

What can corpora tell us about centre and periphery in Czech morphology?

NEIL BERMEL, Sheffield
n.bermel@sheffield.ac.uk

ABSTRACT

This article explores the relationship between corpus data and grammaticality judgements. It takes as its starting point contentions (Hellberg 1992, Oliva and Doležalová 2004, Newmeyer 2003) that data from corpora are useful only in an incidental way to those interested in the grammar of a language; any conclusions that derive directly from the accumulation of corpus data are necessarily suspect. Drawing on observations about the nature of central and peripheral features (Halliday 1991a, 1991b, 1992), I argue that data derived from corpora, if carefully framed and contextualized, can provide information that matches closely with existing observations concerning the grammaticality of certain endings. The corpus data supporting this contention deal with three examples of variation found in the corpus: one where both variants are recognized as standard (the locative singular of masculine and neuter nouns), one where a variant can be stylistically marked (the locative plural of masculine and neuter nouns), and one where one variant is clearly perceived as an error (unexpected syncretism between the genitive plural and other plural cases).

This contribution was supported by the project GA ČR 405/03/0377 *Možnosti a meze gramatiky češtiny ve světle Českého národního korpusu* (*Exploring the Core and Limits of Czech Grammar through the Czech National Corpus*). – Research for this project was carried out in part in Prague under the auspices of an inter-academy exchange sponsored by the British Academy and the Czech Academy of Sciences, and the author gratefully acknowledges their support.

This contribution arose in part as a response to an ongoing debate within the multinational project *Exploring the Core and Limits of Czech Grammar through the Czech National Corpus*.¹ Since the project began in 2002, we have discussed at length exactly how we are using our data to construct an image of Czech grammar, and how scholars mediate and interpret corpus results.

One fundamental question we have been asking is: *what is the relationship between corpus data and linguistic intuition?* From this hang a number of more specific subquestions, including the following:

- *What is the value of corpus data in making linguistic judgments?*
- *To what extent should we be guided by corpus data in evaluating grammaticality?*
- *Is there a way to draw a line between grammaticality and ungrammaticality based on corpus data?*

¹ See <http://mam.ujc.cas.cz>.

1. Textual frequency and grammatical probability

In a 1992 article, Steffan Hellberg of the Swedish Academy Grammar project wrote:

The purpose of this paper is to show, by means of a number of examples, how linguistic corpora can be used as tools outside the field of corpus description proper. It will be demonstrated that the corpora function mainly as a huge collection of instances while the strict definition of a given corpus is of less interest. It is argued that the existence of corpora does not free grammarians from using their intuition as native speakers and their judgement of the representativeness and normality of each instance. As a corollary, exact figures of frequency and dispersion are of limited interest (1992: 311).

Hellberg's approach treats the corpus essentially as a gigantic card file composed of randomly or arbitrarily selected sources. All that corpora do is provide examples for an existing grammar template structured by the linguistic intuition of native speakers.

Corpora have moved on since the early 1990s, and now offer us a sophisticated array of balanced sources and powerful tools for interpreting the data derived from them.² The degree of scepticism that Hellberg expressed in rejecting everything that a corpus offers other than examples is thus no longer warranted. Nonetheless, some of our debates echo this desire to reassure ourselves and others that the corpus is not replacing human reason as the arbiter of the language, and so we warn people sternly against using the corpus mechanically to answer questions of rightness, wrongness, appropriateness, or function.

A convincing version of this sceptical view of corpora appears in Oliva and Doležalová 2004. The authors' argument rests on a fundamental schism between the concepts of *langue* and *parole*, and they warn that if we take ideas about representativity of corpora too literally, we will end up needing to include in our grammar of Czech numerous features that to any native speaker's intuition are simply incorrect and end up neglecting other features that are reasonably frequent. Among the examples presented from the Czech National Corpus were: *v mnoha ze svých dřívějších studiích* and other examples of infelicitous uses of the locative plural in place of the genitive plural. Because these are rare, but not vanishingly so, and because other individual features of *langue* are not attested in the corpus, such as constructions like *snažil jsem se mu ho najít*, Oliva and Doležalová then conclude that:

1. Constructions that are fully possible in *langue* need not be documented at all in a corpus;
2. Constructions that are excluded from *langue*, on the other hand, may be documented in a corpus, or even well-documented.

