This is a repository copy of *Do users' reading skills and difficulty ratings for texts affect choices and evaluations?*.

White Rose Research Online URL for this paper:
http://eprints.whiterose.ac.uk/115302/

Version: Accepted Version

Proceedings Paper:

Bermel, N.H. orcid.org/0000-0002-1663-9322, Knittl, L. and Russell, J.M. orcid.org/0000-0002-4174-7210 (2018) Do users' reading skills and difficulty ratings for texts affect choices and evaluations? In: Fidler, M. and Cvrček, V., (eds.) Taming the Corpus: From inflection and lexis to interpretation. Quantitative Text Analysis for the Humanities and Social Sciences, 08-09 Apr 2016, Providence, USA. Quantitative Methods in the Humanities and Social Sciences . Springer . ISBN 978-3-319-98016-4

https://doi.org/10.1007/978-3-319-98017-1_2

This is a post-peer-review, pre-copyedit version of an article published in Fidler M., Cvrček V. (eds) Taming the Corpus. Quantitative Methods in the Humanities and Social Sciences, Springer. The final authenticated version is available online at: https://doi.org/10.1007/978-3-319-98017-1_2

Reuse

Items deposited in White Rose Research Online are protected by copyright, with all rights reserved unless indicated otherwise. They may be downloaded and/or printed for private study, or other acts as permitted by national copyright laws. The publisher or other rights holders may allow further reproduction and re-use of the full text version. This is indicated by the licence information on the White Rose Research Online record for the item.

Takedown

If you consider content in White Rose Research Online to be in breach of UK law, please notify us by emailing eprints@whiterose.ac.uk including the URL of the record and the reason for the withdrawal request.

eprints@whiterose.ac.uk
https://eprints.whiterose.ac.uk/

Do users' reading skills and difficulty ratings for texts affect choices and evaluations?

Neil Bermel, Luděk Knittl, Jean Russell

University of Sheffield

Abstract

In our contribution, we consider how corpus data can be used as a proxy for the written language environment around us in constructing offline studies of native-speaker intuition and usage. We assume a broadly emergent perspective on language: in other words, the linguistic competence of individuals is not identical or hard-wired, but forms gradually through exposure and coalescence of patterns of production and reaction. We hypothesize that while users presumably all in theory have access to the same linguistic material, their actual exposure to it and their ability to interpret it may differ, which will result in differing judgements and outputs. Our study looks at the interaction between corpus frequency and two possible indicators of individual difference: attitude towards reading tasks and performance on reading tasks. We find a small but consistent effect of task performance on respondents' judgements, but do not confirm any effects on respondents' production tasks.

1. Introduction[1]

Considerable attention has been devoted to whether all native speakers of a language access the same linguistic structures and material in similar ways, and whether, having accessed it, their use of and reaction to language (what we will call linguistic behavior) differ as well in predictable ways. There is accumulating evidence that intra-speaker variation can point to differences in linguistic behavior that are not random or insignificant.

[1] This research 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). The support of the Trust is gratefully acknowledged.

We can propose that speakers' varying backgrounds (i.e. their exposure to language) affect language in use (i.e. their output or their evaluation of input). In other words, if we call these mechanisms a "grammar" for short, each speaker's is subtly different. Corpus data can, if carefully used, be hypothesized to represent this "exposure" to at least the written form of the language, which is the tack we will take in this study. In doing so, we aim to add to the evidence showing how corpus frequency can be useful in detecting and predicting our use of language.

2. Background

Evidence has, at times, pointed to vocabulary size, education, profession, and reading recall abilities as factors differing from subject to subject that affect one's "personal" linguistic behaviour, and these differences have been found in syntax, word-formation and inflectional morphology. While we might try to explain away differences resulting from regional or age variation as the product of language shift and change, it is harder to do so with e.g. educational or professional differences.

In a series of articles, Dąbrowska has tracked some of these differences in speaker backgrounds, which she shows lead to differences in both linguistic performance and linguistic judgments. Dąbrowska 2008 looked at a sample of users stratified by educational background and assessed their performance on a production task. She concluded that "the results... revealed large individual differences in speakers' ability to inflect unfamiliar nouns which were strongly correlated with education" (2008: 941). Having attempted to eliminate some possible confounding factors, she concluded, "We can be reasonably confident... that the observed differences in scores in the other conditions reflect genuine differences in linguistic proficiency" (2008: 945). A logical deduction from that might have been that more educated speakers had larger vocabularies; however, Dąbrowska did not find enough evidence for this, saying, "…the results do not support the hypothesis that the critical variable is vocabulary size, although they do not unequivocally rule it out" (2008: 949). In a later study, she examined judgments of sentence well-formedness given by linguists and non-linguists, and found that:

> Linguists' judgments are shown to diverge from those of nonlinguists. These differences could be due to theoretical commitments (the conviction that linguistic processes apply 'across the board', and hence all sentences with the same syntactic structure should be equally grammatical) or to differences in exposure (the constructed examples of this structure found in the syntactic literature are very unrepresentative of ordinary usage). (2010:1)

While Dąbrowska was cautious in her conclusions about whether educational differences and vocabulary size can be so closely linked, other researchers have made the connection between linguistic behaviour and vocabulary size more directly. For example, Frisch and Brea-Spahn 2010 found that vocabulary size, as measured by the results of a word familiarity rating task, correlates with acceptability scores on a word-formation task. They noted:

> Participants with a larger vocabulary in English were more accepting of low probability nonwords in English. It appears that those with greater vocabulary knowledge are more likely to have experienced improbable phonological constituents, and may also have a lower threshold for "unacceptable" nonwords, if their threshold is based on a likelihood estimate from their individual lexicon. (2010: 345)

Reading abilities also affect judgements: Staum Casasanto et al. 2010 investigated how differences in reading span interact with judgments.[2] Reading span task scores were highly

[2] Reading span tasks ask participants to read unconnected sentences, memorizing the final word of each sentence, which they then must recall later. There is some dispute about what exactly they are measuring (Hupet, Desmette & Schelstraete 1997) but as Conway et al. point out, they have been widely used nonetheless to assess how we tap into our working memory's storage and processing functions: "The task is essentially a simple word span task, with the added component of the comprehending of sentences. Subjects read sentences and, in some cases, verify the logical accuracy of the sentences, while trying to remember words,one for each sentence presented" (2005: 771).

significant predictors of acceptability scores on a task involving the syntax of embedded clauses, e.g. The nurse from the clinic supervised the administrator who scolded the medic while a patient was brought into the emergency room (Staum Casasanto et al. 2010:224). They concluded that

> [P]articipants' reading span scores predict sentence judgments differently for different types of manipulations. Participants with higher reading spans tend to judge ungrammatical sentences as being worse than their low-span counterparts do, yet they tend to judge difficult sentences as being better than participants with lower reading spans (2010: 228).

A further set of factors that have been shown to contribute to analyses of linguistic behavior are those that derive from analyses of the task performance itself. For example, Divjak (2016) demonstrates that ratings given on "filler" items are in fact the best predictor of how a respondent rates the test items (in this instance manipulating the complement of certain verbs). Bermel et al. show that respondents' ratings of the less common of two variants are the best predictor of how they answer on a production task (2015: 304–306).

In summary, then, it seems that a variety of speaker-specific factors can influence linguistic behavior. Some of these, such as educational attainment and profession, masquerade as non-linguistic factors but may in fact be linked to an individual's linguistic abilities. Others, including vocabulary size (either measured via the self-reported familiarity of words or accuracy on a semantics test) and reading span test scores, are more overt measures of reading proficiency. A third group effectively measures the respondent's attitude towards the given features or towards survey data in general.

