the new inanimate pattern *hrad* ‘castle’ they become the default endings for the genitive and locative singular. They also become the default endings in the instrumental singular, genitive and locative plural for all nouns of masculine gender and achieve some prominence in the nominative plural of animate nouns.

However, the old o-stem endings were not excluded completely from these slots. In two cases – the genitive and locative singular – in contemporary Czech there is a minority of nouns, some of them very high-frequency, that either require the use of the old o-stem ending or use it in competition with the u-stem ending. The situation (described in greater detail in Bermel & Knittl 2012a, 93–95) is thus one in which a morph previously associated with two case slots in a small, closed declension class has become the default ending for a much larger, open declension class, without fully replacing the historical endings for that class. We will henceforth refer to these two endings as the expansive endings ({u} in both cases) and the recessive endings ({a} for the genitive sg. and {ě} for the locative sg.).

In the gen. sg. of the *hrad* paradigm, nouns typically have {u} as their ending, but a limited number have {a}. The SYN2005 corpus shows that out of all masculine nouns with a genitive form in {u} or {a}, 98.4% had exclusively {u}, 0.7% had exclusively {a}, and 0.9% showed variation between {a} and {u}.

In the loc. sg. of the same paradigm, nouns typically have {u} as their ending, but a limited number have {ě}. In the SYN2005 corpus, of all masculine nouns with a locative form in {u} or {ě}, 93.9% had {u} only, 0.7% had {ě} only, and 5.4% showed variation between {u} and {ě}.

However, this type frequency is somewhat misleading. Nouns with the recessive endings tend to have significantly higher token frequencies, and in fact the recessive endings constitute 11.9% of all gen. sg. forms in this declension pattern, and 31.1% of all loc. sg. forms. If we look at the median frequency, which is less influenced by outliers than the mean, we see that while nouns with exclusively the expansive endings have frequencies of 4 (gen.) and 3 (loc.), and nouns exclusively with the recessive endings have frequencies of 10 and 1 respectively, the frequency of nouns where both endings occur is respectively 120.5 and 110.[4]

In contrast to Russian, where distinct sub-cases arose that utilized the old u-stem endings, no such clear-cut situation has appeared in Czech. Grammars (see Bermel & Knittl 2012a, 94–95) describe a tendency to use the recessive endings with locational contexts and the expansive endings with non-locational contexts, but this is not entirely borne out in practice (and much less so in the genitive than the locative).

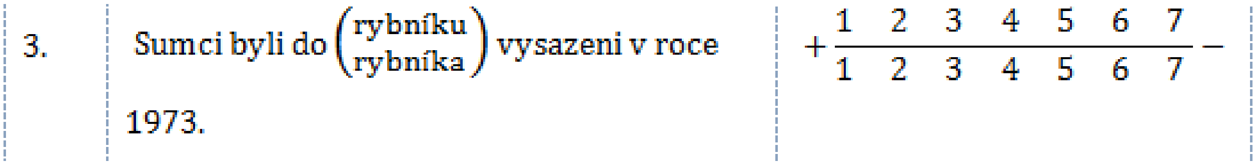
The conclusion is that the recessive endings are well embedded in the system; despite occurring with a tiny minority of types, they constitute a significant portion of the tokens. Fifteen hundred years after it began, the historical change that led to the merger of two declension classes has resulted in continuing variation. While the Russian innovation of sub-cases has been subject to ongoing attrition, with the partitive genitive nearly lost and the locative prepositional restricted to a small number of nouns, the considerably less clear-cut Czech variation has, in contrast, continued to form a prominent part of the system.

4. Methods

Our surveys were structured two types of tasks: acceptability judgements and gap filling. The material consisted of the two case variation studies discussed above and filler entries from verbal morphology.

The triggers consisted of sentence-long contexts drawn from the Czech National Corpus.[5] In the judgement tasks, respondents had to evaluate individual forms with variant endings on a 1–7 Likert scale.[6] In the forced-choice tasks, respondents had to fill in the missing ending for a word.

Acceptability judgements were given in the context of a single trigger, with respondents evaluating both variants:

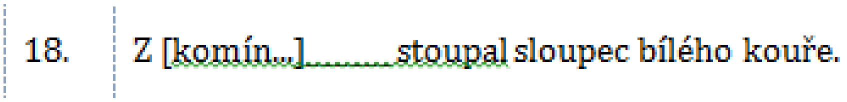


3. Sumci byli do $\binom{\text{rybníku}}{\text{rybníka}}$ vysazeni v roce 1973.

$+\frac{1\quad 2\quad 3\quad 4\quad 5\quad 6\quad 7}{1\quad 2\quad 3\quad 4\quad 5\quad 6\quad 7}-$

‘The catfish were released into the **pond**$_{\textbf{gen.SG}}$ in 1973.’

For the forced-choice task, respondents saw the stem of the word, which is also the nominative sg. or “citation” form; they were to insert the desired ending into the following gap:[7]



18. Z [komín...]______ stoupal sloupec bílého kouře.

‘A column of white smoke rose from the **chimney**$_{\textbf{gen.SG}}$.’

As our main goal in the research was to evaluate various types of word frequency found in the corpus, each survey’s trigger sentences employed a variety of lexemes in 8 frequency bands. The two types of frequency were absolute frequency in a corpus (2 levels: high (1000+)/low (1-999)) and proportional frequency in a corpus (4 levels: 0-5%, 5-50%, 50-95%, 95-100%). This gave eight frequency cells for each of our two features, the genitive sg. and the locative sg. (see Table 1).

Each lexeme was checked twice per survey, in differing syntactic contexts, to reduce the reliance on single examples, as shown in Table 1. For the genitive sg. we used possession for the first context and common prepositions requiring the genitive (*do, od, z, bez, kolem, kromě*) for the second. In the locative case, which only occurs with a limited number of prepositions (*v, na, o, při, po*), we used a locational meaning for the first context and a non-locational meaning for the second.

We included two lexemes in each cell in order to avoid any potential lexical effects that might arise from overreliance on a single lexeme per cell. Due to the amount of material covered and the need to avoid order effects from repeating material in both parts, we structured it in a block design, such that respondents undertook evaluation on one set of lexemes and gap-filling on an interleaved set, with a "complementary" survey looking at the same forms but in the opposite tasks:

- *Survey 1:* gap-filling on 'A:1,3,6,8'; acceptability judgements on 'B:2,4,5,7'
- *Survey 2:* gap-filling on 'B:2,4,5,7'; acceptability judgements on 'A:1,3,6,8'

The second item in each cell was checked in parallel versions that repeated the block design of the first:

- *Survey 3:* gap-filling on 'C:9,11,14,16'; acceptability judgements on 'D:10,12,13,15'
- *Survey 4:* gap-filling on 'D:10,12,13,15'; acceptability judgements on 'C:9,11,14,16'

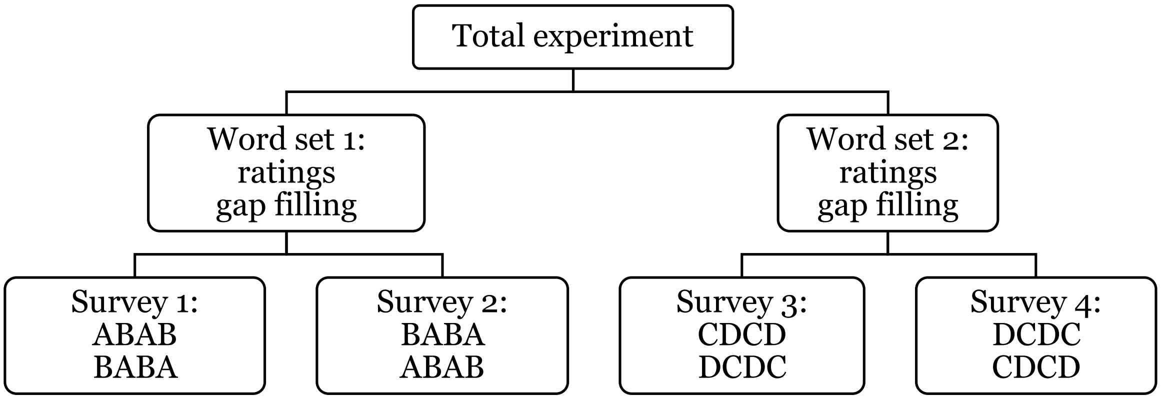
Each of the two lexical sets thus evaluated four lexemes per case (one per cell) for a total of eight lexemes, in two sentences each. In addition, distracter sentences using verbs as targets were added to the survey, so that no feature accounted for more than a third of the total sentences.