The conclusion that derives from this could be succinctly formulated as follows: a large enough corpus (a collection of examples of *parole*) can be an 'objectivization' of that *parole*, but as a matter of principle cannot be an 'objectivization' of *langue*. From this it follows that the fundamental material for grammatical research (i.e. for the explication of *langue*) is that very *langue* itself (an implicit knowledge of the language); for such work, corpora can be 'only' an inspiration and a corrective (Oliva and Doležalová 1994: 10, translation mine).

This point is made as well by Newmeyer in a recent article:

In summary, we have grammar and we have usage. Grammar supports usage, but there is a world of difference between what a grammar is and what we do – and need to do – when we speak.

² See <http://ucnk.ff.cuni.cz> for information about the Czech National Corpus and its corpus query processors.

(...) Because of the divergence between grammar and usage, one needs to be very careful about the use made of corpora in grammatical analysis, and particularly the conclusions derived from the statistical information that these corpora reveal. Now, for some purposes, as we have seen, statistical information can be extremely valuable. Corpora reveal broad typological features of language that any theory of language variation, use, and change has to address. (...) But it is a long way from there to the conclusion that corpus-derived statistical information is relevant to the nature of the grammar of any individual speaker, and in particular to the conclusion that grammars should be constructed with probabilities tied to constructions, constraints, rules, or whatever (Newmeyer 2003: 695).³

But for those of us who research foreign languages rather than our own, there is a powerful impetus to find a way to use corpora more effectively: if you like, to broaden the “interest” that exact figures hold for us, precisely because we do not have access to native-speaker intuition.

Halliday points out that the crucial problem here is relating what he calls *grammatical probability* to *textual frequency*. We can take *grammatical probability* here to mean something like ‘the actual state of the grammar, the range of potential constructions and their standing and function in the language’, while *textual frequency* means, in our given instance, ‘the number of times something appears in the corpus’. Halliday believes that in general, there does exist a set relationship between probability and frequency. To simplify slightly, where two variants are equally probable, or, in Halliday’s terms, members of an *equiprobable system*, both show up with a frequency of .5. An alternative is what Halliday calls a *skew system*, where central and peripheral variants show up with respective frequencies of .9 and .1.

Halliday based this figure on observations of marked and unmarked categories in Chinese and English, done on relatively large-scale collections of data. But this is not simply a statistical shot in the dark. Halliday writes that in information theory, the ratio of roughly 1:9 is that at which redundancy and information balance out:

The skew value of 0.9 / 0.1 seemed rather an unmotivated artefact of decimalism, until I noticed that a possible explanation for it could be found in information theory. A system of probabilities 0.5 / 0.5 is of course minimally redundant. The values 0.9 / 0.1 incorporate considerable redundancy; but this is just the point at which redundancy and information balance out. In a binary system, H (information) = R (redundancy) = 0.5 when the probabilities are 0.89 / 0.11. It seems plausible that the grammar of a natural language should be constructed, in outline (i.e. in its most general, least delicate categories), of systems having just these two probability profiles; rather than, say, having all systems equiprobable, which would be too easily disrupted by noise, or having systems distributed across all probability profiles from 0.5 / 0.5 to 0.99 / 0.01, which would be practically impossible for a child to learn.

I have taken for granted, up to this point, that relative frequency in the corpus is the same thing as probability in the grammar. Let me now put this more precisely: frequency in the corpus is the *instantiation* (note, not realization) of probability in the grammar. But what in fact does this mean? [...] We are so accustomed to thinking about language and text in terms of dichotomies such as the Saussurean *langue* and *parole*, or Hjelmslev’s *system* and *process* that we tend to objectify the distinction: there is language as a system, an abstract potential, and there are spoken and written texts, which are instances of language in use. But the ‘system’ and the ‘instance’ are not two distinct phenomena. There is only one phenomenon here, the phenomenon of language: what we have are two different observers, looking at this phenomenon from different depths in time (1992: 66).

Halliday’s answer to Oliva and Doležalová’s statement is thus to cast doubt on the importance of distinguishing between *langue* and *parole*.