If many of these factors impinge on our ability to read and interpret, it stands to reason that there will be a link between a proxy for the external "textual world", such as a corpus, and the sorts of answers respondents give on surveys. In the next section, we will consider how this relates to our own research data.

3. Corpus data

For a number of years now, we have been looking at places in the Czech conjugational and declensional systems where a syntactic 'slot' has multiple exponents whose usage is not clearly differentiated, a situation described variously as competition (Lečić 2015), variation (Bermel & Knittl 2012a, 2012b, Bermel, Knittl & Russell 2015a, 2017) or overabundance (Thornton 2012).[3]

In common with other Slavic languages, Czech is highly inflected, and thanks to a series of far-reaching phonological changes over the last millennium, the conditions for deploying its broad assortment of inflectional material are not always clear (see Bermel and Knittl 2012b: 93–95) for a fuller discussion).[4] Consequently, while we are able to describe clearly for some syntactic slots what exponent is used there, for others there is considerable variation. Exponents may be described using a list-type approach ("the following lexemes use exponent A; others use exponent B") or using a collection of rules of thumb ("borrowings, multisyllabic stems and labial consonant stems prefer exponent C; others prefer exponent D"). In addition to places where choice is clear-cut, there exists a transitional band of items where both exponents are used in some measure.

In English, with its relatively impoverished inflectional morphology, the best higher-frequency environment in which to study this is the overlap between so-called "strong" and "weak" verb classes in the past tense and the perfect, and it has been studied from various

[3] An example of clearly differentiated usage is e.g. between the exponents {em} and {ou} in the instr. sg.: the former is used with masc. and neut. nouns, while the latter appears with fem. nouns. The only place we get overlap – e.g. s (v)okurkem ~ s (v)okurkou 'with cucumber' – is where the gender of the noun is unstable across dialects. When usage is not clearly differentiated, often some factors or tendencies can be identified that contribute to choice, but none that clearly demarcate it.

[4] A further contributory factor to the persistence of variation in Czech may be the relatively weak position of the standard, which does not function as a common speech variety across the vast majority of the country (see e.g. Sgall 2011: 183, one among many texts that could be cited in this regard). Attempts at standardizing one or another variant tend to be perceived as applying only to formal written texts.

angles over the past several decades (Haber 1976, Bybee and Slobin 1982, Prasada and Pinker 1993, Eddington 2000, Albright and Hayes 2003, Chandler 2010, etc.).[5] In Czech, this phenomenon is widespread across both verbal and nominal morphology (e.g. Bermel 2004a, 2004b, 2010, Bermel et al. 2015b); in particular, nominal morphology, with seven cases, two numbers and between ten and fifteen major declension patterns for nouns, is a fertile area for the study of such competition between variant forms.

Our research involved testing three such slots in Czech where this phenomenon occurs. Two of these are from the so-called “hard masculine” declension pattern (exemplar word hrad ‘castle’). As a result of the merger and reorganization of the dominant o-stem class and the smaller u-stem class that had evidently already begun in proto-Slavic, in Czech the u-stem endings have spread widely across the old o-stem lexical stock in the genitive singular (gen. sg.) and the locative singular (loc. sg.), while the old o-stem endings have also penetrated the much smaller group of nouns that previously formed the u-stem class. The third is the result of a younger innovation in which feminine nouns inherited from the PSl i-stem pattern (exemplar word kost ‘bone’) have acquired to a greater or lesser degree the exponents of the old PSl ja-stem pattern (exemplar word *růže* ‘rose’) in the gen. sg. and most plural cases, forming a new if not always well-defined pattern (exemplar word *píseň* ‘song’).

3.1. The Czech National Corpus

Our main area of interest was to see whether exposure had an impact on the way Czechs perceived these variant forms, as well as how they used them. Our proxy for exposure was the Czech National Corpus (CNC), specifically the frequency with which forms occur in it.

[5] Latinate nouns (octopi~octopuses, etc.) are another area where variation can be looked at in English, but it has been an area of more research in derivational morphology, where variation is more widespread (normality~normalcy, etc.). However, derivational morphology is not seen as having the same impact on our understanding of utterance structure and the creation of “grammatical” meaning as does inflectional morphology.

By CNC, we mean specifically its layer of synchronic representative corpora (SYN2000, SYN2005, SYN2010, SYN2015).[6] Each of these corpora contain roughly 100 million tokens (excluding punctuation) and are representative in that they contain a mixture of text types, broken down at top level into publicistika 'journalistic texts', odborná literatura 'specialist texts' and beletrie 'imaginative texts'.[7] Attempts at producing balanced corpora based on research into reading habits gave a variety of results, summarized in Table 1.

Table 1. Text type breakdown (top level) in the SYN corpora

	SYN2000	SYN2005	SYN2010	SYN2015
Journalistic texts	60%	33%	33%	33.33%
Specialist texts	25%	27%	27%	33.33%
Imaginative texts	15%	40%	40%	33.33%

It is hard to tell without access to the comparative research underlying these changes, but there is a clear shift in favour of a more equal balance of text types.[8] Whether this is in fact representative of what Czechs read, it simplifies the task of comparing results from various text types within the corpus.

Our results drew on both the SYN2010 and SYN2005 corpus. Our goal was to identify nouns that exhibit variation in usage in the cases targeted. We conducted targeted searches in SYN2005 using the corpus query processor to retrieve all word forms with a particular shape and grammatical tag, e.g. ending in <u> and tagged as a masc. inanimate gen. sg. noun, or

[6] On our proxies for perception and use, see section 4 below.

[7] This term is more often translated as 'fiction', but in the corpora, it includes examples of the genre literatura faktu: creative non-fiction such as memoirs, travelogues, etc.

[8] A programmatic explanation for this shift away from "real world balance" towards "text type balance" is given in Cvrček et al. (2016).

ending in <a> with the same tag.[9] We then compared the resulting lists to find variant forms of a word, e.g. jazyku / jazyka, which represented the variation sought.

For each case, then, the lists of lemmas (with each ending and with both endings) ran to many thousands of items, so a manageable process was needed for verifying the data and catching potential errors. Our method is described in detail in Bermel & Knittl (2012b: 97–98), but in brief: all examples of use of the less frequent ending were verified manually, token by token, as were examples of the more frequent ending when it appeared in variation. We also removed all "non-words" from the lists and looked at any errors in the lemmas, which are often a sign that mistagging may have occurred.

These measures did not remove all erroneous forms retrieved, which would have been a much larger job, but they eliminated a large number of them. Even so, the effect on our overall statistics was not all that evident: for most lexemes, the proportions remained roughly constant. We thus arrived at three lists of lexemes where there was variation between two forms in the cases in question.

One early outcome of this work is that variation is a gradient feature. Looked at in absolute terms, we find variation with very high-frequency lexemes as well as very low-frequency lexemes. The proportion of case exponents in one vs. another form is also distributed along a scale: for one word, ending {1} may predominate, whereas for another word it might be ending {2}, and that dominance might be overwhelming or less strong. The only consistent observation is that few lexemes, other than those of low frequency or those where there is some sort of semantic motivation, exhibit equipollent distribution, e.g. both endings {1} and {2} occur in roughly even proportions. Where the variation is unmotivated or only partly motivated, there is almost always some sort of skew to the dominance of one exponent.