Table 1. Survey design

Proportional freq. Absolute freq.	0-5% {a} (G) {ě} (L)	5-50% {a} (G) {ě} (L)	50-95% {a} (G) {ě} (L)	95-100% {a} (G) {ě} (L)
0 – 1000 G. possessive L. local	A/C Word G1/G9 Word L1/L9	B/D Word G2/G10 Word L2/L10	A/C Word G3/G11 Word L3/L11	B/D Word G4/G12 Word L4/L12
G. w/preposition L. non-local	Word G1/G9 Word L1/L9	Word G2/G10 Word L2/L10	Word G3/G11 Word L3/L11	Word G4/G12 Word L4/L12
1000+ G. possessive L. local	B/D Word G5/G13 Word L5/L13	A/C Word G6/G14 Word L6/L14	B/D Word G7/G15 Word L7/L15	A/C Word G8/G16 Word L8/L16
G. w/preposition L. non-local	Word G5/G13 Word L5/L13	Word G6/G14 Word L6/L14	Word G7/G15 Word L7/L15	Word G8/G16 Word L8/L16

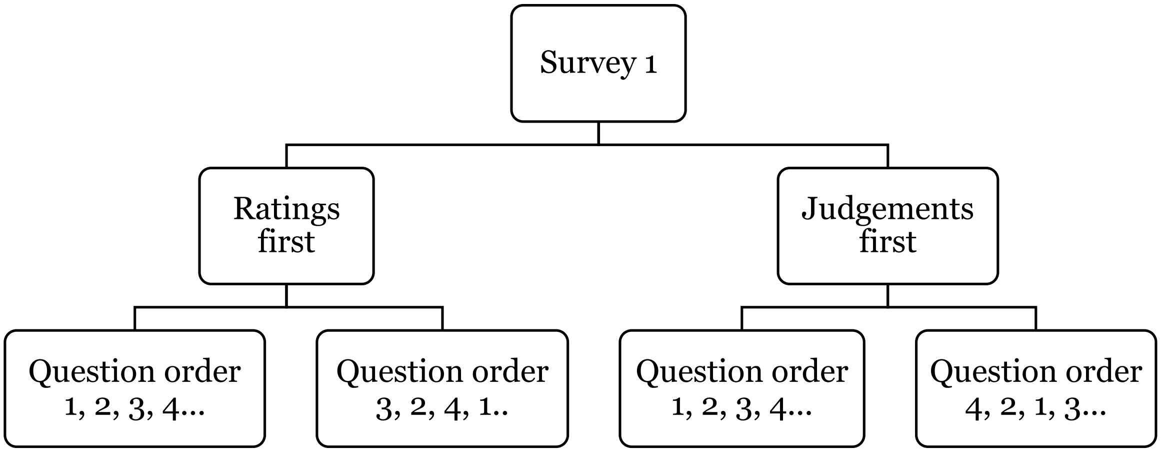
The overall structure of our experiment, as shown in Figure 1, thus consists of four basic surveys constructed out of two basic word sets:

Figure 1. Structure of overall survey



In part because of our use of a block design, we needed to ensure that the order of items or tasks was not a confounding factor. Each survey was thus split into versions, which differed only in the order of presentation of material. The order of tasks was varied from version to version, as was the order of presentation of triggers, as can be seen in Figure 2.

Figure 2. Design of each version


Survey 1
Ratings first
Judgements first
Question order 1, 2, 3, 4...
Question order 3, 2, 4, 1..
Question order 1, 2, 3, 4...
Question order 4, 2, 1, 3...

In total, then, we had 16 different versions distributed to respondents (2 word sets x 2 block designs x 2 task orders x 2 question orders).

5. The respondents[8]

We aimed to obtain 500 responses (250 per survey), which would have given a minimum of 30 responses per version, enough to ensure that significant comparisons could be done between them. After failed attempts and non-native speakers were removed, we had a total of 552 responses, meaning each version was completed by 35-40 respondents.

Surveys were gathered across the Czech Republic in a number of locations in Prague, Mladá Boleslav, Olomouc, Brno, and Přerov. They were either directly administered by the project research associate at universities and colleges, or the RA's contacts were instructed on how to administer them in workplaces.

Surveys were distributed on paper and each administration contained a mixture of versions. Participants were not given a time limit, but most completed the survey within 15-20 minutes.

In addition to answering the survey questions, participants were asked for basic biographical data, including their age (in ten-year ranges), gender, the region (*kraj*) from which they come,[9] and their level of education.

Table 2. Age and gender of respondents

Age range	*N =*	*gender*	*N =*
18-25	341	male	222
26-35	78	female	329
36-45	61	not stated	1
46-55	46		
56-65	18		
66-75	8		
Total	552	Total	552

Distribution by age and gender was skewed, as can be seen in Table 2: lower age groups predominate, and women predominate.

The predominance of respondents in the 18-25 group can be explained by the fact that universities and gymnasia were used for collecting data. We nonetheless had enough respondents in each group to make it possible to look at age as a variable, although in the two highest age bands the numbers do not meet our criteria for reliability.

The predominance of female respondents is more surprising, given that we recruited heavily in technical subjects where we would have expected a better gender balance. If the study had been balanced according to the Czech population, we would have had 271 men (49.1%) to 281 women (50.8%). Other studies where university students predominate have also noted that female respondents are more numerous (see Lečić, this issue, and Golubović, this issue). It may be that women are more likely to complete such surveys, or are more likely to complete them correctly so that the results are usable; and this tendency may be amplified by the gender balance at universities in general.[10]

Table 3 shows the geographical distribution of respondents.

Table 3. Region of origin and other places lived

Region	*N =*	*Expected*	*Stay*	*N =*
Karlovarský (Bohemia)	10	16	none	504
Ústecký (Bohemia)	11	44	abroad	14
Liberecký (Bohemia)	7	23	in Moravia	25
Plzeňský (Bohemia)	10	30	in Bohemia	9
Jihočeský (Bohemia)	15	33		
Středočeský (Bohemia)	39	67		
Praha (Bohemia)	89	65		
Královéhradecký (Bohemia)	24	29		
Pardubický (Bohemia)	22	27		
Vysočina (mixed)	31	27		
Jihomoravský (Moravia)	98	61		
Zlínský (Moravia)	25	31		
Olomoucký (Moravia)	145	34		
Moravskoslezský (Moravia)	26	65		
Total	552		Total	552

As seen in Table 3, there was a reasonable geographical spread of responses, although this was slanted towards certain areas. The “Expected” column shows what the number of respondents would be if split according to current population levels (Český statistický úřad, 2013). Due to stronger recruitment in that area, the Olomouc region is overrepresented, while several other areas are underrepresented.

Respondents gave free responses to the question on other places they had lived, which we summarized into three categories (Bohemia, Moravia, abroad) according to what would be most likely to affect their responses. Few respondents reported having lived abroad or in another part of the Czech Republic, so this factor was in the end not taken into account.

Data on education can be found in Table 4.[11]

Table 4. Education and field

education	N =	Expected	*field*	N =
primary	15	97	general	199
secondary	278	354	natural sciences	5
higher	259	69	technical/engineering	190
			social sciences	95
			humanities	48
			Czech	15
Total	552		Total	552

As can be seen in table 4, those with a university education (final year or completed) are the most numerous group, despite constituting only 12.5% of the Czech population (Český statistický úřad 2014). This is a side-effect of our recruitment methods. Those with only a primary-school education are underrepresented in the survey. The spread by field reflects our avoidance in general of humanities subjects, and specifically those who may have a good knowledge of linguistics.[12]

6. Main results of the study

The main results of this study are reported elsewhere (see e.g. Bermel et al. 2014). However, a quick summary of them is required before we consider the remaining data.

Our first task was to control for order effects. The question was whether the order of items or tasks affected the responses given. We were particularly interested in whether the task order had influenced responses, as it seemed plausible that, for example, one sort of task might exert a priming effect on the other.

We first looked at the results for our ratings task. For each version in which participants were answering the same questions, the mean of the participant judgements was calculated across all variables and a t-test was done with a contrasting version (e.g. ratings - forced choice vs. forced choice - ratings) to establish whether there were order effects. In none of the eight tests did the results reach the level of significance ($p < 0.05$). Results ranged from $0.07 < p < 0.84$, with most over 0.3.[13] Judgements are thus unaffected by their position in the survey; it does not matter whether they are completed before or after the forced-choice task.

We then examined the results for the forced-choice task. For each version in which participants were answering the same questions, the mean of the participants' choices of the expansive ending was calculated across all variables and a t-test was done with a contrasting version as above, to establish whether there were order effects. Out of eight tests conducted, none reached the level of significance, with results ranging from $0.29 < p < 0.75$. Forced choices are thus unaffected by their position in the survey; it does not matter whether they are completed before or after the judgement task.

The lack of order effects allowed us to combine data from all the versions for which respondents were reacting to the same triggers. This means we usually report four results per feature: 2 lexeme sets x 2 block designs.

Our second task was to identify factors that could be put into our analysis. A primary components analysis, performed as an initial diagnostic tool, indicated that most of our factors were textual, i.e. the frequency with which forms appear in the corpus, or the contexts in which the forms can be used. However, one factor – region – was identified as possibly relevant and was thus included in the main analysis. Other features, such as age, education, and gender, were looked at individually and separately (see section 7 below).

Our results showed that corpus frequency is a good predictor of responses, both on the judgement tasks and the forced-choice tasks. We looked at two ways of operationalizing frequency, and came to the conclusion that proportional frequency of items in a corpus is a consistently significant and large factor in the ratings given to endings and the choice of endings. The absolute frequency of a form in the corpus, whether operationalized in “bins” as high and low frequency, or using actual values, is not always a significant factor, and tends to show a smaller effect size than proportional frequency (Bermel et al. 2014, 223–225).