Halliday carefully restricts this statement to large-scale systems, which is disappointing for morphologists, given that our interest tends to focus on small-scale detailed systems. In

³ For further discussion of the issues in this article, see Clark 2005, Laury and Ono 2005, Meyer and Tao 2005, and Newmeyer 2005.

other works, though, he applies this idea to more specific categories at more detailed levels of language.

Halliday suggests that many frequency statistics do not reflect a simple oscillation between two possibilities in a system, but rather the intersection of two systems, each of which has terms that can be marked and unmarked, and which can favour one or the other term in the intersecting system. In morphological terms, we could see one system being, for example, the type of stem, the verb class, or semantic and etymological groupings, while the other system would consist of a pair of possible grammatical endings.

Halliday says that when these two systems interact in various ways, they produce a variety of frequencies of results. In the table below, adapted from a 1992 article, he proposes a series of possibilities in which 400 examples are distributed across two simultaneous systems. In chart (a), neither system has a marked or unmarked term, so the result is an equal distribution across the four possibilities. In chart (b), both systems have marked and unmarked terms; the other way this could fall out is as in chart (g). Chart (c) proposes that only one system has an unmarked term, and charts (d), (e) and (f) propose a kind of conditionality, in that there is some favouring of one or another term, but not a marked-unmarked opposition.

Table 1: Halliday's table expressed in percentage terms

(a) Both systems lack an unmarked term

System I	m	n	T
System II			
p	25%	25%	50%
q	25%	25%	50%
T	50%	50%	100%

(b) Both systems have one unmarked term

System I	m	n	T
System II			
p	81%	9%	90%
q	9%	1%	10%
T	90%	10%	100%

(c) System I has no unmarked term,
System II has one term unmarked

System I	m	n	T
System II			
p	45%	45%	90%
q	5%	5%	10%
T	50%	50%	100%

(d) Both have no unmarked term,
but *m* favours *p*, *n* favours *q*

System I	m	n	T
System II			
p	35%	15%	50%
q	15%	35%	50%
T	50%	50%	100%

(e) Both have no unmarked term,
but *m* favours *p*

System I	m	n	T
System II			
p	35%	25%	60%
q	15%	25%	40%
T	50%	50%	100%

(f) Both have no unmarked term,
but *p* favours *m*

System I	m	n	T
System II			
p	35%	15%	50%
q	25%	25%	50%
T	60%	40%	100%

(g) Both have 1 term unmarked,
but marking is reversed

System I	m	n	T
System II			
p	45%	5%	50%
q	5%	45%	50%
T	50%	50%	100%

2. Grammaticality and corpus frequency: a test case

Returning to the whole issue of *langue* and *parole*: Oliva and Doležalová view *langue* as a set of infinitely equal possibilities, with no item in the system privileged over any other. Privilege comes only in our analysis of *parole*, and it does not affect our description of *langue*. Halliday, on the other hand, seems to be arguing that built into linguistic structures is an innate bias towards or away from certain forms and constructions, and that we can't fully describe *langue* without recourse to information about its instantiation in *parole*.⁴

A convincing argument advanced in Oliva and Doležalová concerns the 168 examples of the locative plural form used erroneously for the genitive plural form that they found in the Czech National Corpus. While native speakers who are listening or reading closely would never accept locative plural forms where a genitive plural is expected, such clear errors are nonetheless found in the corpus, and not merely once or twice. This presents a problem for the codifier who attempts to use the corpus as a source of information on acceptability. Need we accept these examples as part of the grammar, or if you like, as part of *langue*?

Let us approach the question from a different angle. Locative plural/genitive plural confusion may be ungrammatical, but it is not unattested. What about the sort of mistakes that no native speaker would make? One such example is the substitution of dative plural forms where the genitive plural is expected. I searched in the Czech National Corpus for these sorts of constructions in the contexts where Oliva and Doležalová found the locative plural:

Table 2: In search of a genitive plural that looks like a dative plural

```
[word="kolem"] [word="jejich"] [word=".*m"]
[word="kolem"] [word="jejích"] [word=".*m"]
[word="kolem"] [word="jeho"] [word=".*m"]
[word="ze"] [word="všech"] [word=".*m"]
[word="ze"] [word="svých"] [word=".*m"]
[word="z"] [tag="C.*"] [word=".*m"]
[word="ze"] [tag="C.*"] [word=".*m"]
[word="mimo"] [word=".*m"]
[word="většině"] [word=".*m"]
```

