



# Overabundance and inflectional classification: Quantitative evidence from Czech

**MATÍAS GUZMÁN NARANJO** 

**OLIVIER BONAMI** 

*\*Author affiliations can be found in the back matter of this article*

**RESEARCH**

**]u[ubiquity press**

## Abstract

Overabundance is the situation where two or more distinct word forms fill the same cell in an inflectional paradigm (Thornton 2011). While this topic has received renewed attention in recent years, there are still several open questions regarding its properties and status. In this paper we present a new take on the matter. On the basis of a case study of the locative singular and instrumental plural of Czech nouns, we argue that there are at least two kinds of overabundance phenomena which should be distinguished, depending on whether overabundant behavior integrates in the inflection system or is orthogonal to it. The evidence for the distinction comes from a quantitative study of the way phonological, morphosyntactic, semantic, and sociolinguistic factors contribute to partially predicting whether a lexeme is overabundant and which form is used in different contexts.

**CORRESPONDING AUTHOR:**

**Matías Guzmán Naranjo**

Eberhard Karls Universität  
Tübingen, Wilhelmstr. 19,  
72074 Tübingen, Germany  
[mguzmann89@gmail.com](mailto:mguzmann89@gmail.com)

**KEYWORDS:**

overabundance; analogical  
classification; inflection  
classes; Czech declension

**TO CITE THIS ARTICLE:**

Guzmán Naranjo, Matías  
and Olivier Bonami. 2021.  
Overabundance and  
inflectional classification:  
Quantitative evidence from  
Czech. *Glossa: a journal of  
general linguistics* 6(1): 88.  
1–31. DOI: [https://doi.  
org/10.5334/gjgl.1626](https://doi.org/10.5334/gjgl.1626)

Overabundance is the situation where two or more distinct wordforms fill the same cell in an inflectional paradigm (Thornton 2011). In (1) we see examples for the imperfective subjunctive in Spanish, which can be realized by either *-se* or *-ra* markers (DeMello 1993).

- (1) a. *cantara*  
sing.SBJV.IMP.3SG
- b. *cantase*  
sing.SBJV.IMP.3SG  
'That I sing.'

While the phenomenon is well known and documented in many if not all languages with inflectional morphology, overabundance was mostly ignored by theoretical morphologists until the pioneering work of Thornton; it is telling that prominent theoretical works such as Anderson (1992) and Stump (2001) define architectures for inflectional morphology that presuppose overabundance not to exist, without any explicit discussion (Bonami & Boyé 2010). Although Thornton's efforts in the last decade (Thornton 2011; 2012; 2019a;b) succeeded in putting the problem on the agenda, leading to a number of theoretical discussions (Stump 2016; Bonami & Crysmann 2018; Guzmán Naranjo 2019; Beniamine 2021) and renewed interest in detailed empirical studies (see among many others Bošnjak Botica & Hržica 2016; Cappellaro 2013; Lečić 2015; Rosemeyer & Schwenter 2019; Santilli 2014; Thornton 2012), some more general questions still remain unanswered. The clarified empirical landscape allowed Thornton (2019b) to start laying out a typology of overabundance. She identifies four main dimensions of variation in how overabundance manifests itself, which we may describe as follows.<sup>1</sup>

- (2) a. **Lexical prevalence:** an overabundance phenomenon may affect a set of lexemes of any size, from a single lexeme to all members of the same part of speech.
- b. **Paradigmatic prevalence:** an overabundance phenomenon<sup>2</sup> may affect a set of paradigm cells of any size, from a single cell to the whole paradigm.
- c. **Balance:** the statistical distribution of rival forms may vary anywhere from a balanced distribution to a situation where the use of one of the two forms is barely attested.
- d. **Conditions:** the use of rival forms may be subject to various kinds of conditions:
- (i) **Usage conditions:** geographical, sociolinguistic, and/or stylistic factors affect the preference for one or the other form.
- (ii) **Grammatical conditions:** the semantic, syntactic, morphological and/or phonological environment affects the preference for one or the other form.

One aspect of the typology of overabundance that Thornton does not discuss in detail is its interaction with the system of inflectional classification. Inflectional systems of any complexity exhibit differential inflectional behavior, where lexemes of the same part of speech use different marking strategies to contrast the forms filling cells of their inflectional paradigm. Systems of inflection classes are the tool of choice to explicate such variability, and recent research has highlighted how such systems are organized (Corbett & Fraser 1993; Dressler & Thornton 1996; Brown & Hippisley 2012; Beniamine, Bonami & Sagot 2017; Beniamine 2021) and how they tend to be partially but not fully motivated by other lexical properties (Aronoff 1994; Baayen & Moscoso del Prado Martín 2005; Guzmán Naranjo 2019). Overabundance may interact with

<sup>1</sup> Thornton's typology is stated in terms of canonical criteria (Corbett 2007; Brown, Chumakina & Corbett 2013), and focuses on endpoints of the dimensions rather than describing the dimensions directly. We took the liberty of rephrasing Thornton's distinctions in terms that highlight the gradual nature of the scales rather than the endpoints.

<sup>2</sup> This discussion is affected by what exactly one calls a single 'overabundance phenomenon'. A strict definition classifies two instances of overabundance as the same phenomenon only if they exhibit the same form alternation, modulo regular morphophonology. Under this definition, Czech LOC.SG pairs *listu~listě* 'page' and *bazénu~bazěně* 'swimming pool' are instances of the same phenomenon, but the pair *hostu~hostovi* 'host' represents a distinct phenomenon. Using this strict definition, paradigmatic prevalence will generally be low, because it is rare for the same alternations to occur in multiple cells. Thornton however uses a more permissive definition when discussing paradigmatic prevalence, and just counts how many cells in the paradigm of the same lexeme are overabundant, whether the alternation is the same or not. This is clearly an area where the typology would benefit from being refined. In this paper we will alternate between these two definitions depending on context, hoping that it will make the text more readable without introducing much confusion.

inflectional classification in a variety of ways. In the extreme case of systematic overabundance in Spanish imperfective subjunctives illustrated in (1), there is no interaction to speak of, since all lexemes are overabundant and overabundance manifests itself through the use of the exact same exponents across the lexicon. However this is not the only possibility. Even where overabundance is systematic, it may rely on different marking strategies depending on the inflection class. Where overabundance is found with a restricted set of lexemes, it interacts by definition with inflectional classification (it leads to differential inflectional behavior), but there are different conceivable ways in which it may do so. In particular, we may ask whether overabundant classes have the usual properties of inflection classes in terms of partial motivation.

In this paper we present a case study of two situations of overabundance in Czech nominal declension: occasional overabundance in the locative singular, and systematic overabundance in the instrumental plural. We deploy various quantitative techniques applied to lexical and corpus data to show how overabundance is embedded in the inflection class system in the first case, but orthogonal to that system in the second.

The structure of the paper is as follows. In Section 2 we present background information on the Czech declension system and how it is affected by overabundance. Section 3 presents a first study arguing for a qualitative difference between the two cases of overabundance: building on previous work on inflectional classification, we show that overabundant lexemes exhibit a specific pattern of partial motivation in the locative singular, suggesting that overabundant lexemes constitute a *mixed* class sharing properties with two classes of non-overabundant lexemes. By contrast, no such effect can be found in the instrumental plural. Section 4 presents a complementary study of the relationship between overabundance and case government in the locative singular. We document the fact that governing prepositions have preferences as to which variant of an overabundant lexeme they combine with, although no such effect can be found with non-overabundant lexemes. This indicates that, despite their mixed status in terms of motivation, overabundant lexemes form a class whose properties are not reducible to those of its non-overabundant neighbors. Hence they constitute a robust member of the inflection class system. Section 5 concludes the paper.

## 2 Overabundance in Czech nominal declension

### 2.1 The nominal declension system

For the purposes of this paper, we will follow the description of Czech inflection in Cvrček et al. (2010), a careful revision of traditional descriptions based on extensive corpus evidence. This grammar uses evidence from corpora of edited text vs. spoken corpora to document in parallel the two language standards otherwise known as ‘Literary Czech’,<sup>3</sup> mostly used in formal writing, and ‘Common Czech’, mostly used in speech or informal contexts. These differences are only marginal in nominal declension, except in the case of the instrumental plural, as discussed below.

The Czech nominal system distinguishes four grammatical genders (masculine inanimate, masculine animate, feminine and neuter),<sup>4</sup> seven cases (nominative, accusative, genitive, dative, vocative, locative, and instrumental) and two numbers (singular and plural). Nouns are divided up into declension classes which characterize distinct inflectional behaviors. [Table 1](#) illustrates the 12 most prominent classes of Czech nouns according to Cvrček et al. (2010).<sup>5</sup> The

<sup>3</sup> Sometimes also called ‘Standard Czech’, e.g. by Bermel (2000). See this monograph for a useful history of the codification of the distinction, and discussion of its complex relationship to actual sociolinguistic variation.

<sup>4</sup> Masculine animate and masculine inanimate are separate genders since they trigger different agreement patterns; cf. *vidím star-ého muže* ‘I see an old man’ vs. *vidím star-ý kříž* ‘I see an old cross’. Whether they should be considered subgenres of a superordinate masculine gender, in the sense of (Corbett 1991; 2012), is a separate issue. We note that the evidence for this is weaker than in other Slavonic languages, with multiple case-number combinations in agreement targets distinguishing masculine animate from masculine inanimates (ACC.SG, for all adjectives, NOM.PL and VOC.PL for hard adjectives, PL for past verbs), and, in those cells, systematic syncretism between masculine inanimate and feminine and/or neuter. Hence, while masculine inanimate agreement is more similar to masculine animate agreement than to feminine or neuter agreement, it is not entirely dissimilar to those, a fact that is not captured by the notion of a subgender.

<sup>5</sup> Deciding on an exact number of inflection classes depends on the details of criteria for inflection class membership, a notoriously thorny issue; see Beniamine, Bonami & Sagot (2017) for recent discussion. In particular Czech has a number of alternation phenomena straddling the morphology-phonology interface (epenthetic *e* insertion, *û*–*o* alternations, different varieties of palatalization) whose treatment within or outside the inflection system affects the number of postulated classes.

		MASCULINE ANIMATE			MASCULINE INANIMATE		
		hard	soft	hard	hard	soft	
		HOST 'host'	MUŽ 'man'	TÁTA 'dad'	MOST 'bridge'	KŘÍŽ 'cross'	
SG	NOM	host	muž	tát-a	most	kříž	
	GEN	host-a	muž-e	tát-y	most-u	kříž-e	
	DAT	host-ovi~host-u	muž-ovi~muž-i	tát-ovi	most-u	kříž-i	
	ACC	host-a	muž-e	tát-u	most	kříž	
	VOC	host-e	muž-i	tát-o	most-e	kříž-i	
	LOC	host-ovi~host-u	muž-ovi~muž-i	tát-ovi	most-ě~most-u	kříž-i	
	INS	host-em	muž-em	tát-ou	most-em	kříž-em	
PL	NOM	host-é~host-i	muž-ové~muž-i	tát-ové	most-y	kříž-e	
	GEN	host-ů~host-í	muž-ů	tát-ů	most-ů	kříž-ů	
	DAT	host-ům	muž-ům	tát-ům	most-ům	kříž-ům	
	ACC	host-y	muž-e	tát-y	most-y	kříž-e	
	VOC	host-é~host-i	muž-ové~muž-i	tát-ové	most-y	kříž-e	
	LOC	host-ech	muž-ích	tát-ech	most-ech	kříž-ích	
	INS	host-y~host-ama	muž-i~muž-ema	tát-y~tát-ama	most-y~most-ama	kříž-i~kříž-ema	

		FEMININE			NEUTER			
		hard	soft	soft	neither	hard	soft	neither
		BOTA 'shoe'	RŮŽE 'rose'	KRÁDEŽ 'theft'	KOST 'bone'	MĚSTO 'city'	MOŘE 'sea'	STAVENÍ 'building'
SG	NOM	bot-a	růž-e	krádež	kost	měst-o	moř-e	stavení
	GEN	bot-y	růž-e	krádež-e	kost-í	měst-a	moř-e	stavení
	DAT	bot-ě	růž-i	krádež-i	kost-í	měst-u	moř-i	stavení
	ACC	bot-u	růž-i	krádež	kost	měst-o	moř-e	stavení
	VOC	bot-o	růž-e	krádež-i	kost-í	měst-o	moř-e	stavení
	LOC	bot-ě	růž-i	krádež-i	kost-í	měst-ě~měst-u	moř-i	stavení
	INS	bot-ou	růž-í	krádež-í	kost-í	měst-em	moř-em	stavení-m
PL	NOM	bot-y	růž-e	krádež-e	kost-i	měst-a	moř-e	stavení
	GEN	bot	růž-í	krádež-í	kost-í	měst	moř-í	stavení
	DAT	bot-ám	růž-ím	krádež-ím	kost-em	měst-ům	moř-ím	stavení-m
	ACC	bot-y	růž-e	krádež-e	kost-i	měst-a	moř-e	stavení
	VOC	bot-y	růž-e	krádež-e	kost-i	měst-a	moř-e	stavení
	LOC	bot-ách	růž-ích	krádež-ích	kost-ech	měst-ech	moř-ích	stavení-ch
	INS	bot-ami~ bot-ama	růž-emi~ růž-ema	krádež-emi~ krádež-ema	kost-mi~ kost-ma	měst-y~ měst-ama	moř-i~ moř-ema	stavení-mi~ stavení-ma

**Table 1** Main Czech inflection classes.

Czech grammatical tradition classifies inflectional behavior in terms of two dimensions: gender, and the morphophonological status of the noun stem as being ‘hard’ or ‘soft’ (see e.g. Naughton 2005). The hard vs. soft distinction mostly boils down to a distinction between two classes of stem-final consonants, with some exceptions. Most importantly, as Cvrček et al. (2010) note, listing nouns such as KOST ‘bone’ or STAVENÍ ‘building’ as soft, as is traditional, makes little

sense: /t/ is clearly a hard consonant, and i-stem neuters do not end in a consonant: we will follow their lead and take these two inflection classes to be outside the hard/soft opposition. In addition, the traditional bipartition does not exhaust inflection class distinctions; see e.g. the distinct behavior of HOST ‘host’ and TÁTA ‘dad’ in the masculine animate, or RŮŽE ‘rose’ and KRÁDEŽ ‘theft’ in the feminine. Overall, inflection class fully determines gender (no two nouns of different genders inflect in exactly the same fashion), and correlates strongly with stem-final consonant identity, but inflection class assignment is not fully predictable from gender, morphophonology, or a combination of the two. As a result of these and other observations, the Czech inflection class system is not readily describable in terms of an inheritance tree, and is best viewed as a multiple inheritance hierarchy; see Beniamine & Bonami (submitted) for elaboration of this point.