7. Personal data (age, gender, region, education)

The assumptions underlying the hypothesis under consideration here were that although they had not figured in our main survey, differences between native speakers (see section 2) might help explain why some very low-frequency options are maintained in the language, i.e. that some people are more prone to maintain these low-frequency forms than others. Given the basic personal data held on all respondents, our subsequent hypothesis for this section was:

Hypothesis 1: personal data

There will be significant differences in how people rate items and make choices depending on their age, education, gender or region of origin.

To test the hypothesis, one-way ANOVAs were performed to explore the effect of age, educational level, educational field, and gender on the choice of form and on the ratings given. Because we had four surveys in which all the respondents were answering the same questions, regardless of order, this resulted in eight one-way ANOVAs: 2 cases x 4 cohorts (2 word sets x 2 block designs).

Results for region were drawn from a set of complex repeated-measures ANOVAs and generalized linear model regressions (this is the “main analysis” referred to in section 6). The data from these surveys were explored using a complex model to combine the paired “complementary” surveys and thus we had only four analyses for these data: 2 cases x 2 word sets.

7.1 Personal data and ratings

We calculated the average response of ratings by each respondent and subtracted the recessive ending’s combined rating from the expansive ending’s combined rating. This yielded a single rating measure that could be used to measure the responses of different population groups.[14]

Looking at age as a factor, we found two significant results out of eight ($p < 0.05$), one for the genitive case ($F(5, 134) = 2.77$, $p = .02$) and one for the locative ($F(5, 134) = 2.56$, $p = .03$). These resulted in each instance from one significant difference between two age groups; the remaining results were insignificant. Age thus does not seem to play a consistent role in people’s ratings.

Looking at gender as a factor, we found one significant result for the genitive case ($F(1, 135) = 5.05$, $p = .03$) and one for the locative case ($F(1, 131) = 4.86$, p = .03). Our judgement was that this was not reliable enough to label gender as a consistent factor in people’s rating.[15]

Looking at education, we found no significant results out of eight. We also checked to see whether the respondent’s area of specialisation had any effect, and again found none.

We thus had 32 results (4 factors x 4 surveys x 2 cases), of which four significant results were found. If we accept a 5% chance result, then we would expect 1–2 false positives. It does not seem unthinkable that we would have 3–4 false positives, especially given that two occurred in the age category, where the cells are of very unequal size and some cells have only a few respondents.

Turning to our repeated-measures ANOVAs, region did not show a significant effect by itself for any of our data (0.14 < p < 0.78).[16]

Having found no consistent effects with region other than some small interactions with other features, we now turn to the data on the forced-choice task.

7.2 Personal data and forced choices

Initial explorations of our forced-choice data using Primary Components Analysis and regressions suggested that of the between-subjects factors, only the speaker's region of origin was likely to influence our model. We nonetheless performed one-way ANOVAs on age, gender, level of education and field of education to check that our analyses were not hiding anything. We found no significant effects for our first word set. For the second word set, there were three significant results out of 16: one for age (F(4, 128) = 3.19, p = .02), one for gender (F(1, 131) = 6.46, p = .01) and one for level of education (F(2, 130) = 3.17, p = .05), but these were very small. Post-hoc tests show them to be the effect of isolated differences in individual words.

We will now go on to look at some regressions run on these same data. Regression is a statistical technique used to determine which of a variety of possible factors contribute most significantly to our results. It does so by starting from an assumed null model, in which we always choose the most common answer. It measures the answers given by an overfitted model, in which all possible factors are entered, and a more closely fitted model, in which we select the factors that we deem most relevant and specify an entry order for them. The best model is judged to be the one that brings us closest to the actual results or the overfitted model (depending on the measurements) while incorporating the least number of factors. It thus attempts to balance accuracy against simplicity. It thus follows that one can usually find at least some way to marginally improve one's results by adding further factors, but in the case of very small improvements in accuracy, the model may appear worse than a less accurate but simpler one.

The personal characteristic that shows up as significant most often in these analyses is the region our respondents come from. This was thus the one between-subjects factor that we introduced to our general analysis, which was performed using a generalized linear mixed model for each set of data (2 cases x 2 sets of lexemes = 4 analyses). Besides the regional factor, the model used: the proportional frequency of items in the corpus; the absolute frequency of items in the corpus; the syntactic context; and certain combinations of the above factors.

The effects of region on the choice of forms were evident in the first word set (p = .003 for the gen. sg. and p = .02 for the loc. sg.), but not in the second (p = .41 for the gen. sg. and p = .85 for the loc. sg.). Still, small F values (8.77 for the gen. sg., 5.40 for the loc. sg.) mean that these differences are far from a major factor. (By comparison, the

F values for the largest factor, proportional frequency of the recessive form in the corpus, range from 90.43 to 157.52.)

For the genitive case, region was a significant factor in our first data set, but not in our second. However, region did show up as a significant factor in the second set in combination with proportional frequency (*p* = .007, F = 4.09): the words in each frequency band were thus treated differently by people from Bohemia vs. Moravia. The results are shown in Figure 3.

Figure 3. Estimated means of {a} endings chosen in the gen. sg. for set 1 & 2

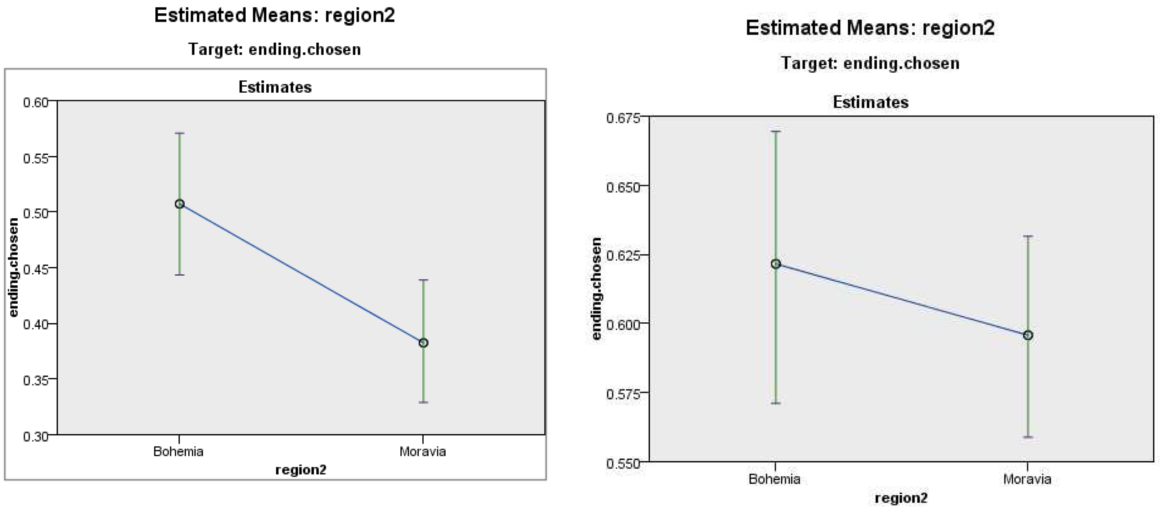


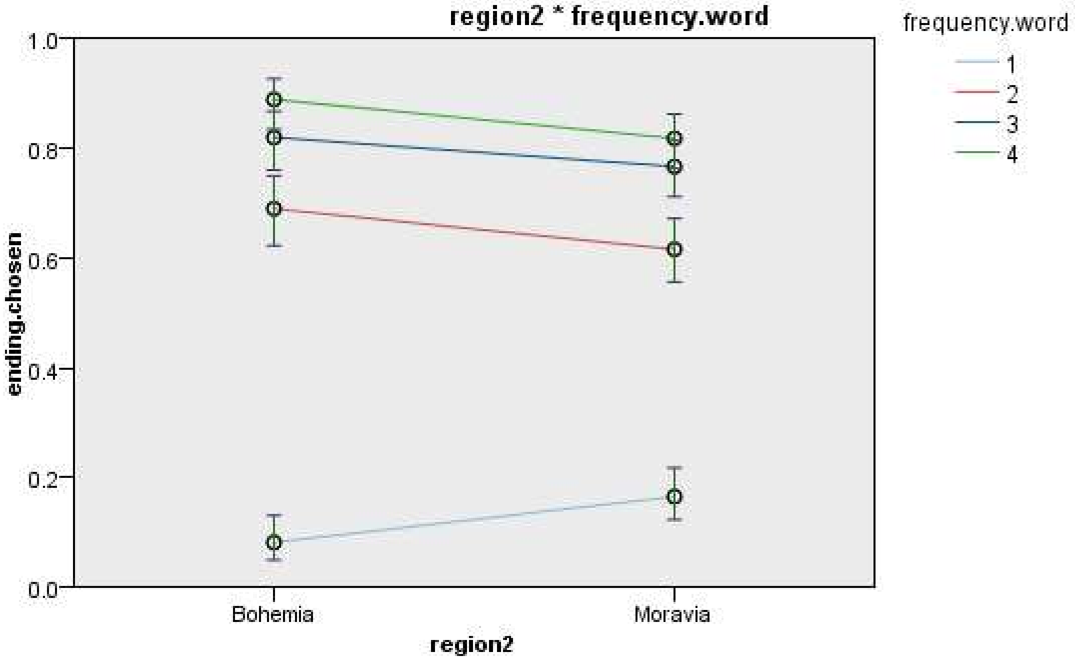
Figure 3 shows that Bohemians are estimated to use of the {a} ending more than Moravians do, although for the words in set 1 this difference is significant (while for the words in set 2 the difference is not statistically significant). The figures are given in Table 5.