Having run these and other searches and manually removed extraneous results, I came up with only one example:

zkušenostmi řadového myslivce . Tím jsou jeho práce nad jiné cennější , což <většině vědeckým> pracovníkům chybí . Autor má velké zkušenosti s chovem a výskytem

This is clearly a marginal construction, and here, rather than having a dative morph used to indicate genitivity, we have a case of mistaken governance, where *vědeckým pracovníkům* is governed by *chybí* instead of by *většině*. In contrast, the genitive and locative plural share adjectival morphs and some nominal morphs (e.g. *tř-ech*, substandard *čtyř-ech*), so perhaps the most accurate conclusion here is that some sorts of errors are not as erroneous as others. By not drawing precise boundaries between *langue* and *parole*, we allow ourselves to admit that some sorts of errors are more *langue*-like than others, just as some bits of *langue* are possibly less *langue*-like than others.

⁴ If we believe what Halliday is suggesting, then corpus data correctly interpreted do not *override* linguistic intuition so much as *reflect* it. If intuition and data clash, then there are two possibilities. On the one hand, our interpretation of the data could be in some way faulty or could fail to take account of all the features involved – in other words, it's not an accurate reflection of intuition. On the other hand, it could be that our interpretation of the data is correct, and instead we need to look again at our linguistic intuition, under the assumption that perhaps there is a difference between a native speaker's individual perceptions and those held or used more generally.

It is worth looking at some larger-scale morphological data to see if Halliday's suppositions bear fruit. It is beyond the scope of this paper to go deeply into the sort of possibilities that Halliday enumerates, but we can look at one overall contention. In Halliday's model, features are either opposed as *core vs. peripheral* (unmarked vs. marked), or *equivalent and competing* based on their frequency, and we can at least see whether these frequencies match those posited in grammars that are constructed on the basis of intuition.

Does the actual situation bear this out? We have already seen that corpus data on forms perceived as *probable* or *conceivable* errors differs from those perceived as *improbable* or *inconceivable* errors. Now we will look at two instances of morphological variation: the locative singular and the locative plural. The first displays variation that is not readily perceived as having a predictable semantic or stylistic function; both endings are found, and while there are distinctions that function at the level of individual words and smaller groups of words, no overall distinction holds. The second displays variation that tends to be perceived as stylistic. When one form is held to represent a neutral style or be used in neutral discourse, the other is perceived as bookish, formal, colloquial, or informal.

3. The locative singular

In the locative singular of masculine and neuter nouns, we have two possible endings, *-e* (sometimes written *-ě*) or *-u*. These tend to be described in grammars as alternatives. There is no attempt to privilege one ending over the other; instead, the grammars make observations about which ending is more likely with particular groups of nouns (see, for example, PMČ 253). If we wanted to talk in terms of intersecting systems, the way Halliday does, then the intersecting systems would consist of either semantic groupings of nouns, various formal structural groupings, such as linguistic origin or phonological structure, or possibly even certain syntactic constructions. Not all of these can be easily followed in the corpus, so I will focus on a few more general points.

Examining the overall distribution of the two possible noun endings in the locative singular (see table 3), we find that there are 1,886,941 locative singular forms with these two endings across all three genders. Of those, 60% have the *-e* ending, and 40% have the *-u* ending. So overall, this is somewhere between an equiprobable system (50:50) and a system that to some degree favours one or another variant (70:30). This sort of distribution would seem to favour intersecting systems, and is probably the case here.

If we leave aside the feminine and the masculine animate genders and limit our investigation to the two genders where there is variation between the two endings, the situation changes. Here there are just over 63% of forms with the *-u* ending, and 37% with the *-e* ending. Again, this broadly fits the parameters for slight favouring of one system over another. Within each of the two genders, there is some variation, but broadly speaking, this is the realm of Halliday's equiprobable systems, with the strong possibility that the equiprobable system under consideration is under the influence of some concurrent system, or that it is composed of balanced and competing skew systems. All this matches up with the description in the grammars.

Table 3: Locative singular: variation between *-e*, *-ě*, *-u*

All genders: hard declensions

1,906,410 tokens of *-e/-ě/-u*

774,060 tokens of *-u* (40.6%)

1,132,350 tokens of -e/-ě (59.4%)

fem. loc. sg.

