---

**Article:**
Bermel, N. orcid.org/0000-0002-1663-9322, Knittl, L., Alldrick, M. et al. (1 more author) (2024) Ideal and real paradigms: language users, reference works and corpora. Cognitive Linguistics, 35 (2). pp. 177-219. ISSN 0936-5907

https://doi.org/10.1515/cog-2023-0032

---

Neil Bermel*, Luděk Knittl, Martin Alldrick and Alexandre Nikolaev

# Ideal and real paradigms: language users, reference works and corpora

https://doi.org/10.1515/cog-2023-0032
Received February 26, 2023; accepted January 18, 2024; published online March 7, 2024

**Abstract:** This article approaches defective and overabundant paradigm cells as an opportunity and pitfall for usage-based linguistics. Through reference to two production tasks involving native speakers of Czech, we show how definitions of these two categories are problematized when multiple forms per context are entrenched, or when pre-emption seems to occur in the absence of entrenchment: in other words, pre-emption occurs via entrenchment of uncertainty. We explain the results by adopting a broader, usage-based perspective. We examine the relationship between frequency (as proxy for exposure) and reference-work information (as proxy for *a priori* structure) to assess their connection with our experimental results. We assign a role to frequency as helping to form perceptions of "suitable" and "unsuitable" forms, but also note places where non-frequency factors predominate. "Structure" as represented by reference-work recommendations appears to have no significant connection to our experimental results; we discuss reasons for this.

**Keywords:** morphology; inflection; Czech; prescriptivism; frequency; entrenchment

## 1 Introduction

Cases of what have come to be known as paradigmatic 'defectivity' (or 'defectiveness') and 'overabundance' (as per Thornton 2012) present an opportunity and a challenge to usage-based approaches. On the one hand, cells with variant forms (*overabundance:* 'I have *striven/strived* for justice') and paradigmatic gaps (*defectivity:* 'I have *strid … ?* across the road') can, in principle, be handled in emergent language models, if those models assume speakers share a set of common learning biases that may yield small differences in learning outcomes in response to

***Corresponding author: Neil Bermel**, University of Sheffield, Sheffield, UK,
E-mail: n.bermel@sheffield.ac.uk. https://orcid.org/0000-0002-1663-9322
**Luděk Knittl,** University of Sheffield, Sheffield, UK
**Martin Alldrick,** University of Surrey, Guildford, UK
**Alexandre Nikolaev,** University of Eastern Finland, Joensuu, Finland. https://orcid.org/0000-0001-8634-5947

variation in the primary linguistic data. In nativist models that deny a significant role to learning, it is less clear what factors would motivate an excess or shortfall of forms. On the other hand, approaches that aim to provide cognitively plausible models of actual patterns of usage must draw on a wider range of sources than corpora or the idealized descriptions in prescriptive or descriptive reference works. In this paper, we address this challenge by constructing an experiment that investigates the status of 'defectivity' and 'overabundance' in inflectional paradigms in Czech.

We begin in Section 2 with working definitions of defectivity and overabundance, and how the problems they pose for theories of language use have been tackled. In Section 3, we describe how we located and described defective and overabundant slots in Czech nominal and verbal paradigms to construct a linguistic experiment and consider the consequences of our approach. Section 4 examines visualizations of our results and reports findings problematizing our initial definitions. Finally, in Section 5, we return to our sources for categorising lexemes as defective or overabundant and assess their relationship to the native-speaker data gathered. Our findings suggest some tweaks to traditional Cognitive Linguistics (CL) concepts, such as acknowledging the possibility of *multiple entrenched forms* and *entrenchment of uncertainty* to explain these two phenomena, through the lens of what we call *opportunistic suppletion*.

## 2 Defectivity and overabundance in theory and practice

In any theory of language that adopts some version of the 'one meaning-one form' principle, defective and overabundant paradigm cells are unexpected.

Within the generative tradition, the treatment of defectivity and overabundance will tend to reflect more general assumptions about the factors that constrain the output of a grammar. Broadly speaking, syntactic models will have greater success with overabundance than defectivity and contemporary phonological models will handle overabundance more easily than defectivity. The types of 'filtering' mechanisms that have been incorporated in syntactic models since Chomsky and Lasnik (1977) can be adapted to block defective forms, though this analysis largely relocates the problem to one of motivating the blocking (but see Yang 2016). Overabundance can be resolved by invoking a paradigmatic version of the Principle of Contrast (Clark 1987: 2): on such an account, variant forms are assumed to vary along some nuanced linguistic or sociolinguistic dimension, or are disregarded as ephemeral artifacts of historical change. Against that, there is evidence (Bermel and Knittl 2012;

Bermel and Knittl 2023; Nichols and Timberlake 1991) that much variation is stable and not clearly attributable to specific features. Within generative frameworks, Optimality accounts (OT, Prince and Smolensky 1993) can treat overabundance as cases in which multiple candidates are equally optimal. Defectivity, on the other hand, presents a special case of 'ineffability', which advocates of Rule-Based-Phonology have argued challenges the fundamental assumptions of OT models (Vaux 2008).

At some level, cognitive approaches face the complementary challenge of accounting for the acquisition of relatively uniform and densely-populated paradigms from the sparse and biased input that learners encounter (Blevins et al. 2017; Janda and Tyers 2021). To account for defectivity, the cognitive mechanisms proposed for morphological acquisition must be general enough to extrapolate from partial exposure to the morphological forms of a language; they must also distinguish *inherent* gaps (as per Chuang et al. [2022], i.e., always unused or underused due to some sort of avoidance of them) from the far more numerous forms that will be *contingently* absent from the language sample that any speaker encounters (but can be readily produced by native speakers when needed). Overabundance presents less of a synchronic challenge: alternate forms may exhibit at least some distributional variation that speakers will, again in accordance with a general Principle of Contrast, associate with differences in meaning or communicative function, or frequency itself (or some proxy for it, see Nikolaev and Bermel [2022]) can be invoked as a factor that differentiates two variants.

Data from reference works have traditionally provided the most detailed descriptions of defectivity and overabundance, based on philological analyses of authoritative (usually written) sources. These materials provide a useful point of departure for the investigation of defectivity and overabundance, but, as in many domains of linguistics, they incorporate idealizations that are largely unconcerned with cognitive plausibility and may not be fully representative of actual patterns of usage. One example of suppressed overabundance is the Russian word *dogovor* 'agreement'. According to language manuals and dictionaries, the only sanctioned forms stress the third syllable, whereas in colloquial Russian forms with initial stress have long co-existed with it. Eventually, some contemporary grammars and dictionaries started acknowledging this competing ("wrong") form. The intrinsic correctness of historically authoritative sources and the intrinsic problematicality of non-sanctioned usage (diverging from past patterns) could be explained by a faithfulness constraint: language community members should avoid innovative language use, as it distorts the message and compromises its communicative efficacy. This faithfulness constraint is, therefore, a cornerstone for prescriptive linguistics and often a target of critique in descriptive linguistics.

For corpus data, the challenge of *contingent* versus *inherent* defectivity is represented by *corpus lacunae* (Kováříková et al. 2020): on purely numerical terms, the Czech verb *zavraždit* 'murder' lacks a feminine past tense form, as corpus examples overwhelmingly attest *zavraždil* 'murdered-MASC.SG', while *zavraždila* 'murdered-FEM.SG' is nearly absent, and yet native speakers readily produce the latter with no difficulty or hesitation. The lack of a feminine form is unlikely to stem from homophony avoidance or from language changes in diachrony, two reasons often advanced for inherent defectivity, but rather from the fact that disproportionally more males are convicted of murder than females. However, *corpus lacunae* as such do not, without further analysis, reveal whether a paradigm is defective for some language-internal or language-external reasons (if we can in fact reasonably make such a distinction). The challenge of distinguishing inherent defectivity from contingency is also discussed using Finnish corpus data in Nikolaev and Bermel (2023). Cases of overabundance present particular challenges for corpus analyses. The contexts that influence the choice between variants may be non-local, spanning varied discourses or genres. Variation may also be conditioned by a range of social factors that cannot readily be extracted from corpora.

# 3 Methodology

We set out to identify lexemes in Czech that have inherently defective cells, which we then wished to compare to lexemes with overabundant cells, as well as to those that are exclusively composed of biunique cells – in other words, are nonvariant throughout. We then tested native speakers' reactions to these different "conditions" as established by our data, and finally circled back to check whether the native-speaker data shed light on the descriptions originally gleaned from corpora and reference works.

## 3.1 Reference-work data

Reference works, whether a grammar manual or a dictionary, have traditionally made a virtue of economy due to the restrictions of printed formats, resulting in a sort of generative-lite approach in which we describe exceptions that "block" the implementation of higher-level rules before they can be applied. In a dictionary entry in a Slavonic language, we expect a citation form, noun gender where applicable, and possibly one other form indicating the declension or conjugation pattern: the gen. sg. of nouns or a non-past tense form of a verb. Other forms are adduced if they represent deviations from the "ideal" paradigm, i.e., "blocking" the expected form we would

otherwise generate.[1] Descriptions in grammars follow similar assumptions: a rule or table precedes lists of exceptions and the overall categories or individual lexemes to which they apply. We can compare this tradition to that of Finnish, a language with a more complex inflectional morphology. In Finnish, the two largest dictionaries (Nykysuomen sanakirja [Dictionary of Modern Finnish] 1951–1961, and Suomen kielen perussanakirja 1990–1994) use a number referring after each lexeme to a table in the appendix with an example paradigm. Nykysuomen sanakirja (1951–1961) has 82 such paradigms (and hence 82 inflectional types/classes) for nouns. However, *Suomen perussanakirja* (1990–1994) has 49 paradigms (inflectional types/classes) for nouns. This does not mean that in a generation Finnish has lost 33 inflectional types; rather, it means that the criteria for assigning inflectional types have changed.[2]

Without the practical constraints of print, and with the goal of providing a quick answer for individual lexemes rather than a set of easily assimilable rules, online resources can present the full paradigm for any given lexeme. The Internet Language Reference Book (ILRB) of the Institute for the Czech Language Jazyková poradna Ústavu pro jazyk český (2008–2024) has over 112,000 entries drawn from a variety of contemporary sources published by the Institute and lists full paradigms; we took this as the basis for our further investigations.

How in fact did these reference works identify the forms displayed? Tradition and inherited description play a role. Earlier works were underpinned by copious excerption from cultural landmarks; they were prone to recommend forms sanctioned by "good authors" (Ertl 1929) regardless of current status. However, the progressive program of the Prague School functionalists bolstered the inclusion of more commonly used forms (Bermel 2007: 136–142). The status of standard Czech as a "superdialectal" standard, with numerous morphological peculiarities absent from spoken varieties, enabled this approach. Czech codificatory works consequently contain many doublet and triplet forms, as is common in many of Europe's medium-sized languages (cf. Estonian, Croatian, Norwegian and others): some of these forms may have fallen out of common use or be posited for the sake of systemic consistency, rather than being evidenced in the contemporary language. These reference works can thus overestimate the presence of overabundance in a language.

**1** The size of these reference works would have made full paradigmatic listings impractical. The largest Czech dictionaries, *Slovník spisovného jazyka českého [Dictionary of the Czech Literary Language,* hereafter DCLL] and *Příruční slovník jazyka českého [Reference Dictionary of the Czech Language]* were published in eight and nine volumes respectively and contain 197,200 and 250,000 entries, plus derived words listed under those top-level entries.

**2** Similar approaches are rare in the Czech tradition: one is the set of bilingual Czech-Russian dictionaries produced in the 1970s. These list, for example, 261 noun declension types for Russian, with some of those also having subtypes and exceptions (Kopecký and Leška, 1978: II, 563–632). We are grateful to Laura Janda for pointing us to this source. A more recent attempt to describe the number of paradigms based purely on identical form sets can be found in Strossa (2015).

Defectivity is harder to identify in Czech reference works. Functionalist language planning rarely identified gaps as a phenomenon worthy of attention, preferring instead to fill them with projected forms. Occasionally a source notes that a form is "rarely used", but sources do not always agree on the specifics. In languages such as French, Spanish and Russian, there is a tradition of marking certain tenses or persons as unused for particular verbs (and in Russian for certain nouns),[3] but this tradition seems absent in Czech. We thus did not look for widespread agreement as to defectivity, but instead took isolated mentions as indicative.

## 3.2 Corpus data

Corpus data offer insight into the variability of linguistic forms in a cell. For Czech, we employed a balanced 100m-token corpus of standard Czech: SYN2015 (Křen et al. 2015), filtered through the GramatiKat tool (Kováříková and Kovářík 2021). This proved sufficient for our proposed nonvariant and overabundant lexemes but was insufficient in most instances for defective lexemes, as more data are required to suggest the absence of a form (for some of the issues with this, see Bermel et al. [2023]). We thus also turned to the csTenTen 10bn-token corpus (csTenTen 2017). Many of the proposed defective lexemes were colloquial and thus appear infrequently in written corpora, but oral corpora of Czech, while comparatively large (6m tokens), are not large enough to contain granular data on case and tense forms of mid-frequency lexemes.

To get around the problem of equating contingent defectivity with inherent defectivity, we looked not only at a form's presence or absence in a corpus, but also how its frequency compared to those of other forms of the same lexeme. A starting point was to ascertain how common the paradigmatic cell(s) in question were in the language, and then compare our suspected defective cells to them.

