Frequency data from corpora partially explain native-speaker ratings and choices in overabundant paradigm cells[1]

Neil Bermel, Luděk Knittl and Jean Russell

Abstract:
If we can operationalize corpus frequency in multiple ways, using absolute values and proportional values, which of them is more closely connected with the behaviour of language users? In this contribution, we examine *overabundant* cells in morphological paradigms, and look at the contribution that frequency of occurrence can make to understanding the choices speakers make due to this richness. We look at ways of operationalizing the term *frequency* in data from corpora and native speakers: the proportional frequency of forms (i.e. percentage of time that a variant is found in corpus data considered as a proportion of all variants) and several interpretations of absolute frequency (i.e. the raw frequency of variants in data from the same corpus).
Working with data from unmotivated morphological variation in Czech case forms, we show that different instantiations of frequency help interpret the way variation is perceived and maintained by native speakers. Proportional frequency seems most salient for speakers in forming their judgements, while certain types of absolute frequency seem to have a dominant role in production tasks.

Key words: corpus linguistics, frequency, morphology, empirical research, surveys, questionnaires, Czech, overabundance

1. Introduction

Frequency data are familiar territory for any linguist who works with corpora. We cite the number of times a feature appears, or its normalized frequency if we are comparing corpora; we cite percentages to show structure within categories or to demonstrate change over time. Hidden behind the way we deal with these data is an implicit *operationalization* of our questions about language. We have chosen to let the corpus stand in for a particular language, type of language, genre, etc., but at the same time we have also chosen representations of frequency that give us the best chance of answering our research questions. It is worth interrogating these differing operationalizations of frequency to see how the same data, approached in different ways, can shed a different light on the way native speakers apprehend and use language.

The term *frequency* is elastic, and once we start looking at *frequency data* there are few limits to the number of ways we can treat it. Divjak (2016) considers, among other meanings, the traditional *relative frequency* (incidence per million), *construction frequency* (which itself covers various ways of relating the frequencies between related items), *family frequency* (incorporating various ways of looking at the size and composition of a class of words) and *measures of probability and association*. These will largely be beyond the scope of this study, which is focused on how we understand and manipulate the numbers that arise from simple counts of individual forms.

Our material comes from three sources. We have data from the Czech National Corpus (CNC) on the frequencies of forms occupying a single morphological "slot". We selected items for inclusion in structured offline experiments, in which native speakers undertook both ratings tasks and gap-filling tasks. We will attempt to show how variation in the corpus data can be used in several different ways to look at the data from our questionnaires, shedding light on various aspects of how we assimilate linguistic data and how we produce it.

2. The research question

We were interested in how native speakers of inflectionally rich languages produce and handle forms in situations where more than one form is available for use. Czech, the language we will work with, is a highly inflected language. Its nominal system has six syntactic cases and a vocative form. Czech also distinguishes singular from plural forms in nominals, and has anywhere between 10 and 20 major declension patterns for nouns, depending on which grammar book you turn to. It distinguishes three or possibly four genders (masculine animate, masculine inanimate; feminine; neuter). The phonological system has undergone significant change and reorganization over the last thousand years, with the result that assignment to genders and declension patterns is no longer easily predictable from phonological shape.

These two features – ample numbers of inherited desinential suffixes and a loss of clear patterns for assignment of lexemes to particular genders and declension patterns – are ideal preconditions for morphological *overabundance* (Thornton 2012), a situation in which a single functional slot is occupied by two or more inflectional morphs or where there is stem allophony. This is a subset of a larger issue in morphosyntactic and morpholexical near-synonymy, whose exponents Baayen et al. (2013: 254) refer to as *rival forms;* their analysis includes not only the "classical" examples from inflectional morphology cited above, but also examples of word-formational near-synonymy.[2] Overabundance is in fact rife throughout the Czech declensional and conjugational systems, and poses issues in terms of how we describe those systems, as well as for how we explain the acquisition and maintenance of such a complex situation.

Studies of frequency of competing morphological variants have played a role in mapping out nativist vs. emergentist views of language, and our work also contributes to this discussion. Dąbrowska (2005, 2006) and Dąbrowska and Szczerbiński (2006) explored type frequency and token frequency as contributors to children's acquisition of case forms in Polish. In a later paper, Dąbrowska (2008) considered how performance on inflection tasks correlates with type frequency and neighbourhood density effects on adult performance. She also found correlations with education and vocabulary size, which she later followed up with a study demonstrating that professional exposure also influences judgements of syntactic well-formedness (Dąbrowska 2010). These results, taken together, suggest that morphosyntactic generalizations or "systems" emerge from usage and vary between individuals and social groups depending on the input they have had.

These approaches sit uncomfortably with a previously held tenet of language acquisition, represented here by the Principle of Contrast (Clark 1987: 2): "Every two forms contrast in meaning." Reading meaning broadly to include various functional and sociolinguistic markers, this represents a long-held consensus in the field. Corbett, for example, incorporates it in his concept of the canonical paradigm, in which each cell is occupied by a single exponent, and this serves as his starting point for the description of the actual situation in linguistic paradigms. But emergentist approaches suggest instead that mismatches and overgeneration of potential forms are natural features of language and of the ways in which individuals' language capabilities develop throughout their lives.

The account we will propose fits well in a cognitively-oriented approach. It presupposes that these case markers form grammatical elements of a composite structure profiling a particular case function. Taking our particular set of lexemes that in other cases nicely fit into a category of hard masculine inanimate nouns, we expect that here one element will emerge as the definitive marker of our two categories (the locative singular and the genitive singular) for all speakers in all situations. Instead, however, due to varied input, forms with two different grammatical elements can be entrenched, leading to conflicting schematizations (in the sense of Langacker 2008: 17). While in each case a single element {u} does emerge as a default for the category, a large number of elements continue to have variable marking. In Schmid's (2015:

10) model, we might say that the varied input individuals receive causes them to experience *entrenchment* of both forms to varying degrees, which is reflected in both forms being *conventionalized* for a significant number of lexemes.

Let us start with a simple example of morphological variation that is analogous to those mentioned in works such as Čech (2012), Cvrček and Kodýtek (2013).. We have data on variation in case forms from a large representative corpus of synchronic Czech, SYN2010.[3] The current example concerns the locative plural of so-called "soft" feminine nouns like *kost* 'bone', *růže* 'rose' and *píseň* 'song'We can present this material in at least three ways. The first is in rank order by form, as in Table 1.

Table 1. Tokens of locative plural forms in the SYN2010 corpus

nocích 'nights' (427), *pamětech* 'memories' (268), *nemocech* 'illnesses' (108), *nemocích* 'illnesses' (24), *pěstích* 'fists' (9), *nocech* 'nights' (5), *pěstech* 'fists' (4), *pamětích* 'memories' (3).[4]

We could also present these data paired as to the morphological 'cell' they occupy, as in Table 2, on the grounds that "morphological variation" presupposes that in a given environment one of a limited number of variants will occur (generally not more than three).

Table 2. Tokens of locative plural forms presented as competing in morphological 'cells'

pamětech (268)	<>	*pamětích* (3)
nemocech (108)	<>	*nemocích* (24)
nocech (5)	<>	*nocích* (427)
pěstech (4)	<>	*pěstích* (9)

The rows in Table 2 are ordered by frequency of the -ech variant, but could of course have been ordered by the -ích variant as well. Reworking Table 2, we can express these figures using percentages of each ending chosen, to arrive at Table 3.

Table 3. Tokens of locative plural forms presented in morphological oppositions (percentages)

pamětech (98.9%)	<>	*pamětích* (1.1%)
nemocech (81.8%)	<>	*nemocích* (18.2%)
pěstech (30.8%)	<>	*pěstích* (69.2%)
nocech (1.2%)	<>	*nocích* (98.8%)

Table 3 is ordered by proportion of the two variants, and looks at first glance more transparent: it makes clear that the dominant forms in the corpus are *pamětech, nemocech, pěstích, nocích* rather than *pamětích, nemocích, pěstech, nocech.* However, in replacing actual form counts with percentages we have eliminated some information. In Table 1 we can see that there are far fewer tokens of the "majority" form *pěstích* than there are of the "minority" form *nemocích*. Would it therefore be justified to assume that *nemocích* is a "dispreferred" form, when in fact it is more frequent in absolute terms than the supposedly "preferred" form *pěstích*?[5]

Each of these approaches captures at least one important truth about our data and disregards another. The arrangement in Table 1 displays the data extracted from the corpus: in this format all the data are measured using the same yardstick and we do not rely on *a priori* knowledge of grammatical structures – in other words, we do not regard it as relevant which forms are "in competition" with other forms and which paradigm (*píseň* or *kost*) a particular form belongs to. The arrangement in Table 3 sets the corpus data within previous linguistic structures of which we claim to have knowledge, and we then use these structures as the basis for our analysis. In this approach, we foreground the relationship between forms in a single

morphological "cell" and also enable direct comparisons between "competing" forms more widely. Table 2 occupies an intermediate position, offering a comparison of "competing" forms within their morphological slots but with the corpus data in their original format.

Starting in Section 3, we focus on the following research question: if we can operationalize corpus frequency in multiple ways, using absolute values and proportional values, which of them is more closely connected with the behaviour of language users? To look at behaviour in sections 4 and following, we will consider two types of offline experimental tasks: linguistic ratings and gap filling (the latter, in the context of cells with only two common possibilities, amounts to a forced choice between those two options).

3. The broader context: absolute and proportional frequency

We often find we have to describe and analyze our data by partitioning it in some fashion. This might be dividing a continuous scale into discrete zones or looking at distinct categories. Some types of statistical analysis (for example, ANOVA or chi-squared) require the characteristics of our data to be expressed as categories or "bins" within a category rather than being arranged on a scale. This grouping of results can help the researcher to discern the relevance or significance of frequency in his or her data.

In our case we considered allocating our data to bins for both absolute and proportional frequency, which here and elsewhere we have called *bands*. Ideally the cut-off points between these bands would be determined by some shift in behaviour, distribution or use of the forms in question, although as we will see, this requirement is not easy to meet.

We will start with the question of how to operationalize the notion of "absolute frequency".

Corpus linguists have steered clear of stating a single cut-off point between high and low absolute frequency. In part this has been because the overall frequency of the features studied is so variable that such attempts would seem entirely arbitrary.

Bybee (2007: 16), for example, works primarily with absolute frequencies and suggests setting the boundaries between bands individually for each feature. Her criteria are: (1) the existence of "frequency gaps" that divide the scale into two parts, for which (2) each part contains 30 to 70 percent of the lemmas.[6] Implicitly, then, the approach is like that in Table 2 above. Absolute frequency is cited, but in making our division into *sorts* of absolute frequency, we operate with the notion of items in competition for a slot. Our assumed prior structural knowledge is not discarded, but used to organize our data.

The data we will use in this study come from variation in the so-called "hard" declension pattern of Czech masculine inanimate nouns. This variation occurs primarily in two cases, the genitive singular and the locative singular, where it is found with many common nouns; hence, the genitive singular of the word *jazyk* 'tongue, language' can be either *jazyku* or *jazyka,* and the locative singular of the same word can be either *jazyku* or *jazyce.* Our corpus data come from the SYN2010 corpus of the CNC, which has something over 100m word forms and is thus 50 to 100 times larger than the corpora used in Bybee's studies.[7] For frequencies equivalent to those she uses, we would set boundaries in our 100-million-token corpus of between 850 and 9090 tokens.

