

A pilot study on the relationship between corpus data and acceptability judgments of competing forms

Neil Bermel

n.bermel@sheffield.ac.uk

Introduction

This study contributes to answering the question: can data from a corpus be interpreted as faithfully representing the acceptability of the features examined for native speakers of the language?¹

The question arises at least in part because in the course of this research, I made use of the Czech National Corpus, in which two extensive components are said to be “representative corpora” of synchronic written Czech: SYN2000 and SYN2005. These corpora were compiled on the basis of research into reading habits conducted in the 1990s and in 2000–2001 (see e.g. Šulc 2001, Králík 2001). A corpus so constituted should give a faithful picture of the written world surrounding the average reader. It is possible, even probable, that language users are to a certain extent influenced by this picture: that it forms or modifies their “aesthetic” reactions and judgments as to the acceptability of linguistic features. Such an opinion underpins (often only implicitly) many a corpus-based study. This forms the basis for our first hypothesis:

Hypothesis 1 (the “strong” hypothesis): The frequency of items in the written language, as reflected in a representative corpus, expresses or forms native speakers’ preferences for one or another form or structure, and thus the proportions of forms represented in the corpus agree with the expressed preferences of native speakers.

This hypothesis most frequently arises from introspection on the part of individual researchers. Their corpus data may be reliable, but the acceptability of the feature studied tends to be based on the intuition of a single scholar, rather than on soundings into the opinions of various users.²

There have been several larger-scale analyses to date, but none of them provides all that much support for the strong hypothesis. Halliday (1991a, 1991b, 1992) suggests a frequency ratio of 9:1 is significant from an informational and structural point of view, and proposes frequency boundaries of 10% and 90% as especially significant, delineating ‘marked’, ‘variant’ and ‘unmarked’ alternatives. Research by Kempen and Harbusch (2005) and by Divjak (2008) points to the asymmetricity of the correspondences between corpus data and tests of native speakers’ opinions in the area of syntax. Their conclusions suggest a less categorical vision of this hypothesis:

Hypothesis 2 (the “weak” hypothesis): Frequency in a representative corpus and preferences are related, but we should not overestimate the extent of this relationship. The link between the ratio of forms in a corpus and the preferences of native speakers is statistically significant, but nonetheless the connections between ratios and preferences are not strict and do not apply in all instances.

Of course, a third possibility is that corpus data do not bring us any insights into the state of a language as such:

¹ This research arose as part of the “Chapters in a Grammar of Czech” project, led by F. Štícha of the Czech Language Institute, Czech Academy of Sciences.

² I do not claim that a correspondence between corpus data and the reactions of a single linguist tells us nothing about the system of a language, but there are well founded methodological reservations to such an approach. Even the most experienced linguist knows what he has already found in the corpus and has a detailed knowledge of the area he is researching, and therefore cannot by himself serve as a reliable source of information. This approach points out the dangers of both the observer’s paradox and the well-known problem of data gleaned from linguists.

Hypothesis 3 (the “null” hypothesis): Frequency in a corpus does not relate to the preferences of native speakers. There are no significant correlations between the frequency of features in a corpus and their acceptability for native speakers.

The third hypothesis seems at first glance to be an extreme position, but it is possible to find well-founded arguments in support of similar propositions, see e.g. Oliva and Doležalová (2004).

Outline of the issue

I decided to attempt an empirical sounding into the relationship between corpus data and the acceptability of morphosyntactic forms. An interesting example of variation presented itself in the Czech nominal paradigm *hrad* ‘castle’, which was suitable for several reasons.

Czech is a language with a rich inflectional morphology: most descriptions ascribe to it four genders or subgenders, two numbers and seven cases. It has a considerable variety of noun paradigms: we can identify at least ten to twelve common declensional patterns. *Hrad* is a high-frequency paradigm with notable variation in case forms. To sketch the extent of this declension type, I looked at forms of the genitive singular (hereafter Gsg), which clearly shows the differences between declension patterns of masculine inanimate nouns (*hrad-u* vs. *stroj-e*). In the corpus SYN2005, there are c. 5.5m tokens marked as inanimate masculine nouns. Of these, 1.5m tokens belong to the paradigm *hrad*, or c. 27%. The frequency of lemmas (lexemes) is similar: roughly 26,000 lemmas (18%) belong to the paradigm *hrad* of a total 142,000 lemmas for inanimate masculine nouns.

Variation within the paradigm *hrad* has been the subject of several studies since the 1950s. Several attempted an overview of the issue based on explanations found in prescriptive and descriptive manuals of the time (Cummins 1995, Rusínová 1992, Sedláček 1982) while others utilised excerption from books (Klimeš 1953); still others relied on soundings into native speakers’ usage (Kasal 1992) or on their evaluations of variant forms (Bermel 1993). Later studies examine the extent of variation in data drawn from the Czech National Corpus (Bermel 2004, Štícha 2009). No study, however, has yet focused on the possible connections between corpus data and soundings into native-speaker judgments.

From the various possibilities offered within the paradigm *hrad*, I chose the Gsg, in which forms end in *-u* or *-a*, with both endings being possible for some nouns. An advantage in this case was the fact that in the latter instance, the choice of formant does not seem to be strongly motivated.³ Lexical conditioning factors are very weak: borrowed words rarely have *-a* and homonymy with animate words can influence the choice of formant. There are no phonological conditioning factors.