[(word=".*[eě]") & (tag="NNFS6.*")] – 695,543 tokens

masc. anim. loc. sg.

[(word=".*u") & (tag="NNMS6.*")] – 19,469 tokens (tagging is v. problematic)

genders with competing variants -e/-ě and -u

1,191,398 tokens of -e/-ě/-u

754,591 tokens of -u (63.3%)

436,807 tokens of -e/-ě (36.7%)

masc. inan. loc. sg.

total: 937,112 tokens in 9198 attested forms

[(word=".*u") & (tag="NNIS6.*")]

618,535 tokens (66.0%) in 7826 lemmas (85.1% of AFs)

[(word=".*[eě]") & (tag="NNIS6.*")]

318,577 tokens (44.0%) in 1372 lemmas (14.9% of AFs)

neut. loc. sg.

total: 254,286 tokens and < 2041 lemmas (or exactly 2041 attested forms)

[(word=".*u") & (tag="NNNS6.*")]

136,056 tokens (53.5%) in 1748 lemmas (> 85.6% of AFs)

[(word=".*[eě]") & (tag="NNNS6.*")]

118,230 tokens (46.5%) in 293 lemmas (> 14.4% of AFs)

4. The locative plural

The distribution of endings in the locative plural is considerably different. Again we will begin with a bird's-eye view of the entire category and then focus in on one part of it in more detail.

In this part of the Czech declensional system, variation occurs in the masculine and neuter genders for nouns ending in a velar consonant. Examining intuitively structured grammars, we find that they treat the two genders differently. For instance, the PMČ gives the *-ich* ending as standard for the masculine gender, and says that it alternates with the ending *-ách*, which it calls 'colloquial to neutral' (249-251). For neuter nouns, by contrast, the *-ách* ending is given as standard, while *-ich* is termed 'bookish'.

Similar treatments occur in descriptions of spoken Czech; Sgall and Hronek call the *-ách* ending "generally widespread" for masculine nouns, and note that "this ending is moving into the realm of standard expression, both in stems ending in *čk* (*domečkách*), where for inanimate nouns we can consider it stylistically neutral (as several grammars note) as well as for those animate nouns that in and of themselves are stylistically marked as colloquial; for example *o klukách* is often found in standard printed texts (fiction, narration, etc.)." For neuter nouns, they note that "Common Czech has not only forms like *kolečkách* (where it is the only possibility, even if one is striving mightily to use standard Czech), but also *na střediskách*, *po pravítkách*, *v sedátkách*, which are of class (1)(a). In the standard language the forms with softening (*střediscích* etc.) are understood as variant (i.e. not the only ones possible)" (1992: 39–41).⁵

In an overview of the entire locative plural, we find three endings that are distributed across the various declension classes. Since we will not be particularly interested in the *-ech* ending – few nouns show alternation between it and another ending – we will leave it aside for

⁵ Class (1)(a) refers to a classification system meaning: "Forms widely used in ordinary conversation across the entire territory where Czech is spoken; forms often used in conversation which is overall evaluated as being standard in character" (Sgall and Hronek 1992: 28).

now and focus on the other two. Once again, what we find here is an equiprobable system, with a rough balance between the number of tokens whose loc. pl. ends in *-ách* and those whose loc. pl. ends in *-ích*.

Table 5: Locative plural overall

619,940 tokens total, 368511 tokens of *-ích/-ách*
 [(word=".*ích") & (tag="NN.P6.*")]
 192,341 tokens (31.0%)
 [(word=".*ách") & (tag="NN.P6.*")]
 176,170 tokens (28.4%)
 [(word=".*ech") & (tag="NN.P6.*")]
 251,429 tokens (40.6%)

The same will hold if we focus in further on stems of all genders that end with a velar consonant, which is where the variation occurs. Here the corpus acknowledges two of the three possible endings, with almost 48% of tokens having the *-ích* ending and just over 52% having the *-ách* ending.⁶ Again, this is either an equiprobable system or a skew system with complementary marking.