The features we tested were less frequent than Bybee's, however – especially if we take into account the fact that each Czech word has forms for seven cases and two numbers. Their distribution in the corpus corresponds to Zipf's law, i.e. there is a small number of words with very high frequency, but the majority of words with such variation have low frequency. In the case of variation in the masc. gen. sg. of the type *jazyka/jazyku*, 52 of the 112 words that display variation have fewer than 100 tokens, and in the case of the masc. loc. sg. of the type *hradě/hradu*, 186 of the 391 lexemes have fewer than 100 tokens. To avoid the uncertainty that comes from relying on low-frequency data, we decided only to test lexemes where we had over 100 tokens for the given case forms attested in the corpus.[8]

In both cases there exists a "frequency gap" (see Bybee 2007 above) at around 1000 tokens, dividing the lexemes into two unequal groups. The high-frequency group contains only 52 of the 391 Lsg forms (13.3%) and 23 of the 112 Gsg forms (20.5%). We could, of course, look for a lower boundary so that the groups were more evenly balanced, but we decided to keep this 1000-token dividing line. The primary reason was that the term *high frequency* is rarely taken to mean fewer than 7–8 in a million-token corpus (i.e. 700–800 in a 100-million-token corpus) and we wished to retain the possibility that our results would have broader relevance beyond this single instance.

The second part of this operationalization concerns *proportional frequency*. In corpus-based studies this concept is quite common, but as with absolute frequency, we meet with various approaches to what it means. Halliday (2005 [1992]: 68–70) proposes boundaries of roughly 9:1 and 1:9, outside which the probabilities are "skew": one term is unmarked or default and one constitutes a choice that adds information value. In instances where the proportions are more evenly balanced (e.g. 4:1, 2:1, 3:2) he says we are dealing with an "equiprobable" distribution, where stylistic, semantic or other functional variations come to the fore.[9] Hare et al. (2001) propose a division into three bands with the boundaries 1:2 and 2:1. Some other systems proposed for Czech can be seen in Table 4:

Table 4. Dividing examples of morphological variation into bands

Source: Bermel and Knittl (2012)
Target: Frequency bands for morphological competition (percentage of "recessive" endings {a}, {ě})
Aim: Describe empirical results of acceptability studies as related to original corpus data

isolated	*marked*	*minority*	*equipollent*	*majority*	*unmarked*	*dominant*
0–1%	1–9%	10–29%	30–69%	70–89%	90–99%	99–100%

Source: Hebal-Jezierska (2008)
Target: Frequency bands for each variant of the Npl. ({i}, {ové}, {é})
Aim: Corpus frequency bands that match with usage in context (corpus data)

sporadic	*variant*	*dominant*
0–1%	1–14%	15–100%

Source: Cvrček et al. (2010), further elaborated in Cvrček and Kodýtek (2013)
Target: Frequency bands for each morphological variant
Aim: Description of corpus frequency for a comprehensive user's grammar of Czech

never / almost never	*rarely*	*sometimes*	*same*	*frequent*	*regularly*	*always / almost always*
0–1%	1–10%	10–35%	35–65%	65–90%	90–99%	over 99%

Source: Šimandl 2010
Target: Frequency bands for each morphological opposition
Aim: Description of variation based on corpus results

marginal - monopolistic	*minority - majority*	*equipollent*	*majority - minority*	*monopolistic - marginal*
≤ 5% > 95.0%	5.1–39.9% 94.5–60.1%	40–60% 40–60%	60.1–94.5% 39.9–5.1%	> 95.0% ≤ 5%

Source: Hebal-Jezierska and Bermel (2011)
Target: Frequency bands for morphological variants
Aim: Delimit bands that determine variants' usage in context and predict users' reactions to them

sporadic	*minority*	*majority*
0–2%	2–49%	50–100%

These scales all have one thing in common: bands characterizing a more even distribution of variants (i.e. the "middle" bands) are broader than the ones showing a distribution skewed towards one variant or the other (i.e. the "outside" bands), with the low-frequency variant having a very narrow range. There are two reasons for this that are taken up in Hebal-Jezierska's study and in our own: the first relates to the analysis of results and the second to the way the research question is operationalized.

As far as results go, the available literature offers two methods of analysis. Either we proceed purely on the basis of corpus data, or we add data from other sources, such as questionnaires. Hebal-Jezierska researched corpus data and did not find any substantial differences in the way items in her broad middle band functioned: by this she meant that there were no limitations, stylistic or otherwise, in the usage of forms appearing more than 15 percent of the time in the given slot. For forms appearing less than 15 percent of the time in a given slot, she found a range of limitations in the way they could be used. The greater the difference in their relative frequency, the greater the divergence in the functionality of competing members of an opposition – in other words, between a low-frequency ending and its competitor which is used far more regularly. The results of acceptability questionnaires (see e.g. Bermel and Knittl 2012a, 2012b) confirmed that native speakers sense a difference between forms that are rarely used and those that they meet more regularly in written texts.

As far as the operationalization of corpus data in research goes, for us to have enough material, we need to have enough sample words in each of our bands. In the two cases we tested, however, lexemes tend towards a bimodal distribution – that is, there are more of them in the outer frequency bands (proportional frequency of 0–10% and 90–100%). In the middle bands there are fewer of them, and in order to have a sufficient choice of suitable lexemes, we had to broaden out the central band in our research.[10] We will return to this point later.

4. Methodology

This section gives an overview of the set-up of our experiments. In our project "Acceptability and forced-choice judgements in the study of linguistic variation" we have been exploring factors influencing native speakers as they evaluate morphological variants and choose between them. We hypothesized that there is a relationship (correlation) between corpus frequency and the reactions of native speakers, both in their evaluations and their choices. What is more, we assume that corpus frequency, because it stands in for the sort of linguistic experience that users have, can be used as a proxy for experience as a determining factor that influences them in their actions.

Scholars of Czech have used qualitative studies to point out the variation found in the Gsg and Lsg masculine (the so-called *hrad* paradigm, which has examples such as *toho jazyka/jazyku*, *na hradě/na hradu*).[11] These works point to further factors, including meaning (for polysemous lexemes), syntactic context and regional features; our research, however, was focused not on the detailed study of texts, but rather on the overall frequency of relevant examples.

To measure the relationships between competing forms, we first had to operationalize the concept of *corpus frequency,* that is, to decide whether absolute or relative frequency was to be used for this measure. We chose words in two absolute-frequency bands and four proportional-frequency bands, and in this way, we arrived at eight frequency "cells" to be tested:

Table 5. Structure of the questionnaire with the lexemes used

prop. of {a/ě} / abs. freq.	**A**: 0–5%	**B**: 5–50%	**C**: 50–95%	**D**: 95–100%
1: 0–999 tokens	**A1**	**B1**	**C1**	**D1**
Gsg	*kožich* ‘fur coat’ *šuplík* ‘drawer’	*obdélník* ‘rectangle’ *velín* ‘control room’	*čtvrtek* ‘Thursday’ *komín* ‘chimney’	*oběd* ‘lunch’ *ocet* ‘vinegar’
Lsg	*stadion* ‘stadium’ *výraz* ‘expression’	*list* ‘sheet’ *kanál* ‘canal’	*fotbal* ‘football’ *strom* ‘tree’	*klášter* ‘cloister’ *nos* ‘nose’
2: 1000+ tokens	**A2**	**B2**	**C2**	**D2**
Gsg	*podzim* ‘autumn’ *zákoník* ‘law code’	*sen* ‘sleep, dream’[12] --	*kout* ‘corner’ *rybník* ‘pond’	*kostel* ‘church’ *národ* ‘nation’
Lsg	*pád* ‘fall’ *parlament* ‘parliament’	*koncert* ‘concert’ *obvod* ‘district’	*les ‘forest’,* *úřad ‘office’*	*okres* ‘district’ *stát* ‘state’

For each cell in Table 5 we chose lexemes that we tested in two different ways: ratings of the two variants (e.g. *koncertě–koncertu*) and gap filling (e.g. *koncert*____) in sentence-long contexts.

Gap-filling triggers were presented with the ending missing as follows:

Na slavnostním [koncert...]_____ zahraje Pražská komorní filharmonie s dirigentem Jiřím Bělohlávkem.

‘The Prague Chamber Philharmonic under the direction of Jiří Bělohlávek will play at a gala ***concert****.’*

Ratings triggers were presented with both forms and a scale on which to rate them:

Nemůžu psát o (koncertu / koncertě) a zamlčet, za jakých podmínek se udál.

\+ 1 2 3 4 5 6 7 / 1 2 3 4 5 6 7 −

‘I can’t write about the ***concert*** *without mentioning the conditions in which it took place.’*

Respondents saw each lexeme twice in different sentences and different syntactic contexts that are characteristic for the given case (above: the locative case used with locational and non-locational prepositions respectively). The use of two lexemes in each band was designed to moderate any unintended lexical effects; with a second lexeme in each category, any peculiarities about the usage of a single lexeme would become clear when the data were analyzed.

Each respondent answered both types of questions: gap filling and rating. Ideally, of course, all respondents would complete all tasks for all lexemes, allowing us to eliminate between-group features as a factor. However, this presented two problems. First, an experiment in which they completed both tasks on all sentences would have been unfeasibly long (32 lexemes × 2

contexts × 2 tasks plus distractor sentences). Second, it would have exposed us to the risk that respondents would react differently the second time they saw the same trigger sentences in a different task.

Two steps were taken to skirt these difficulties. First, the lexemes were divided into two groups, with each containing one lexeme from each cell. Respondents were thus divided randomly into groups that filled in parallel questionnaires using different lexical data. Second, they completed different tasks on different words: the sentences were divided amongst different versions of the questionnaire in a block design such that those who evaluated forms in cells A1, B2, C1 and D2 (shaded grey in Table 5) would fill in gaps in cells A2, B1, C2 and D1 (shaded white in Table 5) and vice-versa.[13] These two adjustments allowed us to keep the questionnaire to a manageable length for both parallel versions.

We obtained 587 completed questionnaires through recruitment at universities, academic high schools and work places in various parts of the Czech Republic.[14] Once incorrectly completed papers and those by non-native speakers were excluded, we had 551 usable returns. Using t-tests we were able to compare the results of individual versions of the questionnaire and confirmed that the ordering of questions and tasks had had no significant effect on the respondents' answers. In the same way, we found no significant differences between groups of respondents as to their composition regarding age, education or gender. A fuller discussion of the gender, age and educational profile of respondents can be found in Bermel et al. (2015: 290–292).[15]

5. Description of corpus data

In the following pages, we will examine how the lexemes we used in our experiments appear in a large, balanced synchronic corpus of Czech. Once again, our data can be reported in various ways, which may influence the findings. We consider a few of them in this overview. For corpus frequency, as previously, indicated, we had selected words from a variety of proportional frequency bands, but for both cases the proportional frequency is unevenly distributed (see Section 3 on the bimodal distribution of these forms), as is evident in Table 6.