Table 5. Estimated means of {a} endings chosen in the gen. sg. by word set

	Bohemia	*Moravia*
Set 1	.51	.38
Set 2	.62	.60

Closer examination of the data from set 2 suggests that the reason for this is an outlier word in the second set, which explains why the interaction between frequency band and region has shown up, as shown in Figure 4:

Figure 4. Interaction of region and proportional frequency in set 2 gen. sg.



The figures for the interaction between region and proportional frequency are given in Table 6.

Table 6. Estimated means of {a} by Region Proportional Frequency*

	Bohemia	*Moravia*
Band 4 (95-100% {a})	.89	.82
Band 3 (50-95% {a})	.82	.77
Band 2 (5-50% {a})	.69	.62
Band 1 (0-5% {a})	.08	.17

For the loc. sg., region again shows up as a significant factor in our first data set. The second data set shows similar trends but the difference between the regions is so small as to be insignificant:

Figure 5. Estimated means of {ě} endings chosen in the loc. sg. for set 1 & 2

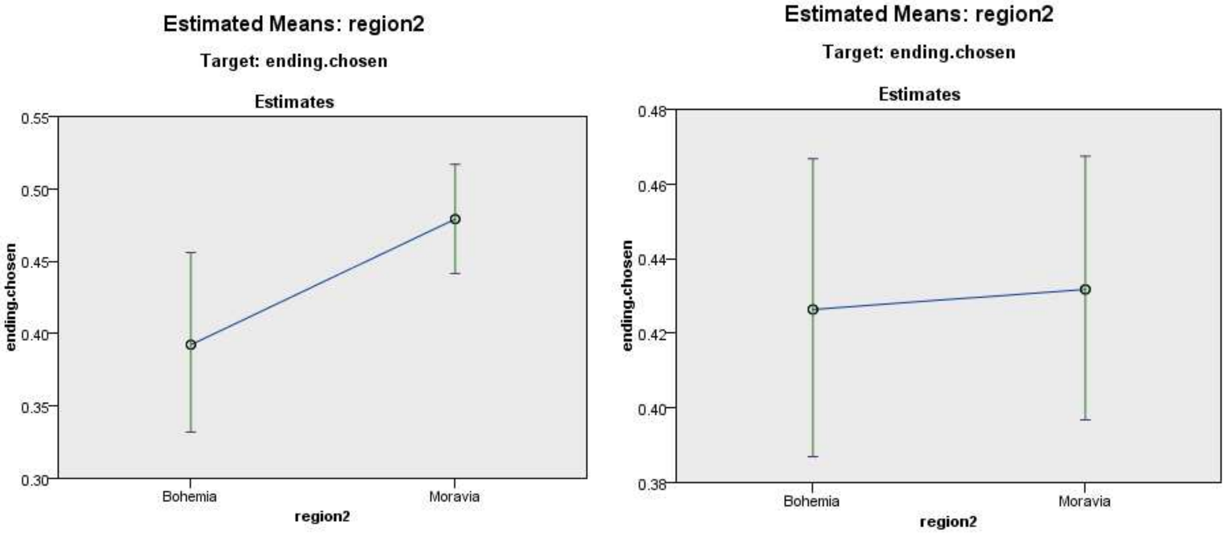


Table 7. Estimated means of {ě} endings chosen in the loc. sg. by word set

	Bohemia	*Moravia*
Set 1	.39	.48
Set 2	.426	.432

Another way to explore our forced-choice data is through the use of classification trees. As Baayen et al. (2013) demonstrated, this method of analysis can be an effective complement to linear regression, as it often shows more clearly how choices emerge in variant systems. A classification tree graphically distributes significant factors in choices according to their influence. If we start from the top, the graph first splits according to the factor where the differences are largest and most significant, and then at each node looks again to split based on the same criterion. We had it make three decisions of this sort, at which point we were down to individual lexical items. Reading down the tables, the first node gives the factor (variable) where the largest differences were found. For instance, in figure 6, the highest-order node is split by the proportional frequency of forms found in the corpus; at the next level, context plays a significant role for two of those groups and absolute frequency for one. It is not until we get to the bottom node that we find a difference in region, but this amounts to a rather large difference (74.3% vs. 94.5% {a}) for one lexeme, so the results are not generalizable.

Figure 6. Classification tree for gen. set 1

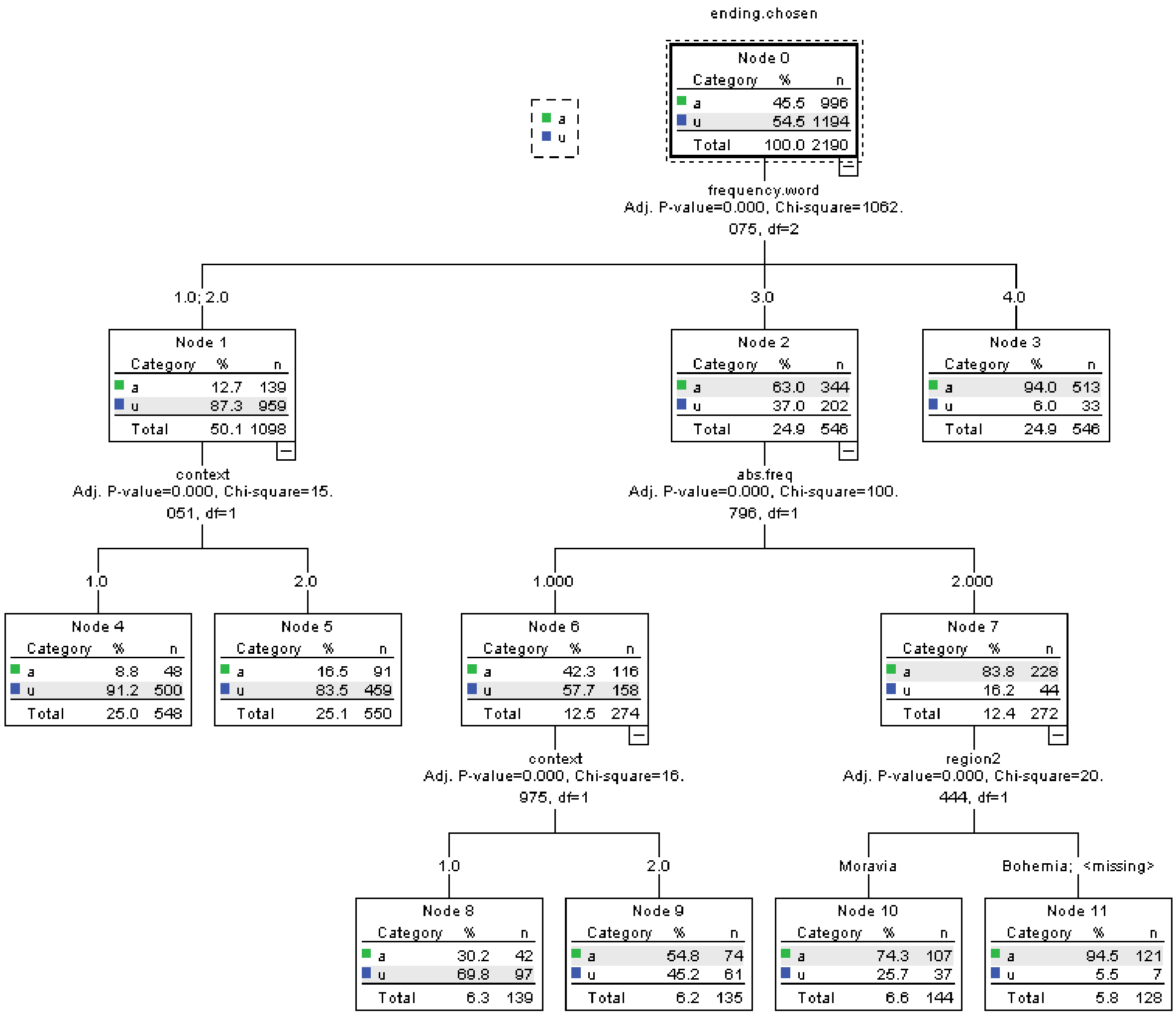
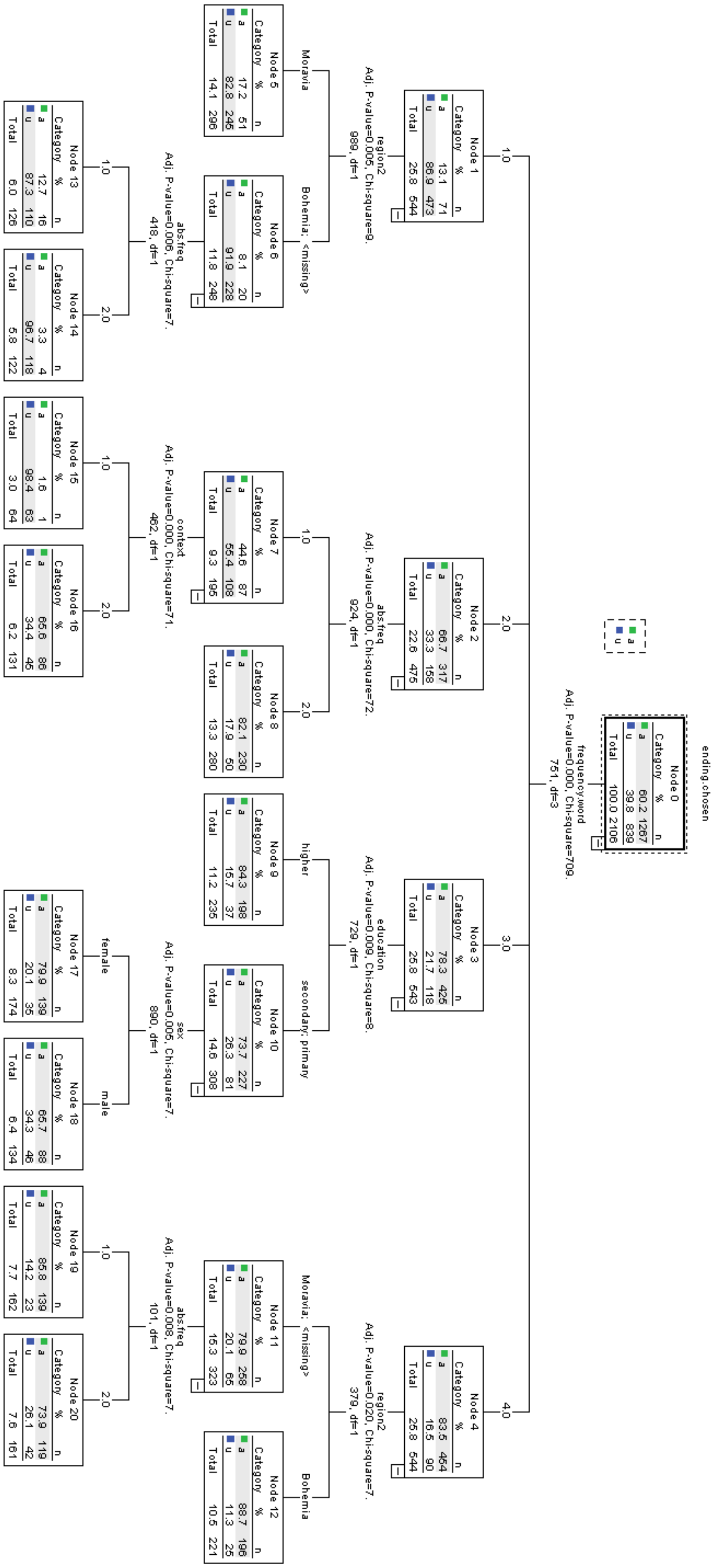