In corpora, we benchmarked the distribution of forms in medium-to-high frequency lexemes. We drew 200 lexemes (100 verbs and 100 nouns) from the Frequency Dictionary of Czech Čermák and Křen (2011) and calculated the frequency of each case/number combination. For nouns, the gen. pl. forms that constitute most of our examples make up 11.2 % of attested tokens for a fem. or neut. noun (median 9.2 %, STD 9.1). This is similar to proportions cited in GramatiKat, which displays quartile ranges for nouns with $N > 100$. For verbs, our examples are from the nonpast tense and passive participles. Non-past forms make up 42.9 % of attested tokens in our verbal sample (median 41.8 %, STD 26.1), and passive participles make up 6.8 %

**3** Nominal inflectional morphology in French and Spanish is not complex enough for this to be an issue.

of attested tokens for transitive verbs in our sample (median 1.9 %, STD 10.9). The gen. pl. is thus a relatively frequent number/case slot, third in frequency after the nom. sg. and acc. sg. Non-past verb forms make up a significant proportion of the total; only passive participles are relatively infrequent.[4]

## 3.3 Intersections between corpus data and reference-work data

Corpus and reference-work data are less cleanly separated than we have proposed here. Corpora of Czech draw their tagging and lemmatization in part from the Czech tradition of grammatical description, and thus some lexemes and forms not appearing in standard reference works are either incorrectly lemmatized to other lemmas and tagged incorrectly or returned as "unrecognized", meaning only manual searches, form by form, will retrieve them. Newer reference works like the *Mluvnice současné češtiny [Grammar of Contemporary Czech,* Cvrček et al. 2010*]* and the *Internetová jazyková příručka [Internet Language Reference Book,* hereafter ILRB] are respectively based on corpus data, and corrected and supplemented by the interpretation of corpus data, meaning that evidence of frequency has made its way into normative manuals.

## 3.4 Identifying defective, overabundant and nonvariant cells

As noted above, Czech has considerable reported and attested overabundance (Guzmán Naranjo and Bonami 2021); the status of defective slots appears to have attracted less attention in Czech and we are not aware of any prescriptive or empirical descriptions of this phenomenon. We thus began by identifying defective cells.

For defective items to appear in our survey, a cell had to be described as defective in a core reference work or meet criteria for non-appearance in a representative corpus, while lacking a semantic motivation for this absence.[5] We describe

**4** We take it as read that each lexeme's distribution of forms is unique, and thus "average" distributions are theoretical constructs: a noun might have a lower-than-average percentage of gen. pl. forms simply because it is overwhelmingly used in the loc. sg. as part of a fixed expression; nothing can thus be read into such deviations *per se*. However, such underrepresented or absent cells are worthy of investigation to determine whether their low frequency is a result of *contingence* (as in the above scenario) or *inherence*.

**5** As examples: our preliminary investigations led us to be interested in missing gen. pl. forms, and thus we removed nouns representing non-pluralizable abstract concepts; and because passive participles form another group of defectives, we thus removed intransitive verbs, including unergatives, which, in Czech, neither form the basis for subjectless passives (as in German) nor for

our methods for identifying cell types in greater detail in Bermel et al. (2023): to summarize briefly, cells identified as potentially defective in Czech came from lexemes with 2+ stem shapes, and occurred in places in the paradigm where the choice of stem is not automatic. For nouns, this turned out to be the gen. pl. of fem. and neut. nouns with stem-final consonant clusters; and the oblique cases of masc. nouns with stem-final consonant clusters. For verbs it was the non-past tense or the passive participle in certain smaller verb classes.

With overabundance, we had a much broader choice of material both in our core reference works and in corpora, so we selected items paralleling the criteria and contexts for defectives. In other words, we also chose gen. pl. forms of fem. and neut. nouns where there were two attested or recommended forms; verbs where there were two attested or recommended non-past tense stems; and two attested or recommended passive participles. To qualify as a case of 'overabundance', a cell had to be identified as such by at least one standard reference and also display variation in a representative corpus that could not be attributed to polysemy or local contextual factors.[6]

Our biunique or nonvariant cells were chosen almost entirely from those same criteria and contexts, but as opposed to the defective and overabundant cells, showed minimal or no variation in representative corpora and in reference works. In keeping our three conditions parallel, we hoped to minimize the possibilities of covarying factors that might come into play if, for example, only the defective group had multiple stem shapes, or if one condition used a more frequently-encountered case/number combination than another.

Where corpus data and reference works differed, we prioritized corpus data, on the grounds that reference works' descriptions are often impressionistic, reflecting the intuitions of specialist authors rather than generalizing over samples drawn from a speech community. We will return to this decision in Section 5 to ask whether that decision was a valid one.

## 3.5 Survey design

We designed two surveys for Czech native speakers to examine how reactions to defective and overabundant cells differed from reactions to nonvariant paradigm cells. Due to the pandemic, our surveys were delivered online and in written format,

impersonals (as in Estonian). We wanted speakers to engage with possible, plausible contexts in which no form was adequate.

**6** An example of differentiation by polysemy would be the lexeme *západ* 'turn, west', which has two distinct loc. sg. forms supposedly determined by sense: *o západu klíče* 'about the turn-LOC.SG of the key' versus *na západě Ameriky* 'in the west-LOC.SG of America').

using standard written Czech.[7] Each survey tested equal numbers of defective, overabundant and nonvariant paradigm cells.

Respondents filled in a survey constructed in Gorilla with either nominal or verbal forms in it. The nominal survey had 51 items; the verbal survey had 33 items. More participants were therefore recruited for the verbal survey ($N = 84$) than for the nominal survey ($N = 60$). After the introductory material, the order of the experimental slides was randomized.

After viewing an example slide, participants saw sentences where a word form in a common and uncontentious slot (the trigger) appeared in bold, followed by a second sentence where they were to insert a missing form of that same word (the target). The target context pointed unambiguously to a single, potentially problematic cell in the paradigm (see example (1)).

(1) Dal bych si jedno **pivčo.** Nebo uvidíme, ale spíš těch [… …] bude víc.
'I'll have one **beer-**DIM-ACC.SG**.** Or we'll see, there might be more [ … .]GEN.PL than that.'[8]

Respondents typed their answer and saved it by clicking forward. They could skip over the question by clicking forward without answering. We recorded the answer produced (or lack thereof, marked as NA), the length of time taken to start typing an answer and to complete an answer.

Survey results were analyzed in R (R Core Team 2021) and utilized both qualitative and quantitative methods. The quantitative results are reported in (Bermel et al. 2023). The present study reports the qualitative findings. These are derived from visualizations of the data produced using the package *igraph* (Csardi and Nepusz 2006). The corpus and Internet Language Reference Book predictors are analysed using the function *estimateNetwork* from the package *bootnet* (Epskamp et al. 2018).

# 4 Visualizing overabundance and defectivity

Our data visualizations take the form of plots centred around the "trigger" form (a nom. or acc. noun form using its citation-form stem; or a past-tense verb form using

7 The written medium led us to use standard Czech (spisovná čeština), a code with significant morphological divergences from commonly spoken Czech dialects: its hallmark forms can sound "affected" in speech. This quasi-diglossic dichotomy is treated extensively elsewhere (see *inter alia* Sgall et al. 1992) and is unavoidable for Czech. Our survey deliberately omits features where the standard/nonstandard dichotomy in Czech is highly salient. Ethical approval was received from the University of Sheffield.

8 The complement of *víc* 'more' appears in the genitive case. A singular reading is improbable, but a plural demonstrative *těch* 'these' is nonetheless inserted to guide the respondent towards a gen. pl. form.

the primary infinitive/past stem): this is the form in sentence 1 of each screen. Clustered nearby and connected by arrows are the forms our respondents produced. Forms closer to the trigger were produced more frequently; distant forms were produced only occasionally. All gaps (no answer, NA) are represented by an NA response circle. The arrangement of individual plots within the graphic is not meaningful. The red text next to each cluster indicates the citation form of the noun or verb.

Each form has a colour based on the type of modification that our respondents made, as presented in Figure 1. This classification was developed exploratorily after the experiment in response to the answers collected. Pink represents the closest available stem to the trigger form (fewest changes required). Orange represents the next closest available stem.[9] The remaining colours represent other modifications found:

- Blue-green: novel stem changes (*ostropeřc-ů* 'milk thistle-GEN.PL' as a novel shortening of the stem *ostropestřc-)* or the adoption of an ending novel for the class (e.g., *jařm-ů* 'yoke-GEN.PL' with an ending borrowed from the masc. class);
- Bright green: modification in person, case or tense (*rolb-ách* 'snowcat-LOC.PL' for an expected gen. pl. form);
- Light green: a modification in number or aspect (e.g., *vyvléká* 'weasels out of-IMPF' for an expected perfective verb form)
- Light blue: substitution of a (near-)synonym or (near-)synonymic phrase (e.g., *zrnek* 'grain-DIM-GEN.PL' for an expected form of the non-diminutive *zrno* 'grain').

Effectively, the pink and orange forms almost always represent licensed or near-licensed choices, while the remaining colours represent more profound deviations from the task. The variety of results described below is reminiscent of the demonstration in Dąbrowska (2018) that no two speakers' grammars are identical.[10]

A summary of these Figures (2, 3, 7, 8, 12, 13) can be found in Tables 7 and 8 later in this article.

## 4.1 Visualizing nonvariant paradigm cells

For nonvariant (biunique) paradigm cells, we expected respondents to converge on a single answer. In a few instances, primarily participle forms, there was

**9** Sometimes the morphologically closest stem is not the phonotactically most likely one, and even when the closest stem is plausible, the second-closest may in fact be the licensed form. It was rare but not unheard of for a respondent to produce a form with serious phonotactic violations.

**10** For online readers, larger versions of the graph can be accessed by clicking on the form or by looking at the supplementary materials.

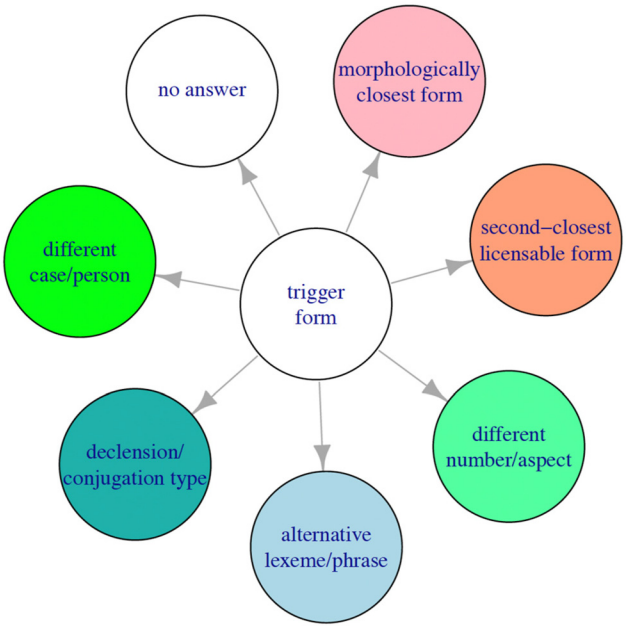


**Figure 1:** Legend for visualizations.

no way to rule out doublets entirely and so there were two or more forms available.[11]

For our 17 nominal triggers seen in Figure 2, we expected up to 17 forms to be produced. For our 11 verbal triggers seen in Figure 3, we expected up to 24 forms to be produced. In the event, as Figures 2 and 3 show (online readers can zoom in to see nodes and edges clearly), our respondents used substantially more forms in both tests.[12] A total of 59 noun forms (including 9 null responses) and 44 verb forms (including 11 null responses) were produced. For nouns, 24 represented *potentially licensable* forms (i.e., forms that appear well-formed by the conventions of Czech, even if they are not licensed in this particular instance or acceptable to most native speakers); in four instances, respondents manipulated the construction to produce a

**11** A participle in Czech takes the form of a short or long adjective (*zvládnut – zvládnutý* 'MASTERED-MASC.NOM|ACC.SG' from *zvládnout* 'to master-INF'), and the long adjectives can have additional colloquial forms (*zvládnutej*). In the contexts given, the short form was most appropriate, but we accepted long forms and colloquial versions of the long forms as licensed.

**12** In these figures, some arrows loop back on themselves; this is an indication that a respondent produced the trigger form as the target. There are occasional arrows connecting unrelated clusters; these must have been the result of people reusing material from a previous answer.

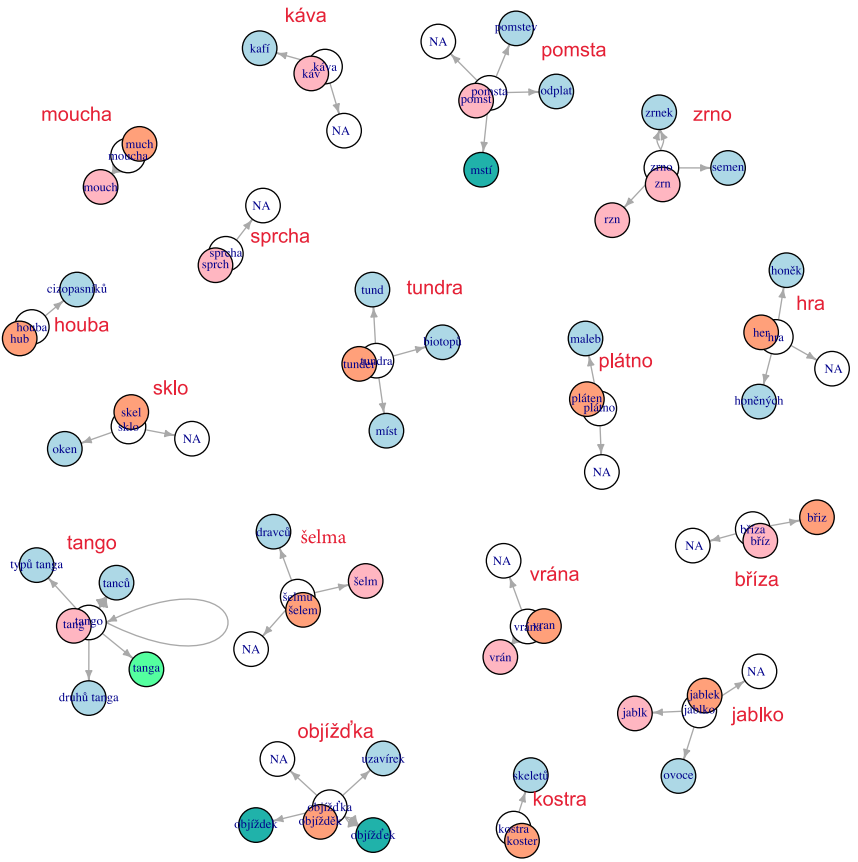


**Figure 2:** Noun forms produced in nonvariant cells.

different case or number form, and in 20 instances, a form of a different lexeme or phrase had been produced. For verbs, many of the available options were not used, but the responses still included 22 potentially licensable forms, three grammar manipulations and eight synonymic or phrasal substitutions.