Table 6. Corpus frequency by percentage, divided into quartiles (grey shading)

category	form	gloss	corpus	category	form	gloss	corpus
{u} Gen	kostelu	'church'	0.02%	{e} Loc	stadioně	'stadium'	1.49%
{a} Gen	podzima	'autumn'	0.12%	{e} Loc	pádě	'case'	1.69%
{u} Gen	obědu	'lunch'	0.25%	{u} Loc	nosu	'nose'	2.68%
{a} Gen	zákoníka	'law code'	0.45%	{e} Loc	výraze	'expression'	3.16%
{a} Gen	šuplíka	'drawer'	0.63%	{u} Loc	kláštera	'cloister'	3.46%
{u} Gen	octu	'vinegar'	0.75%	{u} Loc	okresu	'county'	5.46%
{a} Gen	kožicha	'fur coat'	1.56%	{e} Loc	parlamentě	'parliament'	7.30%
{u} Gen	národu	'nation'	3.75%	{u} Loc	státu	'state'	12.57%
{u} Gen	čtvrtku	'Thursday'	6.65%	{u} Loc	ledu	'ice'	13.56%
{a} Gen	obdélníka	'rectangle'	8.55%	{u} Loc	stromu	'tree'	15.11%
{u} Gen	rybníku	'pond'	11.86%	{u} Loc	fotbalu	'football'	18.20%
{a} Gen	sna	'sleep'	14.28%	{e} Loc	obvodě	'district'	18.56%
{a} Gen	sna	'sleep'	14.28%	{e} Loc	listě	'sheet'	31.38%
{u} Gen	koutu	'corner'	14.73%	{e} Loc	kanále	'canal'	39.66%
{u} Gen	komínu	'chimney'	18.67%	{u} Loc	úřadu	'office'	40.80%
{u} Gen	velínu	'control room'	50.00%	{u} Loc	koncertu	'concert'	43.94%

{a} Gen	velína	‘control room’	50.00%	{e} Loc	koncertě	‘concert’	56.06%
{a} Gen	komína	‘chimney’	81.33%	{e} Loc	úřadě	‘office’	59.20%
{a} Gen	kouta	‘corner’	85.27%	{u} Loc	kanálu	‘canal’	60.34%
{u} Gen	snu	‘sleep’	85.72%	{u} Loc	listu	‘sheet’	68.62%
{u} Gen	snu	‘sleep’	85.72%	{u} Loc	obvodu	‘district’	81.44%
{a} Gen	rybníka	‘pond’	88.14%	{e} Loc	fotbale	‘football’	81.80%
{u} Gen	obdélníku	‘rectangle’	91.45%	{e} Loc	stromě	‘tree’	84.89%
{a} Gen	čtvrtka	‘Thursday’	93.35%	{e} Loc	ledě	‘ice’	86.44%
{a} Gen	národa	‘people’	96.25%	{e} Loc	státě	‘state’	87.43%
{u} Gen	kožichu	‘fur coat’	98.44%	{u} Loc	parlamentu	‘parliament’	92.70%
{a} Gen	octa	‘vinegar’	99.25%	{e} Loc	okrese	‘county’	94.54%
{u} Gen	šuplíku	‘drawer’	99.37%	{e} Loc	klášteře	‘cloister’	96.54%
{u} Gen	zákoníku	‘law code’	99.55%	{u} Loc	výrazu	‘expression’	96.84%
{a} Gen	oběda	‘lunch’	99.75%	{e} Loc	nose	‘nose’	97.32%
{u} Gen	podzimu	‘autumn’	99.88%	{u} Loc	pádu	‘case’	98.31%
{a} Gen	kostela	‘church’	99.98%	{u} Loc	stadionu	‘stadium’	98.51%

Quartiles can be helpful in seeing the distribution of the data in Table 6. If our data were evenly spread, then each quartile would cover 25% of the range and the interquartile range would be 50%. For normally distributed data, the middle quartiles would cover less of the scale, as there would be more items around the 50% mark.

It is obvious that our data are neither evenly spread nor normally distributed. For the genitive case forms, the first quartile goes from 0.0% to 5.9%[16]; the second quartile from there to 50%; the third quartile to 94%; and the fourth quartile to 100%. The interquartile range covers 88.14% of the data. The vast bulk of the forms here appear in the corpus either in a highly *dominant* position vis-à-vis the competing form, or in a highly *recessive* position vis-à-vis the competing form. For the locative case forms, the first quartile goes from 1.5% to 13.3%; the second quartile from there to 50%; the third quartile to 86.7%; and the fourth quartile to 98.5%. The interquartile range covers 73.4% of the data. The locative case forms thus are thus somewhat more evenly distributed, but still do not approach a normal distribution of forms.

In sum, the corpus data shows a distribution favouring Halliday’s “skew” types: a predominance of competing forms have a clear “majority” and “minority” form. This tendency is more pronounced in the gen. sg. than the loc. sg. and reflects the overall distribution of lexemes showing variation in these slots.

We said that the second way of reporting frequency was to take the absolute values – in other words, the actual number of times a particular form is found in the corpus. If we look at our corpus data from this angle, as seen in Table 7, the results are on a different scale.

Table 7. Corpus frequency by form, divided into quartiles

category	**form**	**gloss**	**corpus**	**category**	**form**	**gloss**	**corpus**
{a} Gen	šuplíka	‘drawer’	1	{e} Loc	výraze	‘expression’	8
{a} Gen	kožicha	‘fur coat’	2	{e} Loc	stadioně	‘stadium’	9
{a} Gen	podzima	‘autumn’	4	{u} Loc	nosu	‘nose’	20
{u} Gen	obědu	‘lunch’	5	{u} Loc	kláštera	‘cloister’	28
{u} Gen	octu	‘vinegar’	5	{e} Loc	pádě	‘case’	32
{a} Gen	zákoníka	‘law code’	8	{u} Loc	stromu	‘tree’	81

{a} Gen	obdélníka	‘rectangle’	14	{u} Loc	okresu	‘county’	93
{u} Gen	kostelu	‘church’	33	{e} Loc	parlamentě	‘parliament’	116
{u} Gen	velínu	‘control room’	33	{u} Loc	fotbalu	‘football’	125
{a} Gen	velína	‘control room’	33	{e} Loc	kanále	‘canal’	161
{u} Gen	čtvrtku	‘Thursday’	36	{e} Loc	obvodě	‘district’	243
{u} Gen	komínu	‘chimney’	98	{u} Loc	kanálu	‘canal’	245
{u} Gen	obdélníku	‘rectangle’	107	{u} Loc	výrazu	‘expression’	245
{u} Gen	národu	‘nation’	110	{e} Loc	listě	‘sheet’	266
{u} Gen	rybníku	‘pond’	125	{u} Loc	ledu	‘ice’	306
{u} Gen	kožichu	‘fur coat’	126	{u} Loc	státu	‘state’	355
{u} Gen	koutu	‘corner’	149	{e} Loc	stromě	‘tree’	455
{u} Gen	šuplíku	‘drawer’	158	{e} Loc	fotbale	‘football’	562
{a} Gen	sna	‘sleep’	291	{e} Loc	koncertě	‘concert’	572
{a} Gen	sna	‘sleep’	291	{u} Loc	listu	‘sheet’	581
{a} Gen	komína	‘chimney’	427	{u} Loc	koncertu	‘concert’	584
{a} Gen	čtvrtka	‘Thursday’	552	{u} Loc	stadionu	‘stadium’	596
{a} Gen	octa	‘vinegar’	659	{u} Loc	úřadu	‘office’	716
{a} Gen	oběda	‘lunch’	811	{e} Loc	nose	‘nose’	726
{a} Gen	rybníka	‘pond’	929	{e} Loc	klášteře	‘cloister’	781
{a} Gen	kouta	‘corner’	938	{e} Loc	úřadě	‘office’	1039
{u} Gen	snu	‘sleep’	1320	{u} Loc	obvodu	‘district’	1066
{u} Gen	snu	‘sleep’	1320	{e} Loc	ledě	‘ice’	1422
{u} Gen	podzimu	‘autumn’	1395	{u} Loc	parlamentu	‘parliament’	1473
{u} Gen	zákoníku	‘law code’	1776	{u} Loc	pádu	‘case’	1548
{a} Gen	národa	‘people’	2821	{e} Loc	okrese	‘county’	1609
{a} Gen	kostela	‘church’	4039	{e} Loc	státě	‘state’	2470

The values in Table 7 run from 1 to past 4000. The distribution here is an artefact of our questionnaire structure: as described earlier, we selected half the *lexemes* to have total frequencies in the relevant case of over 1000 tokens, and half under 1000. The distribution of *forms,* however, is skewed towards the lower end.

For the genitive, the first quartile contains forms with 1–33 tokens; the second quartile contains forms with 33–126 tokens; the third quartile contains forms with 149–811 tokens, and all of our “high-frequency” forms are in the fourth quartile, whose forms have 929–4039 tokens. There is a relatively low median (138 tokens) and an interquartile range of 808 tokens.

For the locative, the distribution is not quite as broadly spread. The first quartile runs from 8 to 116 tokens; the second quartile from 125 to 355 tokens; the third quartile from 455 to 726 tokens; and the fourth quartile covers the range from 781 to 2470 tokens. There is a higher median (405) than for the genitive, and the interquartile range, at 617, is smaller as well.

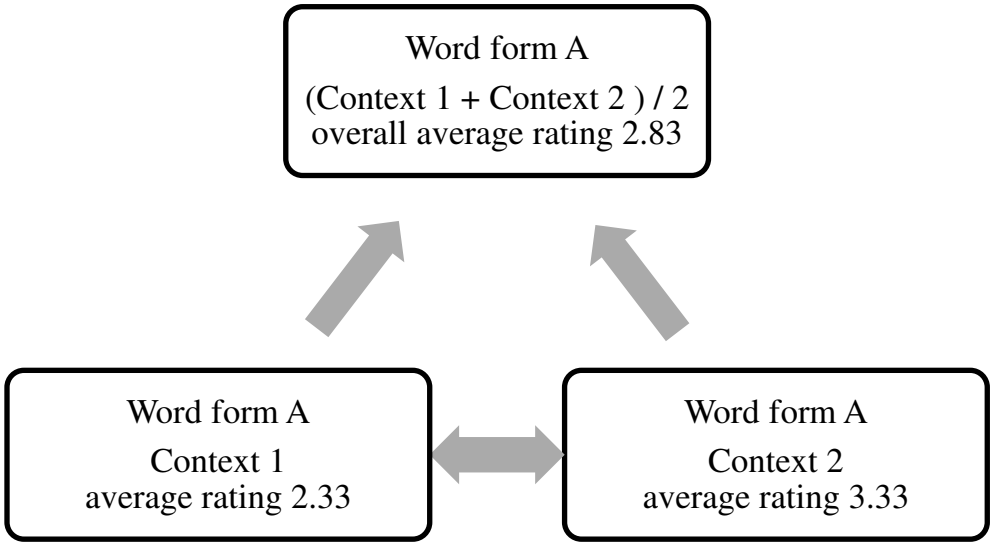
The absolute-frequency data is thus also broadly reflective of the distribution of these forms in the corpus. However, we required that any given slot have a minimum of 100 forms to appear in the questionnaire, and therefore the long “tail” of *hapax legomena* characteristic of the absolute distribution of forms is lacking here.

6. Description of acceptability data

We now turn to the descriptive statistics that allow us to draw some preliminary conclusions about our data. Section 6 reports on the acceptability ratings; Section 7 reports on the gap-filling task.

There are several ways to report the acceptability ratings. We could conceivably look at *average ratings* for each word form. With this method, we have two contexts in which each word form appears, so we take the average rating of the first context and the average rating of the second context and then take the midpoint between them, as in Figure 1.

Figure 1. Calculating average ratings of word forms



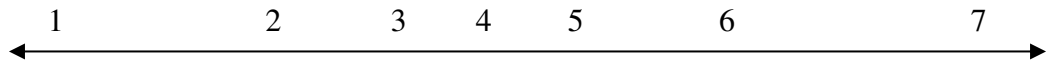
A second way would be to look at the *median rating*. Our respondents rated on a 7-point scale, so we take as the median rating the point at which we have 50% of the ratings accounted for (the "cumulative percent" column). In the case of the two contexts in Table 8, we reach the threshold at point 3 for the first context and point 2 for the second context. The value we assign to this word form's rating is the average of the two, i.e. of 3 and 2 = 2.5.

Table 8. Median values for word forms (example).