Figure 7 shows the breakdown for the second gen. set. Here once again the highest level factor is the proportional frequency of forms in the corpus; however, at the second node region of origin plays a significant role in two bands, and education in one band. Region thus affects four out of eight lexemes. However, the next split is by absolute frequency of the lexeme in the corpus, which means that the difference may be due to the way people in one particular region rate two different lexemes, thus somewhat reducing its impact.

Figure 7. Classification tree for gen. set 2



For the two locative sets, a similar picture emerges. In figure 8 (loc. set 1), textual characteristics dominate the top nodes: proportional frequency in the corpus is again the top node, with context and absolute frequency in the corpus following. The only personal characteristic to register on the tree is region at the third node, meaning it affects only two lexemes. In figure 9 (loc. set 2), proportional frequency in the corpus is the top node, with education at the second node and age at the third node. These effects again seem to concern only two words out of eight, so they are relatively limited.

The data from classification trees can be triangulated with that from the regressions to give a fuller picture of what is happening here.

In every instance, we get furthest fastest when we start with variables for linguistic data. The proportional frequency of a form in the corpus is always at the first node in the tree, and absolute frequency of a form in the corpus always makes an appearance as well, in all but one instance at the second node, and usually at more than one node, meaning it affects 2-6 lexemes out of 8. This correlates well with our regression data, which show that proportional frequency is always the most significant factor in a model.

Variables for personal data appear lower on the tree than textual variables. Region is a lower-order factor in 3/4 trees, and education and age appear sporadically (in 2/4 and 1/4 trees). Gender does not appear at all on the tree, despite having come up as significant in some ANOVAs; this does not negate the significance rating, but does suggest that it plays a very minor role compared to other factors.

Figure 8. Classification tree for loc. set 1

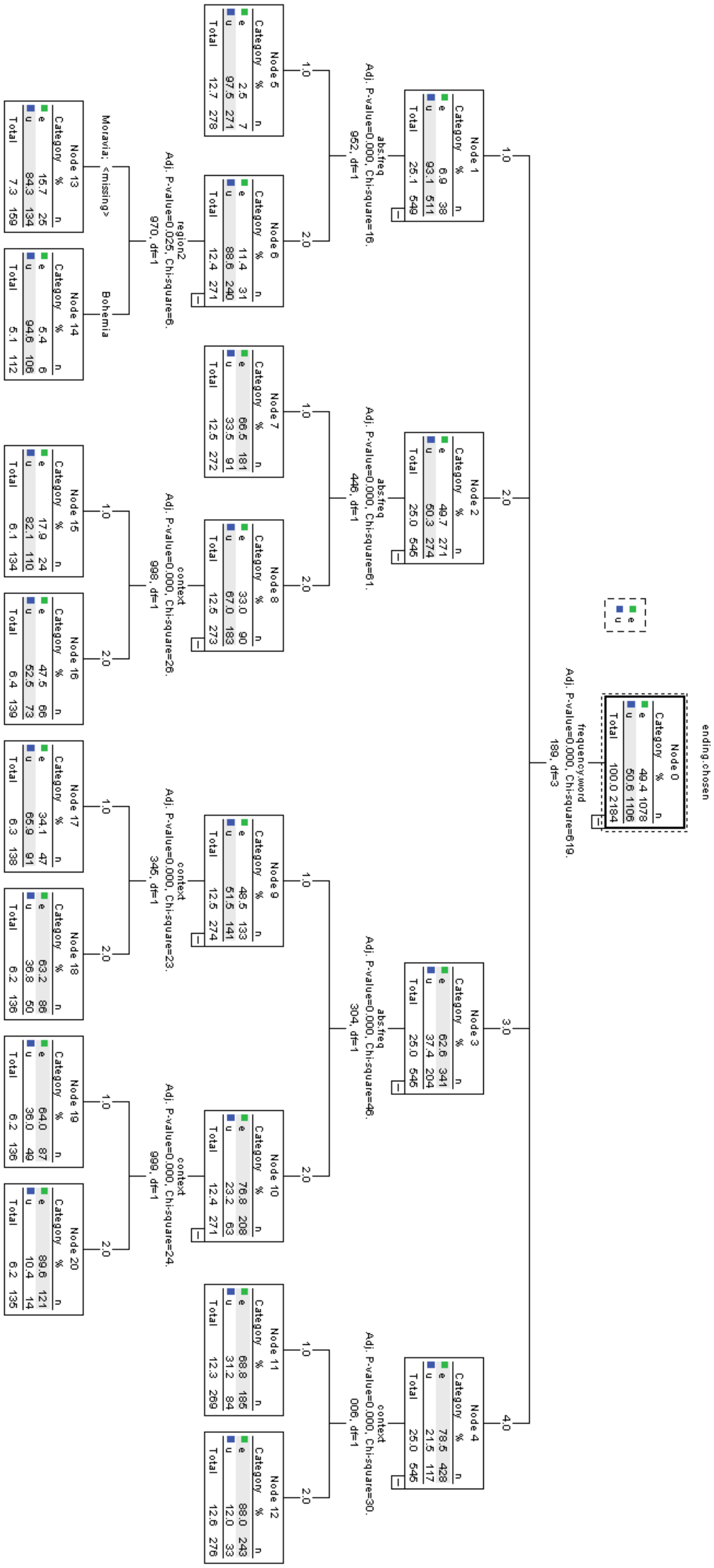
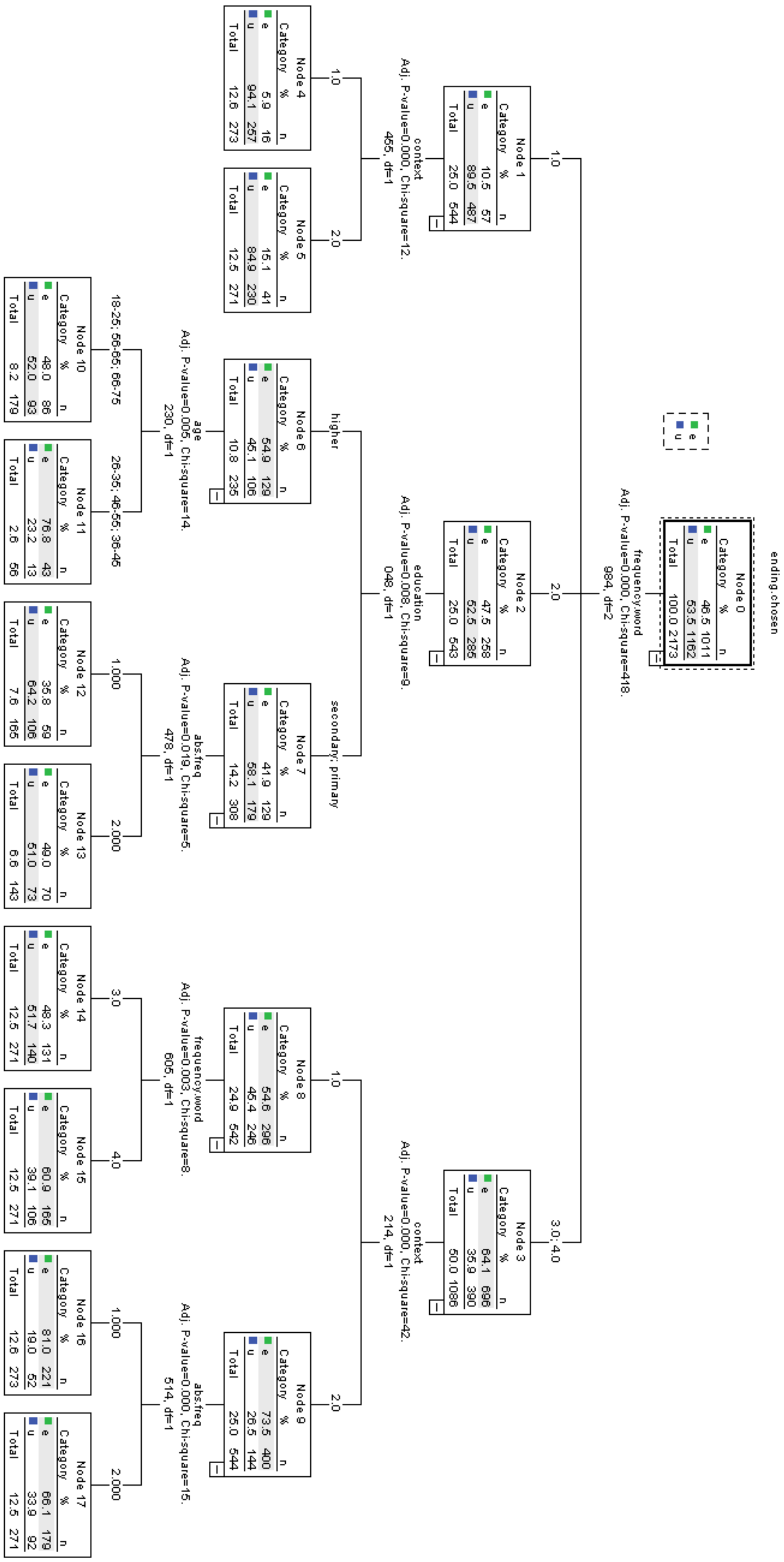