Table 6: All locative plural nouns with stems ending in a velar

93,518 tokens total
 [(word=".*ích") & (tag="NN.P6.*") & (lemma=".*[kgh][ao]?")]
 44,668 tokens (47.8%)
 [(word=".*ách") & (tag="NN.P6.*") & (lemma=".*[kgh][ao]?")]
 48,850 tokens (52.2%)

There are three categories in table 6 of interest: masculine animate, masculine inanimate, and neuter. When we compare and evaluate the data, we find a strong indication that the *-ách* ending is in general peripheral, with *-ích* being central. The ratio is 19:1 in favour of the *-ích* ending.

Table 7: All masculine and neuter locative plural nouns with velar stem⁷

42,645 tokens total
 [(word=".*ích") & (tag="NN[IMN]P6.*") & (lemma=".*[kgh]o?")]
 40,268 tokens (94.4%)
 [(word=".*ách") & (tag="NN[IMN]P6.*") & (lemma=".*[kgh]o?")]
 2377 tokens (5.6%)

masculine animate locative plural with velar stem

2732 tokens total
 [(word=".*ích") & (tag="NNMP6.*") & (lemma=".*[kgh]")]
 2609 tokens
 [(word=".*ách") & (tag="NNMP6.*") & (lemma=".*[kgh]")]
 123 tokens

⁶ For consistency's sake I also searched on the ending *-ech*, and came up with a similarly large class of tokens (41,705). These, however, were all accounted for by tokens of *letech*, *lidech*. The corpus query processor retrieved them because I had done a search partially based on lemmas, and the forms *léta*, *lidé* are lemmatized according to the suppletive sg. forms *rok*, *člověk*.

⁷ These numbers do not exactly match those from individual categories (which would give a total 42,658), because they are based on different searches, and the masculine inanimate searches involved some manual sifting of data to remove ballast.

Now let us consider two of the three specific subgroups I just mentioned: the masculine inanimate and the neuter.

If we look at the masculine inanimate with velar stem, the *-ích* ending seems clearly central, with *-ách* being well below the frequency for peripheral or marked items. The frequency of the peripheral item in the corpus is just under 3%, and that value drops further once we exclude mistakenly tagged place names like *Jeseníky*.

We can use the subcorpus function of the corpus query processor to drill down even further and look at specific text types. My subcorpus was structured to capture all of what could be called ‘creative writing’, and included the following text types:

Table 8: Structure of subcorpus SYN2000:lit

(txtype="NOV" | txtype="COL" | txtype="SCR" | txtype="VER" | txtype="SON" | txtype="IMA")
 NOV - novel
 COL – collection of short stories or individual short story
 SCR – dramatic text or screenplay
 VER - poetry
 SON – song lyrics
 IMA – general notation for creative work, inc. some non-fiction

When we look at the data from the subcorpus, a different picture emerges. We shall take the data excluding proper names as our basis. For these, we notice that the creative subcorpus has a much higher percentage of forms with the peripheral ending. The peripheral ending thus has a more solid position, occupying 1 out of every 7 instances of the locative plural.

Table 9: Masculine inanimate locative plural nouns with velar stem

-including proper names

total: 37,899 tokens in 1211 attested forms

[(word=".*ích") & (tag="NNIP6.*") & (lemma=".*[kgh]")]
 36,836 (97.2%) in 968 lemmas (79.9% of all AFs)
in creative subcorpus 3262 (8.6% of total)

[(word=".*ách") & (tag="NNIP6.*")]
 1063 (2.8%) in 243 lemmas (20.1% of all AFs)
in creative subcorpus 529 (49.8% of total)

-excluding proper names

total: 37,811 tokens

in creative subcorpus 3713

[(word="[aábcčdeéfgghijklmnoóppqrřsstt'úúvwxyzž].*ích") & (tag="NNIP6.*") & (lemma=".*[kgh]")]
 36,832 tokens, or 97.4%
*in creative subcorpus 3227, or 86.9%:
 not quite 1/10 of the forms in -ích are in 1/6 of the corpus*

[(word="[aábcčdeéfgghijklmnoóppqrřsstt'úúvwxyzž].*ách") & (tag="NNIP6.*")]
 979 tokens, or 2.6%
*in creative subcorpus 486, or 13.1%:
 half of the forms in -ách are in 15% of the corpus*

Comparing the creative subcorpus with the entire SYN2000, we find a total of 3713 forms. 3227 (86.9%) have the standard form and 486 (13.1%) have the non-standard form.