Rating	Frequency	Percent	Cumulative Percent	Rating	Frequency	Percent	Cumulative Percent
0	1	.7	.7	0	0	0.0	0.0
1	36	26.1	26.8	1	59	42.8	42.8
2	26	18.8	45.7	**2**	**30**	**21.7**	**64.5**
3	**24**	**17.4**	**63.0**	3	15	10.9	75.4
4	19	13.8	76.8	4	18	13.0	88.4
5	14	10.1	87.0	5	6	4.3	92.8
6	8	5.8	92.8	6	5	3.6	96.4
7	10	7.2	100.0	7	5	3.6	100.0

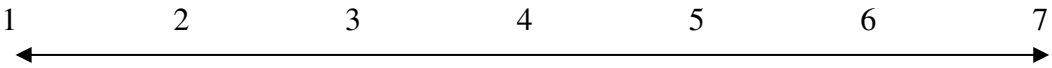
The method in Table 8 has the advantage of not treating the ratings as interval values, i.e. it allows the possibility that respondents have a scale that looks, for example, like the one in Figure 2, with greater differences between the ratings at either extreme than the middle values on the scale.

Figure 2. Ordinal, non-interval scale



Averages, on the other hand, assume a scale like the one in Figure 3, with equal spacing between all ratings.

Figure 3. Ordinal interval scale



In the statistical analysis in Section 8, we will adopt the averaging method, treating the scale as an interval variable for reasons that will be explained therein. For this simple descriptive task, however, without the benefit of statistical corrections, we adopt the median method as a more conservative approach to the data, as seen in Table 9.

Table 9. Acceptability ratings, by quartile

category	form	gloss	rating	category	form	gloss	rating
{a} Gen	velína	'control room'	1	{e} Loc	fotbale	'football'	1
{a} Gen	komína	'chimney'	1	{e} Loc	stromě	'tree'	1
{u} Gen	snu [1] [17]	'sleep'	1	{u} Loc	parlamentu	'parliament'	1
{a} Gen	rybníka	'pond'	1	{e} Loc	okrese	'county'	1
{a} Gen	čtvrtka	'Thursday'	1	{e} Loc	klášteře	'cloister'	1
{a} Gen	národa	'nation'	1	{u} Loc	výrazu	'expression'	1
{u} Gen	kožichu	'fur coat'	1	{u} Loc	pádu	'case'	1
{a} Gen	octa	'vinegar'	1	{e} Loc	koncertě	'concert'	1.5
{u} Gen	šuplíku	'drawer'	1	{e} Loc	úřadě	'office'	1.5
{u} Gen	zákoníku	'law code'	1	{e} Loc	ledě	'ice'	1.5
{a} Gen	oběda	'lunch'	1	{e} Loc	státě	'state'	1.5
{u} Gen	podzimu	'autumn'	1	{e} Loc	nose	'nose'	1.5
{a} Gen	kostela	'church'	1	{u} Loc	stadionu	'stadium'	1.5
{u} Gen	snu [2]	'sleep'	1.5	{u} Loc	ledu	'ice'	2
{u} Gen	koutu	'corner'	2	{e} Loc	obvodě	'district'	2
{a} Gen	obdélníka	'rectangle'	2.5	{u} Loc	kanálu	'canal'	2
{u} Gen	obdélníku	'rectangle'	2.5	{u} Loc	listu	'sheet'	2
{u} Gen	rybníku	'pond'	3	{u} Loc	obvodu	'district'	2
{u} Gen	komínu	'chimney'	3	{e} Loc	kanále	'canal'	2.5
{a} Gen	kouta	'corner'	3	{e} Loc	stadioně	'stadium'	3
{a} Gen	zákoníka	'law code'	3.5	{u} Loc	okresu	'county'	3
{u} Gen	octu	'vinegar'	3.5	{u} Loc	státu	'state'	3
{a} Gen	sna [2]	'sleep'	3.5	{e} Loc	listě	'sheet'	3
{u} Gen	čtvrtku	'Thursday'	4	{u} Loc	úřadu	'office'	3
{u} Gen	velínu	'control room'	4	{u} Loc	koncertu	'concert'	3
{u} Gen	obědu	'lunch'	4.5	{u} Loc	nosu	'nose'	3.5

{u} Gen	národu	‘nation’	4.5	{e} Loc	výraze	‘expression’	3.5
{a} Gen	podzima	‘autumn’	5	{u} Loc	stromu	‘tree’	3.5
{a} Gen	šuplíka	‘drawer’	5	{u} Loc	fotbalu	‘football’	3.5
{a} Gen	kožicha	‘fur coat’	5	{u} Loc	klášteru	‘cloister’	4
{a} Gen	sna [1]	‘sleep’	5	{e} Loc	parlamentě	‘parliament’	4
{u} Gen	kostelu	‘church’	5.5	{e} Loc	pádě	‘case’	6

On average, the median rating in Table 9 for the genitive forms is more generous than that for the locative forms at the top end of the scale (remembering here that our rating system uses a high mark of 1 and a low mark of 7).

For the genitive, the first quartile is entirely made up of median marks of 1, i.e. the highest mark; the median rating overall is 2.5 out of 7, meaning that the overall ratings were quite positive. The third quartile goes only as far as 4, the midpoint of the scale, and the lowest rating is 5.5, a reasonable distance off the bottom point of 7. The interquartile range covers three marks.

For the locative, the ratings are more evenly spread, with fewer very positive and very negative ratings. The first quartile comprises ratings of 1 and 1.5. The median rating overall is slightly more positive than the genitive, at 2.5, and the fourth quartile, with the most negative ratings, stretches from 3.5 to 6. The interquartile range covers just 1.5 marks, showing a concentration of ratings in the central area.

We note that the forms *sna/snu* ‘sleep’, which appear in both questionnaires, got differing results in each questionnaire. The first group were more categorical in their evaluations, giving much higher ratings to *snu* than to *sna.* The second group also rated *snu* highly, but were much happier with *sna* in one of the two contexts, giving it an overall higher rating. We will return to this point in a moment.

7. Description of gap-filling data

For the gap-filling exercise, we counted the number of answers for one ending vs. the total number of answers provided: with *x* examples of ending *A* and *y* examples of ending *B*, our percentages are derived by calculating $x/(x+y)$ and $y/(x+y)$, as in Table 10.

Table 10. Percentage of times this form filled in, by quartile

category	form	gloss	chosen	category	form	gloss	chosen
{u} Gen	kostelu	‘church’	4.3%	{e} Loc	pádě	‘case’	2.5%
{u} Gen	obědu	‘lunch’	7.4%	{e} Loc	parlamentě	‘parliament’	10.4%
{a} Gen	kožicha	‘fur coat’	8.3%	{e} Loc	výraze	‘expression’	10.7%
{a} Gen	sna [1]	‘sleep’	9.1%	{e} Loc	stadioně	‘stadium’	11.2%
{a} Gen	šuplíka	‘drawer’	9.5%	{u} Loc	okresu	‘county’	19.6%
{u} Gen	národu	‘nation’	12.8%	{u} Loc	státu	‘state’	21.7%
{a} Gen	podzima	‘autumn’	13.5%	{u} Loc	klášteru	‘cloister’	22.3%
{a} Gen	zákoníka	‘law code’	16.4%	{u} Loc	fotbalu	‘football’	23.2%
{u} Gen	čtvrtku	‘Thursday’	16.7%	{e} Loc	listě	‘sheet’	32.3%
{u} Gen	velínu	‘control room’	17.9%	{u} Loc	koncertu	‘concert’	33.7%
{a} Gen	obdélníka	‘rectangle’	18.8%	{u} Loc	stromu	‘tree’	38.2%
{u} Gen	octu	‘vinegar’	20.1%	{u} Loc	úřadu	‘office’	38.9%
{u} Gen	rybníku	‘pond’	21.6%	{e} Loc	obvodě	‘district’	42.3%
{u} Gen	komínu	‘chimney’	21.9%	{u} Loc	nosu	‘nose’	44.4%
{a} Gen	sna [2]	‘sleep’	33.6%	{e} Loc	ledě	‘ice’	47.2%

{a} Gen	kouta	‘corner’	41.1%	{u} Loc	kanálu	‘canal’	47.5%
{u} Gen	koutu	‘corner’	56.7%	{u} Loc	ledu	‘ice’	50.7%
{u} Gen	snu [2]	‘sleep’	66.4%	{e} Loc	kanále	‘canal’	52.5%
{a} Gen	komína	‘chimney’	78.1%	{e} Loc	nose	‘nose’	55.6%
{a} Gen	rybníka	‘pond’	78.4%	{u} Loc	obvodu	‘district’	57.7%
{a} Gen	octa	‘vinegar’	79.9%	{e} Loc	úřadě	‘office’	61.1%
{u} Gen	obdélníku	‘rectangle’	81.2%	{e} Loc	stromě	‘tree’	61.5%
{a} Gen	velína	‘control room’	82.1%	{e} Loc	koncertě	‘concert’	65.2%
{a} Gen	čtvrtka	‘Thursday’	82.6%	{u} Loc	listu	‘sheet’	65.2%
{u} Gen	zákoníku	‘law code’	83.6%	{e} Loc	fotbale	‘football’	75.7%
{u} Gen	podzimu	‘autumn’	86.2%	{e} Loc	klášteře	‘cloister’	77.0%
{a} Gen	národa	‘nation’	87.2%	{e} Loc	okrese	‘county’	77.9%
{u} Gen	kožichu	‘fur coat’	89.5%	{e} Loc	státě	‘state’	78.3%
{u} Gen	snu [1]	‘sleep’	89.9%	{u} Loc	stadionu	‘stadium’	87.7%
{u} Gen	šuplíku	‘drawer’	90.5%	{u} Loc	výrazu	‘expression’	89.3%
{a} Gen	oběda	‘lunch’	91.8%	{u} Loc	parlamentu	‘parliament’	89.6%
{a} Gen	kostela	‘church’	93.5%	{u} Loc	pádu	‘case’	96.8%

The gap-filling exercise is essentially a forced-choice situation, as there are in practice only two options in each environment. When we know that for lexeme *A,* a respondent has chosen ending *x* over ending *y*, then the percentage of *A+x* increases, while the percentage of *A+y* decreases. The profile of these data, like the corpus data, is thus always roughly symmetrical (any differences being accounted for by the occasional missed answer). That said, the genitive data in Table 10 is clearly bimodal while the locative responses are more evenly distributed.

For the genitive, the first quartile runs from 4.3% to 16.6%, mirrored by a fourth quartile running from 82.8% to 93.5%. The median is 48.9%, with an interquartile range of 66.2%. The narrow bands of the outside quartiles mean they contain considerably more items than the inner ones.

For the locative, the quartiles cover ranges of more similar sizes. The first quartile runs from 2.5% to 30%. The median is 49.1%. The third quartile ends at 67.9% and the fourth finishes at 96.8%, with an interquartile range of 37.9% - meaning more of an even balance in the middle range than we might have expected given the corpus results.

Again in this chart we note differing responses on the forms *sna/snu,* which appeared in both questionnaires. The difference is if anything more pronounced than with the ratings. A pressing question, then, is whether these results make any intuitive sense, i.e. correlate to any other likely properties of the words, and whether there is any statistical support for their usefulness. The first task will be taken up below, and the second in the following sections.