## 2.2 Overabundance

**Table 1** already illustrates the pervasive presence of overabundance in Czech declension, with 6 out of 14 paradigm cells having multiple forms for at least some nouns in this very small sample. It also illustrates the important fact that overabundant cells typically exploit case-number suffixes also found with non-overabundant lexemes. For instance, the dative singular of HOST has two forms, combining the inflection strategies independently found with TÁTA on the one hand (-ovi) and MOST on the other hand (-u). Finally, it is worth noting that while some cases of overabundance are inflection class dependent, overabundance is systematic in the instrumental plural: all nouns exhibit two distinct marking strategies, one of them involving the vowel /i/ (written as <y> or <i>) potentially preceded by some material, the other the sequence -ma, also potentially preceded by some material.

To get a better grasp of the importance of the phenomenon, we quantified the overall lexical prevalence of overabundance using attestations in corpus data. We used version 4 of the SYN corpus (Křen et al. 2016), a tagged and lemmatized 4.3 billion token corpus of edited text published between 1989 and 2014; see Hnátková et al. (2014) for a detailed description.<sup>6</sup> Note that, this being a corpus of edited text, more informal Common Czech forms are underrepresented in the corpus, although by no means absent, as we will discuss in Section 2.4.

We proceeded as follows. First, for each paradigm cell, we collected all wordforms attested in the corpus and tagged as belonging in that cell, and we noted which lemma they correspond to and the token frequency of that wordform filling that cell of that lemma. **Table 2** reports as ‘attested’ the number of lexemes that are attested at least twice in the relevant cell. Second, we used simple pattern matching to identify the casenumber suffix in each word (if any), relying on the description of exponence provided by Cvrček et al. (2010). A lexeme is counted as ‘overabundant’ in a cell if at least two wordforms are found in that cell ending

	NOM	GEN	DAT	ACC	VOC	LOC	INS
<b>SINGULAR</b>							
Attested	303476	219770	143303	191901	33902	121676	171209
Overabundant	1176	2636	13125	755	88	9027	433
Proportion	0.39%	1.20%	9.16%	0.39%	0.26%	7.42%	0.25%
<b>PLURAL</b>							
Attested	106816	98653	46884	83509	2461	42289	58063
Overabundant	6192	826	1404	650	35	1518	7066
Proportion	5.80%	0.84%	2.99%	0.78%	1.42%	3.59%	12.17%

**Table 2** Overall counts of noun lexemes attested in the SYN corpus in each paradigm cells. The ‘Attested’ row reports the overall number of distinct lexemes that are found in that paradigm cell with a token frequency of 2 or more. The ‘Overabundant’ row reports the number of lexemes among these that are attested in forms using at least two distinct suffixes.

<sup>6</sup> The SYN corpus is the concatenation of a number of smaller corpora that are either representative of the overall production of Czech publishers in a given time period (SYN2000, SYN2005, SYN2010, SYN2015) or exclusively journalistic (SYN2006PUB, SYN2009PUB, SYN2013PUB). Unlike e.g. Bermel & Knittl (2012a) we chose to use the larger, less balanced corpus in the interest of a larger coverage, which is important if we want to be able to assess small proportions of use of alternate forms for a large number of lexemes.

in different suffixes (including the zero suffix). The proportion of overabundant lexemes among attested lexemes arguably provides a lower bound on the actual lexical prevalence of overabundance.<sup>7</sup>

We thus find evidence for overabundance in all paradigm cells, with a proportion of overabundant lexemes varying between 0.25% in the INS.SG and 12.17% in the INS.PL. It is also striking that overabundant lexemes number in the thousands for 8 out of 14 paradigm cells, with the lowest numbers coinciding with paradigm cells that are also the least frequently attested in the corpus (e.g. vocative is barely found in a corpus that contains little dialogue). We conclude that overabundance is overwhelmingly attested in Czech, doing away with any doubt that one would be dealing with a minor phenomenon. It might be that Czech is unusual in that respect, and that the high prevalence of overabundance is linked to the language's particular diglossic history (see Bermel 2000 for a useful discussion). However, since to our knowledge the prevalence of overabundance has never been evaluated on a large scale for any other language at this date, there is currently no evidence to support such a claim.

With these very general observations in hand, we turn to two more specific case studies.

### 2.3 Hard masculine inanimate nouns in the LOC.SG

The locative singular is home to a number of overabundance phenomena. We focus our presentation on the situation of hard masculine inanimate nouns, although similar points could be made about other parts of the system, and we will present some relevant analysis in section 3. Hard masculine inanimate nouns may use two different endings in the LOC.SG: *-u* or *-ě*.<sup>8</sup> Some nouns are attested with both, and are hence overabundant.<sup>9</sup>

- (2) a. DUB 'oak tree', LOC.SG *dubu*  
DŮM 'house', LOC.SG *domě*  
ÚŘAD 'office', LOC.SG *úřadu~úřadě*

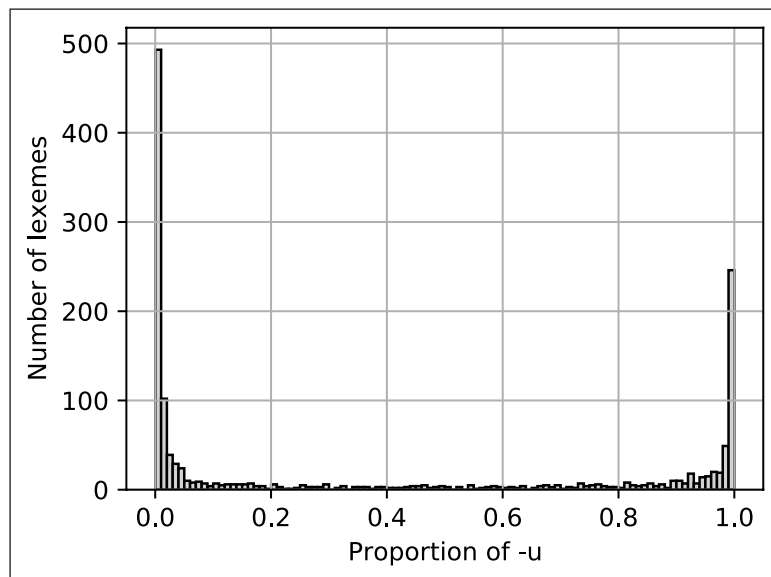
In the SYN corpus we find that, among masculine inanimate nouns attested at least twice, 15959 nouns only appear with *-u*, 1056 only appear with *-ě*, and 2041 are found with both endings. Hence about 10% of the relevant nouns are undisputably overabundant. Sampling accidents may have led to finding attestations of only one of the two forms for other lexemes that are indeed overabundant: hence this 10% proportion should be taken as a lower bound to the true proportion of overabundant lexemes.

7 Our calculations are conservative in at least two ways. First, we only considered (lexeme,cell) pairs with at least two attestations, because if a pair is attested only once, there is no possibility of it having been seen in two distinct forms. But still, the relative frequency of alternants in cases of overabundance is very variable (see Bermel & Knittl 2012a and below). For lexemes for which we have enough attestations to document this, the typical situation is that one of the alternants is much more frequent than the other. As a result, lexemes with a smaller number of attestations that are actually overabundant are likely to be found with only one form in the corpus. Given Zipf's law, this situation is expected to be very common. Second, we purposefully refrained from counting as overabundant all (lexeme,cell) pairs found with two distinct wordforms, and used the more restrictive condition of having distinct suffixes. If we had done the former, we would have counted as cases of overabundance many instances of minute orthographic variation (e.g. the INS.SG of ANALÝZA 'analysis' spelled either *analýzou* or *analyzou*) that are unlikely to be morphologically relevant, and many of which are just spelling errors. Our reliance on the latter strategy avoids that pitfall, but may also lead to excluding some true cases of overabundance involving stem allomorphy.

8 More precisely, one of the options for the exponence of LOC.SG is a morphophonological process that (i) palatalizes the stem-final consonant if that consonant enters palatalization alternations; and (ii) suffixes /e/. Since most consonants end up being palatalized, and Czech orthography mostly notes /e/ preceded by a palatalized consonant as <ě>, a *-ě* ending is the most frequent orthographic reflex of the relevant morphophonological process; the ending may also be *-e*, e.g. after a non-palatalizable consonant (e.g. KOSTEL 'church', LOC.SG *kostele*), or where orthography notes palatalization on the consonant rather than the vowel (e.g. JAZYK 'tongue, language', LOC.SG *jazyce*). All these cases are taken into account below, and for simplicity will be labelled as instances of the *-ě* ending.

9 A reviewer points out the interesting connection between the Czech situation and the phenomenon of second locatives in Russian (Brown 2007; Corbett 2012). While most Russian nouns have a single locative (also known as prepositional) form, some class 1 nouns have a form in *-ú* in addition to their ordinary locative in *-e*, with specialization as to which preposition selects which of the two locative cases. Although the two phenomena may have a common origin and have a strong family resemblance, it is worth pointing out the crucial differences: in Czech, both *-ě* and *-u* are the single available exponent for some nouns of the relevant subclass of hard masculine inanimates; and, where both forms are available with a single noun, there is no complementary distribution in terms of prepositional government, although there are interesting tendencies that we will discuss in Section 4.

To get a better grasp of this situation of overabundance, we examine how the proportion of use of *-u* vs. *-ě* varies across lexemes. [Figure 1](#) shows the distribution of these proportions for lexemes attested at least 100 times in the corpus in the LOC.SG,<sup>10</sup> and at least once with each of the two exponents.



**Figure 1** Histogram of the by-lexeme proportion of use of *-u* in the locative singular for overabundant hard masculine inanimate nouns with a token frequency of 100 or more, in the SYN corpus. Proportions of exactly 0 and exactly 1 are excluded; the first bar (resp. the last bar) hence shows the number of lexemes with strictly more than 0% and at most 1% (resp. at least 99% and strictly less than 100%).

The distribution is strikingly u-shaped: the vast majority of overabundant lexemes exhibit a strong preference for either *-u* or *-ě*. In fact, about half of the relevant lexemes are found 95% of the time or more with one of the two endings, and only 108 (about 5%) have no strong preference, with a proportion of *-u* between 40% and 60%.

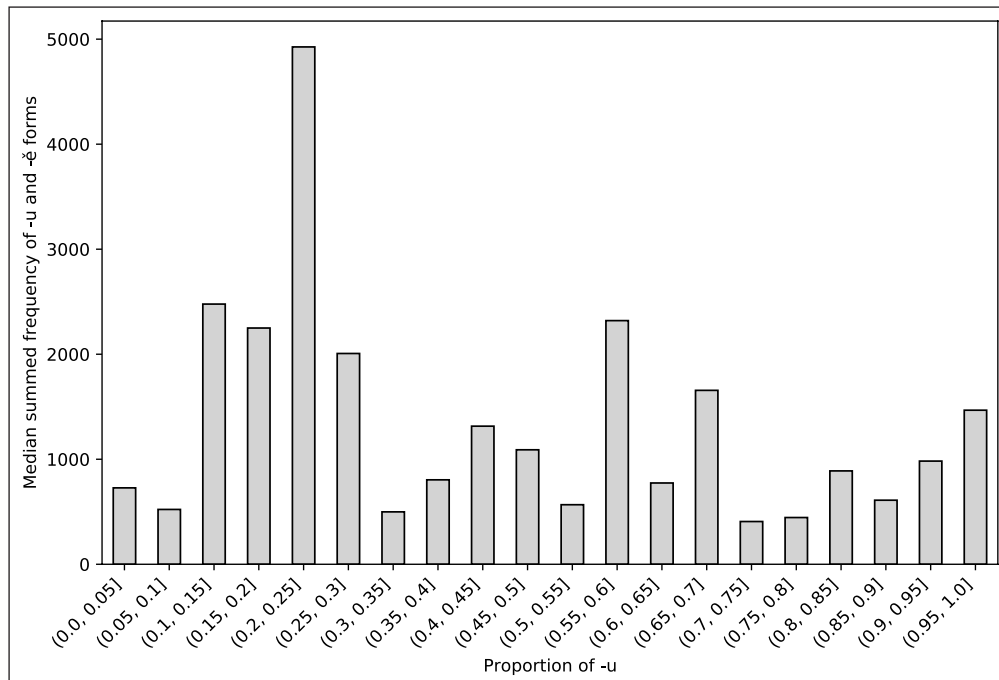
What could be the source of this distribution? Two alternative hypotheses need to be considered. First, the corpus distribution could reflect true lexical variability: each lexeme has its own probabilistic level of preference for *-u* or *-ě*, with most lexemes exhibiting a strong preference for one or the other, whatever the cause of that preference. Alternatively, it could be that the observed distribution is a consequence of noisy data. Although we will conclude that the former is true, it is necessary to take the time to examine the latter hypothesis.

Suppose that each relevant noun truly has a single LOC.SG form, but production errors introduce a bit of random noise: sometimes a speaker will incorrectly inflect a noun with the wrong suffix. For concreteness let us assume that such errors happen 1% of the time. Under that scenario, the observed proportion of use of *-u* for each lexeme corresponds to a different sample of one of two underlying processes. For each of the two processes, most samples will exhibit a proportion of use of *-u* close to the true proportion—by hypothesis, either 1% or 99%—but a few will by chance end up containing a disproportionate proportion of the ‘wrong’ form.

Such a story is appealing, as it explains away apparent overabundance. However, it makes a clear prediction that happens to be falsified. If the hypothesis was true, then the likelihood of a lexeme being seen with a balanced distribution of forms should decrease with the frequency of the lexeme. In other words, lexemes with balanced proportions of *-u* or *-ě* should have a markedly lower frequency than lexemes on the borders. As [Figure 2](#) shows, this is not the case: the median frequency of lexemes with a more balanced distribution is not noticeably lower than that of lexemes with an imbalanced distribution.

In addition to this corpus evidence, Bermel & Knittl (2012a) provide experimental evidence for the same conclusion. In their experiment, speakers were asked to rate the acceptability within a series of syntactic contexts of the two locative singular forms, for nouns with different proportions of *-u* in a corpus. They found that the acceptability of an ending correlates positively with its proportion of use. In particular, lexemes with a balanced use of *-u* and *-ě* do not exhibit a marked preference in acceptability for one or the other ending.

<sup>10</sup> This restriction to higher frequency lexemes is necessary to get a true grasp of the distribution: with a low number of attestations, the estimation of the true proportion is by necessity likely to be false. For instance, an item deemed overabundant and attested twice in the corpus will have a 50% proportion of *-u* in the corpus, but the true proportion in a larger dataset might be vastly different.