This is more variation than anticipated with nonvariant cells. As can be seen in Figure 4, some nonvariant cells behave as expected, having only one target response (although sometimes there were also missed answers): the trigger *bál jsem se* 'I was afraid-MASC.SG', in (3) resulted in *bojím se* 'I am afraid-MASC.SG' with 82 responses and two skipped answers. (Response numbers are shown on the arrow leading from trigger to response).

However, there were two common types of deviations from this pattern. In the first, shown in Figures (5a) and (b), a different lexeme substitutes for the trigger

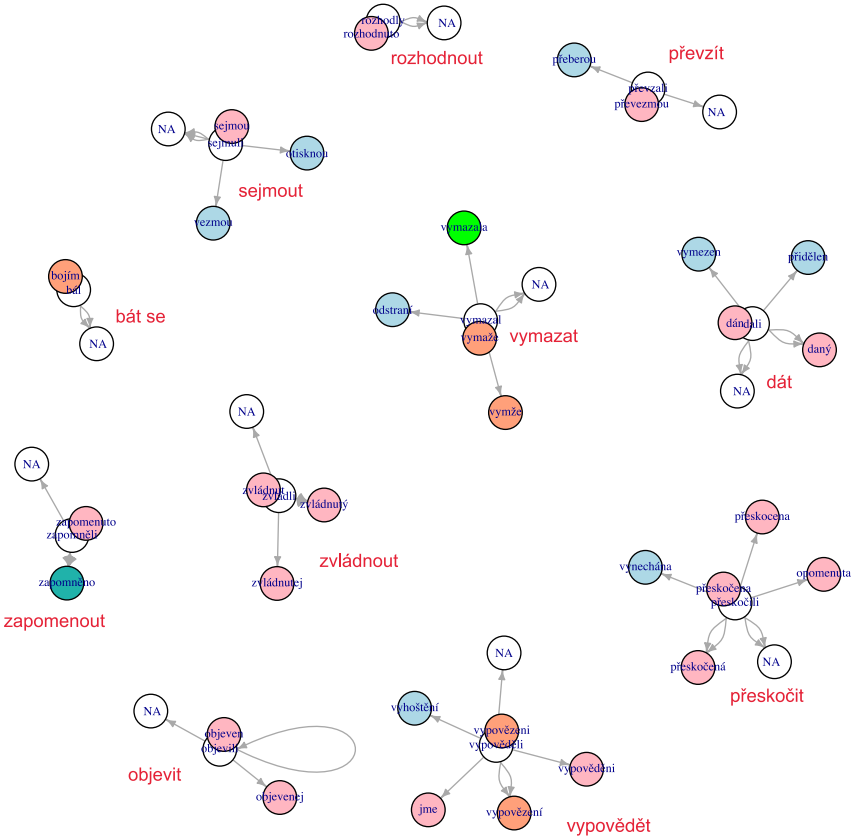


**Figure 3:** Verb forms produced in nonvariant cells.

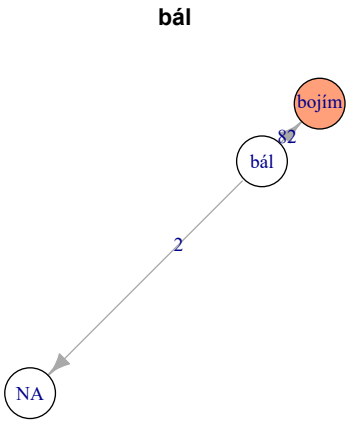


**Figure 4:** Nonvariant responses as expected.

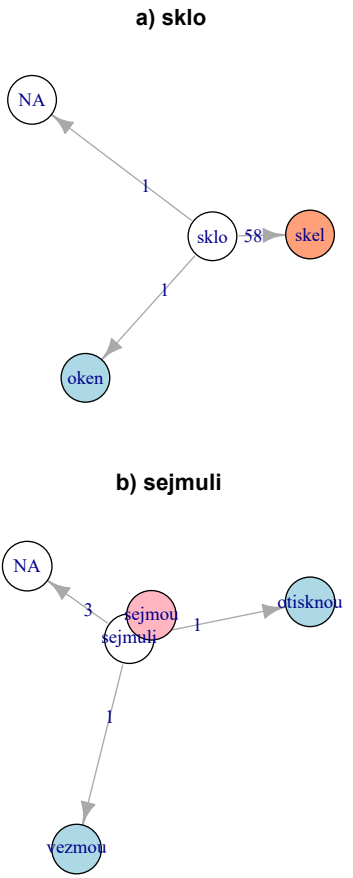


**Figure 5:** Opportunistic suppletion in non-variant cells. (a and b) Opportunistic suppletion in nonvariant cells.

lexeme. In (5a), the respondent may hesitate to give a plural of *sklo* 'glass [material]', although Czech does permit a plural meaning 'kinds of glass'. Instead, s/he substitutes for expected *skel* 'glasses-GEN.PL' the form *oken* 'windows-GEN.PL'. In (5b), one substitution (*vezmou* 'take-NONPAST.3PL') makes use of a less specific and more frequent lexeme from core vocabulary, but the second substitution (*otisknou* 'imprint-NONPAST.3PL') is neither more frequent in a corpus nor less specific; it is possible the speaker was influenced by the collocate *otisky* 'fingerprints'.[13]

**13** The verb *sejmout* has a frequency of 193 non-past tense forms in the SYN2020 100m-wordform balanced corpus of written Czech: the collocation with the lemma *otisk* 'fingerprint' is the eighth strongest (logDice 6.72). This compares to 8606 non-past forms for *vzít* and 134 for *otisknout*. The lexeme *otisk* is collocation no. 313 for *vzít* (logDice 3.13), and does not appear in the list of collocations

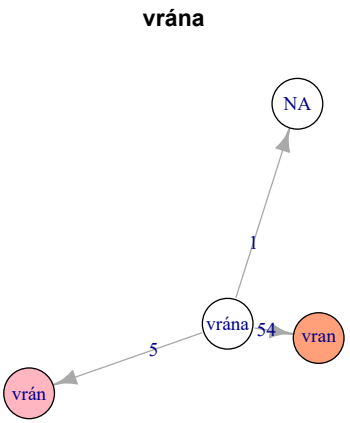


**Figure 6:** New stems and endings in nonvariant cells.

In another substitution type, shown in Figure 6, respondents created a morphologically different form linked to the trigger form. We found 5 out of our 60 respondents did not produce a form with the expected gen. pl. shortening: *vran* from nom. sg. *vrána*. Instead, they produced a form regularizing stem length: *vrán*, which is unattested in SYN2020 (Křen et al. 2020) and unacknowledged in normative reference works. However, this tendency towards *stem unification* in nominal and verbal paradigms in Czech, reducing quantitative and qualitative variants that arose through historical change, is well-attested.

For nonvariant cells, then, we observe low-level individual variation even though no variation is expected or attested. Some respondents substituted other lexemes, creating an *ad hoc* opportunistically suppletive effect, while others produced novel forms or borrowed from related lexemes.[14] In a CL-type model of language, these results are explicable. The form/meaning mapping of schemas (conceived as per Langacker [2019: 349]) provides a path for speakers to borrow similarly structured items at moments of hesitation. The production of similar but not identical forms shows the individuality of language structure, emerging with

for *otisknout*. Search term e.g.: [lemma="sejmout" & tag="V …. … [FP].*"], with collocations set to lemma, context of -5 to +5 and minimum frequency of 3.

**14** Juge (2000) gives a historical perspective on types of suppletion in Romance languages where this sort of substitution can occur in sanctioned cells: he terms it *overlapping suppletion* (2000: 191–193). We distinguish our *opportunistic suppletion* from Juge's in that ours takes in examples that seem to be spontaneously generated and possibly one-offs, as opposed to those that enter circulation and presumably are then transmitted to further speaker groups. We are grateful to an anonymous reviewer for pointing us to this connection.

slight variations within the experience of each individual. The "background noise" so created will get louder as we turn to overabundant and defective cells.

## 4.2 Visualizing overabundant paradigm cells

For overabundant paradigm cells, we expected our respondents to split their answers between the 2–4 variants available; multiples beyond two were most frequent with passive participles as discussed above.

For our 17 nominal triggers in this condition, we expected up to 34 forms to be produced. For our 11 verbal triggers, we expected up to 38 forms to be produced. As Figures 7 and 8 show, our respondents produced more forms in the

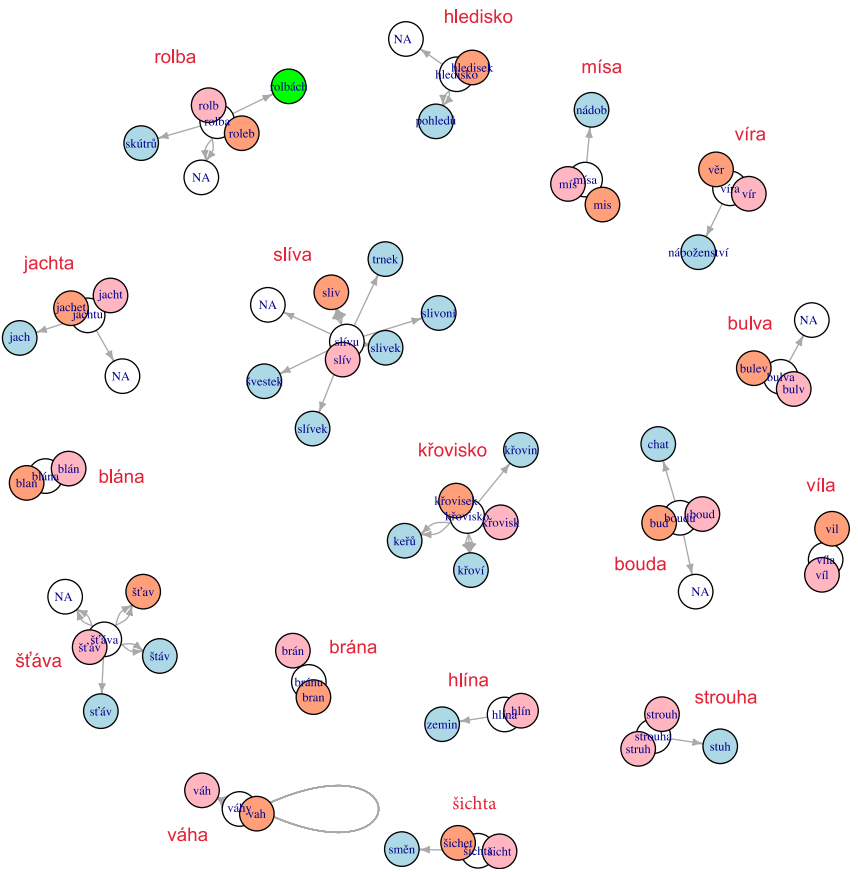


**Figure 7:** Noun forms produced in overabundant cells.

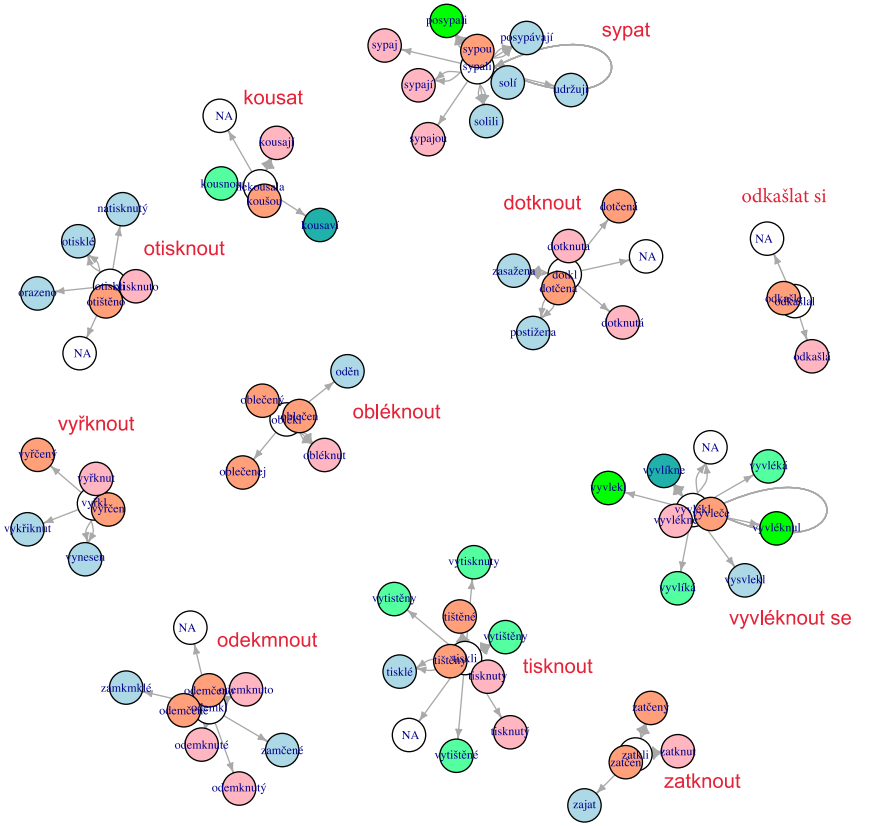


**Figure 8:** Verb forms produced in overabundant cells.

overabundant verb condition than in the nonvariant verb condition but produced the same number of noun forms in both conditions; the distribution of these forms was, however, more evenly split in the overabundant condition. A total of 59 noun forms (including 7 null responses) and 73 verb forms (including 7 null responses) were produced. For nouns, 31 represented potentially licensable forms; in one instance, the grammar was manipulated to produce a different case form, and in 19 instances, a form of a different lexeme or phrase was produced. For verbs, many of the available options were not used, but there were still 35 potentially licensable forms, 12 manipulations of grammar and 17 synonymic or phrasal substitutions.