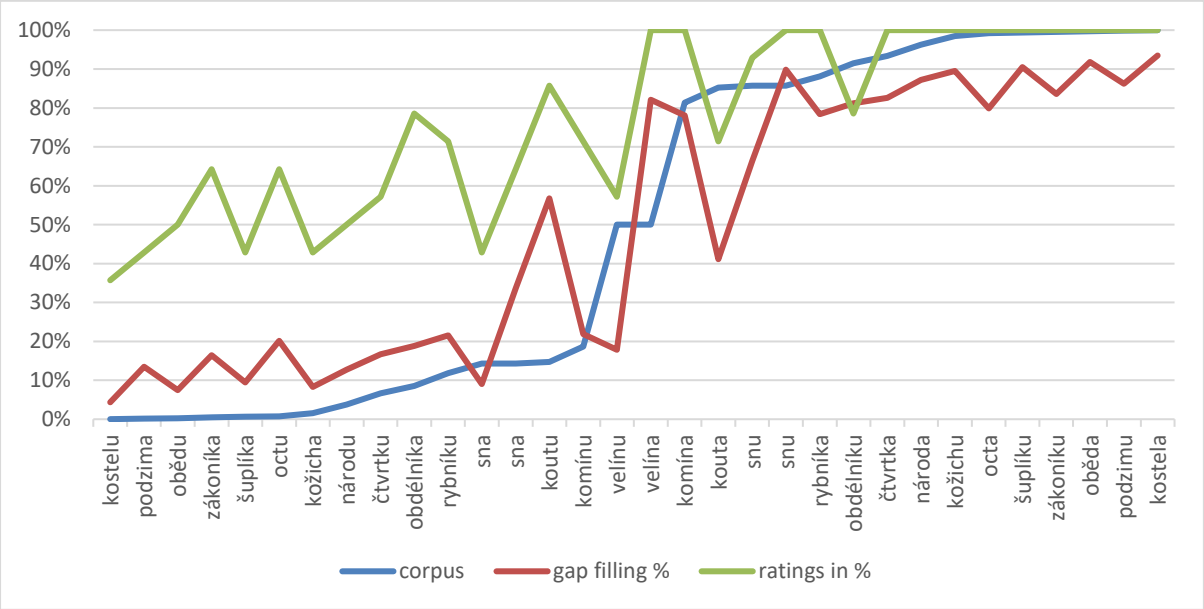
In figures 4–6, we show how corpus frequency, given in terms of the percentage of items or in terms of absolute frequencies, maps onto forced choice and rating data. Again, as all three of these “frequencies” can be treated in different manners, to avoid multiplying the results to nine sets of diagrams we look at two scenarios: the first maps our questionnaire data to proportional frequencies in the corpus; and the second maps them to absolute frequency from the corpus.

In the first scenario, we map the proportional frequency of an ending in the corpus against the rating of that ending and the percentage of time that ending is chosen in a gap-filling task. As two of our data sets are already in percentages, we plot these directly onto the graph. The third data set, acceptability ratings, goes in the opposite direction and is on a different scale (1 is a “high” rating, 7 a “low” rating), so it has to be converted for easier apprehension. To plot it onto the chart in Figure 4, we first reverse the orders to go from 7 to 1 and then convert to a

0-100 scale using the following formula, where V_1 is the original rating and V_2 is the chart value.

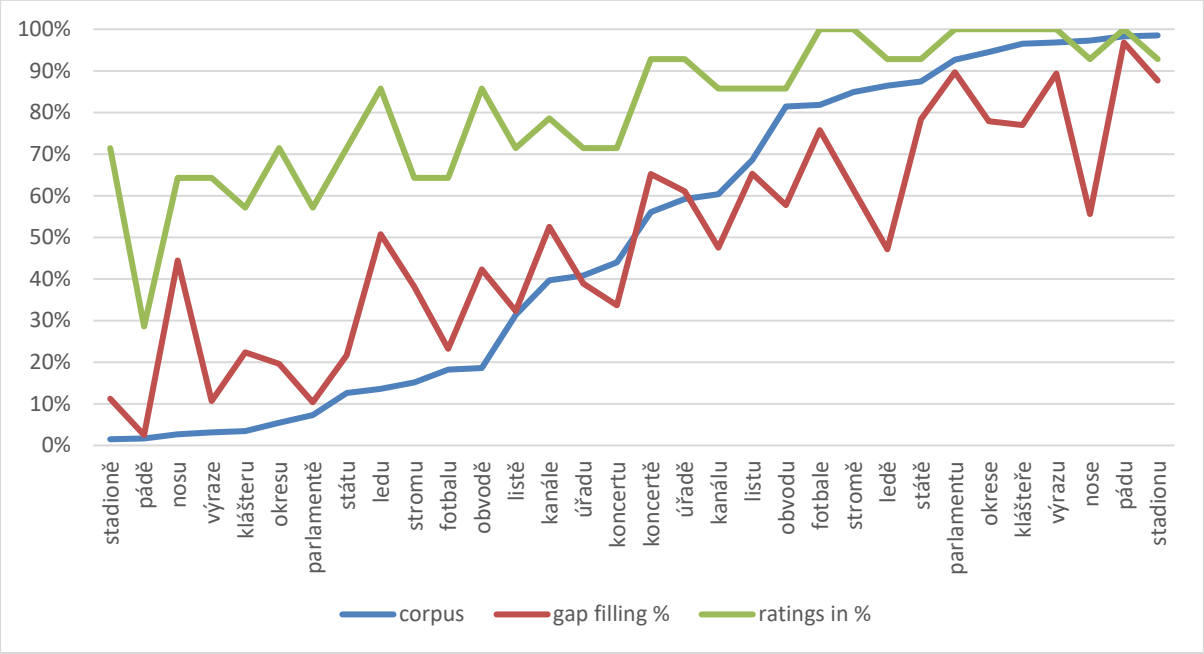
$$V_2 = \frac{V_1}{7}$$

Figure 4. Genitive case forms by proportional frequency in the corpus



The data in Figure 4 are arranged by corpus frequency. There is a clear correlation between them and the gap-filling percentages (r = 0.92), as well as a less robust correlation between them and the ratings medians (r = -0.85).[18] Gap-filling percentages and ratings medians also correlate well (r = -0.94). The peaks and troughs that look so dramatic in the graph are shown to be artefacts of deviations in a relatively small number of lexemes (each is, of course, represented twice on the graph, once for each rival form). We note that while gap-filling tends to result in a fuller use of the 0–100 scale, ratings tasks seem to have a "floor" that expresses more openness to rarely-found forms, and they reach ceiling more quickly, i.e. moderately frequently used forms tend to have full acceptability.

Figure 5. Locative case forms by proportional frequency in the corpus



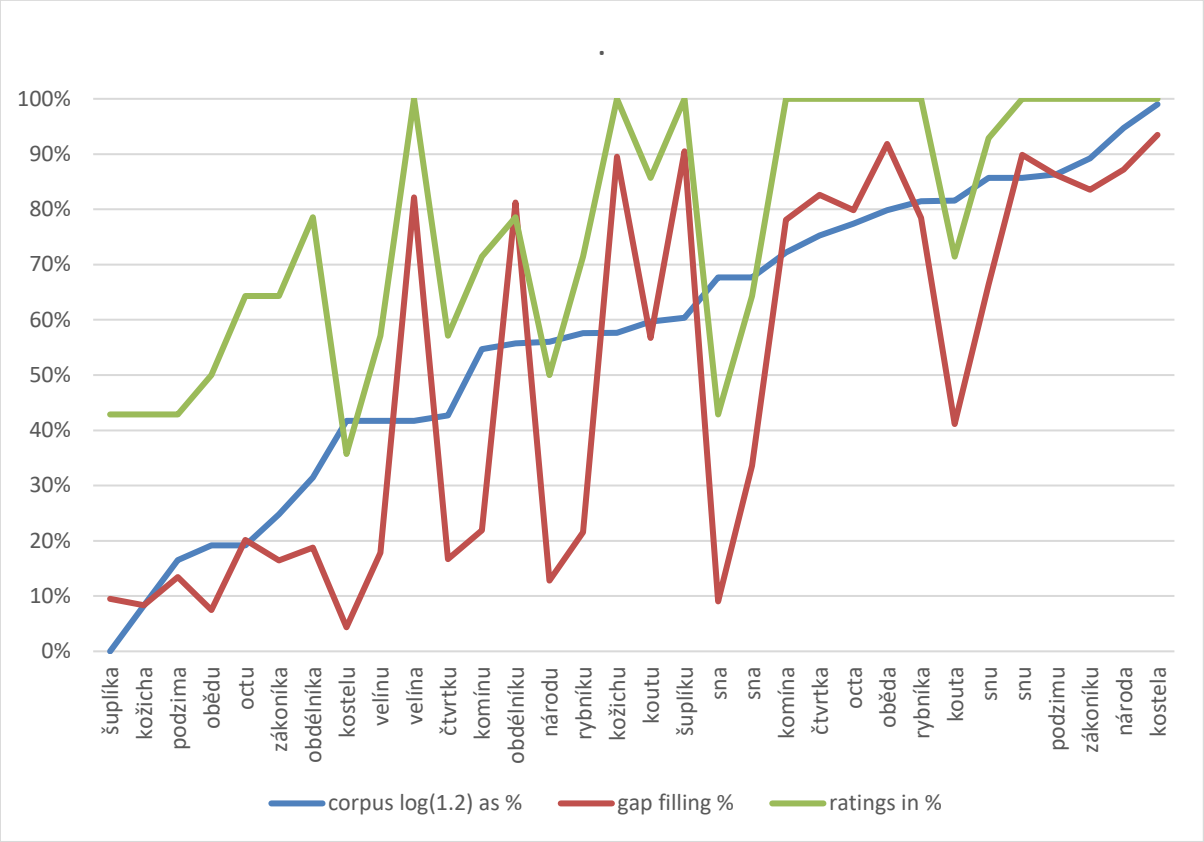
The correlations in Figure 5 are also easily visible, but with more deviant points than in the genitive case. Corpus data correlate with the gap-filling percentages (r = 0.88), and there is a slightly less robust correlation between them and the ratings medians (r = -0.85). Gap-filling percentages and ratings medians also correlate reasonably well (r = -0.87). The same caveat goes for the locative as for the genitive as regards the dips and peaks in the ratings. If anything, the high acceptability of little-used forms is more evident in the locative than it was in the genitive, and there is a similar tendency for ratings to reach ceiling once a critical threshold of frequency is passed. These results confirm those found in Bermel and Knittl (2012a).