As for syntactic conditioning factors, we read in the contemporary grammar manual *Příruční mluvnice češtiny* (1995, p. 253):

“Nouns with the adverbial meaning of place (or less possibly of time or manner) are often connected with the ending *-a*. As a genitive of object (or in other meanings) the same nouns tend to have the ending *-u*: *do kalicha/u* ‘into the chalice’ – *dotkl se kalichu* ‘he touched the chalice’, *do roka* ‘in/within a year’ – *dožil se jednoho roku* ‘he lived a year’, *do kouta* ‘into the corner’ – *nebylo jediného koutu* ‘there was not even a corner’, *do rybníka* ‘into the pond’ – *nebylo jednoho rybníku* ‘there was not a single pond’” (my translation).

The examples cited above concern two syntactic constructions that are relatively rare in contemporary Czech: a verb taking a direct object in the genitive case and the use of the genitive of negation. To check this assertion, however, it was therefore necessary for the

³ The Czech academic grammar *Mluvnice češtiny* (1986, vol. II, p. 305) claims: „The distribution of forms depends on various factors: word-formational, syntactic, semantic, in places on the phonetic structure of the word” (my translation). This, however, refers to the general distribution of forms, not to the distribution of forms specifically in words where variation occurs.

analysis to contain a sufficient number of varied syntactic constructions, including the genitive of negation and direct government by the verb of a genitive object.

I started by determining the extent of variation in forms in the Gsg of this paradigm.

Table 1. Extent of variation in the Gsg

	total	-u	% u	-a	% a
Gsg. tokens of paradigm <i>hrad</i> in1 SYN2005 ⁴	374 025	1 193 709	86,9	180,316	13,1
Lexemes of these tokens	13 688	13 017	95,1	804	5,8

Lexemes with the formant *-a* in the Gsg form only 5.8% of all lexemes, but represent 13.1% of all instances of the Gsg. The overall frequency of forms in *-a* is thus higher than of forms ending in *-u* (primarily because the formant *-a* is not found with such a large number of low-frequency nouns).

Parameters of the questionnaire

In setting up a pilot study, I focused on lexemes with variation between the formants *-u/-a* in the Gsg found in the corpus SYN2005. We find this variation in the corpus in 112 nouns. Some are of high frequency (1000+ occurrences), e.g. *rok* ‘year’, *zákon* ‘law’, *zákoník* ‘codex’, *jazyk* ‘language’, *kostel* ‘church’, *sen* ‘dream’, although there are also numerous nouns in this group with low frequency (< 10 occurrences). The proportion of the formants *-u* vs. *-a* differs significantly within this group (i.e. there is no fixed ratio of forms with one formant to forms with another); however, we notice very few nouns with a balanced distribution between the two endings.

I selected seven words altogether that were to represent different proportions of the formants *-a* and *-u* in the corpus. At either end of the scale (*les* ‘forest’, *průchod* ‘passageway’) were the control words, where only one Gsg formant was actually represented in the corpus: hence we have *lesa*, never **lesu* and *průchodu*, never **průchoda*. Between these extremes lay five other words: *sýr* ‘cheese’, *rybník* ‘pond’, *týl* ‘nape/rearguard’, *dvorek* ‘small courtyard’, *pokojík* ‘little bedroom’. Their Gsg forms are represented in SYN2005 in differing proportions (see table 2).

Table 2. Questionnaire words

form in -u	examples	proportion	gloss	form in -a	examples	proportion
1 <i>lesu</i>	0	0,00 %	‘forest’	<i>lesa</i>	4316	100,00 %
2 <i>sýru</i>	118	9,10 %	‘cheese’	<i>sýra</i>	1185	90,90 %
3 <i>rybníku</i>	125	11,90 %	‘pond’	<i>rybníka</i>	929	88,10 %
4 <i>týlu</i>	115	34,60 %	‘nape/rearguard’	<i>týla</i>	217	65,40 %
5 <i>dvorku</i>	150	74,60 %	‘little courtyard’	<i>dvorka</i>	51	25,40 %
6 <i>pokojíku</i>	199	93,00 %	‘little bedroom’	<i>pokojíka</i>	15	7,00 %
7 <i>průchodu</i>	194	100,00 %	‘passageway’	<i>průchoda</i>	0	0,00 %

In choosing words, I was guided by several principles. The lexemes chosen had to be of higher frequency (>150 tokens in the Gsg in SYN2005), so that the proportion of formants would be sufficiently reliable. It had to be possible to construe the word as having a

⁴ In these figures, proper names are excluded. The tagger did not recognize some rare animate nouns, whose Gsg form ends in *-a*, and mistakenly tagged these as inanimate; however, excluding these nouns would only strengthen the tendency seen in the table for inanimate nouns with a Gsg in *-a* to have a higher average frequency.

locational sense, so that I could test the words in the same sorts of contexts and have the contexts sound reasonably natural; I therefore avoided lexemes with primarily abstract meaning. I looked for words where there were minimal possibilities for lexical effects, meaning differences caused by multiple meanings within a lexeme (with the exception of the word *týl* ‘nape/rearguard’, where I found no such effects). Furthermore, I had to exclude word where the data were coloured by corpus effects – for example, the repeated appearance of a single expression in newspaper advertisements.