**Figure 2** Median frequency of overabundant lexemes in the LOC.SG for different bands of proportion of -u.

Both macroscopic corpus evidence and microscopic experimental evidence thus lead us to conclude that the U-shaped distribution documented in [Figure 1](#) reflects true lexical preferences: most overabundant lexemes have a marked preference for one or the other suffix, but some have a more balanced distribution. Note that, by saying that lexemes have lexical preferences, we do not exclude the possibility that these follow at least in part from general tendencies. The literature on Czech is replete with observations on phonological, morphological, syntactic, semantic, and sociolinguistic factors purported to have an influence—see Cummins (1995) for a review and Bermel & Knittl (2012a;b) as well as Bermel, Luďek Knittl & Russell (2015) and Bermel, Luďek Knittl & Russell (2018) for empirical evidence. In fact, the remainder of this paper will further document such conditioning. The important conclusion for the time being is that overabundance cannot be explained away as a consequence of such factors: there is a robust class of lexemes exhibiting variable inflectional behavior in the locative singular.

## 2.4 The instrumental plural

In Czech, all nouns may occur in two forms in the instrumental plural. Examples of both forms can be seen in (4):

- (3)
- a. MUŽ ‘man’: *muži*~*mužema*
  - b. ŽENA ‘woman’: *ženami*~*ženama*
  - c. MĚSTO ‘town’: *městy*~*městama*

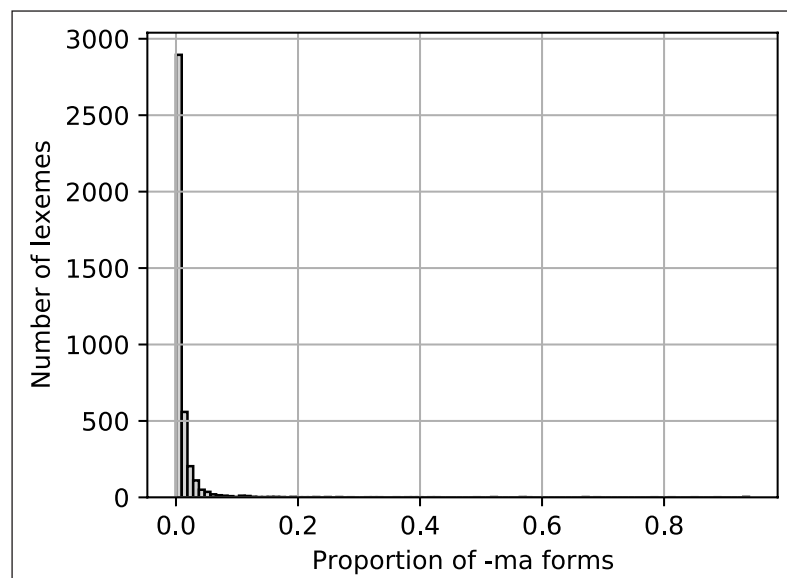
As [Table 1](#) indicates, actual endings vary quite a bit, but can always be distinguished on the basis of whether the ending contains the sequence *-ma* (full ending *-ama*, *-ema*, or *-ma*) or not (*-y*, *-i*, *-ami*, *-emi* or *-mi*). For simplicity we will collectively refer to those as the *-ma* and non-*-ma* endings.

Cummins (2005) provides a useful overview of the historical causes of that situation, from the emergence of *-ma* endings in dialects of Czech in the sixteenth century (a reanalysis of an old dual ending), through its condemnation by early normative grammars and nineteenth century language revivalists, to its role in the codification of Literary Czech and Common Czech in the twentieth century. The alternation between *-ma* and non-*-ma* forms is clearly sociolinguistically conditioned. The *-ma* form is felt as informal, unexpected in writing, and frowned upon in school; it is not listed in most resources providing declension tables, including the *Internetová jazyková příručka* maintained by the Czech Language Institute of the Academy of Sciences of the Czech Republic.<sup>11</sup> On the other hand, the non-*-ma* form is felt as formal, bookish in speech, and the preferred form in schooling. Cvrček et al. (2010) labels the former as spoken forms and the latter as written forms.

<sup>11</sup> <https://prirucka.ujc.cas.cz/>, consulted on February 5, 2021.



The distribution of overabundant instrumental forms the SYN corpus follow the pattern that one would expect given these general observations. Remember that this is a corpus of edited text, comprising press, nonfiction books, and literature. This leads to four expectations. First, we expect *-ma* forms to be rare in that corpus. This is clearly borne out: the overall token frequency of *-ma* forms in the corpus (73,255) is two orders of magnitudes lower than than of non-*ma* forms (17,273,831). Second, we expect most lexemes to be found much more frequently with non-*ma* forms. Again, this is clearly borne out: only 3% of lexemes use the *-ma* form more than 1% of the time, leading to an L-shaped distribution of the proportion of use of the *-ma* forms, shown in [Figure 3](#), that contrasts sharply with the U-shaped distribution found in the locative singular (see [Figure 1](#)).



**Figure 3** Histogram of the by-lexeme proportion of use of *-ma* forms of the INS.PL for overabundant nouns with a token cell frequency of 100 or more, in the SYN corpus

Third, we expect the proportion of use of *-ma* forms to correlate with lexeme-level sociolinguistic properties: lexemes that are more likely to be used in an informal context are also more likely to be used in a *-ma* form. Testing this prediction in detail is beyond the scope of this paper. However it is striking to look at the few lexemes with more than 1000 attestations and a proportion of use of the *-ma* forms above 10%. Of these 11 lexemes, 5 are frequent colloquialisms (KLUK ‘boy’, HOLKA ‘girl’, CHLAP ‘man’, ŽENSKÁ ‘woman’, KÁMOŠ ‘friend’), two are polysemous terms whose relevant attestations have a colloquial secondary meaning (PRÁŠEK ‘powder’, also colloquial term for ‘pill’; KOZA ‘goat’, also colloquial term for ‘breast’), two refer to concepts overwhelmingly discussed in informal settings (CHLUP ‘body hair’, ŠKVAREK ‘greaves’), one is a false positive (SCHOD ‘step’: the vast majority of attestations of *schodama* are in the collocation *Galerie pod schodama*, litt. ‘gallery below the steps’, a fixed proper name). In the end, CHVTLKA ‘short moment’ is the only case where the higher proportion of *-ma* forms does not obviously relate to a lexically-conditioned restriction to informal contexts.

Fourth, we expect the proportion of use of *-ma* forms to correlate with textual genres. Our reference corpus gives us limited access to relevant information in the form of a broad classification of texts. The breakdown is shown in [Table 3](#). As one might expect, literary texts gives rise by far to the highest proportion of *-ma* forms, as these may contain dialogue and/or writing in an informal or speech-like style. Nonfiction, which in this corpus consists mostly of academic writing, is at the other end of the spectrum. Press stands in the middle, with magazines more informal than daily newspaper.

	non- <i>ma</i>	- <i>ma</i>	proportion
Literature	478,363	16,718	3.4%
Magazines	2,082,797	15,981	0.76%
Daily press	14,106,461	43,393	0.31%
Nonfiction	1,511,815	2,788	0.18%

**Table 3** Token frequency of instrumental plural forms by genre in the SYN corpus.

Overall then, the broad distribution of instrumental plural forms in the corpus confirms the observations from the literature.

## 2.5 Summing up

After establishing that overabundance is highly prevalent in Czech nominal declension, we have focused on two particular cases that contrast in multiple dimensions. In the locative singular of hard masculine inanimate nouns, a minority of nouns are overabundant, while in the instrumental plural all nouns are overabundant. Proportions of use of the two forms in the corpus follows a U-shape for the former, and an L-shape in the latter. This is linked to the fact that the choice of form in the instrumental plural is clearly subject to sociolinguistic conditioning, while this is not obviously the case in the locative singular.

In the remainder of this paper we turn to our main topic: how does overabundance interact with inflectional classification? In section 2, we examine the predictability of overabundance: we show that, in the locative singular, the overabundant character of a noun is predictable from its stem shape and distribution, while this is not the case in the instrumental plural. In section 4, we examine the relationship between syntactic usage and overabundance: we show that overabundant locative singular nouns exhibit singular properties that are not found with their non-overabundant counterparts. Both studies lead to the conclusion that some, but not all, overabundance phenomena should be treated in terms of the postulation of a specific overabundant inflection class.

## 3 Predicting overabundance

A well-established property of inflection class systems is that they tend to be partially motivated: while the postulation of inflection classes is justified by the fact that it is not strictly predictable which lexeme will belong to which class (Aronoff 1994), there are typically striking correlations between inflection class assignments and phonological, (morpho)syntactic, and semantic properties of lexemes. This is evident in the traditional description of the Czech declension system above, where stem phonology and grammatical gender were seen as partial predictors of inflection class, with grammatical gender itself being partially predicted by semantic properties such as animacy and social gender.

In this section we rely on this property to explore whether classes of overabundant lexemes should be considered to constitute inflection classes. The reasoning is the following: if the existence of variation (vs. absence of variation) between two exponents for a particular lexeme can be partially predicted from examination of the lexeme's stem phonology, this counts as evidence for this lexeme belonging to a distinct inflection class, as this is the behaviour that is usually seen for non-overabundant classes. We will examine three classes of predictors: aspects of the phonology of the stem, aspects of the distribution of the word in a corpus, and, where relevant, grammatical gender.

### 3.1 Methodology

#### 3.1.1 Computational models

To explore the usefulness of potential predictors, we rely on the notion of analogical classification (Albright & Hayes 2002; 2003; Albright 2009; Arndt-Lappe 2014; Bybee & Slobin 1982; Skousen 1989; Guzmán Naranjo 2019; 2020).

Analogical classification consists in finding the class of some new item, based on the surface similarity of that item to other items whose class is known. The basic idea is that items that look similar on the surface belong to the same class (Blevins, Milin & Ramscar 2017).

From a computational perspective, there are several different techniques one could use for analogical classification. Although these have considerable mathematical differences, and may better or worse performance on different types of data, the final product is conceptually the same: an analogical classifier sees a set of lexemes and their class, and tries to learn the regularities in the surface form of those lexemes which best correlate with that lexeme's class.

In this study we make use of Extreme Gradient Boosting Trees with the package XGBoost (Chen & Guestrin 2016). A boosting tree classifier fits many weak tree classifiers (similar to

decision trees) and then combines them to form a stronger classifier. The principle is similar to that of Random Forests (Breiman 2001), but while a Random Forest fits many small classifiers randomly, a boosting tree classifier fits many small tree classifiers in a guided manner trying to achieve the best accuracy possible.

Our choice of classification method is purely pragmatic. Alternatives like Analogical Modeling (Arndt-Lappe 2011; 2014; Skousen 1989; 2013), or the Minimal Generalization Learner (Albright & Hayes 2002; 2003; Albright 2009), or other machine learning frameworks such as neural networks (Bechtel & Abrahamsen 2002; Churchland 1989; McClelland & Rumelhart 1986; Rumelhart et al. 1986), could be used to the same effect. On a practical level though, boosting trees have several advantages. The main advantage is that boosting trees of this kind can easily handle the type of data in this problem, i.e., categorical predictors with a large number of different levels, while at the same time being computationally efficient. Simpler models like logistic regression tend to over or underestimate the importance of low frequency levels in this kind of data.

Because we want to know whether the model can predict new items instead of just remembering the items it has seen during training, we perform ten-fold cross-validation on every model. This is done by first splitting the dataset into ten groups. The general model is then fitted using nine of the groups as training data, and testing the predictions of the model on the group not used for fitting it. The process is repeated for each of the ten subgroups. This way prediction on all the datapoints is examined while preventing any kind of overfitting (Kohavi 1995).

Although it is possible to look inside the models and see what each predictor is doing with respect to the output classes, this is a tedious process that is not crucial to our purposes in this paper. We are more interested in knowing how well we can predict the inflection classes of the items, rather than exactly knowing which segments correlate with which classes and how.

Instead, we focus on three metrics to evaluate the models: accuracy, no information rate, and kappa score. These metrics are calculated based on a confusion matrix of the model. As an example consider the fictional confusion matrix in [Table 4](#), exhibiting the performance of a classifier on a dataset of 67 items belonging to three classes A, B, and C. The confusion matrix compares predictions of the classifier to the actual, reference classification, by indicating how many members of each actual class (in columns) were predicted to belong to which class (in rows). So for instance, the table reports that, among the twelve items that are truly members of class A, 10 were correctly classified, while 2 were incorrectly classified in B and 1 was incorrectly classified in C.

Reference			
Prediction	A	B	C
A	10	4	5
B	2	20	0
C	1	4	21

**Table 4** Example of confusion matrix.

Accuracy is the number of correct predictions (the sum of the numbers in the diagonal) divided by the total number of items: in our example the accuracy is  $\frac{51}{67} = 0.76$ . The No Information Rate (NIR) is equal to the proportion of the data that belongs to the largest class: it is the best guess one could make in the absence of any predictive information. In our example the largest class is class B with 28 members, hence the NIR is  $\frac{28}{67} = 0.42$ . Comparing the accuracy and NIR is crucial to assessing performance: the same accuracy value may be very impressive if it is much higher than the NIR, or not at all if it is close to (or even smaller than) the NIR. For a statistically meaningful comparison, we report a 95% uncertainty interval for the accuracy value,<sup>12</sup> which reflects uncertainty about the estimation of accuracy related to the size of the

<sup>12</sup> We calculated all uncertainty intervals with a Bayesian Binomial model with mildly informative priors, using Stan (Carpenter et al. 2017; Gelman, Lee & Guo 2015) and the brms interface (Bürkner et al. 2017). Uncertainty intervals (also called credible intervals) are similar to confidence intervals but their interpretation is more straightforward and intuitive: the 95% uncertainty interval is the interval within which the value of a parameter of interest (in this case the accuracy of the classifier in the whole population) falls with 95% probability.

dataset: for a given accuracy value, the larger the dataset, the smaller the uncertainty interval. In our example, the uncertainty interval for accuracy is (0.64, 0.84): hence while we should not be confident about the accuracy value up to a percentage point, we can be confident that it is higher than the NIR, that is, that the classifier performs better than chance.

The kappa statistic gives a value between 0 and 1 measuring the performance of a classifier by comparing the observed accuracy with the expected accuracy (under random chance). The reason for using kappa in addition to raw accuracy, is that accuracy can be skewed in cases with unbalanced classes. In our example, the kappa is 0.64, indicating good though by no way perfect performance of our classifier.<sup>13</sup>

Since our aim is not to test a hypothesis regarding any specific predictor, we do not perform any sort of significance testing. The evaluation metrics we use tell us how well the model performs as a whole, and not whether any specific predictor had a measurable impact. Comparing the observed accuracy to the no information rate lets us know that our model is performing above simply guessing the largest class, and the kappa statistic tells us how much better than random chance our model is doing.

### 3.1.2 Predictors

Our goal is to assess whether and how a lexeme's inflection class can be predicted from other properties of that lexeme. In this paper we use two types of predictors: stem phonology, and distributional vectors.