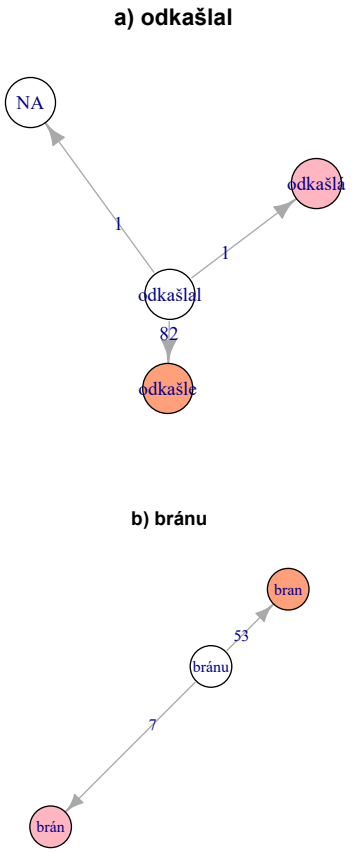


**Figure 9:** Predicted overabundant responses. (a and b) Predicted overabundant responses.

Variation in overabundant cells extended beyond the licensed variants found in corpora and reference works.[15] Figure 9 shows that some overabundant cells were populated only by licensed variants, but these were a small minority (although sometimes there were also missed answers). This is the case with the trigger *odkašlal (si)* 'he coughed-MASC.SG' in (9a), and with the trigger *bránu* 'gate-ACC.SG' in (9b), which resulted in the forms *brán/bran* 'gates-GEN.PL'.

Deviations from this pattern are shown in Figure 10. With the lexeme *kousat* 'bite' in (10a), five users substituted a semelfactive verb from the same root (*kousnou*,

**15** This despite the fact that arguably the task given to our speakers was not all that different from those faced by handbook authors, or even, to some extent, by writers featured in a corpus: to produce a recognizable form of a particular lexeme suited to a fixed context.

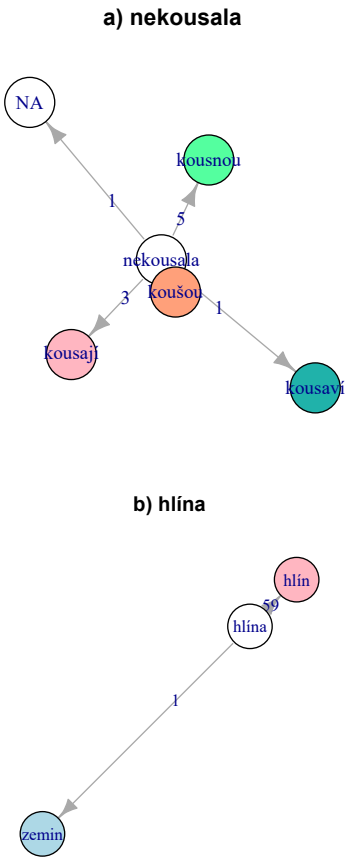


**Figure 10:** Unexpected variation in overabundant cells. (a and b) Unexpected variation in overabundant slots.

from *kousnout* 'take a bite') and one used an adjective *kousaví* 'biting-NOM.PL.ANIM'. The trigger *hlína* 'clay-NOM.SG' in (10b) should, on the other hand, have produced both *hlín/hlin* as possible variants, but respondents only produced the first form, so the overabundance signalled in some reference works and corpora was not forthcoming; *paradigmatic unification* here seems to have overtaken an earlier paradigm with optional length variation.[16]

Another type of substitution, shown in Figure 11, is the use of closely related lexemes instead of the lexeme requested. For *křovisko* 'bush', six out of 60 respondents did not produce either licensed gen. pl. *křovisk* or *křovisek*. Instead, they

**16** The alternative lemma offered by one respondent, *zemin* 'type of earth-GEN.PL', probably reflects a similar issue to that seen above for *sklo*.

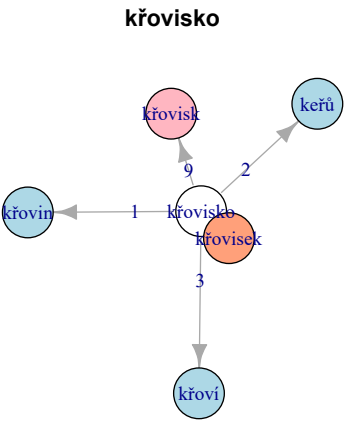


**Figure 11:** Avoiding hard choices in overabundant slots.

produced another word form with the *k-ř* root: *křoví* 'shrubbery-GEN.SG', *keřů* 'bushes-GEN.PL, *křovin* 'shrubs-GEN.PL'. This allowed the respondent to avoid choosing between the two forms of *křovisko*, representing an escape from the choice presented by an overabundant slot. Usually in Czech, suffixation decreases the chance of an irregular, unpredictable inflection pattern, and thus suffixed words play an outsize role in, e.g., language acquisition by increasing the predictability of inflection. However, in a few instances (e.g., the nominal suffixes *-isk-, -ic-, -išť-*), the suffix itself provides multiple inflectional possibilities: our results show that this escape from uncertainty can function "in reverse" from the derived word to the simplex.

For overabundant cells there is no enormous increase in the variety of forms produced, as there was low-level variation even for nonvariant cells. However, in overabundant cells, low-level variation shows a more even split between the two main variants of the licensed forms and more participants choosing non-licensed forms. Neither was there much hesitation on the part of speakers: almost all simply opted for a single form, only rarely citing two forms in their answer (2× for nouns, 2× for verbs). Approaches available with nonvariant paradigms (opportunistic appropriation of a closely related lexeme or another near-synonym) seem to be an effective way of avoiding choice for some speakers. What appears as occasional deviations in the nonvariant cells looks more substantial in this set.

## 4.3 Visualizing defective paradigm cells

For defective paradigm cells, we expected our respondents to have difficulty producing forms; they could theoretically have refused to produce any. In a model of

language where we can mark defectivity in the lexicon to block the appearance of a form, that (along with complete avoidance of the context) is one expected outcome.[17] The results of our survey suggest that something different is going on: while some respondents did opt to skip certain questions, many produced a form regardless of its potential felicity or not.[18]

Figures 12 and 13 show that, although respondents failed to agree on any single form for these items, resulting in a *lack of entrenchment*, they nonetheless managed to produce a wide variety of forms, especially with verbs. We refer to this phenomenon as *entrenchment of uncertainty*. A total of 104 noun forms (including 6 null responses) and 178 verb forms (including 10 null responses) were produced. For nouns, 30 represented potentially licensable forms; in 20 instances, the grammar was manipulated to produce a different case form, and in 48 instances, a form of a different lexeme or phrase was produced. For verbs, many available options were not used, but there were still 20 potentially licensable forms, 51 manipulations of grammar and 97 synonymic or phrasal substitutions.

No defective cells showed respondents settling on a single variant. The simplest ones, as in Figure 14, have a licensed variant or variants in evidence, but with other forms also well represented. The colloquial noun *ségra* 'sis' in (14a) elicited the expected forms *séger, seger* 'sisses-GEN.PL', but also the highly frequent *sester*, which is the gen. pl. of the standard lexeme *sestra* 'sister'. Verb forms were more complex; the active past form of the verb *znát* 'know-INF' in (14b) triggered a variety of forms, among them the expected *znána* 'known-FEM.NOM.SG' but adjectives derived from the same root and aspectual modifications that create new verbs, sometimes with novel forms, were more frequent.

Some variations on this pattern are shown in Figure 15. With the lexeme *zácpa* 'traffic jam', in (15a), users produced both potential licensed gen. pl. forms: *zácp* and *zácep*, as well as a third form *zácpí*, whose ending is borrowed from a different class of nouns; they also produced two near-synonymous lexemes and phrases: *aut*

**17** Neatly, reference works often use such a strategy to present defective cells: they single out a particular cell as unused, or proffer a single form with the remark that it is in some way deficient.
**18** As one reviewer pointed out, this could be the result of a task effect: respondents have been told to write something and so they do because they wish to be helpful. Nonetheless, we feel the variety and types of responses given are representative of the reactions to defective cells that are often reported. Refusal to speak would be uncooperative in most situations; therefore, speakers do not stop and indicate that they cannot say something, but rather proceed with a form that they are not entirely comfortable with, or they may rephrase or substitute other items in its place (see *inter alia* Sims [2009] and Albright [2003] for examples of the tension between production and evaluation of forms in defective cells).

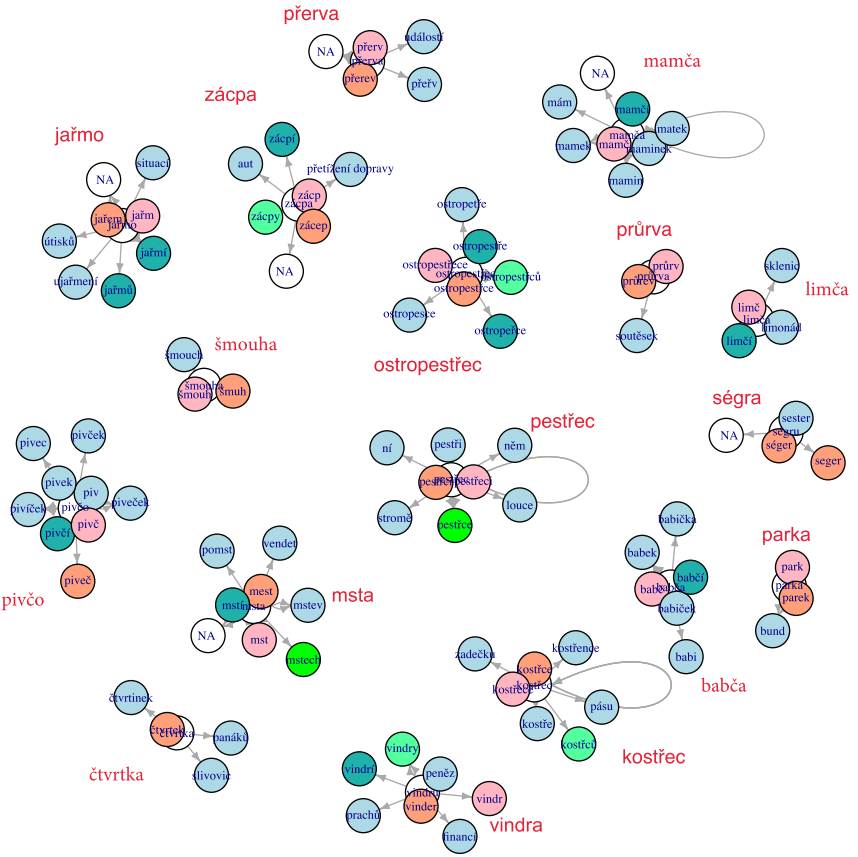


**Figure 12:** Noun forms produced in defective cells.

'cars-GEN.PL' and *přetížení dopravy* 'traffic overload-GEN.SG'.[19] The verb trigger *našli* 'found-PST.MASC-PL' in (15b) resulted in an even richer network of possible forms.

Defective cells had the greatest number of distinct responses, as shown in Figure 16, characterized by numerous grammatical alternatives and near-synonymic substitutions. For the noun *pivčo* 'beer-DIM' in (16a), many of our respondents ($n$ = 20) produced a target form *pivč*, but almost as many ($n$ = 17) entered *piv*, the gen. pl. of the standard word *pivo* 'beer'. A further eight substituted the expansive ending -*í* as above in (15a), avoiding the problem of stem shape in the gen. pl. Only one produced

**19** The ending -*í* seen in morphologically innovative *zácpí* cropped up in several defective lexemes. Because, unlike the traditional fem./neut. gen. pl. ending, it has phonetic material, it obviates the need for a decision about the shape of the gen. pl. bare stem.