Because the *Příruční mluvnice češtiny* had mentioned syntactically conditioned differences in the choice of formant, I selected five syntactic contexts characteristic of the Gsg:

Inherence (ownership, characteristics): *Bydlel v novém domě na kraji <lesa>* ‘He lived in a new house on the edge <of the forest>’

Adverbial phrase of motion (no adjective): *Pustila se do housky a <sýra>*. ‘She set into the roll and <cheese>.’

Adverbial phrase of motion (one adjective): *Z oploceného <dvorku> záhadně zmizela tři nákladní auta*. ‘Three trucks mysteriously vanished from the fenced-in <yard>.’

Other “adverbial” constructions with a preposition: *Auto skončilo uprostřed <rybníka>*. ‘The car ended up in the middle of the <pond>.’

Directly governed genitive: *Nevšimla jsem si <průchodu> v hradbách a tak jsem musela obejít celý hrad*. ‘I did not notice the <passageway> through the fortifications and so had to go round the entire castle.’

The questionnaire thus had seven lexemes with two formants each, tested in five contexts, i.e. in total 70 sentences.⁵ With a few exceptions, the examples were drawn from the corpus and used either verbatim or in simplified form (it was sometimes desirable to leave out excessive details and shorten the sentence). The respondents were asked to evaluate each sentence on a scale from 1 to 7 (where, according to Czech custom, 1 is the best rating and 7 the worst). Only the endpoints of the scale had precise descriptors.

Most questionnaires of this sort measure **active production** of forms, i.e. they offer a gap that the respondent fills with an appropriate form (Kasal 1992 is a typical example). My questionnaire, in contrast, evaluated the acceptability of variants – that is, the **passive reception** of forms. This approach was chosen for several reasons. The representativity of the SYN2005 corpus means that the corpus presents a cross-section of the written world surrounding the average Czech reader, but this fact does not yield any conclusions about this reader’s productive capabilities. In other words, the fact that a reader sees one or another form more or less frequently in written texts does not necessarily imply that he actively uses this form himself to that same extent, but rather that he should find it acceptable in roughly the same degree (see the hypotheses above). A further reason was that filling in gaps creates an “either/or” answer, highlighting only the most common variant and not presenting any possibility for distinguishing degrees of acceptability. An active-production approach therefore proved unsuitable for my study.

Due to the length of the questionnaire, I tested each formant on a different sentence, so that respondents would not fall prey to boredom as they read the same sentence over and over. I thus created two basic versions of the questionnaire: in the first, the formant *-a* appeared in one group of sentences and *-u* with another, while in the second version, each formant appeared with the other group of sentences. This limited the risk of side effects that might arise from individual sentences. Each of these versions was presented in two different

⁵ At seminars where I presented the results of the questionnaire, comments on the choice of words focused either on the fact that there were only a few of them, or on specific characteristics of the lexemes used. Undoubtedly it would be desirable to include more lexemes in the questionnaire, but increasing their number would mean limiting the scope of data gathered on each word, which limits the validity of findings in a different way. The choice of words was somewhat increased in follow-up surveys now being analysed.

orderings of the examples, to limit the risk of respondents being influenced by the order in which the sentences appeared. The first four sentences of each version always consisted of two sentences I hoped would be felt to be normal and acceptable, and two that would be unacceptable. This was to encourage respondents to use the entire scale of possible answers. A total of four versions of the questionnaire were thus distributed randomly to respondents. They filled in the questionnaire in my presence, either individually or in small groups. I chose a scale with precise points to aid in statistical analysis. Having considered a magnitude estimation approach, in the end I did not use it, because it would have proved more demanding and time-consuming for the respondents. Even so, the questionnaire was relatively long, and so I did not make use of distracter sentences. The word in question in each sentence was in bold face; this was meant to avoid side effects (for example, when a completely unrelated feature in the sentence catches a reader's attention, causing him to evaluate the sentence based on that finding rather than the one looked for).⁶

Character of the responses

The pilot study, which took place in Sheffield, UK during January and February 2008, had several characteristics that should be noted.

The number of respondents (N=20) was relatively low. Among them were several students of Czech language and literature, who have a high level of self-awareness regarding their own linguistic usage. Four respondents had emigrated more than 10 years earlier, although in these instances the possibility of interference from English as their second language was minimal, thanks to the complete absence of analogous features in English grammar.

All the respondents knew me personally and thus, despite the length of the questionnaire, evidently put effort into it: all completed questionnaires were usable. Nonetheless, a few respondents did not use the entire scale; they avoided point 7, or used only points 1 and 7.

The average age of my respondents was 29; the median age was 25. The youngest respondent was 21, and the oldest 67. Fourteen of them were from Bohemia, four from Moravia and two from Silesia, with proportions thus coming close to those found in the Czech population as a whole. There were no significant differences between the median or mean marks assigned that would correspond to the different versions of the questionnaire; our measures against side effects were thus either successful or superfluous.