There are many different aspects of stem phonology that could be used as predictors for our classifiers, and many different ways of coding them up. For instance, we could imagine that identification of initial or final segments, initial or final syllables, word length, or the makeup of the word in terms of biphones or triphones (Baayen, Chuang & Blevins 2018) are possibly relevant. In this paper we take a pragmatic approach to the issue, and rely on prior knowledge of the Czech system to guide a choice of simple predictors. First, we rely on orthography rather than an explicit phonemic transcription. This should not lead to any major loss in accuracy, given that the grapheme-to-phoneme relation is fairly transparent in Czech.<sup>14</sup> Second, as segmental predictors we only use the three last characters of the orthographic stem. This is certain to capture the expected main effects of final consonants, and keeps the number of predictors at a manageable size. Note that stems were obtained by cutting off case-number suffixes as documented by Cvrček et al. (2010) from the words under examination. As a result, the stem allomorph used in the word under examination was considered, rather than the stem allomorph of the citation form, where these differ. This should have no major effect on the results, since stem allomorphy is fairly limited in Czech. Finally, in addition to segmental predictors, word length in syllables was approximated by the number of vowels in the stem. Again, this is a fairly reasonable approximation, as diphthongs are not very prevalent, and no vowel is coded in the orthography as a digraph.

In addition to stem phonology, we used distributional vectors to provide information about the context of use of words of interest. Distributional vectors provide a multidimensional representation of the distribution of words in a corpus, such that words with a similar distribution have similar vectors, and different dimensions of the vectors represent different aspects of distributional similarity. Advances in corpus size, computing power, and inference algorithms have made distributional vector spaces a standard tool of the trade in computational linguistics, allowing various systems to take into account lexical properties in a generic manner (Camacho-Collados & Pilehvar 2020). In the context of general linguistics,

---

<sup>13</sup> In the following, all metrics are calculated on the aggregated results of all cross-validation steps.

<sup>14</sup> Rare opacities result from recent borrowings whose orthography was not adapted, e.g. *e-mail*, pronounced [i:mejɫ] instead of the expected [ɛmajɫ]. Note that Czech orthography makes use of digraphs (for instance *ch* notes [x], and palatalization of consonants is often noted on the following vowel), but this does not lead to opacity in the grapheme-to-phoneme direction. Also note that there is a significant amount of opacity in the phoneme-to-grapheme direction, most prominently because of the use of the two letters <*i*> and <*y*>, which note the same sounds (short [ɪ] or long [i:]). After some consonants, <*i*> indicates palatalization of the preceding consonant, but this is not systematic. As a result there are many pairs or words that are orthographically distinct but phonetically undistinguishable, such as masculine and feminine plurals of past verb forms, e.g. *mluvili* 'they (masc.) spoke' vs. *mluvily* 'they (fem.) spoke'. This is not of concern to us here as we are approximating phonology by orthography rather than the other way around.

distributional vectors are typically used as a way of approximating lexical semantics, in accordance with the distributional hypothesis (see Boleda 2020 and references therein), according to which words with similar distributions are semantically similar. In particular, a growing subliterature uses distributional vectors to study the semantic effects of derivational morphology (see e.g. Marelli & Baroni 2015; Varvara 2017; Lapesa et al. 2018; Huyghe & Wauquier 2020). However, in the present context, it is important to remember that lexical semantics *stricto sensu* is only part of what distributional vectors capture. In particular, when using vectors based on wordforms, morphosyntactic contrasts such as grammatical gender or case government will have an effect on what the vectors look like (Bonami & Paperno 2018): for instance, the grammatical gender of nouns will be coded by distributional vectors, even where it has no semantic reflex, because gender will trigger agreement and hence a different distributional environment for the noun. Likewise, sociolinguistic contrasts between lexemes are likely to lead to contrasting vectors, as words used by different speakers in different circumstances are likely to co-occur with other words subject to the same sociolinguistic restrictions.

For the purposes of this paper, we derived a 300 dimension distributional vector space from version 4 of the SYN corpus also used for all other aspects of our study. We used the Gensim (Řehůřek 2010) implementation of the SkipGram variant of the word2vec algorithm (Mikolov et al. 2013), using the following hyperparameters: 9 training epochs, 20 negative samples, window size 20. Importantly, our vectors are based on lexemes rather than individual wordforms: we used the lemmatization provided with the corpus to derive a version of the corpus where individual words are replaced by their lemmas, and then built the vector space on the basis of that version of the corpus, abstracting away from inflectional variation. This is appropriate in the present context: we hope our classifiers to be able to predict whether a lexeme is overabundant, and overabundance is inherently a property involving multiple wordforms, so that it would make little sense to predict that from properties of a single word. However, it is important to keep in mind that a side effect of this decision is to eliminate part of the distributional variation. For instance, the effects of grammatical gender on vectors for nouns will be dampened, as different forms of agreeing adjectives and verbs will be lumped into a single lemma; on the other hand, broader consequences of semantically-motivated gender assignment leading to collocation with different content words may still be captured. Likewise, the vectors cannot capture directly distributional differences between forms of the same lexeme typically used in a formal vs. informal context (as is expected for the contrasting instrumental plural forms), as these will be mapped to the same vector; but they can still capture differences between lexemes that are on the whole used more in collocation with other lexemes that are markers of formality or informality.

## 3.2 Results

### 3.2.1 Hard masculine inanimate locative singulars

**Table 5** shows the type frequency of hard masculine inanimate nouns attested in the corpus at least 20 times with the *-u* ending, the *-ě* ending, or both. Two remarks are in order about these figures. First, there is a strong imbalance between classes, with the *-u* class an order of magnitude larger than the other two. This is a problem for modeling: if classes are too imbalanced, the model will tend to rely on raw frequency rather than predictor variables to make predictions. Second, for lexemes only found in one of the two forms, we cannot be certain that the other form is impossible. The likelihood of such errors is high, given that, as we saw in **Figure 1**, most overabundant lexemes have a strong preference for one or the other variant. To take an extreme example, if the true proportion of use of *-u* for a lexeme is 90% and we have only two occurrences in our corpus, there is an 81% probability that both will be in *-u*, despite the fact that the lexeme is overabundant. To mitigate these two problems, we selected the 600 most frequent lexemes for each class.

	ě	ě~u	u
All nouns	643	1766	7059
Normalized frequencies	600	600	600

**Table 5** Type frequency of *-ě*, *-ě~u* and *-u* classes for the locative singular in the SYN corpus.

We fit three distinct models to this dataset: a model with just the phonological predictors, collectively labelled ‘shape’ predictors, a model with just the distributional vectors as predictors, and a model with both. [Table 6](#) reports the performance of the three models. Note that the No Information Rate (NIR) is  $\frac{1}{3}$  in all cases, as by design the three classes have the exact same type frequency of 600.

Model	Accuracy	95% uncertainty interval	NIR	Kappa
distribution + shape	0.86	(0.85, 0.88)	0.33	0.8
shape + only	0.75	(0.73, 0.77)	0.33	0.63
distribution + only	0.81	(0.79, 0.83)	0.33	0.72

**Table 6** Performance of classifiers for hard masculine inanimate locative singular nouns.

The overall observation is that all three models perform considerably better than chance, although they do not reach a spectacular level of accuracy. Hence it is clear that assignment of lexemes to one of the three classes is not fully arbitrary. It does not look to be fully predictable either—at least the predictors used in this study are far from ensuring fully accurate prediction. We are thus in the typical grey zone of inflection class assignment being partially predictable.

The ‘shape only’ and ‘distribution only’ models reach comparable levels of performance. Combining the two sets of predictors leads to a barely measurable increase in accuracy compared to having one set of predictors only. These observations strongly suggest that, while phonology and distribution both contribute to predicting inflectional behavior, they do not tend to make complementary contributions where one set of predictor helps when the other fail.

To get a more detailed look at what is going on, we now examine the confusion matrix for the combined model, shown in the left hand part of [Table 7](#). Two observations are in order here. First, performance is highest on the *-ě* class (95% correctly classified), followed by the *-u* class (88%), followed by the overabundant class (76%). This suggests that lexemes forming their locative in *-ě* only are more cohesive in their phonological and distributional properties than those that can or must use *-u*. Second, most of the confusion arises between the overabundant class and the two other ones: there are very few situations where the model wrongly assigns *-u* as a unique exponent instead of *-ě* (<1%) or the other way around (0%). The model also rarely assigns to the overabundant class a lexeme found only with *-ě* in the corpus (<4%). However, about 25% of lexemes found only with *-u* in the corpus are wrongly assigned to the overabundant class; and a sizeable subset of undisputably overabundant lexemes are wrongly associated by the model with only *-ě* (2%) or only *-u* (22%).

Shape and distribution				Shape only			Distribution only				
Pred.	Reference			Pred.	Reference		Pred.	Reference			
	<i>-ě</i>	<i>-ě~-u</i>	<i>-u</i>		<i>-ě</i>	<i>-ě~-u</i>		<i>-u</i>	<i>-ě</i>	<i>-ě~-u</i>	<i>-u</i>
<i>-ě</i>	572	10	0	<i>-ě</i>	551	169	6	<i>-ě</i>	570	7	0
<i>-ě~-u</i>	26	457	75	<i>-ě~-u</i>	44	316	108	<i>-ě~-u</i>	27	401	108
<i>-u</i>	2	133	525	<i>-u</i>	5	115	486	<i>-u</i>	3	192	492

**Table 7** Predictions for hard masculine inanimate locative singular nouns.

Examination of the confusion matrices for the two other models reveals a broadly similar picture. Only two differences are worth mentioning. First, the two kinds of predictors seem to differ in how they deal with lexemes that are truly overabundant (middle column): the model based on phonology alone has more of a tendency to conclude that they are instances of *-ě* only, while the model based on distributional vectors alone has more of a tendency to conclude that they are instances of *-u* alone. The combined model manages to build on both kinds of predictors to achieve better performance on this part of the dataset.

How can we explain the patterns of errors we just observed? For this we must distinguish errors on the middle row from errors on the middle column. On the middle row, the errors correspond to cases where the model predicts a lexeme to be overabundant, while it is found

only in one form in the corpus. As we discussed above, that situation is likely partly due to sampling accidents: by chance, some lexemes that are truly overabundant are only found with one of their two forms in the corpus. Unfortunately, there is no direct way of testing what proportion of the errors is due to such accidents: we just do not have a larger sample to make that evaluation. Importantly however, such an explanation does not hold for items in the middle column: for these we do have attestations for both forms, and hence have no hesitation as to what class they belong to. Hence the fact that there is a nontrivial amount of error here is revealing on the nature of the system.

We submit that this pattern justifies seeing the relevant class of overabundant lexemes as a mixed inflection class: an inflection class that is distinct from both the *-u* class and the *-ě* class, but that still has properties that are intermediate between those of its two corresponding single exponent class. This is not a new idea: see in particular Beniamine (2021), Bonami & Crysman (2018) and Guzmán Naranjo (2019) for different takes on overabundant inflection classes as mixes of other classes. What is specific to this study is that we argue for this mixed inflection class status on the basis of partial motivation: overabundant lexemes stand between two inflection classes in terms of predictability of their inflectional behavior from their phonological and distributional properties.

### 3.2.2 The complete system of locative singulars

We now turn to an examination of the complete system of 37656 lexemes attested at least 20 times in the locative singular. The full set of exponents that we expect to encounter is as indicated in (5).

- (5) Locative singular exponents
- a. Ordinary nouns:
    - (i) *-u* with hard masculines and neuters,
    - (ii) *-ě* with hard masculine inanimates, feminines and neuters,
    - (iii) *-ovi* with hard or soft masculine animates,
    - (iv) *-i* with soft nouns.
    - (v) No exponent *í*-stem neuters.
  - b. Nouns converted from adjectives:
    - (i) *-ém* with hard masculines and neuters, in formal contexts,
    - (ii) *-ým* with hard masculines and neuters, in informal contexts,
    - (iii) *-é* with hard feminines, in formal contexts,
    - (iv) *-ý* with hard feminines, in informal contexts,
    - (v) *-m* with soft masculines and neuters.
    - (vi) No exponent with soft feminines.
  - c. No exponent with undeclinable nouns.

As the description suggests, there are multiple situations of potential overabundance: between *-u* and *-ě*, *-u* and *-ovi*, *-i* and *-ovi* for ordinary nouns; between *-ém* and *-ým*, *-é* and *-ý* for converted adjectives; and finally, a small number of nouns exhibit fluidity between genders, hard or soft status, or declinable vs. undeclinable status. As a result, we find evidence in the corpus for 8 non-overabundant behaviors as well as 12 overabundant behaviors, as indicated in [Table 8](#).

A detailed analysis of this complex and heterogeneous dataset is beyond the scope of this paper. However, it is worth examining the overall performance of a classifier applied to this 20 class system. We thus fit classifiers with similar characteristics to those discussed in Section 3.2.1, with two differences. First, we did not perform any type frequency normalization, as the smaller classes just do not have enough members for that to be possible. And second, we also included gender as a predictor, which was irrelevant as long as we were focussing on a subclass of masculine inanimates. Taking into account all possible combinations of three sets of predictors gives us seven models in total.

As indicated [Table 9](#), the performance of these classifiers is remarkably high, given the number of different classes. First, any of the three sets of predictors is highly relevant on its own, moving the accuracy from a baseline of 0.24 to at least 0.56. Second, when taken separately, stem shape is clearly the most relevant predictor, followed by gender and then distribution. Third, shape

Exponents	Frequency	Exponents	Frequency
-u only	9172	-ovi--u	2429
∅ only	7797	-ě--u	2097
-ě only	7012	-i--ovi	449
-i only	5853	-é--ý	143
-ovi only	1475	-ě--ovi	126
-é only	408	-ém--ým	74
-ém only	245	∅~-u	73
-m only	101	-ě~-i	64
		-i~-u	44
		-ím~∅	61
		∅~-ovi	22
		-ě~∅	11

**Table 8** Type frequency of inflectional behaviors in the locative singular for lexemes with at least 20 attestations.

Model	Accuracy	95% uncertainty interval	NIR	Kappa
shape + distribution + gender	0.88	(0.88, 0.88)	0.24	0.85
shape + gender	0.85	(0.85, 0.86)	0.24	0.82
shape + distribution	0.76	(0.76, 0.77)	0.24	0.71
distribution + gender	0.74	(0.74, 0.75)	0.24	0.69
shape	0.69	(0.69, 0.69)	0.24	0.61
gender	0.61	(0.61, 0.62)	0.24	0.53
distribution	0.56	(0.56, 0.57)	0.24	0.46

**Table 9** Performance of classifiers for all locative singular nouns.

and gender together allow a very high level of predictability, with a classification accuracy 0.85. This is what we expect given the traditional description of the inflection class system. However, the addition of distribution to these two predictors still allows for a measurable 0.3 increase in accuracy. This is a strong indication that either lexical semantics or other lexical characteristics reflected in distribution do contribute to predicting inflectional behavior. Overall, the performance of analogical classification on this intricate system of 20 classes confirms inflection class assignment to be highly, although not categorically, predictable.