Our attempt to match questionnaire data to absolute frequencies follows a similar pattern. However, to get our corpus data to fit on the same scale and yield a workable graph, we will need to scale them, otherwise the Zipfian curve (a few items with exponentially higher frequency) will make it hard to visualize. We thus scaled the data logarithmically (by 1.2, which yielded values from 0 to the mid-40s) and divided by the maximum value obtained (43 or 46):

$$V_2 = \frac{\log(1.2)V_1}{\log(1.2)V_n}$$

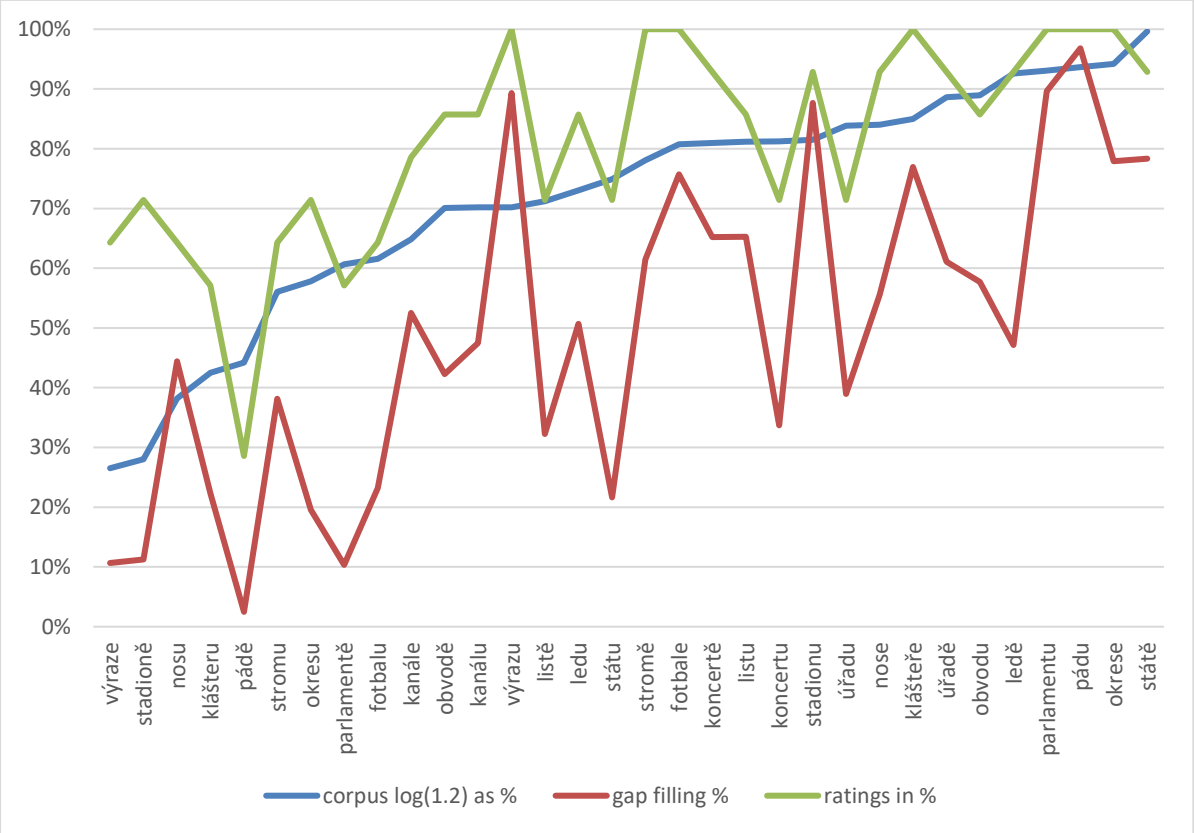
This gives us easily apprehensible charts (figures 6 and 7) graphing the relationships between absolute frequency in the corpus and our questionnaire data.

Figure 6. Genitive case forms by absolute frequency in the corpus



The data in Figure 6 are arranged by corpus frequency (log scaled as above). There are correlations here between the original frequency values, untransformed, and the questionnaire results, but they are not strong ones (with gap-filling $r = 0.57$ and with ratings medians $r = -0.53$).[19] Where $0.5 > r > -0.5$ the correlation is weak to non-existent, and these are just over that limit, i.e. there is a moderate correlation but not enough to be considered a strong or robust one (which would be $r > 0.8$ or $r < -0.8$). Looking at the graph above, we can see where this distribution comes from: there are clearly a higher percentage of gaps filled for frequent items, and a lower percentage for infrequent ones, but there are some strong noticeable exceptions. Similarly, for ratings we can see that highly frequent items tend to be at ceiling while no infrequent items are at ceiling, but again there are numerous exceptions.

Figure 7. Locative case forms by absolute frequency in the corpus



The picture for the locative case in Figure 7 is similar. The correlations are slightly stronger than for the genitive case. Absolute frequency correlates to some extent with the percentage of times the form is chosen ($r = 0.65$) and with the average median rating for the form ($r = -0.61$). The slightly stronger performance for absolute frequency in this case contrasts with what we saw for relative frequency, where the result in the locative case was a slightly weaker correlation than for the genitive case. Again, the graph shows how a general tendency accompanying increases in frequency is offset to some extent by notable divergences.

A look at the descriptive data thus yields two preliminary conclusions:

- Proportional frequency seems to correlate strongly with both ratings and answers supplied, while absolute frequency shows some correlation, but it is not as strong;
- This is true despite the surprising differences between the answers supplied for *sna/snu,* the one lexeme to occur in both questionnaires. While we should thus treat responses to individual items with a certain amount of care, it seems that *overall* the results are indicative of a pattern.

In sections 8 and 9, we will use inferential statistical analysis to examine each set of results more closely and consider what they contribute to our understanding of how we react to and record frequency.

8. Analysis of acceptability ratings

Our goal was to determine whether proportional frequency (PF) or absolute frequency (AF) had a greater effect on the way respondents evaluated forms in competition. We ran three analyses, each representing a different way of looking at these two variables:

1. PF of both forms in bands x AF of both forms in bands
2. PF of both forms in bands x AF with scalar values
3. PF of individual forms in bands x AF of individual forms in scalar values

For the first analysis, we used repeated-measures analyses of variance (ANOVAs), which – if the number of respondents (N > 100) is high enough, can be used to look at ratings made on a Likert scale.[20] The effect of proportional frequency was significant in all instances and accounted for a large part of the variation, but this positive result was not repeated for absolute frequency.

As mentioned above in Section 4, we used a block design in which different versions capture part of the interaction between absolute and relative frequencies and make it possible to combine the results of both versions (Cochran and Cox 1957: 183–185). We then worked with the combined versions using complex analyses of variance. For each set of lexemes, we ended up with a separate analysis that could be compared to a parallel version with different lexemes. The resulting analyses thus report four results each:

- Genitive singular, lexeme set 1
- Genitive singular, lexeme set 2
- Locative singular, lexeme set 1
- Locative singular, lexeme set 2

Example (1) shows the results for proportional frequency in the locative and genitive singular (each parallel questionnaire had separate results and is thus given separately). Attention is called to two of the results: *p* (significance, expressed as a proportional possibility that the effect is down to chance) and Cohen's *r* (a calculation of effect size).

(1) Proportional frequency in the Gsg and Lsg: acceptability ratings
Gsg, questionnaire 1: $F(1, 252) = 1305.97$, $p < 0.001$, $r = 0.92$
Gsg, questionnaire 2: $F(1, 247) = 451.53$, $p < 0.001$, $r = 0.80$
Lsg, questionnaire 1: $F(1, 253) = 489.89$, $p < 0.001$, $r = 0.81$
LSg, questionnaire 2: $F(1, 251) = 223.97$, $p < 0.001$, $r = 0.69$

As the *p* value is consistently well below the cut-off point of 0.05 we confirm that the effects of proportional frequency are significant, i.e. not the result of chance. The value of Cohen's *r* allows us to estimate the effect size: 0.1 is a small effect, 0.3 a medium-sized effect and 0.5 a large effect. These effects are thus all large and we can ascribe a good deal of the variation to them.

Example (2) contains the results for absolute frequency in the Gsg and Lsg:

(2) Absolute frequency in the Gsg and Lsg: acceptability ratings
Gsg, questionnaire 1: $F(1, 252) = 106.66$, $p < 0.001$, $r = 0.55$
Gsg, questionnaire 2: $F(1, 247) = 12.83$, $p < 0.001$, $r = 0.22$
Lsg, questionnaire 1: $F(1, 253) = 16.55$, $p < 0.001$, $r = 0.25$
LSg, questionnaire 2: $F(1, 251) = 223.97$, $p = 0.96$

The *p* values show that the effects of absolute frequency are significant (below 0.05) in three instances, but not for the second locative questionnaire. The value of Cohen's *r* suggests a smaller effect in two out of the three remaining instances.

Proportional frequency thus seems to have a consistent, pronounced effect on users' ratings. Against that the influence of absolute frequency seems less reliable and less pronounced.

One possible reason for the smaller effect of absolute frequency may stem from the breadth of the bands we used (2 vs. 4 for proportional frequency). In our second analysis, we thus looked at the actual values of absolute frequency for each lexeme tested, without the use of bands. This rules out the use of ANOVA, which is a problem for comparability, but we can test

our hypothesis using logistic regression.[21] However, when analysed this way the results were even less significant, as can be seen in Table 11:

Table 11. Results of analysis using exact values for absolute frequency

	Absolute frequency	Abs. frequency * Ending
Gsg, questionnaire 1	F= 0.02, $p = 0.881$	F = 0.46, $p = 0.50$
Gsg, questionnaire 2	F = 1.74, $p = 0.19$	F = 2.79, $p = 0.95$
Lsg, questionnaire 1	F = 91.50. $p = 0.99$	**F = 72.43, $p < 0.001$**
LSg, questionnaire 2	F = 7.97, $p < 0.005$	F = 0.28, $p = 0.63$

We were interested in the effects of absolute frequency on acceptability ratings more generally (independent of its interaction with the ending), as well as in its interaction with particular endings. Crucial here is the *p* value (probability of a chance result) and the effect size, which we can estimate for this sort of analysis using the F value. [22] The results reached the significance threshold only in one instance out of eight, which is given in bold face in Table 13. The lack of significance for the remaining results suggests that while absolute frequency may influence the ratings of native speakers at some overall level (*frequent – infrequent*), the exact figures involved in absolute frequency of the lexeme do not play a role in these ratings.

To this point our analyses had worked with asymmetrical parameters that were not equally informative. Proportional frequency by its nature gives information about *both* variants: to know a percentage, we have to have the totals of both forms. Absolute frequency is a more elastic concept and does not assume this amount of knowledge about a structured category (see Section 2). We had therefore taken total occurrence of the feature as our fingerpost, and so our third analysis allowed us to tease out individual frequency markers.

In it, a linear regression was used to explore the effects of *proportional frequency of each ending*, *absolute frequency of each ending*, *context* and *region* on the average ratings supplied by native speakers. Note that here, as shown in Table 12, we considered the proportional frequency of each ending separately, as well as the absolute frequency of each ending, hence the table doubles in width.

Table 12. Endings considered separately for proportional and absolute frequency

	prop. freq. {a/ě}	abs. freq. {a/ě}	prop. freq. {u}	abs. freq. {u}	F ratio (fit of model)
Gsg 1	p < .001, **F = 286.14**	p < .001, F = 26.91	p < .001, **F = 350.43**	p < .001, F = 104.11	158.96
Gsg 2	p < .001, **F = 253.06**	p < .001, F = 19.95	p < .001, **F = 279.21**	p < .02, F = 5.69	83.18
Lsg 1	p < .001, **F = 189.08**	p < .001, F = 21.35	p < .001, F = 27.17	p < .001, **F = 169.55**	106.48
Lsg 2	p < .001, **F = 149.78**	p = .41	p < .001, **F = 180.35**	p = .94	50.25

In Table 12, the dominant factor influencing the variance (in bold type) is still proportional frequency. With one exception, the values for absolute frequency are much smaller or are not significant. There seems to be some evidence that the model relies more heavily on the use of the emergent {u} ending, rather than the receding {a} and {ě} endings.

9. Analysis of gap-filling tasks

In our gap-filling tasks, the answers measured are not values on a scale but rather choices from a limited range of equivalent answers (i.e. endings). To analyze them, we thus used a binomial logistic regression targeted on the *ending chosen* and included both *proportional* and *absolute frequency* of the given lexemes amongst the possible factors.[23]

We ran two analyses representing different ways of looking at these two variables (fewer than with ratings, where we had two answers per lexeme):

1. PF of both forms x AF of both forms (i.e. one value for PF and one for AF)
2. PF of both forms x AF of individual forms (i.e. one value for PF and two for AF)

The *predictive capability* (R^2) of our model is high for all versions of the questionnaire.[24] The calculation of R^2 here comes from a simple ratio (below) expressing the improvement our model brings over a simple "default" model in which the most frequent ending is always chosen, over the improvement brought by a theoretical "full" model in which every combination of factors proposed is incorporated:

$$R^2 = \frac{\text{Value of our model} - \text{Value of default model}}{\text{Value of full model} - \text{Value of default model}}$$

For our four word sets we obtained the following values for R^2: 76.4% (Gsg 1), 81.0% (Gsg 2), 64.1% (Lsg 1), 91.3% (LSg 2). In other words, in all instances the model was significantly improved, confirming that the factors incorporated in our model (absolute and proportional frequency) are important ones in choosing between these relevant endings.