Figure 9. Classification tree for loc. set 2



To summarize the data for this section: there are individual places where regional differences are in evidence. However, contrary to popular belief, there is no one region that is ‘more conservative’ than another. The region that is ‘more conservative’ for one case is ‘more innovative’ for the other: hence, among Bohemians we see a higher use of gen. {a} and loc. {u}, and among Moravians we see a higher use of gen. {u} and loc. {ě}

Unequal distribution of respondents among groups may explain occasional significant effects, e.g. with educational level and age.

8. Approaches to ratings

Our second hypothesis concerned how the scales provided are used and manipulated by respondents. We were interested to see whether the different sorts of behaviours exhibited by respondents on the acceptability judgement task bore any relation to the way choices were made on the forced-choice task. If so, that would indicate another aspect of inter-speaker variation and possibly help explain how recessive endings are maintained. Our hypothesis for this section was as follows:

Hypothesis 2: behaviour types

The way respondents made use of the ratings scale in the judgement task indicates something about their use of language, and will be a significant and important factor in how they selected one or the other ending in the forced-choice task.

Of the two sorts of tasks respondents performed, judgement scales are the more open to interpretation by the respondent.[17] In our survey, respondents were asked to use a seven-point scale on which only the endpoints were labelled. If respondents behave in the following manner, then our hypothesis will almost certainly be false:

- The defined outer ends of our scale match what respondents would see as the outer ends of their scale
- Respondents make a more or less even, consistent division of the scale between the two endpoints
- Respondents use the intermediate points of the scale in similar ways

Looking at the data, we can see that respondents evidently do not use the scale in the same way, nor do they all interpret the endpoints in the same fashion. The patterns of ratings seen in the data were classified as seen in table 8:

Table 8. Ways in which the rating scale was used

Tag	*Description*	*Points used*	*N=*	*Included in analysis?*
Full scale	All marks used	1, 2, 3, 4, 5, 6, 7	320	Yes
Gaps	1-7 used, but not all midpoints used	e.g. 1, 2, 4, 5, 7	42	Yes
Permissive	Lowest mark(s) not used	1, 2, 3, 4, (5), (6)	175	Yes
Hesitant	Highest and lowest marks are not used	2, 3, 4, 5, (6)	5	No
Categorical	Only endpoints and middle point used	1, 4, 7	8	No

From Table 8 we can see that the majority of respondents (58.1%) used the full scale supplied. A further large group (31.8%) felt that none of the data given was so unacceptable as to be worthy of the lowest marks; most of them simply avoided point 7, but some also avoided 6 or even 5 and 6. We labelled them permissive respondents. The last significant-sized group (7.1%) used the full range from 1-7 but with gaps in the middle.

Two very small groups were hesitant respondents, who never used the scale's endpoints, preferring not to designate any of the forms as 'absolutely fine' or 'unacceptable'; and categorical respondents, who only used the endpoints of the scale and sometimes the middle point, possibly to mean 'I don't know/care'.

A one-way ANOVA was performed to examine whether the respondent's scale use was a significant factor in determining which endings they chose on the forced-choice section. We looked at the three groups with significant enough user numbers, discarding the last two groups. The results were as follows:

$$F\ (2, 258) = 0.91, p = .40$$

This indicates that there is no significant difference between the way respondents treated the ratings scale and the way they answered on the forced-choice section of the questionnaire. We also ran post hoc tests, but failed to find any significant differences between any two groups ($.39 < p < 1.0$).

In summary, our initial finding that people responded in different ways was a piece of evidence in favour of the hypothesis, but in the end there was no evidence to disprove the null hypothesis: although respondents did make use of the scale in different ways, it seems that the range of marks one employs does not tell us much about the linguistic choices one makes.

9. Average ratings and selection of endings

A final way to partition our sample is to look at another aspect of respondents' behaviour, namely the average ratings they gave, and how those relate to the endings they selected.

As was pointed out earlier, the judgement tasks have respondents arriving at two scores for each item (one with each ending), whereas the forced-choice tasks result in one answer per item. To make the data sets comparable, we needed to arrive at a single score per item for our judgement tasks.

We examined five ways of accomplishing this:

- Rating of expansive ending. The average rating of the recessive ending is discarded.
- Rating of recessive ending. The average rating of the expansive ending is discarded.
- Strength of preference. Operationalized as the average ratings for the expansive ending minus the average ratings for the recessive ending.
- Overall permissiveness. Operationalized as the sum of the average ratings for both endings.
- Ratio between scores. The average rating for the recessive ending is divided by the average rating for the expansive ending.

Preliminary investigations suggested that the first three of these would be most likely to yield interesting information, and we thus arrived at our final hypothesis:

Hypothesis 3: gradated behaviour

There will be correlations between:

- *Ratings of individual endings and the selection of one or another ending;*
- *People's overall (mean) strength of preference towards one ending.*

The underlying assumption here was that we can extract information about strength of preference by looking at the difference between the mean scores for the two endings. A high positive score represents a strong preference for the recessive ending; a high negative score represents a strong preference for the expansive ending. A score closer to zero reflects some degree of hesitance or ambivalence.[18]

For the average scores for each ending, we expect the ratings of {a} and {ě} to correlate negatively with the choice of {a} and {ě}. This is because "1" was defined as the most acceptable rating and "7" as the least acceptable, and thus as the rating of the ending moves towards 1 (most acceptable), we expect its usage to increase. Conversely, we expect the rating of {u} to correlate positively with the choice of {a} and {ě}, because as the rating of the {u} ending moves towards 7 (least acceptable), we expect the use of {a} and {ě} to increase.