[9] The CNC always disambiguates and resolves in favor of one assignment for each place in the tag (unlike, for example, the Russian National Corpus, where ambiguities are never resolved and all possible tags are associated with a token). This disambiguation is partially rule-based and partially the result of a heuristic correction based on manual tagging of a portion of the corpus.

Over the past few years, we have used these lists, and a few others compiled in the meantime, to test various hypotheses about frequency. In particular, Bermel, Knittl & Russell 2017 demonstrated that proportional frequency of forms had a consistent effect, at least on the sort of tasks we were asking respondents to perform.

3.2. Using corpus data in surveys

The nature of a survey using native speaker respondents imposes limits on the amount of corpus data that we can test. Respondents fatigue easily; with a high number of short, repetitive tasks, we decided we could not ask them to spend more than 15–20 minutes on the survey without risking their attention flagging. We had the advantage of being able to pay respondents, which proved a useful motivating factor, but even so, the number of factors we could include was constrained. In this round, then, we looked at proportional frequency only. It was operationalized by choosing lexemes that fell into one of six proportional bands. The first questions to address are: why use bands at all; why, if so, do we use six bands; and why those particular boundaries for the bands?

What we are calling bands are often termed bins: all data found in a particular range is treated as having the same value. We might assume that the best option would always be to retain all precise values and thus not use any bands or bins: surely it must be more precise to retain the information that lexeme C has exponent {1} 13.7% of the time, while lexeme D has exponent {1} only 12.5% of the time. However, retaining this level of precision has an impact on the way we test our data. It implies a level of precision that in the real world may not exist, i.e. that because a 100-million-word corpus has those particular values, a native speaker will be more likely to favor exponent {1} in lexeme C than exponent {1} in lexeme D, and will be correspondingly more likely to use it in the first scenario than the second. For this reason, tests using bins may prove to be more realistic if we believe that corpora are best interpreted as a rough guide to the linguistic environment rather than an exact one; and that our abilities to track this linguistic environment may be approximate rather than precise.

To reduce at least one aspect of uncertainty, we limited our choice of nouns to those where at least 100 tokens in the case in question were found in the corpus. While this is admittedly an arbitrary level, we felt it was necessary to ensure validity of results. A set with four tokens of exponent {1} and two tokens of exponent {2} gives a proportional frequency of 67%:33%, but if only two tokens had been different, the proportions would have been reversed. With a sample of N ≥ 100, the chance of this happening is correspondingly reduced.

We set the number of bands and the particular boundaries between them opportunistically. For us, the most important criteria were that we get enough granularity in the results to be able to draw clear conclusions, and that we draw the boundaries around our bins in such a way that each of them represents a meaningful number of items. If we create a bin with few or no items in it, the information it yields will be limited and we will have a severely constrained choice of lexemes to use in our survey. In other words, we are not proposing that these specific bands have any inherent meaning themselves, i.e. that using six bands instead of seven indicates a rougher granularity of response overall, or because a word falls into the fifth instead of the sixth band that its behavior is qualitatively different. Instead, we are testing the usefulness of a scale itself: whether the proportional frequency of items in the linguistic environment makes a difference to people’s judgments and choices.

For our purposes, then, the most important feature of a scale is that the bands each contain adequate numbers of lexemes for us to construct a survey, and that the survey contain enough levels to assess the variation properly. How we assess it has an effect on (and is affected by) the statistical measures chosen.

Previously, for example, we had experimented with seven bands and four bands. The latter had little granularity and thus results were not as clear as we had hoped, while the former presupposed a ‘central’ band with roughly equal proportions of each exponent – which, as it turned out, were very difficult to find. This is because as mentioned in section 3.1, unmotivated and partially motivated variation tends to result in a skew dominance, where one exponent predominates in the vast majority of circumstances. In the end, we went with a division into six unequally-sized bands that allowed us a reasonable choice of lexical items for each band. The middle two bands were much broader (35% each), while the outside bands were very narrow (1% each), as this is where the greatest amount of variation occurs.

Table 2. Proportional bands used in this survey

Feature	0–1%	1–15%	15–50%	50–85%	85–99%	99–100%
{a} vs. {u}	2 lexemes	2 lexemes	2 lexemes	2 lexemes	2 lexemes	2 lexemes
{e/ě} vs. {u}	2 lexemes	2 lexemes	2 lexemes	2 lexemes	2 lexemes	2 lexemes
{i} vs. {e/ě}	2 lexemes	2 lexemes	2 lexemes	2 lexemes	2 lexemes	2 lexemes

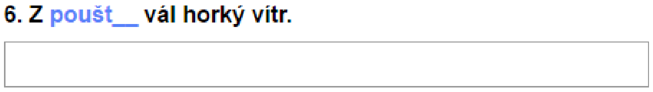
We further restricted our choice of lexemes by checking our findings in both SYN2005 and SYN2010, two corpora with identical high-level structures (see Table 1 above). To warrant inclusion in our survey, a lexeme had to fall into the same proportional f requency band in both corpora.

4. Methodology

Our main hypothesis was that respondents’ performance on production and evaluation tasks would vary depending on speakers’ reactions to reading tasks. However, we know from previous research that other factors have repeatedly shown to be a dominant influence on these sorts of tasks; therefore, we also hypothesize that the effect of reading-task factors will be smaller than those of other known contributing factors, such as the proportional frequency of these forms as observed in e.g. corpora.

Our survey was constructed by drawing sentence-long contexts from the Czech National Corpus wherever possible.[10] Two basic versions of the questionnaire were created: a production variant, where respondents were to input the missing endings of words, and an evaluation variant, where respondents were to rate each ending's acceptability on a scale from 1 (completely normal) to 7 (unacceptable). The same sentences were used as triggers in both basic versions.

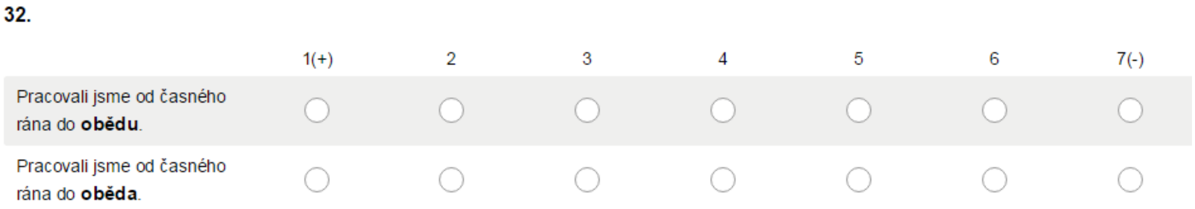
Gap-filling sentences were presented in the following format:



6. Z poušt__ vál horký vítr.

'A hot wind blew from the desert_____'

Ratings tasks were presented in the following format, with both possible forms displayed in context:



32.

	1(+)	2	3	4	5	6	7(-)
Pracovali jsme od časného rána do **obědu**.	○	○	○	○	○	○	○
Pracovali jsme od časného rána do **oběda**.	○	○	○	○	○	○	○

'We worked from early morning through to lunch'.

As can be seen, there was no particular attempt to hide what was being tested. This derived partly from experience and partly from the structure of the survey. In a gap-filling survey it is clear what is being tested, and so to hold conditions constant with the evaluation task, we needed to highlight the word concerned in the same way. On the matter of the naturalness of this sort of task, see e.g. Bermel & Knittl 2012b: 243–245.