Returning to overabundance, we confirm on a larger scale the results already highlighted with masculine inanimates. The full  $20 \times 20$  confusion table for the best classifier can be found in the appendix. However that table is quite hard to read given the number of classes and the diversity of errors made by the classifier. Instead, we extracted from this table the numbers corresponding to the three major overabundant behaviors, with a type frequency above 200 in the corpus. These are shown in *Table 10*. As the reader can check, we see again that there is very little confusion between the two classes with a single exponent, whereas there is a sizable amount of confusion between the overabundant class and each of the other two. Hence in all three cases, the classifier is very efficient at distinguishing two classes of lexemes using a single exponent; it also identifies a mixed class (since presence in this class is predictable above chance), but has a harder time distinguishing overabundant lexemes from non-overabundant ones.<sup>15</sup>

<sup>15</sup> In the full table, all three-way comparisons between an overabundant class and the two corresponding non-overabundant ones lead to qualitatively identical results, except for *-é~ý* and *-ém~ým*, which illustrate the same type of sociolinguistically-conditioned overabundance discussed in greater detail for the instrumental plural below, inherited by converted adjectives from their adjectival source.



<b>-ě vs. -u</b>				<b>-u vs. -ovi</b>				<b>-i~-ovi</b>			
Pred.	Reference			Pred.	Reference			Pred.	Reference		
	-ě	-ě~-u	-u		-u	-u~-ovi	-ovi		-i	-i~-ovi	-ovi
-ě	6727	171	2	-u	8837	91	5	-i	5427	129	5
-ě~-u	207	1017	62	-u~-ovi	170	2051	453	-i~-ovi	30	401	12
-u	22	893	8837	-ovi	5	277	996	-ovi	0	181	996

**Table 10** Predictions of the most accurate classifier for three major cases of potential overabundance in the locative singular.

### 3.2.3 Instrumental plurals

We now turn to the instrumental plural. Grammatical descriptions lead us to expect finding the following set of exponents.

- (6) Instrumental plural exponents
- a. Ordinary nouns:
    - (i) In formal contexts:
      - y with hard masculines and neuters,
      - ami with hard feminines,
      - i with soft masculines and neuters,
      - emi with soft feminines,
      - mi with *kost*-type feminines and *í*-stem neuters.
    - (ii) In informal contexts:
      - ama with all hard nouns,
      - ema with all soft nouns,
      - ma with *kost*-type feminines and *í*-stem neuters.
  - b. Nouns converted from adjectives:
    - (i) In formal contexts:
      - ými for hard nouns,
      - mi for soft nouns.
    - (ii) In informal contexts:
      - ýma for hard nouns,
      - ma for soft nouns.
  - c. With undeclinable nouns:
    - (i) No exponent in formal contexts.
    - (ii) Possibly -ama in informal contexts.

Given that there are 7 formal exponents, 4 informal ones, and only one informal strategy corresponding to each formal one, we expect a maximum of 11 behaviors involving a single exponent and 7 involving two, for a maximum of 18 possible classes. *Table 11* shows how this is implemented for the 16,392 lexemes that are attested at least 20 times in the instrumental plural in the SYN corpus. The left-hand part of the table counts all lexemes found with only

Exponents	Frequency	Exponents	Frequency	Exponents	Frequency
-y only	5307	-y~-ama	2603	-ami~-y	26
-ami only	1858	-ami~-ama	1683	-emi~-i	17
-emi only	1618	-emi~-ema	435	-i~-ama	11
-i only	1205	-i~-ema	401	-ami-∅	11
-mi only	439	-mi~-ma	189		
∅ only	404	∅~-ama	91		
-ými only	77				
-ama only	17				

**Table 11** Distribution of inflectional behaviors for the instrumental plural.

one exponent in the corpus. We find that for all 7 formal exponents, but only the most frequent of the 4 informal exponents, namely *-ama*. This is unsurprising: given the general makeup of our corpus, it is unlikely that a lexeme will be found only with an informal variant. The middle part of the table lists lexemes found in each of the 7 expected combinations of an informal and an formal variant. Finally, the right-hand part lists the few unexpected situations of overabundance, due to hesitations on gender, softness, or declinability. These make up less than 0.4% of the dataset under examination.

Given the general description of overabundance in the instrumental plural above, we do not expect these cases of overabundance to interact with the inflection class system: whether a noun is found in one or two forms in the instrumental plural depends on whether that noun is found in the corpus in both formal and informal contexts. Although this might be predictable to some extent from the noun’s distribution, as there are distributional cues to formality, we do not expect gender or stem phonology to have any predictive power.

To test for this empirically, we fit three separate types of models. The first series of models mimics those shown in Section 3.2.2 for the locative singular, and attempt to predict each of the 19 inflectional behaviors found in the corpus from various combinations of stem shape, gender, and distribution. Accuracy is reported in [Table 12](#). Clearly accuracy is a lot lower than in the locative singular, although it is clearly above chance, and each predictor does make a contribution when added to any other combination of predictors.

Model	Accuracy	95% uncertainty interval	NIR	Kappa
<b>shape + distribution + gender</b>	0.72	(0.72, 0.73)	0.32	0.66
shape + gender	0.66	(0.66, 0.67)	0.32	0.58
shape + distribution	0.61	(0.60, 0.62)	0.32	0.5
distribution + gender	0.6	(0.59, 0.61)	0.32	0.5
shape	0.57	(0.56, 0.57)	0.32	0.45
distribution	0.44	(0.44, 0.45)	0.32	0.25
gender	0.44	(0.43, 0.44)	0.32	0.25

**Table 12** Overall Statistics for instrumental plural predictions.

Examination of the full confusion matrix is crucial to making sense of these numbers. [Table 13](#) shows the confusion matrix for the most accurate model, with rows and columns arranged so that each expected overabundant class is next to the class corresponding to a single, formal exponent. It should be clear from the table that the vast majority of the errors are due to the model being unable to predict whether a lexeme will be found with only a formal exponent (e.g. *-y*) or also with the matching informal exponent (e.g. *-y~-ama*): the model seems to be quite accurate at predicting which formal and which informal exponent can be used with a given lexeme, but quite inaccurate at predicting whether multiple forms are attested in the corpus.

To confirm this, we ran two other series of models that aim at separating these two aspects of prediction. First, we constructed a dataset that neutralizes the effects of overabundance by lumping together lexemes found with only one formal exponent and those found with that exponent and the matching informal exponent. For this experiment we dropped the cases of erratic overabundance documented on the right hand side of [Table 11](#), as these cannot be naturally grouped with other classes. [Table 14](#) reports the accuracy of the models, and the confusion matrix for the most accurate model can be found in the appendix.

We find that all models including stem shape as a predictor are very accurate; gender and distribution also have predictive value, although the strength of gender as a predictor is not as strong as one might have expected. Be that as it may, the high level of accuracy reached by this model confirms that predicting which pair of exponents are available for a given lexeme in the instrumental plural is not hard, while it may be hard to predict whether the two members of the pair are attested or just one of the two.

	-y	-y- -ama	-ami	-ami- -ama	-i	-i- -ema	-emi	-emi- -ema	-mi	-mi- -ma	∅	∅- -ama	-ými	-ami- -y	-emi- -i	-i- -ama	-ami-∅	-ama
-y	4850	1415	2	1	78	11	1	0	49	19	59	59	40	15	4	2	1	8
-y--ama	435	1177	1	0	25	40	0	0	5	7	3	19	2	0	0	1	0	1
-ami	0	1	1395	561	0	0	90	30	24	7	11	0	26	5	2	0	3	7
-ami--ama	0	0	447	1111	0	0	36	84	13	34	5	0	4	5	0	0	1	1
-i	14	4	0	0	1064	290	2	0	8	0	0	0	0	0	7	8	0	0
-i--ema	0	2	0	0	35	58	0	0	0	0	0	0	0	0	0	0	0	0
-emi	1	0	5	2	0	0	1475	269	5	0	2	0	0	0	4	0	0	0
-emi--ema	0	0	3	5	0	1	14	52	1	1	0	0	0	0	0	0	0	0
-mi	5	3	0	0	0	0	0	0	325	98	1	0	0	1	0	0	0	0
-mi--ma	0	1	0	0	0	0	0	0	9	23	0	0	0	0	0	0	0	0
∅	2	0	0	0	1	0	0	0	0	0	323	10	0	0	0	0	6	0
∅--ama	0	0	0	0	2	1	0	0	0	0	0	3	0	0	0	0	0	0
-ými	0	0	0	0	0	0	0	0	0	0	0	0	5	0	0	0	0	0
-ami--y	0	0	5	2	0	0	0	0	0	0	0	0	0	0	0	0	0	0
-emi--i	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
-i--ama	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
-ami-∅	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
-ama	0	0	0	1	0	0	0	0	0	0	0	0	0	0	0	0	0	0

Model	Accuracy	95% uncertainty interval	NIR	Kappa
shape + distribution + gender	0.96	(0.96, 0.97)	0.49	0.95
<b>shape + gender</b>	0.97	(0.97, 0.97)	0.49	0.95
shape + distribution	0.82	(0.81, 0.83)	0.49	0.73
distribution + gender	0.79	(0.78, 0.79)	0.49	0.67
shape	0.82	(0.81, 0.83)	0.49	0.73
distribution	0.58	(0.57, 0.58)	0.49	0.28
gender	0.71	(0.71, 0.71)	0.49	0.51

**Table 13** Confusion matrix for prediction of inflectional behavior in the instrumental plural from shape, distribution, and gender.

**Table 14** Overall Statistics for inanimate instrumental plural predictions.

Our final series of models tests exactly that. This time, instead of grouping lexemes in terms of which formal exponent they may take, the lexemes were grouped according to whether they are found in the corpus only with an exponent of the formal family or with both types of exponents. Accuracy of the relevant models is reported in [Table 15](#), and the confusion matrix for the most accurate model is in the appendix. The performance of this family of models is much lower than the previous one, despite the fact that the number of classes is lower. In addition, gender and shape have no predictive power whatsoever, although distribution does have some.

This result gives a strong confirmation to the hypothesis that overabundance in the instrumental plural is orthogonal to the inflection class system, and uniquely conditioned by sociolinguistic factors. As we suggested above, different lexemes have, either because of their lexical semantics or axiological import, different likelihoods of being used in an informal context, and formality levels are expected to be reflected in a word's distribution, inasmuch as that word's syntagmatic neighbours are subject to the same usage effects. Hence the sociolinguistic conditioning hypothesis does predict that a lexeme's distribution should be predictive of whether it is found in the corpus with informal exponents. On the other hand, potential lexical predictors that are orthogonal to

Model	Accuracy	95% uncertainty interval	NIR	Kappa
shape + distribution + gender	0.76	(0.76, 0.77)	0.67	0.43
shape + distribution	0.76	(0.76, 0.77)	0.67	0.41
shape + gender	0.70	(0.69, 0.70)	0.67	0.23
distribution + gender	0.76	(0.75, 0.77)	0.67	0.41
shape	0.68	(0.67, 0.68)	0.67	0.15
distribution	0.76	(0.75, 0.76)	0.67	0.4
gender	0.67	(0.66, 0.68)	0.67	0.0

**Table 15** Overall Statistics for overbundant vs non-overabundant instrumental plural predictions.

formality distinctions, namely stem phonology and gender, do not have the predictive power we would have expected if overabundance was integrated in the inflection class system.

### 3.3 Taking stock

In this section we developed a sustained argument to the effect that overabundance in the locative singular and instrumental plural interact in different ways with the inflection class system: in the locative singular, there are distinct classes of overabundant lexemes, and these classes are mixed inflection classes;<sup>16</sup> in the instrumental plural, overabundance is orthogonal to the inflection class system, and fully conditioned by sociolinguistic factors.

The exact nature of overabundant locative singular classes remains somewhat confusing at this point: we have argued that they are first class citizens of the inflection class system, but that, in terms of class predictability, they mix and match properties of pairs of other classes. In the next section we develop an independent argument to the effect that overabundant locative singulars have properties of their own, irreducible to those of the neighboring non-overabundant classes.

## 4 The nature of mixed classes

### 4.1 Motivation

The preceding section has shown that overabundance in the locative singular integrates with the inflection class systems. In the case of hard masculine inanimate nouns, we showed that it was partially predictable on the basis of stem phonology and distribution which nouns are overabundant, and further showed that overabundant nouns occupied an intermediate space between ‘-u only’ and ‘-ě only’ nouns in terms of motivation: on average they share properties with both, which makes them easily confusable with both. We concluded that we should think of these overabundant nouns as belonging to a mixed class. We then generalized this result to other cases of overabundance in the locative singular.

In this section we explore in more detail the nature of mixed classes, and start with a quick review of relevant theoretical concepts in the literature. There is a consensus across frameworks in theoretical morphology that inflection class systems should be conceptualized as *inheritance hierarchies*, where classes may have different levels of specificity, and more specific subclasses inherit properties of their less specific superclasses (see among many others Corbett & Fraser 1993; Dressler & Thornton 1996; Koenig 1999; Beniamine, Bonami & Sagot 2017). Beniamine (2021) further argues that inflection class systems are best modeled as MONOTONOUS MULTIPLE INHERITANCE HIERARCHIES of the kind familiar from Head-Driven Phrase Structure Grammar (Pollard & Sag 1994).<sup>17</sup> Under this view, class systems are not trees, but LATTICES, where a node may have more than one parent. Beniamine’s central argument rests on the existence and pervasiveness of HETEROCLITE CLASSES, which have an inflectional behavior intermediate between those of two other classes (Stump 2006).<sup>18</sup> Czech neuter nouns of the class of KUŘE

<sup>16</sup> Except in the case of derived adjectives, which exhibit the same properties found in the instrumental plural.

<sup>17</sup> Although, to the best of our knowledge, this was never discussed in print, this is also implicitly the position adopted in early versions of Paradigm Function Morphology (Stump 2001), where any collection of lexemes may count as an inflection class.

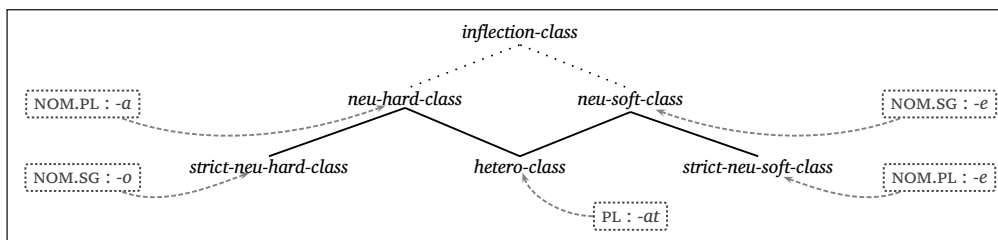
<sup>18</sup> Ironically for present purposes, Stump (2006) uses the Czech masculine inanimate noun PRAMEN ‘spring’ as his primary example of heteroclitisis, while closer examination shows that this is an instance of overabundance instead.