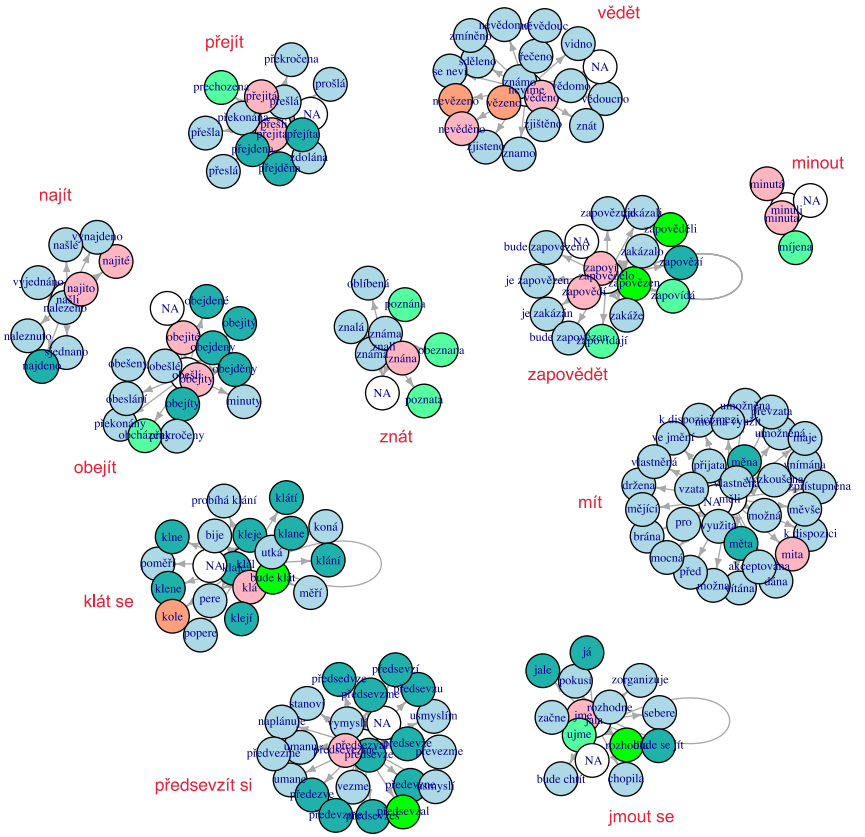


**Figure 13:** Verb forms produced in defective cells.

the potential alternative licensed form *piveč*. The remainder ($n = 14$) substituted other colloquial diminutives and similar-looking occasionalisms. For the verb *předsevzít si* 'resolve' in (16b), respondents produced 23 different forms, including skipped answers. The expected answer *předsevezme* was represented ($n = 18$) but was not the most popular entry.

Finally, many of the answers testified to the respondents' hesitation: rather than list a single form, they temporized or expressed doubt, sometimes explicitly saying that they would not use this word here (22× for nouns, 34× for verbs) instead of just skipping over it unanswered.[20] This suggests that, paradoxically, although there are

**20** Sample comments: *Tomu slovesu bych se vyhnula* 'I would avoid this verb'; *věděno (ale to není česky:)* '*věděno* (but that's not Czech:)'); or simply *???* For the visualizations, where no form was cited,

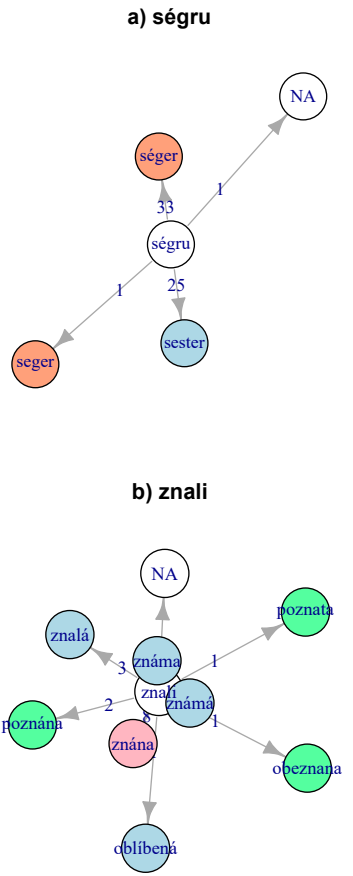


**Figure 14:** A simple response set to a defective cell. (a and b) A simple response set to a defective cell.

no entrenched forms, for many people all potential forms are pre-empted. Again, we explain this as an effect of entrenchment of uncertainty.

In summary, although Czech reference works typically do not explicitly label paradigms as defective, we were able to identify candidates for this category through occasional mentions in reference works that claim certain forms or tenses are 'rarely used', and through cross-checks with corpus frequency.[21] Our list of 17 nouns and 11

we assigned the answer to NA; where 2+ forms were cited despite critical or hedging remarks, we took the first form presented.

**21** Some have sanctioned full paradigms in reference works but nonetheless appeared in our list of potential defectives based on mentions elsewhere and corpus evidence.

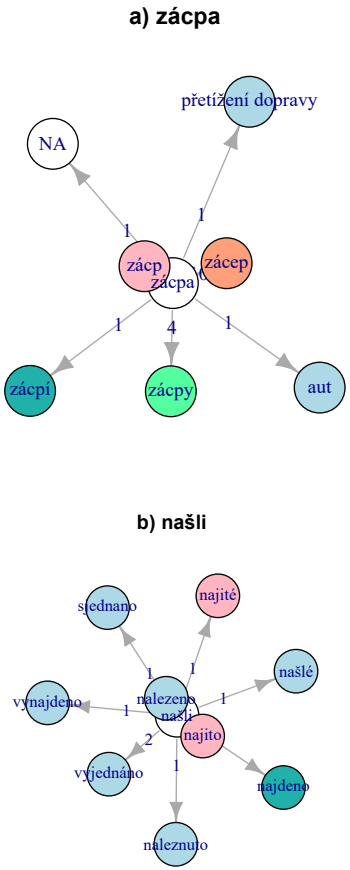


**Figure 15:** Other sorts of variation in defective cells. (a and b) Other sorts of variation in defective slots.

verbs represents the best-known and easiest-to-find lexemes in this group. Our respondents readily produced forms for these lexemes, but the forms were dispersed across a range of expected, near-synonymic and novel forms. There was a strong, although not universal, tendency towards opportunistic suppletion. In CL terms, it is clear what the schema for these forms should be, but they are not firmly associated with a realization. Speakers either attempt to find a match in their experience for the realization, or look for similar or identical schemas from which a suitable realization can be co-opted.

Sims (2009: 6) suggests that paradigmatically there must already be a "defective" paradigm as a model, which mandates avoidance of particular cells as a property of

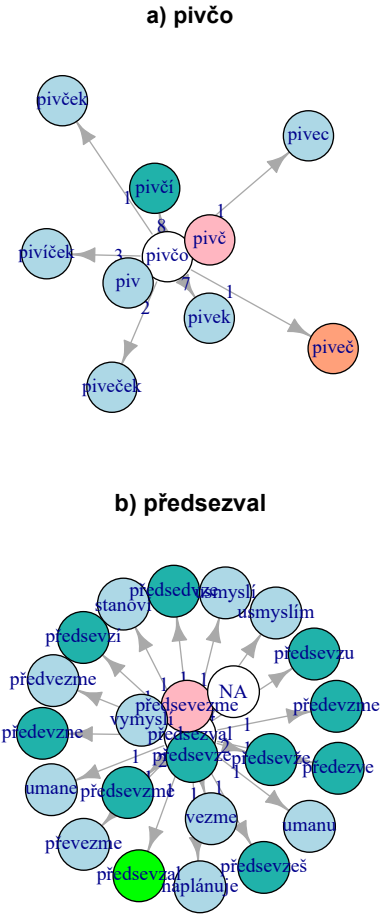


**Figure 16:** Complex defective networks. (a and b) Complex defective networks.

them. Our data suggest that opportunistic suppletion occurs everywhere in morphology as a strategy for managing uncertainty, although its use increases greatly in the defective condition, where uncertainty is greatest. This finding lends weight to Albright's hypothesis about the gradience of variation, but also underlines Sims's contention that defective cells have passed some "tipping point"; we observe that in defective cells, variability then increases dramatically due to the absence of entrenched forms.

# 5 Sources for defective and overabundant paradigms

The previous section offered a qualitative look at the survey results; a comprehensive overall analysis is available in Bermel et al. (2023). It suggests that participants responded differently to defective paradigms than to non-defective paradigms: the difference between non-defective types (nonvariant or overabundant) is visible at a group level but less so at an individual level. Defective cells triggered fewer expected responses than non-defective cells and took our respondents longer to process, both at the point of initiating a response and completing one (Bermel et al. 2023: 274–276). In this section, we revisit our definitions of these categories to see how they held up in the course of our experiment. Our approach included both corpus and reference-work data, although it was not possible to implement this identically for all conditions and lexemes; the reasons for this will be discussed below.

## 5.1 Comparing speakers' responses with corpus and reference-work data

Given that our choice of lexemes had been defined in part by corpus data and reference-work data, we checked whether the relationship persisted once speaker data were fed into it. We ran network analyses to establish relationships between several data points: the relative frequency of the most-produced form in our experiment; the relative frequency of its appearance in a corpus; and a value defined for its appearance in reference works.

The form most frequently produced by our respondents – referred to here as the *target form* – was given as a percentage value of the total forms produced, normalizing the different numbers of respondents in our two surveys ($N$ = 60 for nouns and $N$ = 84 for verbs). For example, for the proposed defective noun *pivčo* 'beer-DIM', 20 of our 60 respondents produced the gen. pl. form *pivč*. This results in a value of 33.33 %.

We then checked the appearance of that form in a corpus versus the total number of forms of that lemma produced for that slot. For overabundant and nonvariant slots, we used the SYN2015 representative corpus of written (Křen et al. 2015) Czech. As many of our defective items did not register in this corpus, we turned

to the much larger (csTenTen 2017) web corpus. To put these all on the same scale, we report our values in terms of the percentage found.[22]

For reference-work data, we took as our starting point the (Jazyková poradna Ústavu pro jazyk český 2008-2024) ILRB. For items unmentioned in this work – either due to being "non-standard" or having low frequency – we used the much larger (DCLL Dictionary of the Czech Literary Language: Slovník spisovného jazyka českého 1960). Our heuristic started from a top value of 100 for non-variant forms: if only the target form was mentioned in the reference work, it received a value of 100. If the reference work mentions multiple items, we assigned a partial value between 0 and 100 based on the number of forms mentioned and the place of the target form in the list. A simple example is the lexeme *rolba* 'snowcat'. The reference works list two forms: *roleb* and *rolb.* If the target were the first cited form (*roleb*), we would assign a value of 66 (i.e., 2:3) but because the target form is the second cited form, we assign a value of 33 (i.e., 1:3). A more complex example was the oblique form of *ostropestřec* 'milk thistle'. This has four sanctioned gen. sg. forms in the ILRB: *ostropestřce, ostropestřece, ostropestrce, ostropesterce*. As the ordering is not alphabetical and there is no statement to the effect that it reflects any distributional, etymological or other principle, we assume, in line with general handbook practice, that it represents a preferential order.[23] Our target form, *ostropestřce,* is the top item out of four and receives a value of 80, with the other items in the list having possible values of 60, 40 and 20. Had it been third in the list (as with the related lexeme *pestřec* 'earthball mushroom'), we would have assigned it a value of 40. If not mentioned at all, it would have received a value of 0. This heuristic thus assigns a weight to the target form vis-à-vis other forms in the reference work.

**22** We acknowledge, as per Baayen (2007), that because vocabulary size and appearance of forms does not increase linearly with corpus size, this heuristic may give greater weight to the percentages estimated from the larger corpus vis-à-vis the smaller one; some of the implications of this are discussed in Nikolaev and Bermel (2023). All these figures are nonetheless at the extreme low end of the frequency scale, so the point still holds.

**23** The modern ILRB and the mid-century RDCL have no information about the ordering of items within an entry. However, the DCLL, which was published in the second half of the 20th century by the same institute, explains: *Existují-li tvarové dublety a nejsou-li rovnocenné, rozlišují se podle frekvence nebo podle stylové příslušnosti. Tento rozdíl bývá naznačen i pořadím tvarů (první je běžnější, popř. stylově neutrální).* 'If morphological doublets exist and are not equal, they are differentiated by frequency or by belonging to a style. This difference tends to be indicated as well by the order of forms (the first is more common or stylistically neutral).' (1989: I:VII).

## 5.2 Description of comparative data

The data for nonvariant slots showed, as expected, that few respondents produced any forms other than those represented in a corpus of standard written Czech and promoted in the ILRB. The results are in Tables 1 and 2.[24]

For overabundant lexemes, the results are in Tables 3 and 4. Corpora reflect the greater variation seen in our respondents' data, but the results are not always in the same proportions. The ILRB often lists two variants, but not necessarily in the order favoured by respondents. Some variation found in other manuals and in native-speaker responses does not find its way into the ILRB's recommendations.

With defective lexemes, the results need some interpreting. For the verbal lexemes, *mít* 'have', *vědět* 'know [of]', *najít* 'find', *znát* 'know [personally]' the form most

**Table 1:** Nonvariant noun forms.

| Lexeme | Survey | | | Corpus | | | Reference works | | |
|---|---|---|---|---|---|---|---|---|---|
| | Target | % | *N* | % | Target | Other | Value | Top form | Other |
| houba 'mushroom' | hub | 98.3 | 59 | 99.9 | 1796 | 1 | 100 | hub | – |
| kostra 'skeleton' | koster | 98.3 | 59 | 100 | 93 | 0 | 100 | koster | – |
| sprcha 'shower' | sprch | 98.3 | 59 | 100 | 149 | 0 | 100 | sprch | – |
| bříza 'birch tree' | bříz | 96.7 | 58 | 100 | 117 | 0 | 100 | bříz | – |
| káva 'coffee' | káv | 96.7 | 58 | 100 | 47 | 0 | 100 | káv | – |
| plátno 'screen' | pláten | 96.7 | 58 | 100 | 161 | 0 | 100 | pláten | – |
| sklo 'glass' | skel | 96.7 | 58 | 100 | 334 | 0 | 100 | skel | – |
| hra 'game' | her | 95 | 57 | 100 | 3197 | 0 | 100 | her | – |
| jablko 'apple' | jablek | 95 | 57 | 100 | 658 | 0 | 100 | jablek | – |
| šelma 'predator' | šelem | 95 | 57 | 100 | 302 | 0 | 100 | šelem | – |
| tundra 'tundra' | tunder | 95 | 57 | 100 | 1 | 0 | 100 | tunder | – |
| moucha 'fly' | much | 93.3 | 56 | 100 | 289 | 0 | 100 | much | – |
| pomsta 'revenge' | pomst | 93.3 | 56 | 100 | 4 | 0 | 100 | pomst | – |
| zrno 'grain' | zrn | 93.3 | 56 | 100 | 242 | 0 | 100 | zrn | – |
| objížď'ka 'detour' | objížděk | 90 | 54 | 100 | 18 | 0 | 100 | objížděk | – |
| vrána 'crow' | vran | 90 | 54 | 100 | 100 | 0 | 100 | vran | – |
| tango 'tango' | tang | 88.3 | 53 | 100 | 12 | 0 | 100 | tang | – |

**24** For Tables 1–6, the column headings are: *Lexeme* gives the citation form. Under *Survey*, the entry *Target* shows the form most frequently produced by respondents; the percentage of responses and the actual number of responses are in the subsequent columns. Under *Corpus*, the percentage shows the share of *Target* forms for this cell over all forms for this cell (the remainder are in the column *Other*). Under *Handbook*, *Value* shows the calculation described above; *Top form* is the first recommended form and *Other* contains additional ones.