[10] Sometimes these sentences needed to be modified – typically shortened to remove extraneous material, but also sometimes substituting lexical items to achieve a more 'neutral' effect for the trigger. This was to avoid respondent reactions directed not at the target feature, but at some other aspect of the text that was irrelevant, which could confound the results. In some instances (esp. with rarer lexemes), no suitable sentence could be found, and so we looked for sentences with synonyms or other lexemes close in meaning and substituted the target word in order to create the trigger.

The survey was supplied to users recruited via colleagues, family and friends on Surveymonkey. Each user completed 36 questions mixing a variety of features. A reading skills test followed, and then a further 36 questions as at the beginning.

Within each basic version (gap-filling and ratings), the test questions were thus divided into two ‘blocks’ (before/after reading skills test). Half the respondents took block A before block B; the other half took block B before block A. Within blocks, the order of questions was randomized.[11]

The reading texts contained two specially written passages containing the test words. We aimed to create one passage that would be comprehensible to ordinary readers, so as not to intimidate respondents and induce them to abandon the task, but we needed at least one passage to be considerably more difficult to ensure that not all respondents were at ceiling on the reading task as a whole.[12]

We tested our passages for “readability” using online tools at readability-score.com and read-able.com. The tests used on these sites (Flesch Kincaid Reading Ease and Grade Score, Gunning Fog Score, SMOG index, Coleman Liau Index, Automated Readability Index) consider factors such as sentence length, word length, and number of syllables per word. For a language like Czech with a relatively “shallow” orthography, they can be predicted to give reasonable results. Our first text was rated “easily understandable by 11-12 year olds”, while the second was rated as having postgraduate-level complexity.[13]

Following each passage there were four questions. The first asked respondents to evaluate, subjectively, their experience of reading. Jak pochopitelný je podle vás tento text? ‘How comprehensible did you find this text?’ Possible answers ranged from 1 - Velmi snadno ‘Very

[11] Surveymonkey does not support randomizing question order across two separate locations in a survey, so the constituent triggers of a block always had to remain in that block.

[12] If all respondents are at ceiling, the task will not serve to isolate relevant factors, as we cannot distinguish amongst the respondents based on performance.

[13] Read-able.com warned us, “Ooh, that's probably a bit too complicated. Have you thought about using smaller words and shorter sentences?”

easy' to 7 - *Velmi špatně* 'Very hard'. The intermediate points were numbered but not named. The remaining three questions were multiple-choice comprehension checks and were designed to test the precision or accuracy of the respondent's reading skills.

In one version of the passages, most test words appeared with the 'expansive' features {u} (masc. gen.), {u} (masc. loc.) and {e/ě} (fem. gen.). In the other version, most test words appeared with the 'recessive' features {a} (masc. gen.), {ě/e} (masc. loc.) and {i} (fem. gen.)[14]

There were thus eight basic possible permutations (task type (2) x block order (2) and reading passages (2). The assignment of respondents to these eight basic versions was done randomly by the software.

5. Results

305 Czech native speakers completed our surveys. Of those, 151 completed the gap-filling task and 154 completed the ratings task.

5.1. Between-subjects variables

Our respondents are from a cross section of Czech society, although they cannot be said to be a proportional representation of it. Younger, more educated, female respondents predominate compared to their numbers in society as a whole. The survey has this in common with others of its type (see Bermel, Knittl & Russell 2015: 291–292). Only the geographic distribution between two major speech regions (Bohemia vs. Moravia/Silesia) is proportional to the populations in those areas. The breakdown is given in Table 3.

Table 3. Biographical details

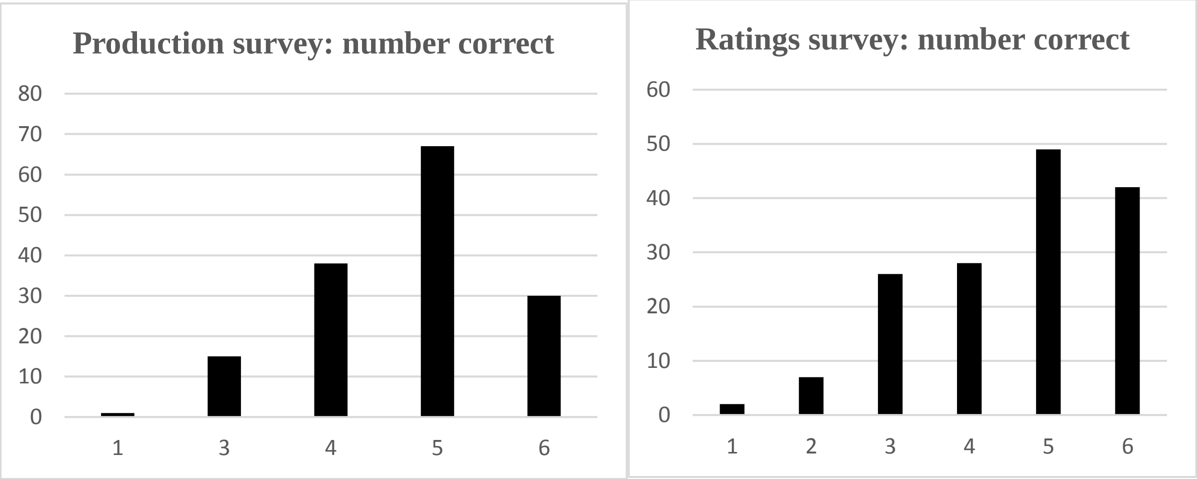
Age and region			Education and gender		
	Group	N		Group	N
Age	18-25	122	Education	Primary school	41

[14] Forms that were unrepresented in the corpus or represented only sporadically were not used, so as not to create the impression of an unnatural text. Instead, for those lexemes the common form was inserted.

	26-35	63		Technical school	7
	36-45	43		Secondary school	106
	46+	77		Tertiary education	151
Region	Bohemia	182	Gender	Male	101
	Moravia	123		Female	204

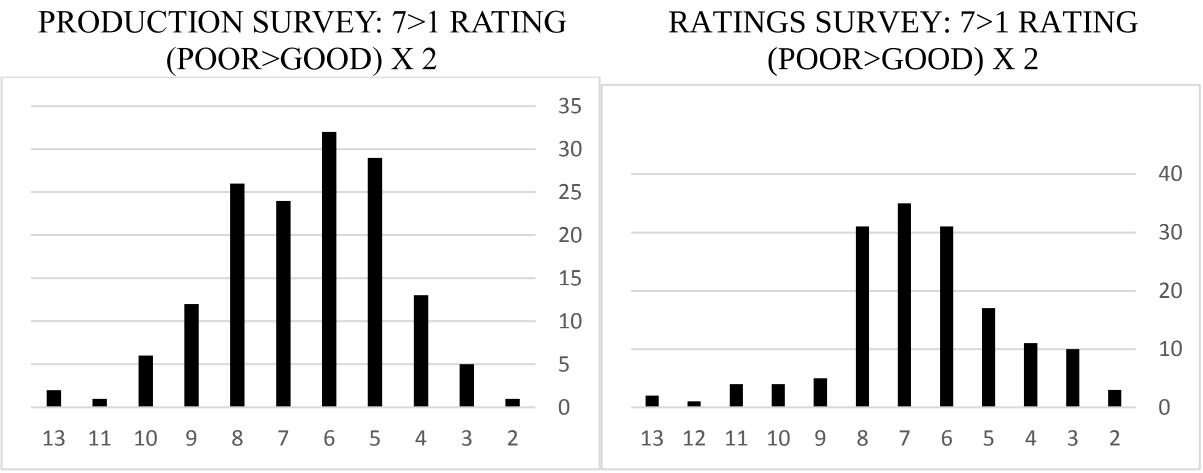
As previously mentioned, the between-speakers variables that interested us most in this study were those that involved reading. The first, given in Figure 1, concerns the accuracy of answers to reading comprehension questions (results are given separately for those completing the production version of the survey and those completing the ratings version of the survey). The second, given in Figure 2, concerns respondents’ perceptions of difficulty of the texts.