		<i>hard</i>	<i>heteroclite</i>	<i>soft</i>
		MĚSTO 'city'	KUŘE 'chicken'	MOŘE 'sea'
SG	NOM	měst-o	kuř-e	moř-e
	GEN	měst-a	kuř-et-e	moř-e
	DAT	měst-u	kuř-et-i	moř-i
	ACC	měst-o	kuř-e	moř-e
	VOC	měst-o	kuř-e	moř-e
	LOC	měst-ě-měst-u	kuřet-i	moř-i
	INS	měst-em	kuřet-em	moř-em
PL	NOM	měst-a	kuř-at-a	moř-e
	GEN	měst	kuř-at	moř-í
	DAT	měst-ům	kuř-at-ům	moř-ím
	ACC	měst-a	kuř-at-a	moř-e
	VOC	měst-a	kuř-at-a	moř-e
	LOC	měst-ech	kuř-at-ech	moř-ích
	INS	měst-y-měst-ama	kuř-at-y-kuř-at-ama	moř-l-moř-ema

**Table 16** Czech heteroclite neuter nouns.

'chicken' are a case in point, as illustrated in [Table 16](#): in the singular they use the same exponents as soft nouns, while in the plural they use the same exponents as hard nouns. Importantly, although this is not the canonical situation, heteroclite nouns may also have properties of their own, not deducible from the properties of the two other relevant classes (Kaye 2015: chap. 2). Again, Czech neuter heteroclites illustrate this: the presence of the stem augments *-et* in some singular cells and *-at* in all plural cells is found with all and only neuter heteroclites.

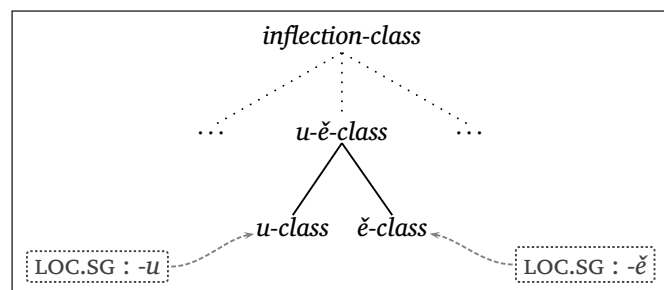
This behavior is easily captured in a monotonous multiple inheritance hierarchy by assuming that heteroclite nouns are assigned to a MEET node in the hierarchy: a node that inherits from two parents rather than one. Bonami & Crysmann (2018) propose an analysis along exactly these lines of the Czech nominal system, within the framework of Information-based Morphology (Crysmann & Bonami 2016). The crux of their analysis is summarized in [Figure 4](#). Here the structure of the hierarchy is represented in the center, while the dotted boxes indicate sample exponence rules that are associated with nodes in the hierarchy. Nouns like MĚSTO use *-o* in the nominative singular by virtue of being assigned to the *strict-neu-hard-class*, but also inherit from the superclass *neu-hard-class* the property of using *-a* in the plural. By contrast, nouns like MOŘE use *-e* in the nominative plural by virtue of being assigned to the *strict-neu-soft-class*, but also inherit from the superclass *neu-soft-class* the property of using *-e* in the singular. Most importantly, heteroclite nouns such as KUŘE inherit their singular exponents from the *soft-class* they share with strictly soft nouns, and their plural exponents from the *hard-class* they share with strictly hard nouns.



**Figure 4** Heteroclite classes as meets in an inflection class hierarchy (Beniamine 2021; Bonami & Crysmann 2018).

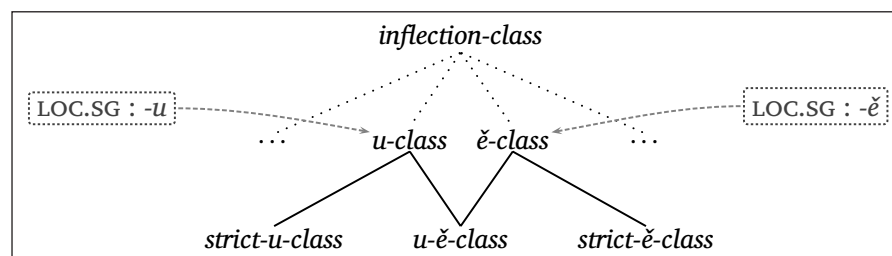
Let us now come back to overabundance against this background. There are two possible conceptualizations of overabundant inflection classes, both of which have been proposed in the literature. On the one hand Bonami & Crysmann (2018: 193–198) argue that overabundant

classes should be analyzed as *joins* in an inflection class hierarchy. This is illustrated in [Figure 5](#). The intuition here is that overabundant lexemes belong to a class that underspecifies the distinction between the two corresponding non-overabundant classes, and is the mirror image of a heteroclite class. The way this is captured is by associating the inflection rules respectively introducing the suffixes *-u* and *-ě* to the two more specific classes, and assigning lexemes to classes as indicated in the figure. The architecture of IbM then ensures that overabundant lexemes will be compatible with both exponents, because any concrete use of an overabundant lexeme has to pick one of the subtypes of the overabundant class.<sup>19</sup>



**Figure 5** Overabundant classes as joins in an inflection class hierarchy (Bonami & Cysmann 2018).

On the other hand, both Guzmán Naranjo (2019) and Beniamine (2021) propose that overabundant classes be analyzed as *meets* in an inflection class hierarchy. This is illustrated in [Figure 6](#). The two authors converge on this solution for different reason. For Beniamine, this is a consequence of deriving the inflection class lattice from individual properties of exponence exhibited by lexemes using formal concept analysis (Ganter & Wille 1998): in this framework, meets represent shared features, while joins represent the absence of features. For Guzman Naranjo, it derives from the decision that meet nodes in the hierarchy inherit all inflection strategies exhibited by their parents. Note that, under this line of analysis, heteroclite and overabundant classes are both represented by meet nodes in the hierarchy, but contrast as to whether the two parents of the meet contribute complementary or competing inflectional strategies.



**Figure 6** Overabundant classes as meets in an inflection class hierarchy (Guzmán Naranjo 2019; Beniamine 2021).

It is worth noting that, although they are conceptually distinct, both views of overabundant inflection classes are compatible with our observations on the motivation of mixed classes: in both cases, we expect overabundant classes to exhibit intermediate behavior between their non-overabundant counterparts, and a higher confusability between overabundant classes and the others than among non-overabundant classes.

There is however one area in which the two approaches make different predictions. Under the join approach, overabundant classes can't have positive properties that are not to some extent shared by their single exponent counterparts. This is a consequence of the monotonous flow of information in the inheritance hierarchy: a higher node in the hierarchy can't have properties that are not shared by its descendants. By contrast, under the meet approach, nothing precludes overabundant classes from having *some* idiosyncratic properties not shared by higher nodes.

<sup>19</sup> This is a consequence of the hypothesis inherited by IbM from HPSG that linguistic objects are *sort-resolved* (Pollard & Sag 1994: 18–21).

In this section we document exactly one such situation: we show that prepositions governing the locative exhibit differential preferences for one or the other exponent of overabundant nouns, a behavior that is not paralleled with non-overabundant nouns. Hence overabundant nouns have irreducible properties, which is not compatible with the join approach to overabundant classes.

## 4.2 Hypothesis: Prepositions exhibit preferences for exponents of overabundant lexemes

In the previous sections we established that there is a class of Czech masculine inanimate nouns that can take both the *-u* and the *-ě* ending in the locative singular. From this it does not follow that the choice of one or the other is entirely free: as discussed at length by Thornton (2019b), there can be both usage and grammatical conditions on overabundance. The existence and strength of such conditions is a recurring topic in the description on the Czech locative singular, reviewed by both Cummins (1995) and Bermel & Knittl (2012a;b). While no factor or combination of factors comes close to predicting categorically which of the two exponents will be used in what context, anecdotal evidence can be found for influences of noun polysemy (different senses of a noun having different preferences) and preposition polysemy (different senses of the governing preposition leading to different preferences), as well as individual idiosyncratic preferences of particular (preposition, noun) collocations. Bermel & Knittl (2012a;b) provide more compelling evidence from corpus and judgment data that types of syntactic environments have an influence: all other things being equal, *-u* is most preferred where the preposition heads a locative adverbial, and least likely when it is an empty preposition governed by a verb.

Elaborating on this literature, we study collocational preferences between prepositions and case-number exponents for overabundant nouns. We start from the observation that different governing prepositions seem to have different preferences as to which locative exponent is used. [Table 17](#) shows the distribution of the two locative singular forms of the two nouns MOST ‘bridge’ and ÚŘAD ‘office’ in the SYN corpus, when they are immediately preceded by one of the five main prepositions governing the locative: *na* ‘on, at’, *o* ‘near, about’, *po* ‘towards, after’, *při* ‘at, around’ and *v* ‘in’.

	<i>na</i>	<i>o</i>	<i>po</i>	<i>při</i>	<i>v</i>	Total
<i>mostu</i>	1006	259	222	13	107	1607
<i>mostě</i>	13823	114	3155	8	277	17377
Total	14829	373	3377	21	384	18984

MOST ‘bridge’

	<i>na</i>	<i>o</i>	<i>po</i>	<i>při</i>	<i>v</i>	Total
<i>úřadu</i>	21012	336	267	816	7482	29913
<i>úřadě</i>	17876	20	113	17	5345	23371
Total	38888	356	380	833	12827	53284

ÚŘAD ‘office’

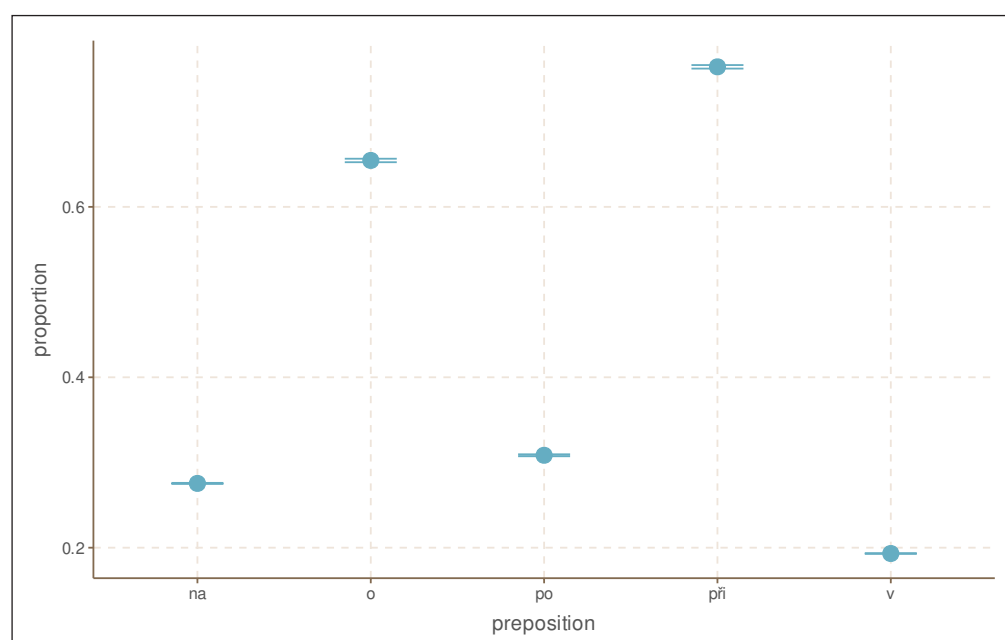
**Table 17** Cooccurrence counts of prepositions and nouns in the locative singular in the SYN corpus.

Three important observations are in order. First, and unsurprisingly, different lexemes have different preferences in terms of combinability with prepositions: for instance ÚŘAD ‘office’ is much more likely to be found in combination with *v* ‘in’ than MOST ‘bridge’. Second, proportions of use of the two variants varies widely across lexemes (MOST has a preference for the *-ě* form, while ÚŘAD prefers the *-u* form), and across combinations of lexemes and prepositions: at one extreme, the *-u* for of MOST is used 6% of the time with *po*; at the other extreme, the *-u* form for ÚŘAD is used 98% of time with *při*. But third and most importantly for us, despite lexeme-dependent variation, there still seem to be tendencies as to which prepositions prefer to co-occur with each exponent: across these two nouns, *o* and *při* have a strong preference for *-u* when compared to the other three prepositions *na*, *po* and *v*. Our goal in this section is to establish whether such contrasts are general.

### 4.3 Model 1: Predicting exponent preference from preposition for overabundant nouns

To this end, we collected from the SYN corpus all 27,768,583 occurrences of a hard masculine inanimate noun in the locative singular immediately preceded by a preposition. We then tabulated how many tokens of each inflectional variant (-*u* vs. -*ě*) was found for each of the 21,830 relevant lexemes in combination with each preposition. Then, among the 1733 overabundant lexemes in the dataset, we selected for study the 481 lexemes that are attested in collocation with all five prepositions under examination.

Our goal is to establish whether the likelihood of using a locative singular in -*ě* vs. -*u* varies across governing prepositions. To approach this question we built a Bayesian binomial model using Stan (Carpenter et al. 2017; Gelman, Lee & Guo 2015) and the brms interface (Bürkner et al. 2017). The predicted variable was the proportion of use of -*u* among uses of a locative singular, and the predictor variable was the identity of the preposition. *Figure 7* shows the conditional effects of the model, with whiskers representing 95% uncertainty intervals; the fact that the whiskers are barely distinguishable shows these intervals to be very narrow and hence uncertainty very low.



**Figure 7** Conditional effects of preposition on proportion of use of -*u* for overabundant lexemes.

The model very confidently establishes that each preposition has specific preferences: despite variability among noun lexemes, at the level of the system it is very clear that the -*u* form is more likely to be used in combination with *o* and *při*, while -*ě* is more likely to be used in combination with *na*, *po* and *v*.<sup>20</sup>

This result indicates that overabundant nouns exhibit properties that can only be found with such nouns: by definition, non-overabundant nouns use only one form, and hence cannot exhibit differential exponence properties in combination with different prepositions.