A two-tailed Pearson's *r* was thus computed on the number of times that {a} or {ě} was chosen compared with three further values: the individual rating for {a} or {ě}; the individual rating for {u}; and the rating of {u} minus the rating of {a} or {ě}.[19]

The results generally upheld our hypothesis. The results for the genitive sets were significant and on every count it seems that the way our respondents rate is a small contributory factor to the choices they make. The results can be seen in Table 9.

Table 9. Correlations between average ratings and choices (gen. sg.)

		No. of times {a} is responded	
		Set 1	Set 2
average rating for {a}	Pearson Correlation	-.17	-.26
	Sig. (2-tailed)	<.005	<.001
average rating for {u}	Pearson Correlation	.19	.15
	Sig. (2-tailed)	<.005	<.05
preference towards {a}	Pearson Correlation	.24	.24
	Sig. (2-tailed)	<.001	<.001

For the locative case, the results were less convincing. In Table 10, we can see that locative set 1 did not have any significant results. However, locative set 2 shows a similar result to the genitive sets.

Table 10. Correlations between average ratings and choices (loc. sg.)

		No. of times {ě} is responded	
		Set 1	Set 2
average rating for {ě}	Pearson Correlation	-.07	-.31
	Sig. (2-tailed)	.28	<.001
average rating for {u}	Pearson Correlation	.08	.27
	Sig. (2-tailed)	.19	<.001
preference towards {ě}	Pearson Correlation	.09	.36
	Sig. (2-tailed)	.13	<.001

We can observe in Table 9 that of the three rating types, the strongest correlation, both in terms of significance and effect size, is reliably the "preference towards {a}" rating. The weakest correlation on both counts is the average rating for {u}. From this we can conclude that the way people rate the {u} ending, which is the default ending for the gen. sg., is the least predictive of their choice of endings. The most predictive is the strength of their preference for {a}. The rating that they assign to {a}, which might be thought to be the most straightforward sort of correlation, tends to be significant but it seems the strength-of-preference measurement is more reliable in this regard.

In Table 10, despite the lack of a significant effect for the loc. sg. in set 1, we can see the same effect, except even more strongly, in the loc. sg. set 2. The results also confirm the reliability of the strength-of-preference measurement, as it is larger in size for set 2 and comes much closer to significance for set 1.

10. Conclusions

In our analysis, we have examined some aspects of inter-speaker variation that might influence respondents' choice of one form or another when competing variants are possible. Our reason for doing this was to ascertain whether inter-speaker variation might have a role in maintaining the use of the less common ending.

Some of our possible factors were not, in fact, contributors. There was little evidence of a significant role for age, education or gender. In a few instances we identified places where one or another lexeme showed a significant difference when the data was partitioned by these factors. We propose that such results be considered type I errors: due to the number of levels in our analyses (3 education levels x 2 genders x 5 age groups x 2 regions), it is highly likely that occasional significant results will be obtained. A difference between, say, the 46-55 age group and the 36-45 age group for one set of lexemes is unlikely to be more than a statistical fluke.

This set of results is not surprising due to the durability of this variation. Although the shift towards the expansive ending {u} in both cases represents a historical change in progress, its timescale – over a millennium so far – means we should not expect to see clear-cut generational differences, and given that these features do not index anything within Czech language culture (high/low prestige, geographical origin, etc.) there was no reason to anticipate that education or gender would play a role either.

The one quasi-variable that regularly scored a significant result in our analyses was the region of origin, for which a slight regional preference could be detected. We noted two points about the distribution of significant differences here.

First, region by itself was not a significant factor at all in the judgement task. It does show up with small but significant differences in the forced-choice task. We propose that this difference is one way of operationalizing the notion of "language" vs. "dialect": we notice some small but significant geographical differences in usage, but these are not reflected in speakers' evaluation of how "normal" a form is.

Second, we noted that the effect was not consistent, in that Moravians were more likely than average to use the recessive, historically older {ě} ending in the loc. sg., but less likely than average to use the recessive, historically older {a} ending for the gen. sg. The traditional view of Moravia as a more linguistically conservative region is thus not consistently upheld.

We were interested to see whether affective, as well as biographical, factors might give us insight into how people make choices, and thus we also partitioned our respondents' behaviour by looking at how they responded to the questionnaire. In doing so, we attempted to relate their performance on the judgement task, which is highly nuanced and offers quite a lot of contributory data, to their performance on the simpler gap-filling task, where there is essentially one and only one measurement available (number of times one or another ending is filled in).

Although our respondents did not all use the 7-point Likert scale in the same way, it nonetheless turned out that their different responses to it and interpretations of it did not have any relation to their choice of forms. This was in some ways a welcome result, as it suggests that the judgement task is not significantly influenced by people's interpretation of the tools provided for it.

The remaining links between people's performance on the judgement task and their choices on the gap-filling task did show some consistent factors. The most reliable one was the strength of their preference for one ending over the other. The weakest factor was their rating of the expansive {u} ending. The logic of this hierarchy is clear. Many people will rate the expansive ending highly because it is the more common ending overall by a ratio of anywhere between 2:1 and 9:1, so in some cases the results will be more due to uncertainty than to any real preference for it. Somewhat more reliable is the rating they give to the ending {a}, but a ranking that combines the two ratings turned out to be the most reliable.

Our conclusion is thus that individual differences play a role in the maintenance of variation, but it must be emphasized that the size of this effect is not great compared to effects visible within the data, which are mostly linked to the frequency of forms or to contextual features. The interpersonal variation can be hard to discern through the overwhelming similarities when compared to frequency effects.

References

Arppe, A. and J. Järvikivi: 2007, 'Take empiricism seriously! In support of methodological diversity in linguistics', Corpus Linguistics and Linguistic Theory 3 (1), 99–109.

Baayen, R. Harald, A. Endresen, L. A. Janda, A. Makarova & T. Nesset: 2013, 'Making choices in Russian: pros and cons of statistical methods for rival forms', Russian Linguistics 37, 253–291.

Bader, M. and J. Häussler: 2009, 'Toward a model of grammaticality judgments', Journal of Linguistics 45, 1–58.

Bermel, N. & L. Knittl: 2012a, 'Morphosyntactic variation and syntactic environments in Czech nominal declension: Corpus frequency and native-speaker judgments', Russian Linguistics 36 (1), 91-119.

Bermel, N. and L. Knittl: 2012b, 'Corpus frequency and acceptability judgments: A study of morphosyntactic variants in Czech', Corpus Linguistics and Linguistic Theory 8(2), 241-275.

Bermel, N., L. Knittl & J. Russell: 2014, 'Absolutní a proporcionální frekvence v ČNK ve světle výzkumu morfosyntaktické variace v češtině', Naše řeč 97, 216-227.

Borschev, V. and B. Partee: 2002, 'The Russian genitive of negation: Theme-rheme structure or perspective structure?' Journal of Slavic Linguistics 10, 105-44.

Brooks, P. J. & M. Tomasello: 1999, 'How children constrain their argument structure constructions', Language 75 (4), 720–738.

Bybee, J.: 2002, 'Word frequency and context of use in the lexical diffusion of phonetically conditioned sound change', Language Variation and Change 14, 261–290.

Bybee, J.: 2006, 'From usage to grammar: The mind's response to repetition', Language 82, 711–713.

Český statistický úřad: 2011, 'Zaostřeno na ženy a muže', available online at http://www.czso.cz/csu/2011edicniplan.nsf/kapitola/1413-11-r_2011-13 . Last accessed 16 December 2014.

Český statistický úřad: 2013, 'Stav a pohyb obyvatelstva v ČR v roce 2012 (předběžné výsledky)', available online at http://www.czso.cz/csu/2012edicniplan.nsf/publ/4001-12-q4_2012 . Last accessed 16 December 2014.

Český statistický úřad. 2014. 'Souhrnná data o České republice (Obyvatelstvo podle dosaženého vzdělání)'. Available online at http://www.czso.cz/csu/redakce.nsf/i/souhrnna_data_o_ceske_republice . Last accessed 16 December 2014.

Cvrček, V., V. Kodýtek, M. Kopřivová, D. Kovářiková, P. Sgall, M. Šulc, J. Táborský, J. Volín, M. Waclawičová: 2010, Mluvnice současné češtiny, Prague.

Dąbrowska, E.: 2008, 'The effects of frequency and neighbourhood density on adult speakers' productivity with Polish case inflections: An empirical test of usage-based approaches to morphology', Journal of Memory and Language 58, 931–951.

Dąbrowska, E.: 2010, 'Naive v. expert intuitions: An empirical study of acceptability judgments', Linguistic Review 27, 1–23.

Divjak, D.: 2008, 'On (in)frequency and (un)acceptability', in Lewandowska-Tomaszczyk, B. (ed.), Corpus linguistics, computer tools and applications – State of the art, Frankfurt, pp. 213–233.

Grepl, M., Z. Hladká, M. Jelínek, P. Karlík, M. Krčmová, M. Nekula, Z. Rusínová, D. Šlosar: 1996, Příruční mluvnice češtiny, Prague.