Figure 1. Accuracy: production vs. ratings



In Figure 1, we can see that the bell curve is shifted right: on average, more people got a question right than wrong, so the top of the curve is at 5/6 correct answers. This compares with Figure 2, where we have more centered bell curves. The score here represents the sum of their answers: thus a score of 7 could represent a judgement that one text was very hard (6) while another was very easy (1). The most common score was either 6 or 7, suggesting that people found at least one text relatively easy.

Figure 2. Difficulty: production vs. ratings



One problem with bell curves like those in Figures 1 and 2 is that some of our data is quite sparse. On the accuracy tests, no one got all questions wrong, and the number of respondents getting 1-2 questions right is also vanishingly low. This was particularly notable in the production survey, where only 1 respondent scored 1 correct question and none scored 2 correct questions.

On the difficulty rating, the scores could run from 14 (both texts maximally difficult) to 2 (both texts maximally easy). For the production cohort, only one respondent rated both texts as maximally easy and few people rated both texts as difficult (only three respondents between 11 and 14 points). For the evaluation cohort, three people rated both texts as maximally easy and a further three gave between 12 and 14 points.

Thus, although the bell curve is evident in all four permutations, the sparseness of data at the ends of the bell curve (points on the scale with 0-2 answers) means that results may not appear significant.

5.2. Results of production task

Repeated measures ANOVAs were carried out to ascertain the influence of proportional frequency (“mixture” of forms) and sentential context on the frequency of choice of the “expansive” ending. ANOVA is a statistical test that shows which of a series of entered factors had a statistically significant effect. It and allied measures such as the partial eta squared (η^2) value can be used to detect how the overall effect is apportioned out amongst the factors.

Region of origin and age groups were entered as between-subjects ("biographical") factors, alongside the self-rated difficulty of the text and the number of correct comprehension-check answers.

In our results, there were occasional significant "biographical" factors, but they differed from feature to feature. For the masc. gen. sg., Region was the only significant feature: $F(1, 132) = 9.85$, $p < .003$, partial $\eta^2 = .07$. For the masc. loc. sg., there was a significant interaction between Region and Proportional Frequency: $F(4.18, 551.26) = 2.65$, $p < .04$, partial $\eta^2 = .02$. For the fem. gen. sg. we found a significant interaction between Perceived Difficulty and Proportional Frequency: $F(41.61, 549.24) = 1.50$, $p < .03$, partial $\eta^2 = .10$, and Reading Accuracy: $F(4, 132) = 3.31$, $p < .02$, partial $\eta^2 = .09$. All these significant results were sporadic and had small effect sizes.

In sum, we found no consistent evidence that reading scores or other biographical data consistently influence the production task.

5.3. Results of evaluation task

Repeated measures ANOVAs were carried out to ascertain the influence of proportional frequency ("mixture" of forms) and sentential context on the acceptability rating of forms.

Region of origin and age group were entered as between-subjects ("biographical") factors, alongside the self-rated difficulty of the text and the number of correct comprehension-check answers.

In examining our analyses, we will be interested in (1) which factors seem to have the largest effects, (2) which factors crop up most consistently across all three features examined.

5.3.1. Masculine genitive singular {a} vs. {u}

In the masc. gen. sg., we found two major effects (based on the F value and the partial η^2 value, which is derived in part from it). These were both connected with the proportional frequency in the corpus of the ending tested:

proportional frequency * ending: F (3.53, 468.95) = 538.45, $p < 0.001$, partial $\eta^2 = .80$

proportional frequency: F (4.30, 571.74) = 63.66, $p < 0.001$, partial $\eta^2 = .32$

These suggest that the largest effect is due to the frequency of the ending tested in the corpus relative to the frequency of the untested ending. A second, medium-sized effect is that of proportional frequency itself, which suggests that e.g. more skewed distributions of endings for a lexeme are treated differently from more equal distributions.

There were a number of minor effects, which are listed in order of decreasing effect size in Table 4.

Table 4: Minor effects in the masc. gen. sg.

Feature	F values	p value	part. η^2
context	F (1, 133) = 20.60	p < 0.001	.13
prop. frequency * ending * age group	F (10.58, 468.95) = 6.38	p < 0.001	.13
context * proportional frequency	F (4.78, 635.26) = 17.04	p < 0.001	.11
age group	F (3, 133) = 4.86	p < 0.004	.10
prop. frequency * reading accuracy	F (21.49, 571.74) = 1.70	p < 0.03	.06
ending * region	F (1, 133) = 5.95	p < 0.02	.04
prop. frequency * age group	F (12.90, 571.74) = 2.05	p < 0.02	.04
prop. frequency * ending * region	F (3.53, 468.95) = 2.61	p < 0.05	.02

These minor effects (where the F value and the partial η^2 value are much smaller) frequently involve interactions with proportional frequency, suggesting that they are not equally distributed across all the lexemes studied. Instead, for example, reading accuracy scores play a role in people's ratings, but only for certain lexemes based on their placement on the proportional frequency scale (again, suggesting that they react differently to words whose alternate endings have a skewed distribution vs. those whose endings have a more equal distribution in the corpus).

Somewhat surprisingly, age group shows up three times here, suggesting that there are more general differences in how people of different ages reacted, as well as specific interactions with corpus frequency.

5.3.2. Masc. loc. sg.

The two major effects in the masc. loc. sg. were identical to those in the gen. sg.:

proportional frequency * ending: F (3.36, 447.13) = 465.63, $p < 0.001$, partial $\eta^2 = .78$

proportional frequency: F (4.21, 560.21) = 79.90, $p < 0.001$, partial $\eta^2 = .38$

The minor effects are listed in Table 5. As with the gen. sg., many of the minor effects also include proportional frequency, indicating that they are not equally distributed across all words but take account of skewed vs. equal distribution of variant forms in the corpus. Reading accuracy showed up again, in interaction with proportional frequency. Age group also showed up, by itself and in two interactions. Difficulty rating showed up twice in the minor effects, both in interactions with features of the sentences presented (context and proportional frequency by ending).

Table 5. Minor effects in the masc. loc. sg.

Feature	F values	p value	part. η^2
context * diff. rating	F (11, 133) = 2.31	$p < 0.02$	.16
prop. frequency * ending * diff. rating	F (36.98, 447.13) = 1.49	$p < 0.04$	.11
prop. frequency * ending * age group	F (10.09, 447.13) = 5.07	$p < 0.001$	.10
prop. frequency * age group	F (12.64, 560.21) = 3.61	$p < 0.001$	.08
context	F (1, 133) = 10.42	$p < 0.003$	.07

age group	F (3, 133) = 3.19	p < 0.03	.07
prop. frequency * reading accuracy	F (21.06, 560.21) = 1.75	p < 0.03	.06
context * proportional frequency	F (4.87, 647.15) = 8.19	p < 0.001	.06

5.3.3. Fem. gen. sg.

The two major effects in the fem. gen. sg. were identical to those seen in both masc. sg. cases:

proportional frequency * ending: F (3.63, 482.91) = 510.25, p < 0.001, partial η^2 = .79

proportional frequency: F (4.18, 555.88) = 73.89, p < 0.001, partial η^2 = .36

The minor effects are listed in Table 6. The continuing significance of proportional frequency is shown here as well. Additional factors in this analysis include reading accuracy, region, context and perceived difficulty.