The significance of the variables we chose is measured once again by the *p* value and we can estimate its relative weight by the value of F in Table 13.

Table 13. Proportional and absolute frequency in the Gsg a Lsg: gap filling

	Proportional frequency	Absolute frequency
Gsg, questionnaire 1	**F= 157.52,** $p < 0.001$	F = 81.10, $p < 0.001$
Gsg, questionnaire 2	**F = 122.62,** $p < 0.001$	F = 21.66, $p < 0.001$
Lsg, questionnaire 1	**F = 90.43,** $p < 0.001$	F = 0.07, $p = 0.80$
LSg, questionnaire 2	**F = 91.50,** $p < 0.001$	F = 1.99, $p = 0.16$

We can see in Table 13 that proportional frequency consistently plays a significant role ($p < .05$). The weight of this variable is higher than e.g. that of any demographic characteristics of respondents, syntactic context, etc. Against this absolute frequency plays a smaller role, which is significant only in the genitive case; it does not reach significance in the locative.

As with the ratings, we then looked at the absolute frequency of each individual form as part of our model, and the results, as seen in Table 14, are quite different (the feature contributing most to the model is given in bold).

Table 14. Proportional and absolute frequency with endings considered separately

Set	**Prop. Freq.**	**Abs. Freq. {a/ě}**	**Abs. Freq. {u}**	**R^2 (fit)**
Gsg 1	p < .002, F = 5.90	p < .001, **F = 37.24**	p < .1	.64
Gsg 2	p < .001, F = 15.24	p < .005, F = 8.46	p < .001, **F = 33.16**	.55
Lsg 1	p < .001, F = 35.52	p < .001, F = 23.87	p < .001, **F = 74.41**	.52
Lsg 2	p < .001, **F = 70.75**	p < .001, F = 28.089	p < .02, F = 6.44	.32

Once the two absolute values are taken independently as variables, they outweigh the input of proportional frequency by a considerable amount in three out of the four models.

For fit, we used a slightly different calculation, in which the "full" model is replaced by a theoretical "perfect" model with 100% accuracy:

$$R^2 = \frac{\text{Value of our model} - \text{Value of default model}}{100 - \text{Value of default model}}$$

Despite the increased stringency of the fit, we still get reasonable results that explain a large portion of the variation.

10. Conclusions

Our analyses of our questionnaire data have shown that proportional frequency of forms in a balanced corpus like those of the SYN set is reliably connected with their acceptability for native speakers and with the frequency with which speakers select such forms in free choice. For ratings tasks, proportional frequency predominates in all analyses, playing a much larger role than absolute frequency, whether the latter is construed as falling into two 'bins' or incorporated with its actual values. Even in the analysis that should be most favourable to absolute frequency – shown in Table 14 – the proportion of variance assigned to proportional frequency is far greater.

However, in the gap-filling task, proportional frequency only dominates the effects when we use a weak measure of absolute frequency: both endings combined. Once we separate out different values of absolute frequency by individual endings, then absolute frequency comes to the fore in three out of the four data sets. For one of these data sets, it is the recessive ending that contributes most to the model, but in two others the frequency of the dominant ending contributes most to the model.

Receptive and productive tasks thus seem to activate different knowledge about the frequency of words. When judging the well-formedness of certain forms (which might involve decisions about whether we think someone is a "good" writer, the stylistic adequacy of a text, what register it belongs to, etc.), we seem to judge forms largely *against each other* – in other words, we access information about a category or cell and the items that potentially occupy it. This suggests that at some point in the process *schematization* has occurred and these two forms are now related within that schema, as suggested in Section 2. However, in production tasks using the same stimuli, we seem to be more influenced by *raw frequency.* This seems to go back to more of a connectionist or exemplar-based model: each reiteration strengthens the connection, entrenching it for that one particular form, even if the relationships between closely-allied forms are retained through proximity:

> In this model, every token of experience is classified and placed in a vast organizational network as a part of the decoding process. The major idea behind exemplar theory is that the matching process has an effect on the representations themselves; new tokens of experience are not decoded and then discarded, but rather they impact memory representations. In particular, a token of linguistic experience that is identical to an existing exemplar is mapped onto that exemplar, strengthening it. Tokens that are similar but not identical (differing in slight ways in meaning, phonetic shape, pragmatics) to existing exemplars are represented as exemplars themselves and are stored near similar exemplars to constitute clusters or categories (Bybee 2006: 716).

We have tried here to focus attention on a well-known problem – defining *frequency* – and have tested three ways of operationalizing this concept on our data to work out which of them explains the data best: proportional frequency in all categories; absolute frequency in two categories (high/low); and absolute frequency as a scale. Our formulation of the question was circumscribed by the data at our disposal (a limited set of lexemes from which to choose); for other types of data it would be possible to consider other operationalizations of the concept. The choice of one or another operationalization appears to have an observable effect on how we answer the question about the relationship between *frequency* and the behaviour of native speakers.

In focusing on various types of definitions of frequency and discovering which does better at predicting respondents' answers, we should avoid claiming that the mind actually stores information in this way: that is, with an exact number or a percentage figure attached to each form. What we have shown is somewhat different: that models using these definitions of frequency turn out in some instances to be adequate, that they explain some of the variation, sometimes much of it, and they show which features do better at that than others. Our models do not show how that frequency figure gets into our minds and is retained there, so it is entirely possible that there are tweaks to the definitions that would improve them, or that a more realistic model of our mind's activity would change the way we see this working. Rather, proportional and actual frequency, in the various guises that we have examined them, stand more as metaphors – intellectual shortcuts, if you like – for the still unconfirmed processes that underlie them.

---

[1] Research for this article was carried out as part of the project "Acceptability and forced-choice judgements in the study of linguistic variation", funded by the Leverhulme Trust (RPG-407). An earlier version of some of this material was published in in Czech in abbreviated form in Bermel et al. (2014). The authors would like to thank Dagmar Divjak and the anonymous reviewers for their comments and suggestions.

[2] For Baayen et al. (2013), rival prefixed forms with similar or overlapping meanings form part of the same continuum of features as rival endings or rival stems, as we can analyze this variation using the same set of techniques. Overabundance, on the other hand, has remained confined to inflectional, rather than derivational morphology, observing (and thereby indirectly reinforcing) a distinction that may not be relevant for our techniques.

[3] On the representativity of the corpora of the Czech National Corpus see e.g. Čermák et al. (1997), Králík and Šulc (2005). According to Baayen et al. (2016: 5–6), "for the assessment of language processing in general, subtitle corpora can be taken as the source for normative measures of lexical frequency", which they put down to the predominance of emotional language in such corpora, following Heister and Kliegl (2012). This observation needs to be borne in mind, although our studies are not processing studies as they were carried out offline.

[4] These are the exact numbers of tokens in the corpus rather than the "relative" or "normalized" frequency, i.e. the frequency with respect to the size of the whole corpus, or the number of examples in a "normal" corpus of 1m tokens (see e.g. McEnery and Hardie 2012). We could thus also express the incidence of *nocech* either as .000427 relative to the whole 100m-token corpus, or as 4.27 incidences per million tokens.

[5] For example, Čech (2012: 210–211) notes that the use of percentage values for forms without reference to their actual frequency can be misleading, as significance is closely linked to the size of the sample.

[6] In a study of the elision of /t/ and /d/ in English, Bybee (2002: 264) set 35 forms as her cut-off point, using a corpus of 1m tokens, while in a further study of the elision of /d/ and /ð/ in Spanish (2002: 265–266), she used 100 forms as her boundary marker in a corpus of roughly the same size (1.1m tokens). In a third study (Bybee and Eddington 2006: 329), using data from two corpora with a total size of 2m tokens, her high-frequency item appeared 17 times, while her low-frequency items appeared 9 times or less, meaning that the boundary was somewhere between 4.5 per million and 8.5 per million.

[7] On the overall frequency of competing forms see Bermel and Knittl 2012a: 249.

[8] In this way, we avoided some of the further problems discussed in Čech (2012: 211) and Cvrček and Kodýtek (2013: 141) concerning the significance of small sample sizes.

[9] This is formulated later in terms of Shannon and Weaver's (1963 [1949]) balance between *information value* and *redundancy*. Halliday proposes that when the proportions are more skewed than 0.89: 0.11, information outweighs redundancy. Between 0.11 and 0.89, redundancy outweighs information (2005 [1992]: 74–75 and 2005 [1993]:138–139).

[10] This was not the case, for example, for Cvrček's grammar, because the scale was not meant to describe functional differences; it aimed simply to offer labels for corpus frequency, and for this reason it made sense to have bands of more equal width.
[11] See e.g. Bermel (1993, 2004, 2010); Cummins (1995), Kasal (1992); Klimeš (1953); Kolařík (1995); Rusínová (1992); Sedláček (1982); Štícha (2009). There are also extensive discussions in some grammars of Czech, e.g. Komárek et al. (1986); Karlík et al. (1995); Cvrček et al. (2010). A historical overview of this variation set in its cross-Slavonic context is given in Janda (1996: 121–153) and a similar feature in Russian is discussed in Brown (2007).
[12] This is the only lexeme that reliably falls in this group, so it occurs at a higher frequency in all questionnaires to ensure that this 'cell' is covered. It thus also appears twice in Tables 6–10 and figures 4–5 on subsequent pages.
[13] We tested various orderings of individual sentences as well as the ordering of tasks: gap filling first, then ratings first. On questionnaire structure see e.g. Cowart (1997), Schütze (1996).
[14] In each group of respondents 16 versions of the questionnaire were distributed randomly, so that the sociolinguistic profile of respondents would be consistent across all groups.
[15] Sample questionnaires are made available on the project website:
http://www.sheffield.ac.uk/russian/research/slal/variation
[16] Here we report the figures direct from SPSS. The figure of 5.9% is halfway between the last value in Q1 and the first value in Q2. The "boundary line" between the quartiles thus does not correspond to specific values in the table.
[17] As mentioned above in the note to Table 5, a lack of words in this particular cell made it necessary for us to use the same word twice (once in each questionnaire). For the sake of transparency – and because the results were not identical – the results are reported separately for each questionnaire and thus each form appears twice here.
[18] Despite the chart's display, this correlation is correctly stated as negative, because in the original data, the more positive the rating, the lower the number.
[19] See note 18 above as to why this correlation is negative.
[20] The Likert scale is used in questionnaires for expressing agreement or preference on a scale (i.e. in a more nuanced manner than merely "yes/no"). Our questionnaires used a seven-point scale, from 1 (high) to 7 (low), with midpoints numbered and evenly spaced on the page but not named. The use of 1 as the high mark is traditional in Czech grading schemes, hence its employment here. As there is disagreement as to whether speakers can apply concepts like "acceptability" and "grammaticality", we added descriptions to assist them. The label 1 was glossed as *completely normal (in the given context I would definitely write it this way* - original: *naprosto normální (v rámci daného kontextu bych to určitě takto napsal/a)),* while the label 7 was glossed as *inadmissible (in the given context something really bothers me about this; I don't regard it as normal Czech* - original: *nepřijatelné (v daném kontextu mi něco hodně „nesedí", nepovažuji to za normální češtinu).*
The goal was to induce users to treat the scale as an *interval measure* to the extent that this is possible, as the assumption of interval results underpins the set-up of the ANOVA. There is, of course, no guarantee that users do precisely this, but few respondents seem to have difficulty with the concept, and with over a thousand questionnaires distributed in paper and electronic form in the last three years, we have never had a user question the meaning or significance of the scale, which at least testifies to the fact that the task is neither unnatural nor problematic for respondents.
[21] In this instance we used a generalized mixed linear model targeted at the rating chosen and amongst our factors we included the absolute frequency of the particular lexeme.
[22] In simple terms, the value of F is calculated by taking the variation explained by our model, divided by the variation not explained by the model. Higher values for F as a rule indicate a larger effect.
[23] Subjects are included in the data structure and thus are accounted for as a random effect, improving the fit of the model.
[24] That is, using these two factors amongst several others, we made a large improvement in our ability to predict which ending was used, and including the remaining possible combinations of factors would not have improved the predictability by a great deal.