Statistical analysis of the questionnaire⁷

We performed an ANOVA (analysis of variability) test on combined scores for each word (level of acceptability of *-a* minus level of acceptability of *-u*). The model had two factors: word (7 levels) and syntactic constructions (5 levels). The main effect was that of the word:

$$F(6, 114) = 36.27, p < 0.001, \text{partial } \eta^2 = 0.66$$

The existence of an interaction between syntactic construction and word was confirmed:

$$F(24, 456) = 2.45, p < 0.001, \text{partial } \eta^2 = 0.15$$

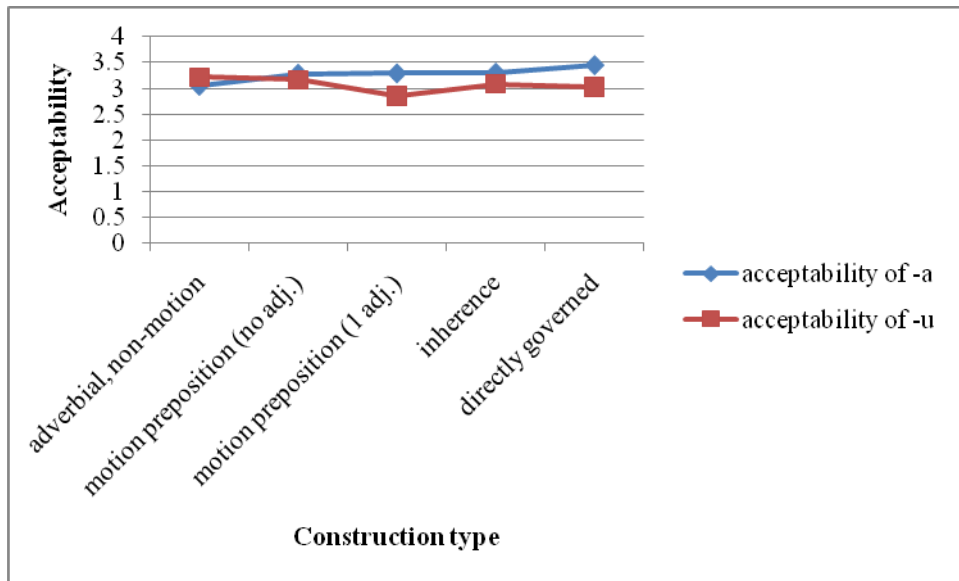
This indicates that choice of word is a very significant factor in assessing the acceptability of one or another formant. It is worth noting the effect of individual constructions for certain words; however, this effect is not as significant.

⁶ I would like to thank Ewa Dąbrowska and Dagmar Divjak for advice on questionnaire compilation and Luděk Knittl for help compiling and editing the questionnaire. I followed where possible the principles set out in Schütze (1996) and Cowart (1997). Although in syntactic questionnaires it is not recommended to call respondents' attention to the feature studied, I contravened this by not including distracter sentences and by highlighting the form in question. I did this in the belief that in morphosyntactic studies we can allow ourselves a greater degree of freedom than in syntactic ones: a morphological form is a clearly delimited item and marking it does not force the respondent to think about structures he may be only dimly aware of. What we lose in the immediacy of his reaction is compensated for by the precision of his response.

⁷ I am grateful to Ewa Dąbrowska for the statistical analysis of these data and for help with its interpretation.

Tests showed that there is a correlation of low significance (-0.53) between the syntactic context and the formant's acceptability. The results are in table 3

Table 3. Syntactic constructions



Finally we conducted t-tests (tests of significance), which measure the differences in acceptability between individual word pairs. We used combined scores, as in the ANOVA tests and noticed significant differences for scores between certain neighbouring words:

les and *sýr*, $t(19) = 4.35$, $p < 0.001$

pokojík and *průchod*, $t(19) = 7.95$, $p < 0.01$

We found no other significant differences between neighbouring words. We found statistically significant differences between some non-adjacent pairs of words:

les and *rybník*, $t(19) = 6.23$, $p < 0.001$

rybník and *dvorek*, $t(19) = 5.58$, $p < 0.001$

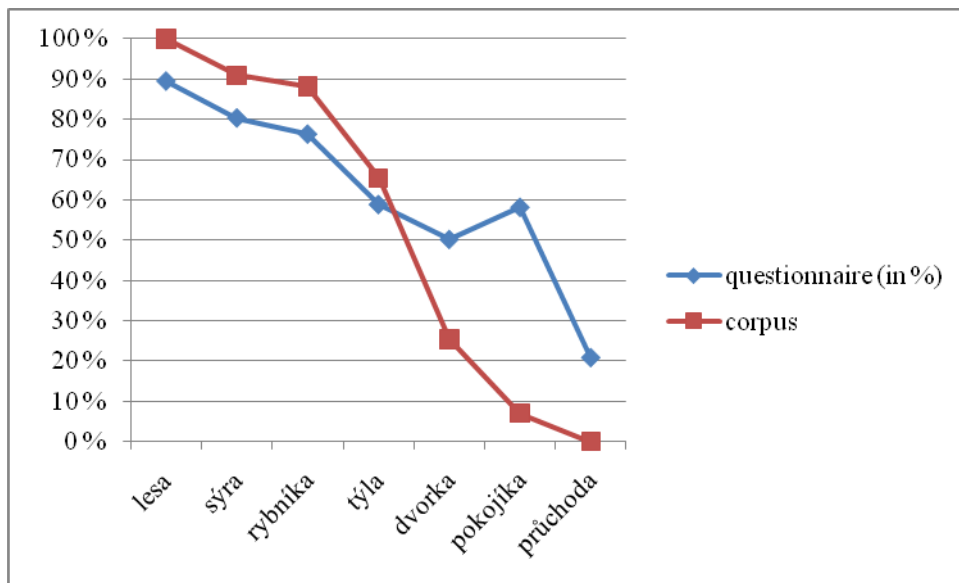
dvorek and *průchod*, $t(19) = 7.01$, $p < 0.001$

It is important to note that in four out of these five instances we were comparing words at either end of our scale, which actually showed no variation at all in the corpus (*les*, *průchod*), with words from the middle of the scale, i.e. there is no regular correlation between words where we do have evidence of variation in the corpus.