Table 6. Minor effects in the fem. gen. sg.

Feature	F values	p value	part. η^2
ending * diff. rating	F (11, 133) = 2.30	p < 0.02	.16
prop. frequency * ending * diff. rating	F (39.94, 482.91) = 1.92	p < 0.002	.14
prop. frequency * ending * age group	F (10.89, 482.91) = 4.10	p < 0.001	.09
context	F (1, 133) = 11.04	p < 0.002	.08
prop. frequency * reading accuracy	F (20.90, 555.88) = 1.79	p < 0.02	.06
context * proportional frequency	F (4.60, 612.30) = 3.62	p < 0.005	.03
prop. frequency * region	F (4.18, 555.88) = 2.37	p < 0.05	.02
prop. frequency * ending * region	F (3.63, 482.91) = 2.59	p < 0.01	.02

5.4. Significant factors in common

Certain features showed up in two or three of our cases. In two cases we found significant effects of the following features or interactions:

- Age group
- Proportional frequency * ending * region
- Proportional frequency * ending * difficulty rating
- Proportional frequency * age group

In all three cases we found significant effects of the following features or interactions:

- Proportional frequency * ending
- Proportional frequency
- Proportional frequency * ending * age group
- Context
- Context * proportional frequency
- Proportional frequency * reading accuracy

6. Discussion

We noted above a difference between the two sorts of tasks completed by our respondents. The production task showed sporadic significant contributions by features or interactions of features, but no sign of consistent, significant effects in any one area. The number of significant features was much greater with the ratings task and the primary problem facing the researcher is to distinguish which of them to single out for further investigation.

6.1. Avoiding Type I errors

A Type I error, or a "false positive" result, occurs when our statistical test reports that the connection noticed is not the result of chance, i.e. is a significant predictor of future behavior. However, the number of apparently anomalous positive results here deserves comment. We can explain them in two ways. One possibility is that there really is an effect here, but it is not general to the category of "morphological overabundance" and we can thus draw no further conclusions from it. For example, there may be a feature of one or two of the words used that we did not account for, and what we are actually looking at is a feature limited to a particular lexeme or small set of lexemes. Another possibility is that the presence of a significant result is a side-effect of having a large number of variables and interactions. Significance is of course nothing more than an estimation of the probability that the results are down to chance, and hence if enough variables and interactions are included, the probability rises that at least one of them will register as significant. The probability of these occasional “false positives” is increased by the fact that our surveys were relatively large, with over 150 participants each; analyses of larger cohorts are more prone to return small effects as significant.

For this reason, we focused our attention on factors that held constant across all three of the features studied. Doing so reduced the chance that we would be committing a Type I error.

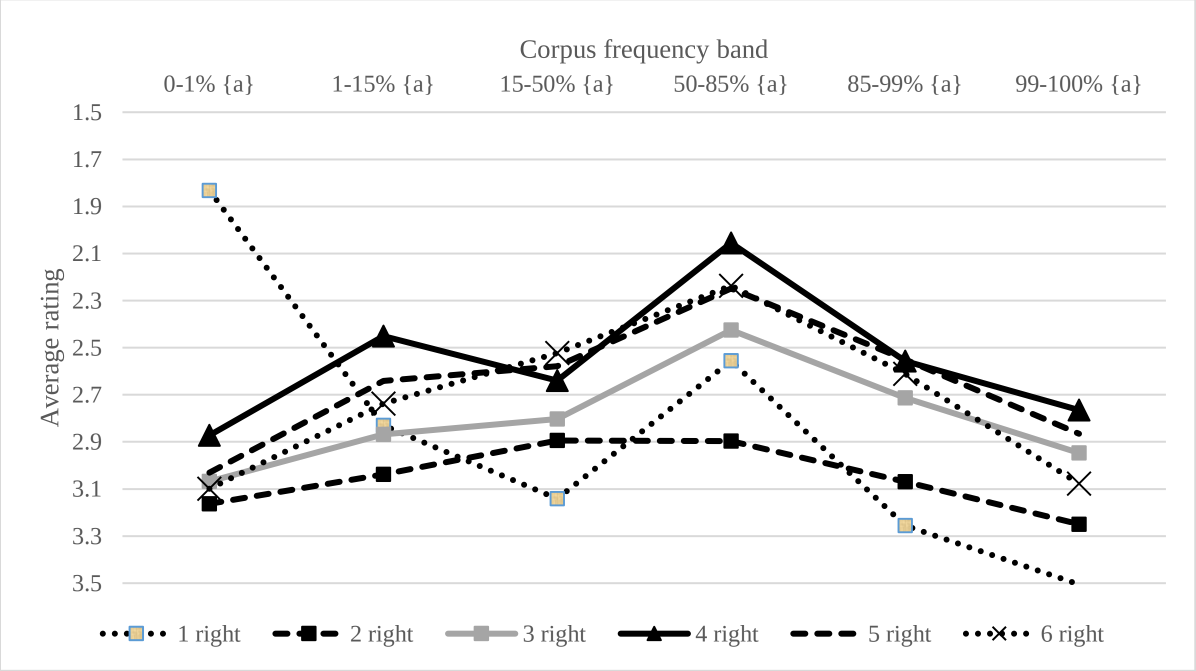
6.2. Explaining variations in ratings

Most of the variation in ratings is accounted for by the effects of the interaction between proportional frequency of forms in a corpus and the specific variant ending used. In other words, the relative frequency with which we see one form vs. another is the largest influence on our ratings of it. A second, medium-sized effect is always proportional frequency by itself, which indicates that, regardless of which variant is involved, different types of balance of variants affect our judgments. A skewed ratio of forms (say, 99:1) is treated differently than a more balanced ratio of forms (say, 5:1 or 3:1), and this operates regardless of which specific variant is in question. These findings are entirely in line with our previous investigations (Bermel & Knittl 2012a, 2012b, Bermel, Knittl & Russell 2015, 2017).

Some variation in our ratings is attributable to the syntactic context in which the lexeme is situated. This again is in line with previous findings. Bermel & Knittl (2012b) had found a larger and more consistent effect of context, but that difference is probably down to the different structure of the study. Our earlier study had focused on two features only: proportional frequency and context, and so tested a wider variety of contexts, allowing for more detailed results. In the current study, the addition of other factors made it impractical to include more than two contexts without the survey becoming unwieldy for respondents. The current analysis is consequently less fine-grained and on only two levels as regards context, so the importance of this factor is suppressed.

Most interestingly for our current purposes, we identified a consistent small effect of the interaction between proportional frequency and accuracy: Better reading scores indicate more positive ratings, with the most positive ratings coming from those who had moderate-to-high scores on the reading accuracy task.

Figure 3. Text comprehension accuracy vs. frequency of -a ending for masc. gen. sg.