If we look at the neuter plural, we find a different distribution, as expected according to the grammars. Here the centrality is clearly reversed, with the predominant ending being *-ách* for this group and the minority ending *-ích*. The rather large number of tokens for *-ích* is

accounted for by a very small number of lemmas. In fact, it works out exactly as posited in the PMČ, that the vast majority of forms – 432 out of 647 – are accounted for by one word, *středisko*.

Table 10: Neuter plural nouns with a velar stem in the locative plural

neut. loc. pl. with velar stem

total: 2027 tokens in 150 attested forms

[(word=".*[šzč]ích") & (lemma=".*[ghk]o") & (tag="NNNP6.*")] – 647 (31.9%) in 10 lemmas (> 6.7%)

[(word=".*[gkh]ách") & (tag="NNNP6.*")] – 1380 (68.1%) in 140 lemmas (> 93.3%)

5. Conclusions

In closing, we will return to the questions I asked at the beginning:

- *What is the value of corpus data in making linguistic judgments?*
- *To what extent should we be guided by corpus data in evaluating grammaticality?*
- *Is there a way to draw a line between grammaticality and ungrammaticality based on corpus data?*

To answer these briefly:

Corpus data can, in bulk, be trusted, if we do not treat them categorically or try to follow them blindly. They always have to be placed in context. Raw numbers as a guide are balanced by other information, including the ARF and distribution in the corpus. Failure to give such information does tend to undermine the validity of the conclusions derived from such data.

The statistics that the corpus query processor provides can help us determine grammaticality, and there are rough guidelines that we can apply. Halliday's rule of thumb seems to function reliably for certain aspects of morphology. In the choice of forms between cases, it gives a rough match between standard descriptions of variation and corpus data. What it does not tell us offhand is whether we are dealing with simple systems or with interacting systems, and that is its major shortcoming as a predictive device.

If we accept the various limitations of its searching functions and the reliability of its tagging, the corpus provides relatively accurate data on usage, but whether that corresponds to grammaticality is an open question. Even Halliday's heuristic does not provide an index of grammaticality, but rather of the popularity of particular choices within a system: a way, if you like, of pinpointing core and periphery of a system, but not of pinpointing what lies completely outside a system.

It seems likely, however, that despite these shortcomings, a corpus can be more than an 'inspiration and corrective', as Oliva and Doležalová propose. Despite the fact that the corpus is a series of instantiations of a system, rather than a description of a system, it nonetheless offers us ample clues about the centrality and peripherality of various constructions, and can be of use in deciding degrees of marginality and peripherality of certain features.

References

- Clark, B. (2005): On stochastic grammar. *Language* 81 (1), 207–217
- Český národní korpus – SYN2000. Ústav Českého národního korpusu FF UK, Praha 2000. <http://ucnk.ff.cuni.cz>
- Halliday, M. A. K. (1991a): Corpus studies and probabilistic grammar. In: Aijmer, K. and Altenberg, B. (ed.), *English Corpus Linguistics*. New York and London: Longman, 30–43
- Halliday, M. A. K. (1991b): Towards probabilistic interpretations. In: Ventola, Eija (ed.), *Functional and Systemic Linguistics: Approaches and Uses*. Berlin and New York: Mouton de Gruyter, 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 and New York: Mouton de Gruyter, 61–77
- Hellberg, S. (1992): Using corpus data in the Swedish Academy Grammar. In: Svartvig, J. (ed.), *Directions in Corpus Linguistics*. Berlin and New York: Mouton de Gruyter, 311–334
- Laury, R. & Tsuyoshi O. (2005): Data is data and model is model: You don't discard the data that doesn't fit your model! *Language* 81 (1), 218–225
- Meyer, Ch. & Hongyin T. (2005): Response to Newmeyer's 'Grammar is grammar and usage is usage'. *Language* 81 (1), 226–228
- Newmeyer, F. (2003): Grammar is grammar and usage is usage. *Language* 79 (4), 682–707
- Newmeyer, F. (2005): A reply to the critiques of 'Grammar is grammar and usage is usage'. *Language* 81 (1), 229–236
- 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, 7–10
- Sgall, P. & Hronek, J. (1992): *Čeština bez příkras*. Praha: H&H
- Ústav českého jazyka FF MU (1995): *Příruční mluvnice češtiny*. Praha: Lidové noviny