Statistical comparisons of questionnaire data with corpus data

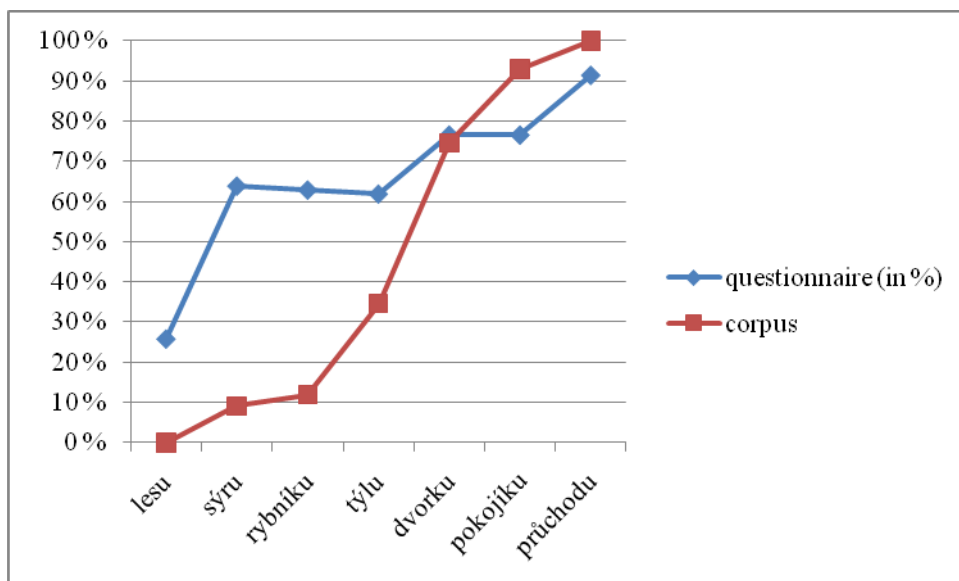
In the next phase of our analysis, we compared relative frequency from the corpus with acceptability judgments from the questionnaire. Purely for the purpose of this exercise, we converted the acceptability judgments from our 1–7 scale into percentages 0 % – 100 %. For the formant *-a* the correlation was 0.88 when endpoint words (*les*, *průchod*) were included, or 0.83 without them; this correlation is considered very significant (see table 4):

Table 4. The formant *-a:* correlation between corpus data and questionnaire data



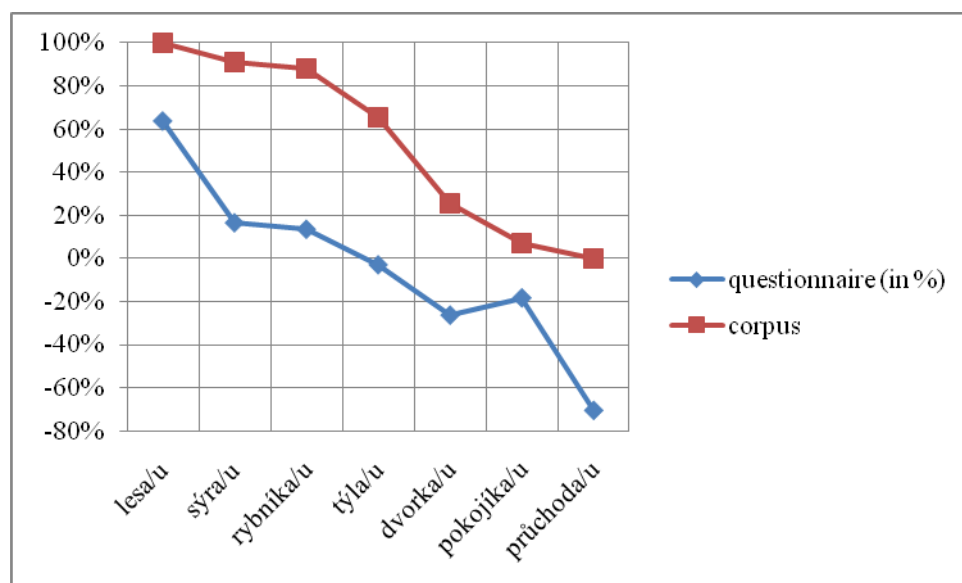
With the formant *-u* we found a correlation of 0.81 including endpoint words, 0.92 without them; again this proved to be a very significant correlation (see table 5):

Table 5. The formant *-u:* correlation between questionnaire data and corpus data



With combined scores, the correlation was 0.89 with endpoint words included, 0.94 without them (very significant), see table 6:

Table 6. Combined scores: correlation between questionnaire data and corpus data



Conclusions: corpus data and acceptability

The results of this pilot study confirm the results of earlier research into syntax, see e.g. Divjak (2008): corpus data point to a significant correlation with data on the acceptability of forms, but the relationship between corpus frequency and acceptability is not a simple one. It is not possible to determine the degree of acceptability of a form purely from corpus data, especially in the case of forms with low relative frequency.