As can be seen in Figure 3, the effect was more noticeable for words where both endings are better attested (middle four bands), as opposed to those where one ending is completely predominant (outer two bands). The same pattern can be observed in Figures 4 and 5, for the masc. loc. sg. and the fem. gen. sg. respectively.[15]

Figure 4. Text comprehension accuracy vs. frequency of -*ě ending for masc. loc. sg.*

[15] The anomalous shape of the “one correct” band has to do with the fact that only two respondents fell into this bracket, so the reactions are highly dependent on individual idiosyncrasies.

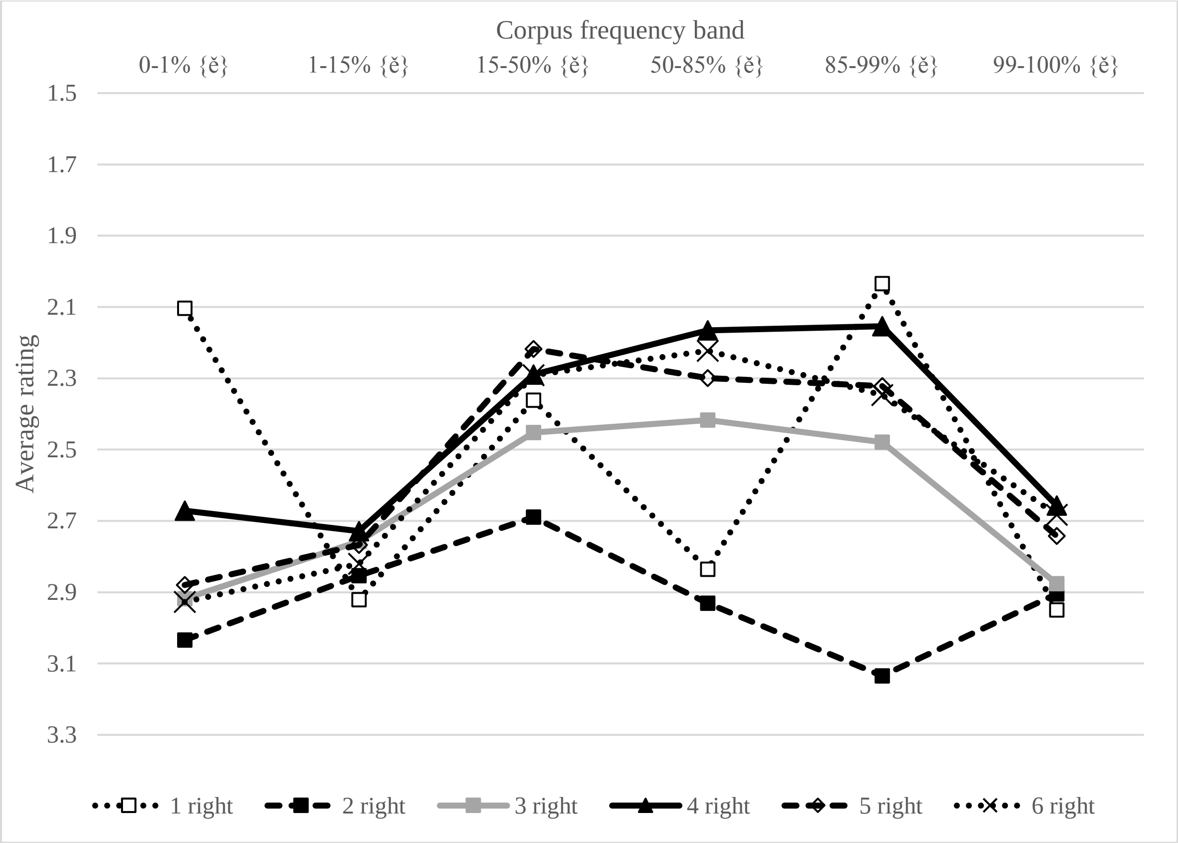
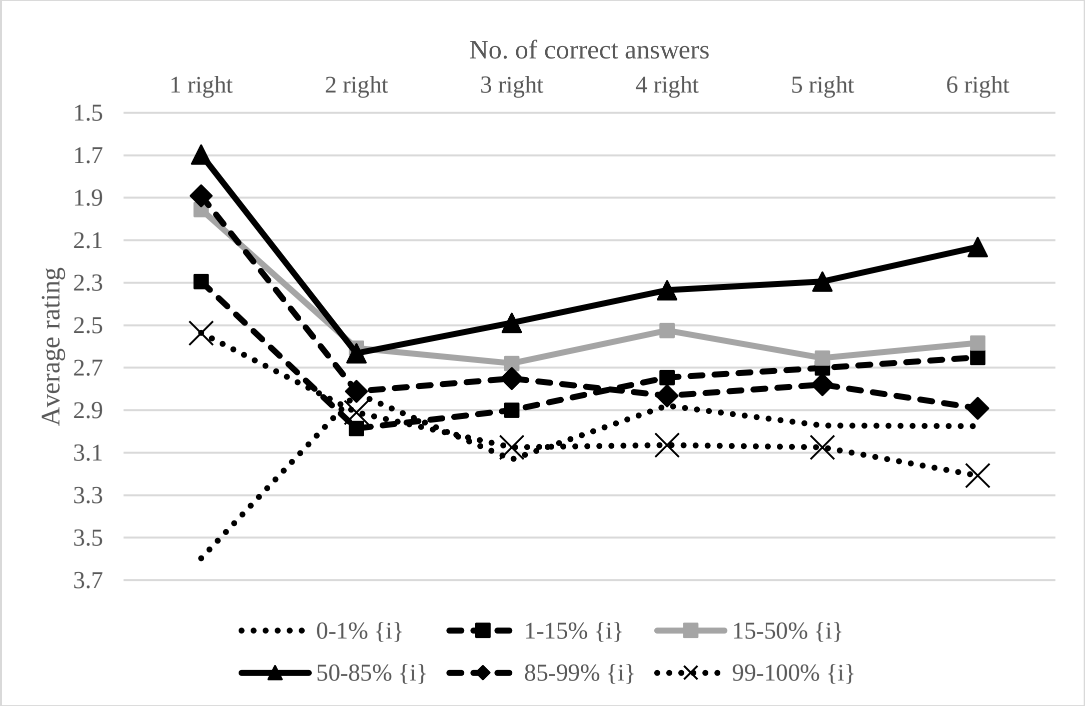
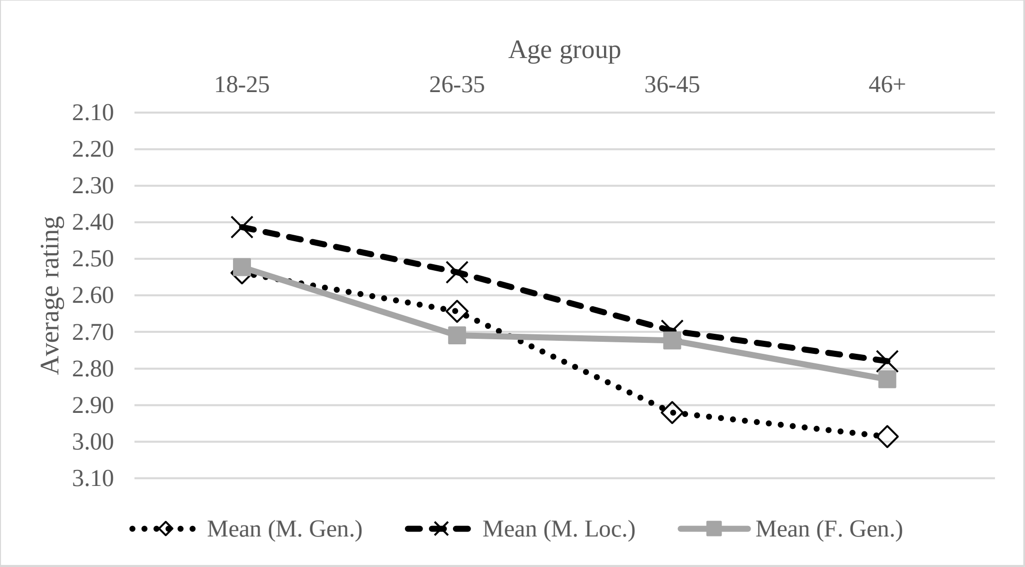


Figure 5. Text comprehension accuracy vs. frequency of -*ě ending for masc. loc. sg.*



Age plays a surprisingly consistent role in choices, as can be seen in Figure 6. Across all features studied, older people are less susceptible to rate items positively

Figure 6. Mean rating by age group



This result was surprising, as age had not emerged in our previous surveys as a consistent and significant factor.