**Table 2:** Nonvariant verb forms.

| Lexeme | Survey | | | Corpus | | | Reference works | | |
|---|---|---|---|---|---|---|---|---|---|
| | Target | % | *N* | % | Target | Other | Value | Top form | Other |
| převzít 'take over' | převezmou | 97.6 | 82 | 100 | 144 | 0 | 100 | převezmou | – |
| rozhodnout 'decide' | rozhodnuto | 97.6 | 82 | 100 | 1103 | 0 | 100 | rozhodnut | – |
| objevit 'discover' | objeven | 96.4 | 81 | 100 | 300 | 0 | 100 | objeven | – |
| bát (se) 'fear' | bojím | 95.2 | 80 | 100 | 1637 | 0 | 33 | bojím | – |
| vymazat 'erase' | vymaže | 94.0 | 79 | 100 | 90 | 0 | 100 | vymaže | – |
| sejmout 'take down' | sejmou | 94.0 | 79 | 100 | 18 | 0 | 100 | sejmou | – |
| vypovědět 'expel' | vypovězeni | 92.9 | 78 | 100 | 12 | 0 | 33 | vypověděni | vypovězeni |
| dát 'give' | dán | 92.9 | 78 | 100 | 928 | 0 | 100 | dán | – |
| zapomenout 'forget' | zapomenuto | 92.9 | 78 | 100 | 140 | 0 | 67 | zapomenuto | zapomněno |
| přeskočit 'skip over' | přeskočena | 91.7 | 77 | 100 | 1 | 0 | 100 | přeskočena | – |
| zvládnout 'master' | zvládnut | 89.3 | 75 | 100 | 8 | 0 | 100 | zvládnut | – |

**Table 3:** Overabundant noun forms.

| Lexeme | Survey | | | Corpus | | | Reference works | | |
|---|---|---|---|---|---|---|---|---|---|
| | Target | % | *N* | % | Target | Other | Value | Top form | Other |
| hlína 'clay' | hlín | 98.3 | 59 | 82.0 | 50 | 11 | 100 | hlín | – |
| šťáva 'juice' | šťáv | 96.7 | 58 | 97.4 | 148 | 4 | 100 | šťáv | – |
| hledisko 'viewpoint' | hledisek | 95.0 | 57 | 99.7 | 339 | 1 | 66 | hledisek | hledisk |
| brána 'gate' | bran | 88.3 | 53 | 82.2 | 370 | 80 | 100 | bran | – |
| víla 'nymph' | víl | 85.0 | 51 | 95.3 | 82 | 4 | 100 | víl | – |
| váha 'scales' | vah | 83.3 | 50 | 99.7 | 294 | 1 | 100 | vah | – |
| jachta 'yacht' | jachet | 78.3 | 47 | 91.5 | 54 | 5 | 66 | jachet | jacht |
| mísa 'bowl' | mís | 78.3 | 47 | 33.1 | 88 | 178 | 33 | mis | mís |
| křovisko 'bush' | křovisek | 75.0 | 45 | 96.2 | 25 | 1 | 66 | křovisek | křovisk |
| slíva 'plum' | slív | 75.0 | 45 | 93.3 | 14 | 1 | 66s | lív | sliv |
| bulva 'eyeball' | bulv | 71.7 | 43 | 62.1 | 18 | 11 | 33 | bulev | bulv |
| rolba 'snowcat' | rolb | 66.7 | 40 | 20.0 | 1 | 4 | 33 | roleb | rolb |
| blána 'membrane' | blan | 65.0 | 39 | 98.9 | 90 | 1 | 100 | blan | – |
| strouha 'gulley' | struh | 63.3 | 38 | 83.3 | 15 | 3 | 100 | struh | – |
| šichta 'shift' | šicht | 60.0 | 36 | 50.0 | 1 | 1 | 66 | šicht | šichet |
| víra 'faith' | vír | 58.3 | 35 | 3.3 | 1 | 29 | 100 | věr | – |
| bouda 'hut' | bud | 51.6 | 31 | 97.0 | 130 | 4 | 100 | bud | – |

often produced by our respondents was a near-synonym. We therefore substituted the most common produced form *of the trigger lexeme*. These figures are therefore

**Table 4:** Overabundant verb forms.

| Lexeme | Survey | | | Corpus | | | Reference works | | |
|---|---|---|---|---|---|---|---|---|---|
| | Target | % | *N* | % | Target | Other | Value | Top form | Other |
| odkašlat 'cough' | odkašle | 97.6 | 82 | 99.3 | 151 | 1 | 100 | odkašle | – |
| zatknout 'jail' | zatčen | 90.5 | 76 | 99.9 | 1033 | 1 | 67 | zatčen | zatknut |
| kousat 'bite' | koušou | 88.1 | 74 | 98.0 | 350 | 7 | 33 | kousají | koušou |
| obléknout 'dress' | oblečen | 81.0 | 68 | 99.8 | 559 | 1 | 50 | obléknut | oblečen |
| dotknout (se) 'touch' | dotčena | 77.4 | 65 | 98.3 | 414 | 7 | 67 | dotčena | dotknuta |
| otisknout 'imprint' | otištěno | 71.4 | 60 | 96.0 | 190 | 8 | 67 | otištěno | otisknuto |
| vyřknout 'pronounce' | vyřčen | 71.4 | 60 | 98.2 | 110 | 2 | 67 | vyřčen | vyřknut |
| sypat 'spread' | sypou | 65.5 | 55 | 99.3 | 429 | 3 | 67 | sypou | sypají |
| tisknout 'print' | tištěny | 59.5 | 50 | 90.7 | 49 | 5 | 67 | tisknut | tištěn |
| odemknout 'unlock' | odemčeno | 54.8 | 46 | 87.1 | 27 | 4 | 67 | odemčeno | odemknuto |
| vyvléct (se) 'wriggle out' | vyvlékne | 42.9 | 36 | 72.7 | 8 | 3 | 50 | vyvlékne | vyvleču |

not exactly parallel, as the corpus and reference work data refer only to the target lexeme *sensu stricto,* whereas the survey data contain all results produced.

As mentioned above, we used the larger (csTenTen 2017) corpus for defective lexemes, but despite a corpus whose size exceeds that of the average speaker's linguistic experience several times over, we failed to find examples of some forms. Homonymy with other lexemes for the target form complicates matters (e.g., the gen. pl. of *čtvrtka* 'quarter-piece, 250 ml bottle' is homonymous with *čtvrtek* 'Thursday-NOM.SG'; the gen. pl. of *parka* 'parka' is homonymous with *park* 'park-NOM.SG'), and some of the other forms produced by respondents were forms of other non-synonymous lexemes expropriated for use here (e.g., *já* 'I' as a present tense of *jmout se* 'sets to', or *kleje* [se] 'swears' as a present tense of *klát se* 'joust'). In a few places we could not manually disambiguate forms due to the amount of data, so the figures are derived from samples.[25]

Interpretation of reference-work recommendations is also problematic. Several nouns and one verb do not appear in the ILRB, thus the data are drawn from the much older (but larger) DCLL. Colloquial words (*pivčo* 'beer-DIM', *mamča* 'mamma', *ségra* 'sis', *limča* 'soda pop') frequently fail to appear in either, so we extrapolated the recommendations for them from the general rules found in the ILRB.[26]

**25** Most examples of *jme se* 'sets to' were misspellings of the common auxiliary verb *jsme* 'are'; most examples of *minut(a)* 'passed' were forms of the lexeme *minuta* 'minute'. The figures are based on samples of 100 random concordance lines, extrapolated to the total results returned.

**26** The lexeme *ségra* 'sis' is a curious case. It does not appear in ILRB, but is in DCLL with no accompanying forms, suggesting that the gen. pl. with zero ending should be *ségr*. However, the

Our respondents produced a multitude of forms in this condition: for a few lexemes, one form dominated, but for most there was a broad variety. Corpus data are less varied; reference-work data is even less varied than corpus data. The tendency in Czech normative manuals to supply at least one item to fill a slot is evident.

## 5.3 Network analysis

Our network analysis uses the information in Tables 1–6 to examine how our respondents' top answers related to the information provided from corpus data and reference works. Network analysis provides an easily apprehensible way of examining complex relationships between variables. Within a network, variables are directly interconnected, allowing for the examination of both direct and indirect relationships. It does not rely on strong parametric assumptions about the data distribution, making it more robust in cases where relationships might be nonlinear or heterogeneous.

We produced two network analyses – for nouns and verbs – in R using the function *estimateNetwork* from the package *bootnet* (Epskamp et al. 2018).[27] The analysis for nouns is shown in Figure 17, with an explanatory graph in Figure 18. Each analysis included all defective and overabundant items. We did not include the nonvariant condition because outcomes were almost all identical, showing 100 % or close to 100 % across all items – this would have overwhelmed any other distinctions in the analysis.

In the network analysis graph (Figure 17), the strength of a connection is shown by the thickness of the line between them. A blue line represents a positive connection (more/higher A = more/higher B). A red line represents an inverse connection (more/higher A = less/lower B).

The graph in Figure 18 shows Centrality indices: Strength (indicating how well a node is directly connected to other nodes), Betweenness (illustrating the importance of a node in the average path between other nodes), and Closeness (indicating how well a node is indirectly connected to other nodes). Nodes are ordered from top to bottom on the graph based on their influence on the network, and the *X*-axis represents z-scores. For example, if the z-score of the node "Type" (defective vs.

lexeme is marked as 'vulgar', and a note in the dictionary introduction explains that declined and conjugated forms for vulgar words are not given. We have thus gone by the general rules in the ILRB suggesting that with this consonant cluster, a vowel should be inserted, e.g., *séger*. It can become difficult to work out what an entry means when layers of instructions collide.

27 Because the number of lexemes was small, we selected a dense regularized network (lambda < 0. 1 × lambda.max) which leads to a possible drop in specificity. As a result, we must interpret the presence of the smallest edges with care.

**Table 5:** Defective noun forms.

| Lexeme | Survey | | | Corpus | | | Reference works | | |
|---|---|---|---|---|---|---|---|---|---|
| | Target | % | *N* | % | Target | Other | Value | Top form | Other |
| čtvrtka ‘quarter-litre’ | čtvrtek | 95.0 | 57 | 100 | 380 | 0 | 100 | čtvrtek | – |
| šmouha ‘smudge’ | šmouh | 83.3 | 50 | 98 | 2753 | 54 | 100 | šmouh | – |
| vindra ‘penny’ | vinder | 75.0 | 45 | 100 | 1 | 0 | 100 | vinder | – |
| zácpa ‘traffic jam’ | zácp | 70.0 | 42 | 92 | 919 | 79 | 100 | zácp | – |
| limča ‘lemonade’ | limč | 68.3 | 41 | 100 | 1 | 0 | 100 | limč | – |
| ostropestřec ‘milk-thistle’ | ostropestřce | 68.3 | 41 | 96 | 3030 | 128 | 80 | ostropestřce | ostropestřece<br>ostropestrce<br>ostropesterce |
| parka ‘parka’ | parek | 60.0 | 36 | 90 | 18 | 2 | 100 | parek | – |
| pestřec ‘earthball’ | pestřci | 60.0 | 36 | 85 | 78 | 14 | 40 | pesterci | pestrci<br>pestřci<br>pestřeci |
| přerva ‘gap’ | přerv | 58.3 | 35 | – | 0 | 0 | 0 | přerev | – |
| jařmo ‘yoke’ | jařem | 56.7 | 34 | 50 | 48 | 48 | 100 | jařem | – |
| kostřec ‘rump’ | kostřce | 56.7 | 34 | 20 | 10 | 41 | 0 | kostrce | – |
| ségra ‘sis’ | séger | 55.0 | 33 | 50 | 67 | 67 | 100 | séger | – |
| průrva ‘cleft’ | průrev | 53.3 | 32 | 86 | 150 | 25 | 100 | průrev | – |
| msta ‘vengeance’ | mest | 50.0 | 30 | 100 | 3 | 0 | 100 | mest | – |
| mamča ‘mum’ | mamč | 40.0 | 24 | 100 | 2 | 0 | 100 | mamč | – |
| babča ‘granny’ | babč | 35.0 | 21 | 100 | 7 | 0 | 100 | babč | – |
| pivčo ‘beer [dim.]’ | pivč | 33.3 | 20 | – | 0 | 0 | 0 | piveč | – |