Divjak points out that low relative frequency of syntactic constructions in the corpus does not correspond to unacceptability for native speakers, and the data from this pilot study support her conclusions for morphosyntax. On the other hand it does seem, again in line with Divjak, that low acceptability (**průchoda*, **lesu*) corresponds to low frequency. High relative frequency (> 50%) testifies to high acceptability, whereas higher acceptability (< 3.5) does not indicate high relative frequency. According to my data, forms with low relative frequency (*sýru*, *rybníku*, *pokojíka* 10%) can be just as acceptable for native speakers as forms with higher relative frequency (*týla* 65%, *týlu* 35%, *dvorka* 25%).

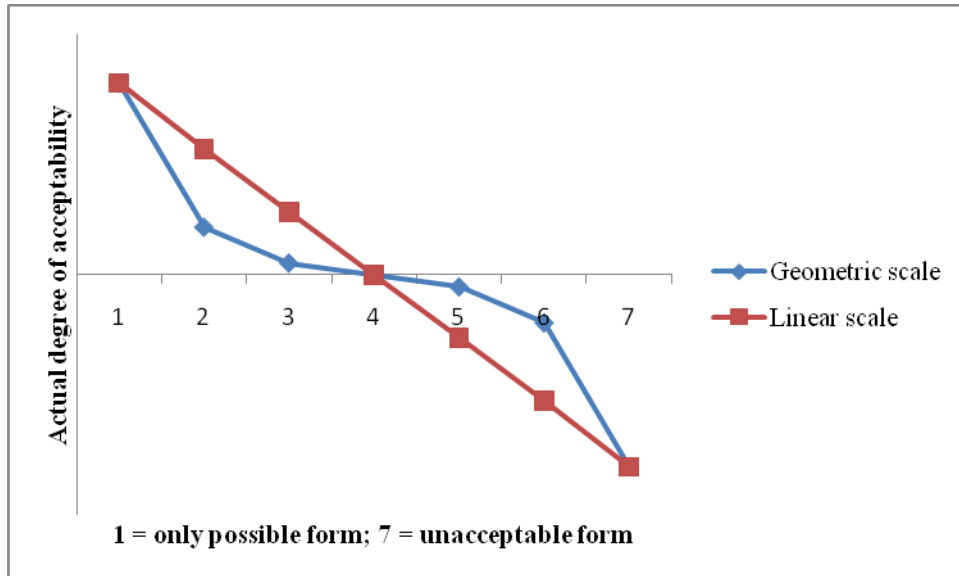
To deduce acceptability, then, the relative frequency of an item in the corpus alone is not enough. It has to be combined with a careful analysis of empirical data. Hypothesis 1 is therefore not confirmed; the degree of acceptability is not closely tied to relative frequency of items in the corpus. Hypothesis 3 is likewise rejected, because a certain relationship (correlation) between corpus data and acceptability judgments was confirmed. Based on these data, Hypothesis 2 seems the most probable: a general correlation does exist, but we should not overestimate the significance of an item's relative frequency in the corpus in attempting to predict acceptability. Only figures over 60%, and the virtual absence of an item in the corpus seem to be significant predictors of acceptability.

First postscript: Are acceptability scales linear or geometric?

It is appropriate to ask whether some of the effects mentioned above could arise as a side-effect of the testing methods used. Ideally, the seven-point scale is linear and yields interval data; in other words, the difference in acceptability between points 1 and 2 is of the same order as the difference between 2 and 3, between 3 and 4 etc. The specified definitions of the end points ("only possible", "unacceptable") and the undefined nature of the points between the two extremes ("more" or "less" acceptable) could, however, point to the existence of a

geometric scale (i.e. yielding ordinal data only), in which respondents perceive a more significant difference between the endpoints and those adjacent to them than they do between adjacent middle points (see table 7).