REFERENCES

Baayen, R. Harald, Anna Endresen, Laura A. Janda, Anastasia Makarova & Tore Nesset. 2013. Making choices in Russian: Pros and cons of statistical methods for rival forms. *Russian Linguistics* 37(3). 253–291.

Baayen, R. Harald, Petar Milin & Michael Ramscar. 2016. Frequency in lexical processing. *Aphasiology* 30. 1–47.

Bermel, Neil. 1993. Sémantické rozdíly v tvarech českého lokálu [Semantic differences in the forms of the Czech locative]. *Naše řeč* 76. 192–198.

Bermel, Neil. 2004. *V korpuse* nebo *v korpusu*? Co nám řekne (a neřekne) ČNK o morfologické variaci v tvarech lokálu [*V korpuse* or *v korpusu*? What the Czech National Corpus will (and will not) tell us about morphological variation in locative case forms]. In Hladková, Zdeňka & Petr Karlík (eds.), *Čeština – univerzália a specifika 5*, 163–171. Prague: Nakladatelství Lidové Noviny.

Bermel, Neil. 2010. Variace a frekvence variant na příkladu tvrdých neživotných maskulin [Variation and the frequency of variants in hard masculine inanimate nouns]. In Čmejrková, Světla, Jana Hoffmannová & Eva Havlová (eds.), *Užívání a prožívání jazyka,* 135–140. Prague: Karolinum.

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

Bermel, Neil & Luděk Knittl. 2012b. Morphosyntactic variation and syntactic environments in Czech nominal declension: Corpus frequency and native-speaker judgments. *Russian Linguistics* 36. 91–119.

Bermel, Neil, Luděk Knittl & Jean Russell. 2014. Absolutní a proporcionální frekvence v Českém národním korpusu ve světle výzkumu morfosyntaktické variace v češtině. *Naše řeč* 97. 216–227.

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

Brown, Dunstan. 2007. Peripheral functions and overdifferentiation: The Russian second locative. *Russian Linguistics* 31. 61–76.

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

Bybee, Joan. 2006. From usage to grammar: The mind's response to repetition. *Language* 82. 711–733.

Bybee, Joan. 2007. *Frequency of use and the organization of language.* Oxford: Oxford University Press.

Čech, Radek. 2012. Několik teoreticko-metodologických poznámek k Mluvnici současné češtiny [Some theoretical-methodological notes on the *Grammar of Contemporary Czech*]. *Slovo a slovesnost* 73. 208–216.

Čermák, František, Jan Králík & Karel Kučera. 1997. Recepce současné češtiny a reprezentativnost korpusu (Výsledky a některé souvislosti jedné orientační sondy na pozadí budování Českého národního korpusu) [The reception of contemporary Czech and corpus representativity: results and some relevant points of a preliminary sounding done during the building of the Czech National Corpus]. *Slovo a slovesnost,* 58. 117–123.

Čermák, František, Drahomíra Doležalová-Spoustová, Jaroslava Hlaváčová, Milena Hnátková, Tomáš Jelínek, Jan Kocek, Marie Kopřivová, Michal Křen, Renata Novotná, Vladimír Petkevič, Věra Schmiedtová, Hana Skoumalová, Michal Šulc & Zdeněk Velíšek. 2005. *SYN2005: A genre-balanced corpus of written Czech.* Prague: Ústav Českého národního korpusu FF UK. www.korpus.cz

Clark, Eve V. 1987. The Principle of Contrast: a constraint on acquisition. In MacWhinney, B. (ed.), *Mechanisms of language acquisition: the 20th annual Carnegie symposium on cognition,* 1–33. Hillsdale NJ: Erlbaum.

Cochran, William G. & Gertrude M. Cox. 1957. *Experimental designs.* [Second edition]. New York: John Wiley and Sons.

Cowart, Wayne. 1997. *Experimental syntax: Applying objective methods to sentence judgments*. Thousand Oaks, CA: Sage Publishers.

Cummins, George. 1995. Locative in Czech: -u or -e: Choosing locative singular endings in Czech nouns. *Slavic and East European Journal* 39. 241–260.

Cvrček, Václav & Vilém Kodýtek. 2013. Ke klasifikaci morfologických variant [On classifying morphological variants]. *Slovo a slovesnost* 74. 139–145.

Cvrček, Václav, Vilém Kodýtek, Marie Kopřivová, Dominika Kováříková, Petr Sgall, Michal Šulc, Jan Táborský, Jan Volín & Martina Waclawičová. 2010. *Mluvnice současné češtiny* [A grammar of contemporary Czech]. Praha: Karolinum.

Dąbrowska, Ewa. 2005. Productivity and beyond: mastering the Polish genitive inflection. *Journal of Child Language* 32. 191–205.

Dąbrowska, Ewa. 2006. Low-level schemas or general rules? The role of diminutives in the acquisition of Polish case inflections. *Language Sciences* 28. 120–135.

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

Dąbrowska, Ewa & Marcin Szczerbiński. 2006. Polish children's productivity with case marking: the role of regularity, type frequency, and phonological diversity. *Journal of Child Language* 33. 559–597.

Divjak, Dagmar. 2016. The role of lexical frequency in the acceptability of syntactic variants: Evidence from that-clauses in Polish. *Cognitive Science* 40. 1–29.

Halliday., M. A. K. 2005 [1992]. Language as system and language as instance: The corpus as a theoretical construct. In: J. Svartvik (ed.), *Directions in Corpus Linguistics,* Berlin: Mouton de Gruyter, p. 61–77. Reprinted in Webster, Jonathan J. (ed.), *Computational and quantitative studies*, 76-92. [Collected Works of M. A. K. Halliday 6]. London: Continuum.

Halliday., M. A. K. 2005 [1993]. Quantitative studies and probability in grammar. In Hoey, Michael (ed.), *Data, description and discourse,* 61–77. London: HarperCollins. Reprinted in Webster, Jonathan J. (ed.), *Computational and quantitative studies*, 130–156. [Collected Works of M. A. K. Halliday 6]. London: Continuum.

Hebal-Jezierska, Milena. 2008. *Wariantywność końcówek fleksyjnych rzeczowników męskich żywotnych w języku czeskim* [Variation in the flexional endings of masculine animate nouns in Czech]. Warsaw: Wydział Polonistyki Uniwersytetu Warszawskiego.

Hebal-Jezierska, Milena & Neil Bermel. 2011. Frequency and oppositions in corpus-based research into morphological variation. In: Konopka, Marek, Jacqueline Kubczak, Christian Mair, František Šticha & Ulrich H. Waßner (eds.), *Gramatik und Korpora* 3, 373–388. Tübingen: Narr

Heister, Julian & Reinhold Kliegl. 2012. Comparing word frequencies from different German text corpora. *Lexical Resources in Psycholinguistic Research* 3. 27–44.

Janda, Laura. 1996. *Back from the brink: A study of how relic forms in language serve as source material for analogical extension.* Munich and Newcastle: Lincom Europa.

Karlík, Petr, Marek Nekula, Zdena Rusínová, Miroslav Grepl, Zdeňka Hladká, Milan Jelínek, Marie Krčmová & Dušan Šlosar. (1995): *Příruční mluvnice češtiny* [A grammar handbook of Czech]. Prague: Nakladatelství Lidové Noviny.

Kasal, Jindřich. 1992. Dublety a jejich užití [Doublets and their usage]. *Philologica*, 65. 107–114.

Klimeš, Lumír. 1953. Lokál singuláru a plurálu vzoru „hrad“ a „město“ [The locative singular and plural of the “hrad” and “město” paradigms]. *Naše řeč* 36. 212–219.

Kolařík, Josef. 1995. Dynamika ve flexi substantiv běžně mluveného jazyka ve Zlíně [Dynamics in the inflection of nouns in ordinary spoken language in Zlín]. In Davidová, Dana (ed.), *K diferenciaci současného mluveného jazyka* [On differentiation in the contemporary spoken language] , 79–83. Ostrava: Universitas Ostraviensis, Facultas Philosophica.

Komárek, Miroslav, Jan Kořenský, Jan Petr, Jarmila Veselková (eds.). 1986. *Mluvnice češtiny. Díl 2: Tvarosloví* [A grammar of Czech, part II: Morphology]. Prague: Academia.

Králík, Jan & Michal Šulc. 2005. The representativeness of Czech corpora. *International Journal of Corpus Linguistics* 10. 357–366.

Křen, Michal, Tomáš Bartoň, Václav Cvrček, Milena Hnátková, Tomáš Jelínek, Jan Kocek, Renata Novotná, Vladimír Petkevič, Pavel Procházka, Věra Schmiedtová & Hana Skoumalová. 2010. *SYN2010: A genre-balanced corpus of written Czech*. Prague: Ústav Českého národního korpusu FF UK. www.korpus.cz.

Langacker, Ronald. 2008. *Cognitive Grammar: A Basic Introduction.* Oxford: Oxford University Press.

McEnery, Tony & Andrew Hardie. 2012. *Corpus linguistics: Method, theory and practice.* Cambridge: Cambridge University Press.

Rusínová, Zdena. 1992. Některé aspekty distribuce alomorfů (genitiv a lokál sg. maskulin) [Some aspects of the distribution of allomorphs: genitive and locative singular masculine]. *Sborník prací filozofické fakulty brněnské univerzity,* A 40. 23–31.

Schmid, Hans-Jörg. 2015. A blueprint of the Entrenchment-and-Conventionalization Model. In: Uhrig, Peter and Thomas Herbst (eds.), *Yearbook of the German Cognitive Linguistics Association,* 3–26. Berlin: Walter de Gruyter.

Schütze, Carson. 1996. *The empirical base of linguistics: Grammaticality judgments and linguistic methodology.* Chicago: University of Chicago Press.

Shannon, Claude E. & Warren Weaver. 1963 [1949]. *The mathematical theory of communication.* Urbana, IL: University of Illinois Press.

Sedláček, Miloslav. 1982. V Záhřebě i v Záhřebu [Both “v Záhřebě” and “v Záhřebu”]. *Naše řeč* 65. 11–15.

Šimandl, Josef. 2010. *Dnešní skloňování substantiv typů* kámen, břímě [The declension of nouns of the type *kámen, břímě* today]. Prague: Nakladatelství Lidové Noviny.

Štícha, František. 2009. Lokál singuláru tvrdých neživotných maskulin (*ve vlaku* vs. *v potoce*): úzus a gramatičnost [The locative singular of hard inanimate masculine nouns (*ve vlaku* vs. *v potoce*): usage and grammaticality]. *Slovo a slovesnost* 70. 193–220.

Thornton, Anna. 2012. Reduction and maintenance of overabundance: A case study on Italian verb paradigms. *Word Structure* 5. 183–207.