**Table 6:** Defective verb forms.

| Lexeme | Survey | | | Corpus | | | Reference works | | |
|---|---|---|---|---|---|---|---|---|---|
| | Target | % | *N* | % | Target | Other | Value | Top form | Other |
| minout ‘pass by’ | minuta | 89.3 | 75 | – | 0 | 0 | 100 | minuta | – |
| zapovědět ‘forbid’ | zapoví | 66.7 | 56 | 89 | 54 | 7 | 100 | zapoví | – |
| obejít ‘avoid’ | obejity | 66.7 | 56 | – | 0 | 0 | 100 | obejity | – |
| jmout (se) ‘set to’ | jme | 50.0 | 42 | 100 | 320.3 | 0 | 100 | jme | – |
| přejít ‘cross over’ | přejita | 46.4 | 39 | – | 0 | 0 | 0 | () | – |
| předsevzít (si) ‘resolve’ | předsevze | 40.5 | 34 | 44 | 15 | 19 | 0 | předsevezme | – |
| klát (se) ‘joust’ | klaje | 29.8 | 25 | 100 | 1 | 0 | 0 | klá | kole |
| najít ‘find’ | najito | 19.1 | 16 | 0 | 0 | 1 | 100 | najito | – |
| vědět ‘know’ | věděno | 16.7 | 14 | 89 | 8 | 1 | 0 | () | () |
| znát ‘know’ | znána | 9.5 | 8 | 100 | 22 | 0 | 100 | znána | – |
| mít ‘have’ | měna | 4.8 | 4 | – | 0 | 0 | 0 | () | () |

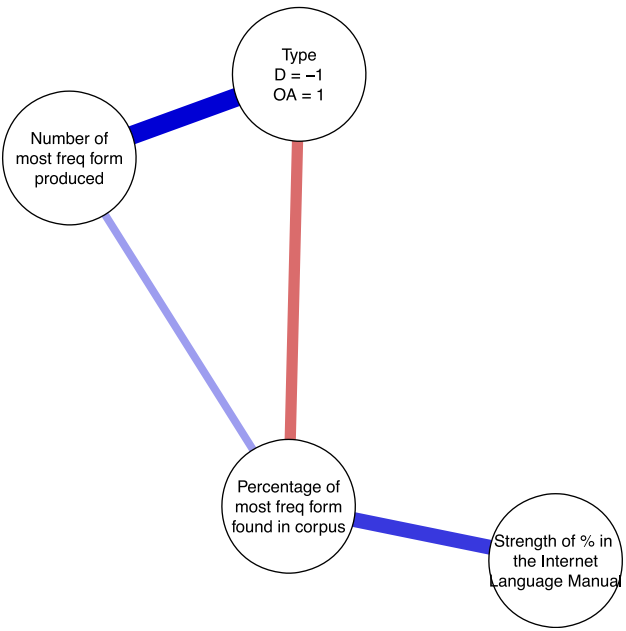


**Figure 17:** Network analysis for nouns.

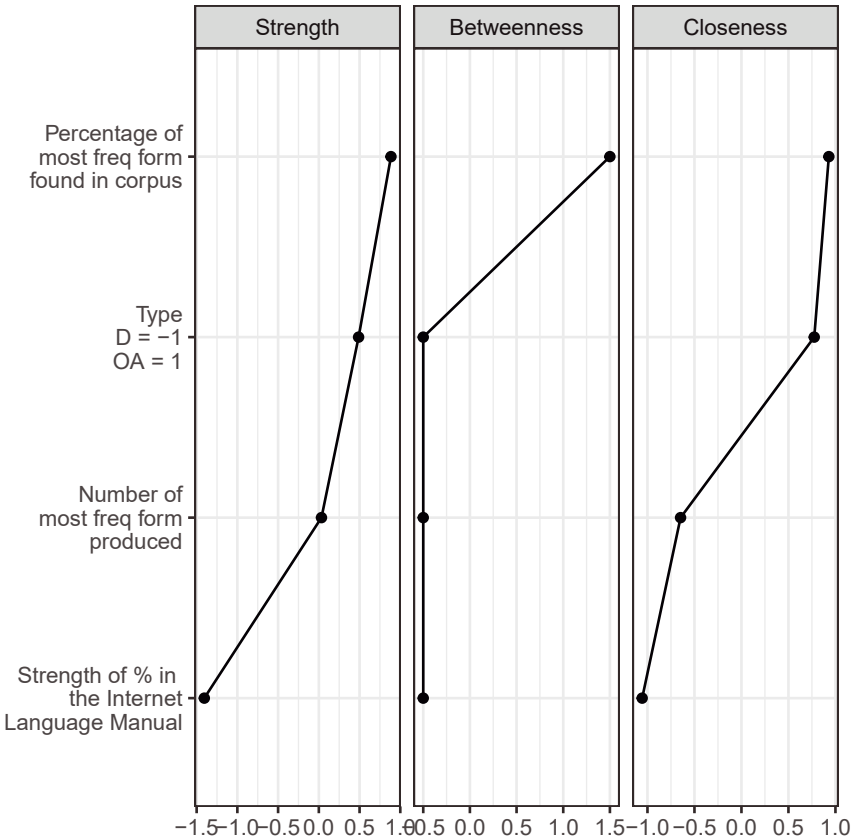


**Figure 18:** Explanatory graph for nouns (*x*-axes represents z-scores).

overabundant) is greater on the *X*-axis than that of the node "Number of the most frequent forms produced," it indicates that the "Type" node has a greater influence on the network when measured according to these three centrality indices.

As can be seen, the thickest blue line in Figure 17, and thus the strongest connection, with the "number of most frequent form produced" is the type of lexeme (defective or overabundant), suggesting that the two types elicit different responses: despite the fact that defective cells have, from a word-formational perspective, the same number of morphologically probable variants as overabundant or non-variant cells, people converge more on a chosen inflectional variant in the overabundant condition than in the defective condition.

In terms of our corpus data, a thin blue line connects it to the percentage of the most frequent form found in csTenTen or SYN2015 (as a percentage of all forms found

in that case). That slightly positive result reflects that fact that this correlation is higher in the defective condition than in the overabundant condition. Corpus data are also positively correlated with our operationalization of strength of preference in the Internet Language Reference Book (and supplemental sources) for the most frequent form. The corpus node is the most crucial node in this network, with connections to all the other nodes; this also surfaces in Figure 18, where it is at the top of the graph.

Our behavioural measure (the number of times the most frequent form was produced out of the total 60 results) is positively correlated with the percentage of the most frequent form found in our corpora as a percentage of all forms found in that case.

The behavioural measure is more likely to be predicted by corpus frequencies rather than by the recommendations in the ILRB, which appear not to be based on distributional criteria. The latter is positioned on the periphery of the network in Figure 17 and at the bottom of the graph in Figure 18.

**Table 7:** Nouns: number of forms produced.

| **Defective nouns** | ***N*** | **Overabundant nouns** | ***N*** | **Nonvariant nouns** | ***N*** |
|---|---|---|---|---|---|
| mamča ‘mum’ | 9 | slíva ‘plum’ | 7 | tango ‘tango’ | 6 |
| pestřec ‘earthball’ | 9 | křovisko ‘bush’ | 5 | objížďka ‘detour’ | 5 |
| pivčo ‘beer [dim.]’ | 9 | rolba ‘snowcat’ | 5 | pomsta ‘revenge’ | 5 |
| jařmo ‘yoke’ | 8 | šťáva ‘juice’ | 5 | hra ‘game’ | 4 |
| kostřec ‘rump’ | 8 | bouda ‘hut’ | 4 | jablko ‘apple’ | 4 |
| msta ‘vengeance’ | 8 | jachta ‘yacht’ | 4 | šelma ‘predator’ | 4 |
| ostropestřec ‘milk-thistle’ | 7 | bulva ‘eyeball’ | 3 | tundra ‘tundra’ | 4 |
| vindra ‘penny’ | 7 | hledisko ‘viewpoint’ | 3 | zrno ‘grain’ | 4 |
| zácpa ‘traffic jam’ | 7 | mísa ‘bowl’ | 3 | bříza ‘birch’ | 3 |
| babča ‘granny’ | 6 | šichta ‘shift’ | 3 | káva ‘coffee’ | 3 |
| přerva ‘gap’ | 5 | strouha ‘gulley’ | 3 | plátno ‘screen’ | 3 |
| čtvrtka ‘quarter-litre’ | 4 | váha ‘scales’ | 3 | sklo ‘glass’ | 3 |
| limča ‘lemonade’ | 4 | víra ‘faith’ | 3 | vrána ‘crow’ | 3 |
| ségra ‘sis’ | 4 | blána ‘membrane’ | 2 | houba ‘mushroom’ | 2 |
| parka ‘parka’ | 3 | brána ‘gate’ | 2 | kostra ‘skeleton’ | 2 |
| průrva ‘gap’ | 3 | hlína ‘clay’ | 2 | moucha ‘fly’ | 2 |
| šmouha ‘smudge’ | 3 | víla ‘nymph’ | 2 | sprcha ‘shower’ | 2 |
| total | 104 | | 59 | | 59 |
| average | 6.12 | | 3.47 | | 3.47 |
| median | 7 | | 3 | | 3 |
| stdev | 2.19 | | 1.33 | | 1.14 |

**Table 8:** Verbs: number of forms produced.

| Defective verbs | *N* | Overabundant verbs | *N* | Nonvariant verbs | *N* |
|---|---|---|---|---|---|
| mít 'have' | 32 | sypat 'spread' | 10 | přeskočit 'skip over' | 6 |
| předsevzít (si) 'resolve' | 23 | tisknout 'print' | 10 | vypovědět 'expel' | 6 |
| klát (se) 'joust' | 21 | vyvléct (se) 'wriggle out' | 10 | dát 'give' | 5 |
| vědět 'know' | 19 | odemknout 'unlock' | 8 | vymazat 'erase' | 5 |
| zapovědět 'forbid' | 17 | dotknout (se) 'touch' | 7 | objevit 'discover' | 4 |
| jmout (se) 'set to' | 15 | otisknout 'imprint' | 6 | sejmout 'take down' | 4 |
| obejít 'avoid' | 15 | kousat 'bite' | 5 | zvládnout 'master' | 4 |
| přejít 'cross over' | 14 | obléct 'dress' | 5 | převzít 'take over' | 3 |
| najít 'find' | 9 | vyřknout 'pronounce' | 5 | zapomenout 'forget' | 3 |
| znát 'know' | 9 | zatknout 'jail' | 4 | bát (se) 'fear' | 2 |
| minout 'pass by' | 4 | odkašlat (si) 'cough' | 3 | rozhodnout 'decide' | 2 |
| total | 178 | | 73 | | 44 |
| average | 16.18 | | 6.64 | | 4.00 |
| median | 15 | | 6 | | 4 |
| stdev | 7.67 | | 2.54 | | 1.41 |

Our network for overabundant and defective verbs failed to show any relations other than a connection between the condition (OA/D) and the outcome (frequency of top answer). Referring back to Figures 12 and 13, however, we can see one obvious reason for this, which is further detailed in Tables 7 and 8. Defective verb slots in Czech resulted in a much higher diversity of forms produced by our respondents. Some of the lexemes included are among the core lexical items of the language, with high frequency and a low degree of lexical specificity: *mít* 'have', *vědět, znát* 'know', *najít* 'find', *obejít* 'avoid', *přejít* 'cross' are easily pre-empted by a variety of synonyms with more specific meanings: *vlastnit* 'possess'; *držet* 'hold'; *obeznámit* 'familiarize'; *poznat* 'recognize'; *nalézt* 'locate'; *zjistit* 'clarify'; *překonat* 'overcome'; *překročit* 'stride past'; *zdolat* 'surmount'. In some instances, the most common form produced in our survey was a near-synonym rather than a form of the target lexeme, suggesting that opportunistic suppletion is a valid and accessible strategy for dealing with intractable defectivity.

# 6 Conclusions

For overabundance, our study confirms the fundamental assumption that entrenchment of a form causes the pre-emption of other forms, but entrenchment

appears to be a gradient rather than a binary phenomenon. Our data show that entrenchment can, in certain circumstances, be shared between two or more forms: individuals settle on "their" entrenched form, but due to differing exposure, etc., a large group of individuals may evidence two or more common outcomes. This preserves the current view of entrenchment, but we have no proof of whether that view is true universally, or only in part – if the latter, then perhaps a single individual can entrench both forms, using them at different times. We can see several places in our data where individuals offered multiple options, although this was only occasional. Results like ours are compatible with CL analyses, although some generative approaches like Optimality Theory capture the competition seen in parts of this study.

Defectivity presents a tougher condition to explain. In a construction-based approach, pre-emption is normally assumed to follow from entrenchment of a specific form in that construction, as summarized neatly by Goldberg: "How is that we know we should use *went* instead of **goed*? Clearly it is because we consistently hear *went* in contexts where *goed* would have been at least as appropriate: this is statistical preemption (2011: 133)."

However, in the case of truly defective cells, we find *pre-emption without entrenchment*: some respondents produce an expected form, but many others avoid it. Our respondents' heavy use of suppletive items, grammatical reworkings, and substitute phrases within the context of a strictly designed lexical task ("produce the form of lexeme A expected in context B") suggests that this is a more extreme version of the double- or triple-entrenchment scenario we saw for overabundance. In the defective condition, multiple alternatives appear in the environment, many of which are poor matches due to distance from the target lexeme through periphrasis, violation of common patterns, etc. Effectively, then, what has become "entrenched" is a sense that the cell is to be avoided. In this case, we argue that uncertainty is entrenched. Sims (2015: 220–231) describes one way that this development can be modelled in the Word and Paradigm approach.