Table 7. Linear and geometric scales

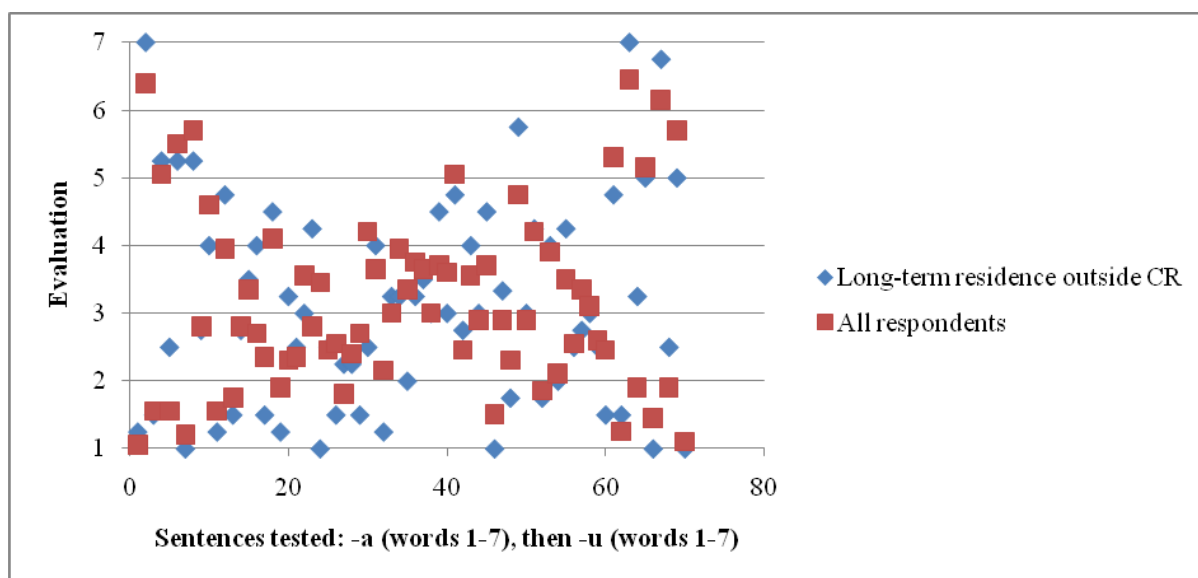


Providing that studies are carefully designed, it is common in linguistics to accept such data as valid for statistical analysis (Cowart 1997: 120-121) and we must simply resign ourselves to this as a potential drawback to this method of assessing acceptability. The fact that these results were confirmed as significant allows us to place some confidence in the discriminative capacities of our respondents.

Second postscript: Long-time residence outside the Czech Republic

Among my respondents were four Czechs who have lived outside the Czech Republic for many years. From the data it seems that long-term émigrés reject some forms more categorically than speakers ordinarily resident in the Czech Republic. In the answers in table 8 we see more “extreme” values (closer to 1 and 7) than is the case for respondents at large. It is possible that émigrés, being to some extent “alone” with their own Czech, are less exposed to the variety of contemporary language use and so tend towards more pronounced decisions, depending on whether or not a given form is in their own vocabulary. This effect is not, however, a regular one and does not seem to have a predictable effect on their evaluation of forms.

Table 8. Long-time residence outside the Czech Republic



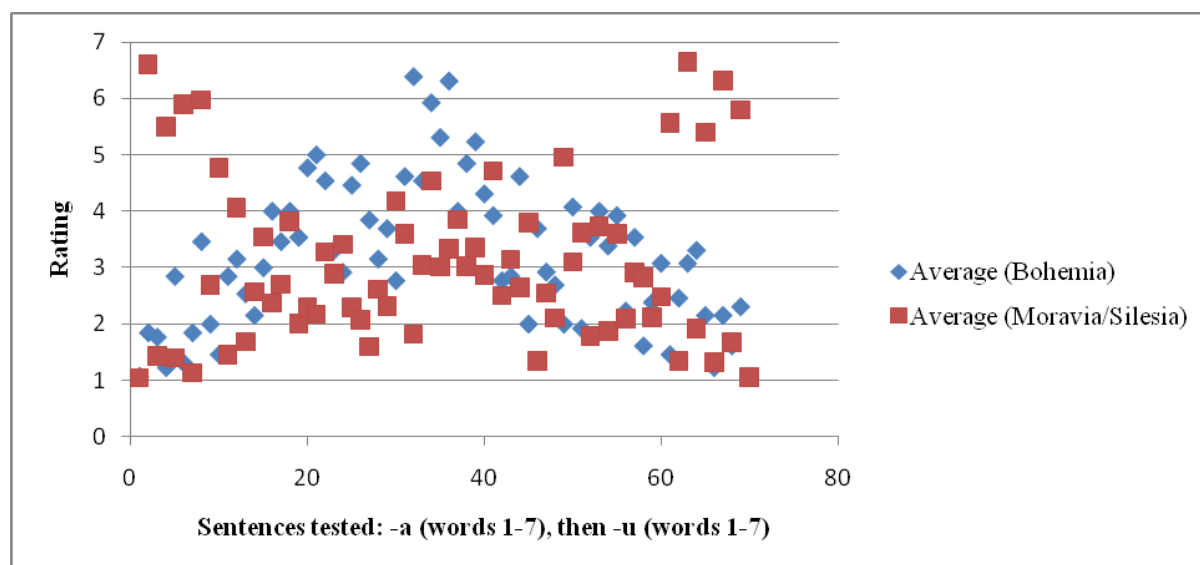
Long-term residence abroad does not seem to have a consistent effect on the use of genitive case forms (although my results and the existence of literature on changes in the syntax and lexicon of émigrés indicate the need for a more detailed study of this phenomenon).

Third postscript: Regional differences within the Czech Republic

Case forms in Czech can vary significantly depending on the region where the speaker lives and where he comes from. The pilot study respondents overall formed a representative sample of the Czech population in that they represented the nation's three major regions (Bohemia, Moravia and Silesia) in appropriate proportions, and so it was possible to at least ask whether there were any perceptible differences between the preferences of Czechs and those of Moravians or Silesians. For some other case forms it is often alleged that Moravians prefer the historically conservative endings while Bohemians prefer the more innovative ones, and if this were the case, we would expect Moravians to favour forms in *-a*, which are for many words the historically older variant and for others would represent the extension in Czech of the less productive formant.