### 4.4 Model 2: Predicting preposition preference from locative singular exponent

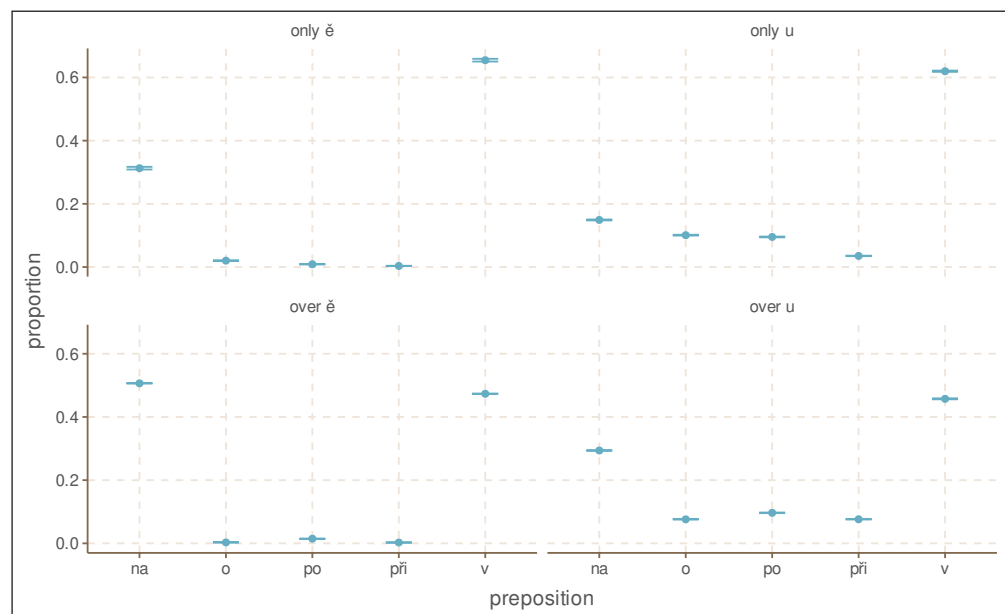
It is tempting to see the differential collocational preferences we just documented as a property characterizing the class of overabundant nouns. Before reaching such a conclusion, however, we must eliminate an alternative hypothesis. Above we have reasoned in terms of properties of overabundant lexemes, as opposed to the individual wordforms that realize these lexemes. But it is conceivable that the observed behavior is a consequence of collocational preferences linking

<sup>20</sup> Note that our model does not directly take into account the preferences of individual nominal lexemes for -*ě* vs. -*u*. Unfortunately, because of strong collinearity between predictors, models using both prepositions and nominal lexemes as predictors on the whole dataset consistently fail to converge. A model based on a smaller sample of 100 overabundant nouns gives results that are qualitatively consistent with what we report here, although the effects are not as clearcut.



prepositions and the individual locative singular exponents *-ě* and *-u*: perhaps the preposition *v* is more likely to be collocated with a *-ě* form than a *-u* form, irrespective of whether that form belongs to an overabundant lexeme or not. To test for that possibility, we need to examine how likely one is to use each preposition, depending on both the identity of the exponent and whether the lexeme is overabundant or not.

To address this question, we sampled from the dataset described in the previous subsection 400 locative singular noun forms: 100 nouns from the ‘*-u* only’ class, 100 nouns from the ‘*-ě* only’ class, 100 *-u* forms of overabundant nouns, and 100 *-ě* forms of overabundant nouns.<sup>21</sup> We then built a Bayesian multinomial model predicting proportion of use of each of the five prepositions from the class of nouns. *Figure 8* reports the conditional effects of the model, with each subplot corresponding to one of the subclasses of nouns.



**Figure 8** Conditional effects of exponent and overabundance on proportion of use of five prepositions.

Observation of the conditional effects suggests the existence of various tendencies grouping the data in both relevant dimensions. On the one hand, the proportion of collocation with three prepositions (*o*, *po* and *při*) is higher for *-u* nouns than for *-ě* nouns, irrespective of whether we are talking about one of the two forms of an overabundant lexeme or the only form of a non-overabundant one. On the other hand, overabundant nouns exhibit a more or less balanced propensity to combine with *na* and *v*, while non-overabundant ones have a marked preference for *v*.

Although making sense of the details of the distribution is well beyond the scope of this paper, this model clearly establishes that collocation preferences with prepositions are partially predicted by the overabundant or non-overabundant character of the noun, and cannot be solely reduced to preferences of collocation with locative singular exponents.

#### 4.5 Discussion

At the beginning of this section we set out to establish whether mixed classes such as that of overabundant hard masculine inanimate nouns should be conceptualized as meets or joins in the inflection class systems. Under the first hypothesis, the mixed class is the superclass of two non-overabundant classes, and should hence exhibit underspecified characteristics: its properties are the disjunction of properties of the non-overabundant classes. Under the second hypothesis, mixed classes share a parent with each of the non-overabundant classes. As such, they will exhibit some properties in common with each of their sister classes, but may also have properties of their own.

<sup>21</sup> The nouns were chosen so that each noun was attested in combination with at least three distinct prepositions, and at least 100 times in combination with at least one of these.

We then presented evidence for the existence of such properties specific to mixed classes. We documented differential collocational preferences between exponents of locative singular and governing prepositions, and showed these not to be reducible to a more general preference of prepositions for one or the other exponent which would also manifest itself for non-overabundant classes: for instance, although the *-u* form of overabundant nouns is barely ever used with *v*, *-u* only nouns show no sign of reluctance to combine with *v*. This behavior provides an argument for conceptualizing mixed classes as joins. Irrespective of whether one sees the relevant collocational requirements as following from selectional requirements of the preposition, as reverse-selection of the governor by its governee (Bonami 2015), or as a non-directional phenomenon, the statement of these requirements needs to reference the class of overabundant lexemes without the relevant property being inherited by their non-overabundant counterparts. This is precisely what is allowed by seeing the mixed class as a join node descending from the non-mixed classes rather than a meet node with the non-mixed classes as descendents. In this instance, external motivation, in the form of collocational preferences, provides crucial evidence for the proper internal organization of the inflectional system.

## 5 Conclusion

As is the case for many variation phenomena in other areas of grammar, variation in inflectional behavior is largely uncharted territory for linguistic theory. Descriptive and typological efforts spearheaded by Thornton in the last decade have led to a recognition of the widespread character of the phenomenon, and to explicit proposals to accommodate the general phenomenon in formal models of inflection (Stump 2016; Bonami & Crysmann 2018). However the traditional toolkit of formal linguistics is arguably ill-equipped to do justice to the richness of the phenomenon: a statement of variation between inflectional strategies is a correct but blunt approach to the question, which does not capture the gradient conditioning of that variation.

In this paper we attempted to improve the state of the art in this area, both by exploring in depth how alternate inflection strategies interact within a single system, and by relying on quantitative modeling to explore the fine properties of overabundance phenomena. We reached two main conclusions.

First, we established a qualitative difference between two types of overabundance. The Czech locative singular nominal declension exemplifies multiple cases where overabundance is embedded in the inflection class system, with overabundant lexemes forming distinct, mixed classes which contrast with their non-overabundant neighbors. These contrast with the situation found in the instrumental plural, where overabundance is fully orthogonal to the inflection class system: exponents come in pairs, and each lexeme is compatible with a pair of distinct exponents. Our arguments in favor of this conclusion are based on the external motivation of inflection classes. Following previous literature, we started from the assumption that inflection class assignment can be partially motivated by inflection-external (phonological, morphosyntactic, and semantic) properties of lexemes. We then showed that overabundant locative singulars exhibit external properties that are intermediate between those of their non-overabundant counterpart, while no such effect is found with overabundant instrumental plurals.

Second, we argued that mixed classes should be conceptualized as truly intermediate between two non-mixed classes, rather than unspecific or underspecified; technically, they should be seen as join nodes rather than meet nodes in the inflection class hierarchy. Our argument again rests on observations on external motivation, but of a different kind. We showed that, where overabundance provides two different locative singular forms for a hard masculine inanimate noun, the two forms exhibit different collocational preferences with governing prepositions. Crucially, these collocational preferences are distinct from those witnessed with non-overabundant nouns; hence they need to be stated as properties of the mixed class that are not shared with their non-overabundant neighbors, which is contradictory with the underspecification view.

We end by going back to the typology of overabundance phenomena. In this paper we examined exactly two cases of overabundance in just one language. On the basis of quantitative

evidence from external motivation, we were able to provide a rather detailed account of the commonalities and differences between these two. These two cases are both informative on the overall typology, each in its own way. On the one hand, the locative singular nicely exemplifies overabundance conditioned by morphological factors, as the possibility of overabundance is linked to the structure of the inflection class system, a strictly morphological notion (Aronoff 1994). Morphological conditioning is a possibility that Thornton (2019b: 248) anticipated, but did not provide a clearcut example of. On the other hand, the instrumental plural situation highlights the fact that conditions on overabundance may associate with whole series of exponents rather than individual ones: each Czech noun has two possible forms for the instrumental plural, and the (mostly sociolinguistic) conditions are the same for all nouns, but the identity of the exponents varies depending on the inflection class. This illustrates the deeply paradigmatic nature of the phenomenon of overabundance.

That being said, we make no claim as to typological, or even language internal generality. As a case in point, consider the fact that, in the dataset under consideration, we observed a combination of three contrasts: overabundance in the LOC.SG is lexically restricted to some corners of the inflection class systems, whereas it is lexically general in the INS.PL. It is subject to grammatical conditions in the LOC.SG and not in the INS.PL; conversely it is subject to usage conditions in the INS.PL and not in the LOC.SG. However, we have no reason to assume that the three contrasts align in this way. Detailed examination of the Czech system already provides evidence that things are not so simple. As we briefly commented on in Section 3, in the particular case of nouns derived by conversion from adjectives, we do find sociolinguistic conditioning in the LOC.SG that is exactly parallel to what we documented in the INS.PL—and hence this is a situation of lexically restricted overabundance subject to usage but to no grammatical conditions.

While we make no claim as to typological generality, we submit that the set of computational methods deployed in this paper constitutes a crucial toolkit to explore the typology of overabundance, by providing operational ways of exploring the graded dimensions of this typology that are largely free of language-particular descriptive biases. We hope this paper to be a useful first step in that direction.

## Abbreviations

NOM = nominative, GEN = nominative, DAT = nominative, ACC = accusative, VOC = accusative, LOC = accusative, INS = accusative, PL = plural, SG = singular

## Additional file

The additional file for this article can be found as follows:

- **Appendix.** Full confusion matrices. DOI: <https://doi.org/10.5334/gjgl.1626.s1>

## Acknowledgements


A very preliminary version of this work was presented at the Grammar and Corpora conference, Mannheim in 2016. We thank the audience for their comments. We also thank Benoit Crabbé, Rob Malouf, Alexandr Rosen, Jana Strnadová, Anna M. Thornton, and two anonymous reviewers for various suggestions that led to significant improvements.

## Funding information

This work was supported by a public grant overseen by the French National Research Agency (ANR) as part of the “Investissements d’Avenir” program (reference: ANR-10-LABX-0083).

## Competing interests

The authors have no competing interests to declare.

Matías Guzmán Naranjo  [orcid.org/0000-0003-1136-6836](https://orcid.org/0000-0003-1136-6836)  
Eberhard Karls Universität Tübingen, Wilhelmstr. 19, 72074 Tübingen, Germany

Olivier Bonami  [orcid.org/0000-0003-0688-3855](https://orcid.org/0000-0003-0688-3855)  
Université de Paris, Laboratoire de linguistique formelle, CNRS, 8, Place Paul Ricoeur, 75013 Paris, France

## References

- Albright, Adam. 2009. Modeling analogy as probabilistic grammar. In James P. Blevins & Juliette Blevins (eds.), *Analogy in grammar*, 200–228. Oxford, New York: Oxford University Press. DOI: <https://doi.org/10.1093/acprof:oso/9780199547548.003.0009>
- Albright, Adam & Bruce Hayes. 2002. Modeling English past tense intuitions with minimal generalization. In *Proceedings of the ACL-02 workshop on Morphological and phonological learning* 6. 58–69. Association for Computational Linguistics. DOI: <https://doi.org/10.3115/1118647.1118654>
- Albright, Adam & Bruce Hayes. 2003. Rules vs. analogy in English past tenses: A computational/experimental study. *Cognition* 90(2). 119–161. DOI: [https://doi.org/10.1016/S0010-0277\(03\)00146-X](https://doi.org/10.1016/S0010-0277(03)00146-X)
- Anderson, Stephen R. 1992. *A-morphous morphology*. Cambridge: Cambridge University Press. DOI: <https://doi.org/10.1017/CBO9780511586262>
- Arndt-Lappe, Sabine. 2011. Towards an exemplar-based model of stress in English noun–noun compounds. *Journal of Linguistics* 47(3). 549–585. DOI: <https://doi.org/10.1017/S0022226711000028>
- Arndt-Lappe, Sabine. 2014. Analogy in suffix rivalry: the case of English *-ity* and *-ness*. *English Language and Linguistics* 18(3). 497. DOI: <https://doi.org/10.1017/S136067431400015X>
- Aronoff, Mark. 1994. *Morphology by itself*. Cambridge: MIT Press.
- Baayen, Harald & Fermin Moscoso del Prado Martín. 2005. Semantic density and pasttense formation in three Germanic languages. *Language* 81. 666–698. DOI: <https://doi.org/10.1353/lan.2005.0112>
- Baayen, R. Harald, Yu-Ying Chuang & James P. Blevins. 2018. Inflectional morphology with linear mappings. *The Mental Lexicon* 13(2). 230–268. DOI: <https://doi.org/10.1075/ml.18010.baa>
- Bechtel, William & Adele Abrahamsen. 2002. *Connectionism and the mind: Parallel processing, dynamics, and evolution in networks*. 2nd edition. Oxford: Blackwell.
- Beniamine, Sacha. 2021. One lexeme, many classes: inflection class systems as lattices. In Berthold Crysmann & Manfred Sailer (eds.), *One-to-many relations in morphology, syntax and semantics*, 23–51. Berlin: Language Science Press.
- Beniamine, Sacha & Olivier Bonami. Submitted. Inflection class systems. In Peter Ackema, Sabrina Bendjaballah, Eulàlia Bonet & Antonio Fábregas (eds.), *Wiley-blackwell companion to morphology*. Wiley-Blackwell.
- Beniamine, Sacha, Olivier Bonami & Benoît Sagot. 2017. Inferring inflection classes with description length. *Journal of Language Modelling* 5(3). 465–525. DOI: <https://doi.org/10.15398/jlm.v5i3.184>
- Bermel, Neil. 2000. *Register variation and language standards in Czech*. Munich: Lincom Europa.
- Bermel, Neil & Luďek Knittl. 2012a. Corpus frequency and acceptability judgments: A study of morphosyntactic variants in Czech. *Corpus Linguistics and Linguistic Theory* 8. 241–275. DOI: <https://doi.org/10.1515/cllt-2012-0010>
- Bermel, Neil & Luďek Knittl. 2012b. Morphosyntactic variation and syntactic constructions in Czech nominal declension: corpus frequency and native-speaker judgments. *Russian Linguistics* 36. 91–119. DOI: <https://doi.org/10.1007/s11185-011-9083-x>
- Bermel, Neil, Luďek Knittl & Jean Russell. 2015. Morphological variation and sensitivity to frequency of forms among native speakers of Czech. *Russian Linguistics* 39. 283–308. DOI: <https://doi.org/10.1007/s11185-015-9149-2>
- Bermel, Neil, Luďek Knittl & Jean Russell. 2018. Frequency data from corpora partially explain native-speaker ratings and choices in overabundant paradigm cells. *Corpus Linguistics and Linguistic Theory* 14(2). 197–231. DOI: <https://doi.org/10.1515/cllt-2016-0032>
- Blevins, James P., Petar Milin & Michael Ramscar. 2017. The Zipfian Paradigm Cell Filling Problem. In Ferenc Kiefer, James P. Blevins & Huba Bartos (eds.), *Morphological paradigms and functions*. Leiden: Brill.
- Boleda, Gemma. 2020. Distributional semantics and linguistic theory. *Annual Review of Linguistics* 6(1). 213–234. DOI: <https://doi.org/10.1146/annurev-linguistics-011619-030303>
- Bonami, Olivier. 2015. Periphrasis as collocation. *Morphology* 25. 63–110. DOI: <https://doi.org/10.1007/s11525-015-9254-3>
- Bonami, Olivier & Berthold Crysmann. 2018. Lexeme and flexeme in a formal theory of grammar. In Olivier Bonami, Gilles Boyé, Georgette Dal, Hélène Giraudo & Fiammetta Namer (eds.), *The lexeme in descriptive and theoretical morphology*, 175–202. Berlin: Language Science Press.