Of our two reading tasks, difficulty ratings registered as influential for ratings on two out of our three features, but only accuracy on the comprehension checks registered as influential for all three features. We noted that this variation is strongest for slots where both forms are represented in the corpus in more than sporadic fashion (> 1%).

7. Conclusions.

In our original hypotheses, we had proposed that performance on production and ratings tasks would vary depending on speakers' reactions to reading tasks. The first part of this hypothesis – concerning the production task – was not confirmed. The second part – concerning the ratings task – was confirmed. We only felt confident proposing one of the two reading tasks – the accuracy test – as a reliable indicator, as the other task only registered significant for two of the three features studied.

We had also proposed that the effect of these factors would be smaller than those of other known contributing factors, such as the proportional frequency of these forms as observed in e.g. corpora. This part of the hypothesis was confirmed.

We noted that neither reading task seemed to influence production tasks in cells where there is overabundance. In retrospect, the ability to comprehend a text and answer questions correctly might not be closely connected with how we produce forms. However, levels of reading skills do seem to influence ratings tasks in cells where there is overabundance: the better one's tested accuracy, the more positively one evaluates the endings. The difference between high-scorers and low-scorers is more marked for items where speakers are regularly exposed to both forms. This made us wonder whether accurate readers might turn out to be broader or more proficient readers, who would be likely to have more exposure to written texts, and thus be more accepting of a variety of forms.

Age showed up in these studies as a significant factor, whereas in our other studies of the same features its effect had not been significant. Users of different ages may not have significantly different mechanisms for judging and producing case endings, but nonetheless they appear to react differently to linguistic stimuli that attempt to influence their behavior, such as our reading passages and tests. It may be that the greater linguistic experience of older speakers results in a different pattern of response.

Our hypothesis regarding wider exposure and higher ratings would lead us to expect, therefore, that older respondents would have had more exposure to a larger number of forms and thus be more positive about a greater variety of them. However, the results were in fact the exact opposite: age group came out as a significant factor in the evaluation tasks, but the older the group, the less positive overall were the ratings. This means that the two variables in question here (reading accuracy and age) are not covariate, as they do not share in producing the same result. Greater exposure over time, as opposed to over quantity of text solely, seems to lead, paradoxically, to a hardening of opinion, giving indirect evidence for pre-emption ("how speakers learn what not to say" (Goldberg 2011: 132)). It suggests that pre-emption, like other cognitive processes, does not finish at some “critical age” but continues to operate through adulthood.

Another way to look at this might be to see age as a counterweight to growing vocabulary and increased exposure. As these rise over time, pre-emption can provide a mechanism for ensuring that our reaction time does not rise in the same degree.

References

Albright, Adam, and Bruce Hayes. 2003. Rules vs. analogy in English past tenses: a computational/experimental study. Cognition 90: 119–161.

Bermel, Neil. 2004a. 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 *Čeština* – univerzália a specifika 5, ed. Zdeňka Hladková and Petr Karlík, 163–171. Prague: Nakladatelství Lidové Noviny.

Bermel, Neil. 2004b. Jak často se vyskytují (vyskytujou) tzv. hovorové tvary 1. os. j. č. a 3 os. mn. č. v Českém národním korpusu? [How often do the so-called colloquial forms of the 1 sg. and 3 pl. occur in the Czech National Corpus]? In Korpus jako zdroj dat *o češtině,* Petr Karlík, ed., 29-40. Brno: Masarykova univerzita.

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

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

Bermel, Neil, Luděk Knittl and Jean Russell. 2015b. From standard to norm through the lens of corpora and native speakers. Prace Filologiczne 67: 21–43.

Bermel, Neil, Luděk Knittl and Jean Russell. 2017. Frequency data from corpora partially explain native-speaker ratings and choices in overabundant paradigm cells. Corpus Linguistics and Linguistic Theory. doi: 10.1515/cllt-2016-0032.

Bybee, Joan L. and Dan I. Slobin. 1982. Rules and schemas in the development and use of the English past tense. Language 58: 265–289.

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, and Zdeněk Velíšek. 2005. SYN2005: A genre-balanced corpus of written Czech. Prague: Ústav Českého národního korpusu FF UK. Available at www.korpus.cz

Čermák, František, Jan Králík and 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.

Chandler, Steve. The English past tense: Analogy redux. Cognitive Linguistics 21: 371–417.

Conway, Andrew R., Michael J Kane, Michael F. Buntin, D. Zach Hambrick, Oliver Wilhelm and Randall W. Engle. 2005. Working memory span tasks: a methodological review and user's guide. Psychonomic Bulletin & Review 12: 769–786.

Cvrček, Václav, Anna Čermáková, and Michal Křen. Nová koncepce synchronních korpusů psané češtiny. Slovo a slovesnost 77: 83–101.

Dąbrowska, Ewa. 2008. The effects of frequency and neighbourhood density on adult speakers' productivity with Polish case inflections. Journal of Memory and Language 58: 931–951.

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

Eddington, David. 2000. Analogy and the dual-route model of morphology. Lingua 110: 281–298.

Frisch, Stefan and Maria Brea-Spahn. 2010. Metalinguistic judgments of phonotactics by monolinguals and bilinguals. Laboratory Phonology 1: 345-360.

Gordon, Peter and Randall Hendrick. 1997. Intuitive knowledge of linguistic co-reference. Cognition 62: 325–370.

Haber, Lyn R. 1976. Leaped and leapt: A theoretical account of linguistic variation. Foundations of Language 14: 211–238.

Hupet, Michel, Donatienne Desmette, Marie-Anne Schelstraete. 1997. What Does Daneman and Carpenter's Reading Span Really Measure? Perceptual Motor Skills 84 (2): 603–608.

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

Lečić, Dario. 2015. Morphological doublets in Croatian: The case of the instrumental singular. Russian Linguistics 39: 375–393.

Prasada, Sandeep, and Steven Pinker. 1993. Generalization of regular and irregular morphological patterns. Language and Cognitive Processes 8: 1–56.

Sgall, Petr. 2011. Perspektivy standardní češtiny [Perspectives on standard Czech]. In: Jazyk, mluvení, psaní, ed. Eva Hajíčová and Jarmila Panevová, 180–204. Prague: Karolinum.

Staum Casasanto, Laura, Philip Hofmeister, and Ivan Sag. 2010. Understanding acceptability judgments: Additivity and working memory effects. In Proceedings of the 32nd Annual Conference of the Cognitive Science Society, 224–229. Austin: CSS.

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