The differences, however, turned out to be insignificant. The average Moravian and Silesian rating of forms in *-a* is 3.02 vs. 3.16 for Bohemians. As for forms in *-u*, the average rating from Moravians and Silesians was 3.18 vs. 3.04 for Bohemians (see table 9).

Table 9. Differences in individual ratings between Moravians/Silesians and Bohemians



The regional differences were not statistically significant. As can be seen in table 9, the distribution of answers is nonetheless interesting and would merit more detailed investigation with a larger number of respondents and words. It seems that despite the lack of significance of the data overall, there may be regional differences in preferences as regards individual word forms.

Conclusions

Data from this large representative corpus do appear to correlate with frequency ratings from native speakers to a certain extent, but not precisely. A large percentual representative of forms in the corpus testifies to high acceptability, but a percentage below 50 did not turn out to be a reliable measure of acceptability. These findings correspond to those in the existing literature on syntactic features and, once tested on larger sample sizes and a more extensive range of words, could be useful as we seek to interpret corpus data.

Sources

- BERMEL, N. (1993): Sémantické rozdíly v tvarech českého lokálu. *Naše řeč* 76, s. 192–198.
- BERMEL, N. (2004): *V korpuse nebo v korpusu? Co nám řekne (a neřekne) ČNK o morfologické variaci v tvarech lokálu.* In: Hladká, Z. – Karlík, P. (eds.): *Čeština – univerzálie a specifika* 5. Praha: Nakladatelství Lidové Noviny, s. 163–171.
- COWART, W. (1997): *Experimental Syntax: Applying Objective Methods to Sentence Judgments.* Thousand Oaks, CA: Sage Publications.
- CUMMINS, G. (1995): Locative in Czech: -u or -e: Choosing locative singular endings in Czech nouns. *Slavic and East European Journal* 39, s. 241–260.
- ČNK: *Český národní korpus – SYN2005.* Ústav českého národního korpusu FF UK. Dostupné na webu: www.korpus.cz
- DIVJAK, D. (2008): On (in)frequency and (un)acceptability. In: Lewandowska-Tomaszczyk, B. (ed.): *Corpus Linguistics, Computer Tools and Applications – State of the Art.* Frankfurt am Main: Peter Lang, s. 213–233.
- HALLIDAY, M. A. K. (1991a): Corpus studies and probabilistic grammar. In: Aijmer, K. – Altenberg, B. (eds.): *English Corpus Linguistics.* New York – London: Longman, s. 30–43.

- HALLIDAY, M. A. K. (1991b): Towards probabilistic interpretations. In: Ventola, E. (ed.): *Functional and Systemic Linguistics: Approaches and Uses*. Berlin – New York: Mouton de Gruyter, s. 39–61.
- HALLIDAY, M. A. K. (1992): Language as system and language as instance. In: Svartvig, J. (ed.): *The Corpus as a Theoretical Construct: Directions in Corpus Linguistics*. Berlin – New York: Mouton de Gruyter, s. 61–77.
- KASAL, J. (1992): Dublety a jejich užití. In: *Philologica* 65. Olomouc: Univerzita Palackého, s. 107–114
- KEMPEN, G. – HARBUSCH, K. (2005): The relationship between grammaticality ratings and corpus frequencies: A case study into word order variability in the midfield of German clauses. In: Kepser, S. – Reis, M. (eds.): *Linguistic Evidence: Empirical, Theoretical and Computational Perspectives*. Berlin – New York: Mouton de Gruyter, s. 329–349.
- KLIMEŠ, L. (1953): Lokál singuláru a plurálu vzoru „hrad“ a „město“. *Naše řeč* 36, s. 212–219
- KRÁLÍK, J. (2001): Vyvážení zdrojů Synchronního korpusu češtiny SYN2000. *Slovo a slovesnost* 62, s. 38–53.
- MČ: *Mluvnice češtiny* (1986). Petr, J. (ed.), Praha: Academia.
- OLIVA, K. – DOLEŽALOVÁ, D. (2004): O korpusu jako o zdroji jazykových dat. In: Karlík, P. (ed.): *Korpus jako zdroj dat o češtině*. Brno: Masarykova univerzita, s. 7–10.
- PMČ: *Příruční mluvnice češtiny* (1995). Karlík, P. – Nekula, M. – Rusínová, Z. (eds.), Praha: Nakladatelství Lidové Noviny.
- RUSÍNOVÁ, Z. (1992): Některé aspekty distribuce alomorfů (genitiv a lokál sg. maskulin). *Sborník prací filozofické fakulty brněnské univerzity A* 40, s. 23–31.
- SEDLÁČEK, M. (1982): V Záhřebě i v Záhřebu. *Naše řeč* 65, s. 11–15.
- SCHÜTZE, C. (1996): *The Empirical Base of Linguistics: Grammaticality Judgments and Linguistic Methodology*. Chicago: University of Chicago Press.
- ŠTÍCHA, F. (v tisku). Lokál singuláru tvrdých neživotných maskulin (ve vlaku vs. v potoce): úzus a gramatičnost. *Slovo a slovesnost*.
- ŠULC, M. (2001): Tematická reprezentativnost korpusů. *Slovo a slovesnost* 62, s. 53–61.