Without a clear entrenched competitor, how do we describe this in CL terms? Our results show that, especially in the case of high-frequency defective lexemes, there is often a workaround in the form of a near-synonymic word or phrase. This does not amount to suppletion in the classic sense of the word (as per Corbett 2007), because the borrowed word or form has its own complete paradigm to belong to, but it seems to be a kind of opportunistic quasi-suppletion based on schematic or phonological similarities. Suppletion itself thus might be a gradient phenomenon, with occasional occurrences in non-variant lexemes at one end, and conventionalized suppletive paradigms like Czech *člověk~lidé* 'person-NOM.SG~people-NOM.PL' or *rok~let* 'year-NOM.SG~year-GEN.PL' at the other.

Here Schmid's (2015) Entrenchment-and-Conventionalization Model can be helpful. Overabundant slots can occur where individuals manage to accumulate enough instances of a particular usage to assign it to a pattern, and thus entrenchment and conventionalization can proceed, even if multiple "strands" of this emerge within a community. In linguistic exchanges, speakers often "take over and repeat linguistic material produced by their interlocutors," a process known as *co-adaptation* (Schmid 2015). When this pattern is replicated across a speech community, it leads to *diffusion*, the cumulative effect of the linguistic process. For a linguistic form to become fully established, it must enter metalinguistic awareness through *normation*, which includes explicit codification and other forms of conventionalization in society (Schmid 2015: 17–18). Should that change, the situation can of course adapt (as documented in Baerman [2008]). However, defective slots, which lack many prerequisites for entrenchment (Schmid 2015: 15–16), do not support these processes. As a result, even if individuals can produce certain forms when directly queried, co-adaptation, diffusion, and normation do not occur in normal use, which suppresses the match between form and meaning.

We also investigated which of our two initial source materials – reference works and corpora – best matched our survey results, operationalized on the most frequently produced form of the lexeme. The link with corpus data was confirmed for nouns, but not for verbs; moreover, it was not confirmed for reference-work data with any of our materials. There are several possible reasons for this.

Dictionaries and grammars distil a mixture of traditional prescriptions, filled in and cross-checked at times against corpus data, to present individual recommendations, rules or tendencies forming a coherent system. Reference-work authors have used analogy and material from previous reference sources to posit potential forms for gaps, sometimes with a note as to the rarity or unusualness of the form. This method can give rise to "false positives": a form posited, perhaps which was in use at one point, but which does not occur in contemporary usage or in our respondents' answers.

The focus in reference works on a specific understanding of "standard" language, and the overrepresentation of standard language in corpora, are further reasons why our survey data failed to corroborate our initial soundings. Defective cells, especially, often seemed to arise under conditions of obsolescence (e.g., *klát se* 'joust'), or a restricted sphere of usage where normation fails to occur (e.g., colloquial words like *mamča* 'mum', *ségra* 'sis'). These lexemes are not well represented in handbooks or most large-scale corpora. In other situations, the presence of a large array of near-synonyms (e.g., for *předsevzít si* 'resolve'), one or more of which have achieved greater popularity, can contribute to the loss of certainty when using a lexeme in a specific context. We found that Czech reference works, possibly as a holdover from the days before large-scale corpora, avoid implying the existence of

"empty" cells, instead describing them as occupied by a "rare" or "not typically used" form.[28] In our survey, however, speakers reached for this specific recommended form relatively rarely, and most often preferred a near-synonym or periphrastic form. This is a cognitively plausible result that manages connections between different sorts of linguistic data on various levels: semantic (schemas), phonological (sound similarity) as well as structural (word forms).

The current study opens further avenues for exploration. We hope to consider whether any of the above features – obsolescence, stylistic limitedness, and available synonymy – are susceptible to not only confirming defective and overabundant behaviour, but also predicting it: given a confluence of the above characteristics, are any of them more or less likely to identify a defective or overabundant lexeme? It also converges on an issue we have begun to consider in Nikolaev and Bermel (2023): the status we should give to unexpectedly low frequencies of forms of a lexeme in new, multi-billion-token corpora, as the sort of "native-speaker intuition" that we previously used reference works to represent does not seem to map clearly onto either what speakers do in online tasks or how they produce texts, as measured in corpora. Finally, the repeated assertions of gradience in defectivity, overabundance and non-variance suggest that we should be able to find subcategories within them that will help us understand the routes that lexemes take in and out of these categories.

**Acknowledgments:** The authors would like to thank Jim Blevins, Laura Janda, Petar Milin, and two anonymous reviewers for their helpful feedback and suggestions.
**Research funding:** This research is supported by grant AH/T002859/1from the UK Arts and Humanities Research Council.
**Data availability statement:** The datasets generated and analysed during the current study are available at https://osf.io/bdgjm/ To cite: Bermel, Neil, and Alexandre Nikolaev. "Morphological Overabundance and Defectivity in Czech." OSF, 15 Jan. 2024. Web.

# References

Albright, Adam. 2003. A quantitative study of Spanish paradigm gaps. In Gina Garding & Mimu Tsujimura (eds.), *Proceedings of the 22nd west coast Conference on formal linguistics*, 1–14. Somerville, MA: Cascadilla Press.

**28** This problem has not, however, disappeared entirely even in the age of multi-billion word-form corpora; see Nikolaev and Bermel (2023). It is also worth noting that this is a particular strategy characteristic of Czech reference works; Russian reference works frequently posit that a form is not used (*ne upotrebljaetsja*) for non-semantic reasons.

Baayen, R. Harald. 2007. 5: Storage and computation in the mental lexicon. In G. Jarema & G. Libben (eds.), *The mental lexicon: Core perspectives*, 81–104. Brill.

Baerman, Matthew. 2008. Historical observations on defectiveness: the first singular non-past. *Russian Linguistics* 32. 81–97.

Bermel, Neil. 2007. *Linguistic authority, language ideology, and metaphor: the Czech spelling wars. Language, Power and Social Process*, vol. 17. Berlin: Mouton de Gruyter.

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

Bermel, Neil & Luděk Knittl. 2023. Trajectories of change in paradigmatic cells in Czech. *Naše Rec* 106. 247–274.

Bermel, Neil, Luděk Knittl & Alexandre Nikolaev. 2023. Uncertainty in the production of Czech noun and verb forms. *Word Structure* 16. 258–283.

Blevins, James P., Petar Milin & Michael Ramscar. 2017. The Zipfian paradigm cell filling problem. In Ferenc Kiefer, James P. Blevins & Huba Bartos (eds.), *Perspectives on morphological structure: Data and analyses*, 139–158. Leiden: Brill.

Čermák, František & Michal Křen. 2011. *A Frequency Dictionary of Czech: Core Vocabulary for Learners*. London: Routledge.

Chomsky, Noam & Lasnik Howard. 1977. Filters and control. *Linguistic Inquiry* 8(3). 425–504.

Chuang, Yu-Ying, Dunstan Brown, R. Harald Baayen & Roger Evans. 2022. Paradigm gaps are associated with weird "distributional semantics" properties: Russian defective nouns and their case and number paradigms. *The Mental Lexicon* 17. 395–421.

Clark, Eve. 1987. The principle of contrast: A constraint on language acquisition. In B. MacWhinney (ed.), *Mechanisms of language acquisition*, 1–33. Mahwah NJ: Lawrence Erlbaum.

Corbett, Greville G. 2007. Canonical typology, suppletion, and possible words. *Language* 83. 8–42.

Csardi, Gabor & Tamas Nepusz. 2006. The igraph software package for complex network research. *InterJournal, Complex Systems* 1695(5). 1–9.

csTenTen. 2017. *Corpus of the Czech web*. https://www.sketchengine.eu/cstenten-czech-corpus/.

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

Dabrowska, Ewa. 2018. Experience, aptitude and individual differences in native language ultimate attainment. *Cognition* 178. 222–235.

DCLL (Dictionary of the Czech Literary Language): Slovník spisovného jazyka českého. 1960–1971, *Second printing with minor corrections 1989*. Prague: Academia. Digitalized version of 2011. Available at: http://bara.ujc.cas.cz/ssjc/.

Epskamp, Sacha, Denny Borsboom & Eiko I. Fried. 2018. Estimating psychological networks and their accuracy: A tutorial paper. *Behavior Research Methods* 50(1). 195–212.

Ertl, Václav. 1929. Dobrý autor. In Ertl Václav (ed.), *Časové úvahy o naší mateřštině*, 42–67. Prague: Náklad jednoty československých matematiků a fysiků.

Goldberg, Adele. 2011. Corpus evidence of the viability of statistical preemption. *Cognitive Linguistics* 22. 131–153.

Janda, Laura A. & Tyers Francis. 2021. Less is more: Why all paradigms are defective, and why that is a good thing. *Corpus Linguistics and Linguistic Theory* 17. 109–141.

Jazyková poradna Ústavu pro jazyk český. 2008-2024. *Internetová jazyková příručka [The Internet Language Reference Book]*. Ústav pro jazyk český.

Juge, Matthew. 2000. On the rise of suppletion in verbal paradigms. *Proceedings of the 25th Annual Meeting of the Berkeley Linguistics Society*, 183–194. Berkeley, CA: Berkeley Linguistics Society.

Kopecký, Leontij Vasiljevič & Oldřich Leška. 1978. *Rusko-český slovník [Russian-Czech Dictionary]*. Prague: Státní pedagogické nakladatelství.

Kováříková, Dominika & Oleg Kovářík. 2021. *GramatiKat: A tool for research into grammatical categories and grammatical profiles*. Prague: Faculty of Arts, Charles University. https://www.korpus.cz/gramatikat.

Kováříková, Dominika, Michal Škrabal, Václav Cvrček, Lucie Lukešová & Jiří Milička. 2020. Lexicographer's lacunas, or how to deal with missing representative dictionary forms on the example of Czech. *International Journal of Lexicography* 33. 90–103.

Křen, Michal, Václav Cvrček, Tomáš Čapka, Anna Čermáková, Milena Hnátková, Lucie Chlumská, Tomáš Jelínek, Dominika Kováříková, Vladimír Petkevič, Pavel Procházka, Hana Skoumalová, Michal Škrabal, Petr Truneček, Vondřička Pavel & Zasina Adrian. 2015. *SYN2015: A representative corpus of written Czech*. Prague: Institute of the Czech National Corpus, Faculty of Arts, Charles University. https://www.korpus.cz.

Křen, Michal, Václav Cvrček, Jan Henyš, Milena Hnátková, Tomáš Jelínek, Jan Kocek, Dominika Kováříková, Jan Křivan, Jiří Milička, Vladimír Petkevič, Pavel Procházka, Hana Skoumalová, Jana Šindlerová & Michal Škrabal. 2020. *SYN2020: A representative corpus of written Czech*. Prague: Ústav českého národního korpusu FF UK.

Langacker, Ronald. 2019. Morphology in cognitive grammar. In Jenny Audring & Francesca Masini (eds.), *The Oxford Handbook of morphological theory*, 346–364. Oxford: OUP.

Naranjo, Matías Guzmán, & Bonami Olivier. 2021. Overabundance and inflectional classification: Quantitative evidence from Czech. *Glossa* 6(1). 88.

Nichols, Johanna & Alan Timberlake. 1991. Grammaticalization as retextualization. In Elizabeth C. Traugott & Bernd Heine (eds.), *Approaches to grammaticalization, vol. I: Focus on theoretical and methodological issues*, 129–146. Amsterdam and Philadelphia: John Benjamins.

Nikolaev, Alexandre & Neil Bermel. 2022. Explaining uncertainty and defectivity of inflectional paradigms. *Cognitive Linguistics* 33. 585–621.

Nikolaev, Alexandre & Neil Bermel. 2023. Studying negative evidence in Finnish language corpora. *Word Structure* 16. 206–232.

Nykysuomen sanakirja [Dictionary of Modern Finnish]. 1951–1961. Helsinki: WSOY.

Prince, Alan & Paul Smolensky. 1993. *Optimality theory: Constraint interaction in generative grammar*. Boulder: Rutgers University and University of Colorado.

R Core Team. 2021. *R: a language and environment for statistical computing*. Vienna, Austria: R Foundation for Statistical Computing. Available at: https://www.R-project.org/.

RDCL (Reference Dictionary of the Czech Language). *Příruční slovník jazyka českého*. 1935–1957. Prague: Various Publishers. Digitalized version of 2007. Available at: http://bara.ujc.cas.cz/psjc/.

Schmid, Hans-Jörg. 2015. A blueprint of the 'Entrenchment-and-conventionalization' model. In Beate Hampe & Anja Binanzer (eds.), *Yearbook of the German cognitive linguistics association*, 3–26. Berlin: Mouton de Gruyter.

Sgall, Petr, Jiří Hronek, Alexandr Stich & Ján Horecký. 1992. *Variation in language: Code-switching in Czech as a challenge for sociolinguistics*. Amsterdam and Philadelphia: John Benjamins.

Sims, Andrea. 2009. Why defective paradigms are, and aren't, the result of competing morphological patterns. *Proceedings of the 43rd annual meeting of the Chicago Linguistic Society* 43(2). 267–281.

Sims, Andrea. 2015. *Inflectional defectiveness*. Cambridge: CUP.

Strossa, Petr. 2015. The text frequency of Czech noun declension patterns. *Journal of Quantitative Linguistics* 22(4). 273–288.

Suomen kielen perussanakirja. 1990–1994. *[Basic Dictionary of Modern Finnish]*. Helsinki: Edita Oyj.

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

Vaux, Bert. 2008. Why the phonological component must be serial and rule-based. In Bert Vaux and Andrew Nevins (eds.), *Rules, constraints, and phonological phenomena*, 20–60. Oxford: OUP.
Yang, Charles. 2016. *The price of linguistic productivity: How children learn to break the rules of language*. Cambridge MA/London: The MIT Press.

**Supplementary Material:** The online version of this article offers supplementary material (https://doi.org/10.1515/cog-2023-0032).