- Bonami, Olivier & Denis Paperno. 2018. Inflection vs. derivation in a distributional vector space. *Lingue e Linguaggio* 17(2). 173–195.
- Bonami, Olivier & Gilles Boyé. 2010. La morphologie flexionnelle est-elle une fonction ? In Injoo Choi-Jonin, Marc Duval & Olivier Soutet (eds.), *Typologie et comparatisme, hommage offert à alain lemaréchal*, 21–35. Leuven: Peeters.
- Bošnjak Botica, Tomislava & Gordana Hržica. 2016. Overabundance in croatian dual-class verbs. *FLUMINENSIA: časopis za filološka istraživanja* 28(1). 83–106.
- Breiman, Leo. 2001. Random forests. *Machine learning* 45(1). 5–32. DOI: <https://doi.org/10.1023/A:1010933404324>
- Brown, Dunstan. 2007. Peripheral functions and overdifferentiation: the Russian second locative. *Russian Linguistics* 31. 61–76. DOI: <https://doi.org/10.1007/s11185-006-0715-5>
- Brown, Dunstan & Andrew Hippisley. 2012. *Network Morphology: a defaults based theory of word structure*. Cambridge: Cambridge University Press. DOI: <https://doi.org/10.1017/CBO9780511794346>
- Brown, Dunstan, Marina Chumakina & Greville G. Corbett. (eds.) 2013. *Canonical morphology and syntax*. Oxford: Oxford University Press. DOI: <https://doi.org/10.1093/acprof:oso/9780199604326.001.0001>
- Bürkner, Paul-Christian, et al. 2017. Brms: an R package for Bayesian multilevel models using Stan. *Journal of statistical software* 80(1). 1–28. DOI: <https://doi.org/10.18637/jss.v080.i01>
- Bybee, Joan L. & Dan I. Slobin. 1982. Rules and schemas in the development and use of the English past tense. *Language* 58(2). 265–289. DOI: <https://doi.org/10.1353/lan.1982.0021>
- Camacho-Collados, Jose & Mohammad Taher Pilehvar. 2020. Embeddings in natural language processing. In *Proceedings of the 28th international conference on computational linguistics: tutorial abstracts*, 10–15. Barcelona, Spain (Online): International Committee for Computational Linguistics. <https://www.aclweb.org/anthology/2020.coling-tutorials.2>. DOI: <https://doi.org/10.18653/v1/2020.coling-tutorials.2>
- Cappellaro, Chiara. 2013. Overabundance in diachrony: a case study. In Silvio Cruschina, Martin Maiden & John Charles Smith (eds.), *The boundaries of pure morphology. diachronic and synchronic perspectives*. Oxford University Press. DOI: <https://doi.org/10.1093/acprof:oso/9780199678860.003.0011>
- Carpenter, Bob, Andrew Gelman, Matthew D. Hoffman, Daniel Lee, Ben Goodrich, Michael Betancourt, Marcus Brubaker, Jiqiang Guo, Peter Li & Allen Riddell. 2017. Stan: a probabilistic programming language. *Journal of statistical software* 76(1). DOI: <https://doi.org/10.18637/jss.v076.i01>
- Chen, Tianqi & Carlos Guestrin. 2016. XGBoost: a scalable tree boosting system. In *Proceedings of the 22nd ACM SIGKDD international conference on knowledge discovery and data mining*, 785–794. DOI: <https://doi.org/10.1145/2939672.2939785>
- Churchland, Paul M. 1989. *A neurocomputational perspective: The nature of mind and the structure of science*. Massachusetts: MIT press.
- Corbett, Greville G. 1991. *Gender*. Cambridge: Cambridge University Press.
- Corbett, Greville G. 2007. Canonical typology, suppletion and possible words. *Language* 83. 8–42. DOI: <https://doi.org/10.1353/lan.2007.0006>
- Corbett, Greville G. 2012. *Features*. Cambridge: Cambridge University Press.
- Corbett, Greville G. & Norman M. Fraser. 1993. Network morphology: a DATR account of Russian nominal inflection. *Journal of Linguistics* 29. 113–142. DOI: <https://doi.org/10.1017/S0022226700000074>
- Crysmann, Berthold & Olivier Bonami. 2016. Variable morphotactics in Information-Based Morphology. *Journal of Linguistics* 52(2). 311–374. DOI: <https://doi.org/10.1017/S0022226715000018>
- Cummins, George M. 1995. Locative in Czech: -u or -ě? Choosing locative singular endings in Czech nouns. *The Slavic and East European Journal* 39(2). 241–260. DOI: <https://doi.org/10.2307/309376>
- Cummins, George M. 2005. Litterary Czech, Common Czech, and the instrumental plural. *Journal of Slavic Linguistics* 13(2). 271–297.
- Cvrček, Václav, Vilém Kodýtek, Marie Kopřivová, Dominika Kovářková, Petr Sgall, Michal Šulc, Jan Táborový, Jan Volný & Martina Waclawičová. 2010. *Mluvince současné češtiny 1*. Prague: Karolinum.
- DeMello, George. 1993. -Ra vs. -se subjunctive: a new look at an old topic. *Hispania* 76(2). 235–244. DOI: <https://doi.org/10.2307/344667>
- Dressler, Wolfgang U. & Anna M. Thornton. 1996. Italian nominal inflection. *Wiener Linguistische Gazette* 55–57. 1–26.
- Ganter, Bernhard & Rudolf Wille. 1998. *Formal concept analysis: mathematical foundations*. Berlin: Springer. DOI: <https://doi.org/10.1007/978-3-642-59830-2>
- Gelman, Andrew, Daniel Lee & Jiqiang Guo. 2015. Stan: a probabilistic programming language for Bayesian inference and optimization. *Journal of Educational and Behavioral Statistics* 40(5). 530–543. DOI: <https://doi.org/10.3102/1076998615606113>
- Guzmán Naranjo, Matías. 2019. *Analogical classification in formal grammar*. Berlin: Language Science Press.

- Guzmán Naranjo, Matías. 2020. Analogy, complexity and predictability in the Russian nominal inflection system. *Morphology* 30. 219–262. DOI: <https://doi.org/10.1007/s11525-020-09367-1>
- Hnátková, M., M. Křen, P. Procházka & H. Skoumalová. 2014. The SYN-series corpora of written Czech. In *Proceedings of the ninth international conference on language resources and evaluation*, 160–164.
- Huyghe, Richard & Marine Wauquier. 2020. What's in an agent? *Morphology* 30(3). 185–218. DOI: <https://doi.org/10.1007/s11525-020-09366-2>
- Kaye, Steven J. 2015. *Conjugation class from Latin to Romance: heterocclisis in diachrony and synchrony*. University of Oxford dissertation.
- Koenig, Jean-Pierre. 1999. *Lexical relations*. Stanford, CA: CSLI Publications.
- Kohavi, Ron. 1995. A Study of Cross-Validation and Bootstrap for Accuracy Estimation and Model Selection. In, 1137–1143. Morgan Kaufmann.
- Křen, Michal, Václav Cvrček, Tomáš Čapka, Anna Čermáková, Milena Hnátková, Lucie Chlumská, Tomáš Jelínek, Dominika Kovářková, Vladimír Petkevič, Pavel Procházka, Hana Skoumalová, Michal Škrabal, Petr Truneček, Pavel Vondříčka & Adrian Jan Zasina. 2016. *Corpus SYN, version 4 from 16. 9. 2016*. Tech. rep. Ústav Českého národního o korpusu FF UK, Praha. <http://www.korpus.cz>.
- Lapesa, Gabriella, Lea Kawaletz, Ingo Plag, Marios Andreou, Max Kisselew & Sebastian Padó. 2018. Disambiguation of newly derived nominalizations in context: a distributional semantics approach. *Word Structure* 11(3). 277–312. DOI: <https://doi.org/10.3366/word.2018.0131>
- Lečić, Dario. 2015. Morphological doublets in Croatian: the case of the instrumental singular. *Russian Linguistics* 39(3). 375–393. DOI: <https://doi.org/10.1007/s11185-015-9152-7>
- Marelli, Marco & Marco Baroni. 2015. Affixation in semantic space: modeling morpheme meanings with compositional distributional semantics. *Psychological Review* 122. 485–515. DOI: <https://doi.org/10.1037/a0039267>
- McClelland, James L. & David E. Rumelhart. 1986. A distributed model of human learning and memory. In James L. McClelland & David E. Rumelhart (eds.), *Parallel distributed processing: Explorations in the microstructure of cognition: Vol. 2 Psychological and biological models*, 170–215. Cambridge: MIT Press. DOI: <https://doi.org/10.7551/mitpress/5237.001.0001>
- Mikolov, Tomas, Kai Chen, Greg Corrado & Jeffrey Dean. 2013. Efficient estimation of word representations in vector space. *CoRR* abs/1301.3781. <http://arxiv.org/abs/1301.3781>.
- Naughton, James. 2005. *Czech: an essential grammar*. Routledge. DOI: <https://doi.org/10.4324/9780203567036>
- Pollard, Carl & Ivan A. Sag. 1994. *Head-driven phrase structure grammar*. Chicago: University of Chicago Press.
- Řehůřek, Radim. 2010. Software framework for topic modelling with large corpora. In *Proceedings of the Irec 2010 workshop on new challenges for nlp frameworks*, 45–50.
- Rosemeyer, Malte & Scott A. Schwenter. 2019. Entrenchment and persistence in language change: the Spanish past subjunctive. *Corpus Linguistics and Linguistic Theory* 15. 167–204. DOI: <https://doi.org/10.1515/cllt-2016-0047>
- Rumelhart, David E., James L. McClelland, David E. Rumelhart & James L. McClelland. 1986. On learning the past tenses of English verbs. In *Parallel distributed processing: Explorations in the microstructure of cognition: Vol. 2 Psychological and biological models*. Cambridge: MIT Press.
- Santilli, Enzo. 2014. *Italian comparatives: a case of overabundance*. doctoral dissertation, University of Aquila dissertation.
- Skousen, Royal. 1989. *Analogical modeling of language*. Springer Science & Business Media. DOI: <https://doi.org/10.1007/978-94-009-1906-8>
- Skousen, Royal. 2013. *Analogy and structure*. Springer Science & Business Media.
- Stump, Gregory T. 2001. *Inflectional morphology. A theory of paradigm structure*. Cambridge: Cambridge University Press. DOI: <https://doi.org/10.1017/CBO9780511486333>
- Stump, Gregory T. 2006. Heterocclisis and paradigm linkage. *Language* 82. 279–322. DOI: <https://doi.org/10.1353/lan.2006.0110>
- Stump, Gregory T. 2016. *Inflectional paradigms*. Cambridge: Cambridge University Press. DOI: <https://doi.org/10.1017/CBO9781316105290>
- Thornton, Anna M. 2011. Overabundance (Multiple Forms Realizing the Same Cell): A Non-canonical Phenomenon in Italian Verb Morphology. In Martin Maiden, John Charles Smith, Maria Goldbach & Marc-Olivier Hinzelin (eds.), *Morphological Autonomy. Perspectives From Romance Inflectional Morphology*. Oxford: Oxford University Press. DOI: <https://doi.org/10.1093/acprof:oso/9780199589982.003.0017>
- Thornton, Anna M. 2012. Reduction and maintenance of overabundance. A case study on Italian verb paradigms. *Word Structure* 5(2). 183–207. DOI: <https://doi.org/10.3366/word.2012.0026>
- Thornton, Anna M. 2019a. Overabundance in morphology. In *Oxford research encyclopedia of linguistics*. Oxford University Press. DOI: <https://doi.org/10.1093/acrefore/9780199384655.013.554>

- Thornton, Anna M. 2019b. Overabundance: a canonical typology. In Franz Rainer, Francesco Gardani, Hans-Christian Luschützky & Wolfgang U. Dressler (eds.), *Competition in morphology*, 223–258. Dordrecht: Springer. DOI: [https://doi.org/10.1007/978-3-030-02550-2\\_9](https://doi.org/10.1007/978-3-030-02550-2_9)
- Varvara, Rossella. 2017. *Verbs as nouns: empirical investigations on event-denoting nominalizations*. University of Trento dissertation.

Guzmán Naranjo and  
Bonami  
*Glossa: a journal of  
general linguistics*  
DOI: 10.5334/gjgl.1626

31

TO CITE THIS ARTICLE:

Guzmán Naranjo, Matías and Olivier Bonami. 2021. Overabundance and inflectional classification: Quantitative evidence from Czech. *Glossa: a journal of general linguistics* 6(1): 88. 1–31. DOI: <https://doi.org/10.5334/gjgl.1626>

Submitted: 10 March 2021

Accepted: 01 June 2021

Published: 02 July 2021

COPYRIGHT:

© 2021 The Author(s). This is an open-access article distributed under the terms of the Creative Commons Attribution 4.0 International License (CC-BY 4.0), which permits unrestricted use, distribution, and reproduction in any medium, provided the original author and source are credited. See <http://creativecommons.org/licenses/by/4.0/>.

*Glossa: a journal of general linguistics* is a peer-reviewed open access journal published by Ubiquity Press.