Janda, L.: 1996, Back from the brink: a study of how relic forms in languages serve as source material for analogical extension, Munich/Newcastle.

Kempen, G. & K. Harbusch: 2008, 'Comparing linguistic judgments and corpus frequencies as windows on grammatical competence: A study of argument linearization in German clauses', in Steube, A. (ed.), The discourse potential of underspecified structures, Berlin, pp. 179–192.

Labov, W., M. Karen & C. Miller: 1991, 'Near-mergers and the suspension of phonemic contrast', Language Variation and Change 3, 33–74.

Matthews, W. K.: 1967, Russian historical grammar (reprinted with corrections), London.

Meillet, A.: 1965, Le Slave commun. Seconde édition, Paris.

Pierrehumbert, J.: 1994, 'Knowledge of variation', in Beals, K., J. Denton, R. Knippen, L. Melnar, H. Suzuki & E. Zeinfeld (eds.), Papers from the Parasession on Variation, 30th meeting of the Chicago Linguistic Society, Chicago Linguistic Society, Chicago, pp.

Rácz, P., C. Beckner, J. B. Hay and J. B. Pierrehumbert: 2014, 'Rules, analogy, and social factors codetermine past-tense formation patterns in English', in Proceedings of the 52nd Annual Meeting of the Association for Computational Linguistics, Baltimore, Maryland, available at http://acl2014.org/acl2014/W14-28/index.html (last accessed 14 January 2014).

Theakston, A.: 2004, 'The role of entrenchment in children's and adults' performance on grammaticality judgment tasks', Cognitive Development 19, 15–34.

Vaillant, A.: 1964, Manuel de vieux slave. Tome I: Grammaire. Seconde edition, Paris.

[1] This article forms part of the project "Acceptability and forced-choice judgements in the study of linguistic variation", funded by the Leverhulme Trust (RPG-407).

[2] As an example, Grepl et al. give 13 basic types (*typy*) named after common nouns: *pán, muž, předseda, soudce, hrad, stroj, žena, růže, kost, město, moře, kuře, stavení*. Cvrček et al. (2010, 144) have 12 patterns (*vzory*) similarly named: *list, město, had, táta, žena, muž, stroj, duše, píseň, moře, kost, stavení*. Both then have lists of subtypes or subpatterns (*podtypy, podvzory*): Cvrček et al. have 10 subpatterns and then a lengthy list of further exceptions, while Grepl et al. do not distinguish strictly between subtypes and other sorts of deviations from the basic types. Tradition evidently plays a significant role in these descriptions: a "basic" pattern or type such as *moře* may have only a handful of items, while a subclass of it such as *letiště* may have many more, and a much more productive class such as *cyklus* with hundreds of items is not even classed as a subtype or subpattern.

[3] Meillet (1965, 347), writing about Common Slavonic, lists only five nouns reliably falling into the u-stem class: *domŭ* 'house', *vrŭxŭ* 'summit', *volŭ* 'ox', *polŭ* 'half', *medŭ* 'honey', to which Matthews, writing about old Russian, adds сынъ 'son', родъ 'clan', рядъ 'row', чинъ 'rank' "and several others" (1967, 106). Vaillant (1964, 90–91), writing about Old Church Slavonic, in addition lists оудъ 'member, (body) part', даръ 'gift', санъ 'post' as largely convergent with this class, and жидъ 'Jew' as convergent in the plural.

[4] Details can be found in Bermel & Knittl (2012a, 99–100). SYN2005 has just over 100 million word tokens, so this equates respectively to .004/.003 per million for types with the expansive ending, .01/.001 for types with the recessive ending, and 1.21/1.1 for types that have both endings.

[5] Since our explicit goal was to relate corpus frequency to user experiment data, one of the ways we aimed to make the two data sets converge was by having users react to data drawn from the corpus. This meant they were dealing with material that came from the stylistic and structural ambit of the corpus data. To reduce the possibility of respondents being distracted by extraneous material, influenced by similar constructions elsewhere in the sentence, or confused by complex syntax, we simplified or modified some of the sentences used. However, we did not always use the simplest possible sentence structures. Our hope was that in a questionnaire of significant length, having people read sentences that caught their attention in some way or exhibited varied structure would increase their attention span for the task.

[6] The endpoints on the scale were labelled: *1=naprosto normální (v rámci daného kontextu bych to určitě takto napsal/a); 7=nepřijatelné (v daném kontextu mi něco hodně „nesedí“, nepovažuji to za normální češtinu)* '1 = absolutely normal (in this context I would definitely write it that way); 7 = unacceptable (in this context something really doesn't feel right; I don't think it's normal Czech)'. Midpoints were not labelled; this encourages respondents to use the scale as equally-spaced points between 1 and 7, although there is no guarantee they will do so. The use of 1 as the high mark conforms to general Czech rating and marking systems.

[7] For the gap-filling, it was important that the context be clear enough to elicit the desired answer. Sometimes this meant inserting an adjective to make sure that a singular form was obtained. In a few places a plural was judged so unlikely that no adjective was inserted; however, in some instances respondents nonetheless used one. For some this may have represented an attempt at avoiding the task, possibly because they were unsure of the "right" answer (there is no choice to be made in the plural form).

[8] Ethical approval for this survey was sought and obtained under the University of Sheffield's Research Ethics Policy.

[9] We were interested here in regional differences, but did not wish to call that fact to our respondents' attention by using traditional terms like "Bohemian" or "Moravian" that might highlight dialect affinity. We therefore used the division into 14 *kraje* – modern administrative regions – which can handily be divided along dialectal lines. The only problematic region was Vysočina, which is bisected by Bohemian/Moravian dialect isoglosses, and thus analyses involving regional variables leave out respondents from this area. Respondents were asked to identify the area "they came from", presupposing that they would select the area for which they have the greatest affinity. They were also asked to indicate if they had lived anywhere else for a year or longer, but most did not indicate that this was the case.

[10] In 2009, there were 129.8 women studying at Czech universities for every 100 men (Český statistický úřad, 2011), meaning that women constituted 56.5% of tertiary students. Even if all our respondents had been current university students, however , this would only have predicted 312 female respondents, compared to the actual 329, so it cannot completely explain the disproportionate response from women.

[11] We asked for three levels: *ZŠ* (primary), *SŠ* (secondary), *VŠ* (tertiary). In addition, respondents were asked to indicate their field of study if they had finished university. The "Expected" column shows how many we might have expected to have in each area if the survey had been weighted to the proportions of the Czech population as a whole.

[12] As some studies have shown that linguists, or even specifically those linguists with specific theoretical training, may answer differently from other respondents due to their level of metaknowledge (Dąbrowska 2010), we tried to limit linguists' participation by specifically targeting students in modules on management, computer science and civics.

[13] The one result of $p = 0.07$ is just outside the conventional threshold for significance (1 in 20, or $p = 0.05$). In the context of seven other non-significant results, it is not worth examining this too closely.

[14] For a discussion of what this measure signifies, see section 9.

[15] As discussed earlier, Dácz et al. (2014), among others, have suggested that in nonce-word tasks, men rely statistically more on analogy and women on inference of general rules. Our data does not provide enough support for this, possibly because neither of our tasks involves unknown or little known lexemes requiring necessary resort to these processes.

[16] It shows up occasionally in combination with other factors, but the effect sizes are very small. In all probability this is a matter of one word that people from different regions judge slightly differently.

[17] By this we mean that the task of inserting a single response in a forced-choice question is a relatively easy and comprehensible task by comparison, familiar from school exercises and tests and from other questionnaires. As researchers we are not thereby absolved of interrogating those results with similar rigor (i.e. is it correct to deduce from the production of one variant that the other variant would not be produced?), but from the respondent's point of view the task is a simpler one.

[18] In combining the two ratings into one, all of the methods proposed strip out some information from the original data, and this one is no exception. Lost here is the absolute value of the ratings (overall "strictness" or "permissiveness" of each user). For example, if speaker S rates the {a} ending as 1 and the {u} ending as 3, the "delta" is 2. This gives speaker S the same score as speaker T, who rated {a} as 3 and {u} as 5. Speaker S is significantly more positive about both endings, but that fact is not captured in the final result. (The overall level of permissiveness is better captured by taking the sum of scores, but that calculation in turn strips out the strength and direction of preference, i.e. scores for competing variants of 1 and 7 give the same result as scores of 3 and 5, or of 7 and 1.)

[19] The direction of correlation is easiest to see on a hypothetical example. If the average rating for {ě} is 1.5 and the average rating for {u} is 3.5, that means respondents rate {ě} as better than {u}, and it results in a score of 2 (subtracting the {ě} score from the {u} score). A positive correlation thus indicates that the more definitively respondents like {ě}, the more they use {ě} (consonant with expectations). If the positions are reversed, then the average score for {ě} is 3.5 and the average score for {u} is 1.5. In this instance, respondents rate {u} as better than {ě}, resulting in a preference score of -2 (negative because we subtract the {ě} score from the {u} score). A negative correlation thus means that the more definitively people like {u}, the more they use {ě} (contrary to expectations). A score close to zero occurs when both forms get a similar rating (inconclusive).