
Article:
Bermel, N. orcid.org/0000-0002-1663-9322 and Brown, D. (2023) Introduction. Word Structure, 16 (2-3). pp. 147-153. ISSN 1750-1245

https://doi.org/10.3366/word.2023.0226

Introduction

Neil Bermel
University of Sheffield

Dunstan Brown
University of York

1. Background to the thematic issue

Overabundance and defectivity appear to be opposing tendencies within morphology. The former involves multiple forms for the same cell in a paradigm (Thornton, 2019a), while for the latter there appears to be no viable option for a paradigm cell (Baerman & Corbett, 2010; Sims, 2015, pp. 26–28). However, our starting point is the observation that descriptions of overabundance and defectivity rely on similar operational definitions, suggesting inter alia that competing analogical processes, evidence of multiple potential and actual forms, and a cluster of reaction types that could be labelled *uncertainty* characterise both phenomena. That is why this special edition is devoted to the joint study of overabundance and defectivity.

2. The language scope of the project

Both overabundance and defectivity are found cross-linguistically. In English, with its relatively impoverished inflectional morphology, overabundance is typically studied as an issue in historical change (see e.g., Denison, 2003: 56); the movement of verbs between the 'strong' and 'weak' patterns has been a focus of enquiry, typically framed as a change in progress from the strong to the weak conjugation in which, at certain times, overabundance is found before the change moves to its final phase. Defectivity is more sporadic in English, possibly due to the paucity of inflectional forms, and here scholars have relied more heavily on evidence from languages with more elaborate inflectional systems, such as French, Greek, Latin, Russian and Spanish, where it is found in predictable places in the nominal and verbal systems. As Thornton (2019b) documents, though, our understanding of the scope and function of overabundance can be improved by looking beyond the larger, better-known European languages, and the same is true for defectivity.

The Feast and Famine project (see below) undertakes to examine data from a number of highly-inflected and typically less-studied languages in the Slavonic and Finno-Ugric groups: Croatian, Czech, Estonian, Finnish, Russian and the non-standardized languages of the Polish/Belarusian/Ukrainian border regions. In contrast to a language like English, where a noun has a maximum of four potential forms (e.g., *mouse, mouse's, mice, mice's*), nouns in a language like Czech can have up to a dozen potential forms, and in languages such as Finnish, there are approximately 150 basic forms per noun, with a further 1850 'peripheral' forms involving agglutinative discourse particles (Nikolaev and Bermel, this issue). Under these conditions, multiple potential stem and inflectional exponents are not so much a curiosity as a fact of life encountered in nearly every sentence. Instead of studying a handful of examples, we can study literally hundreds in each language. This offers fertile ground for comparing and contrasting the differing fates of lexemes with paradigmatic gaps and competition.

3. The project methodologies

We began our research from a position of methodological unclarity, as regards both defectivity and overabundance.

Defectivity has been studied primarily as an inherent feature of inflectional cells. To establish this inherence, scholars working on it have tended to rely on the pronouncements of authoritative reference works, although the reliability of such works, and the changeability of their pronouncements, has been documented (see Baerman, 2008 for examples from Russian). Sims (2015) identified measures of self-reported doubt as a reliable indicator of which cells were defective, opening the possibility of using language tests to measure it. Corpus data have been employed by some scholars to propose that defective cells are detectable through their unusually low frequency, and seek to define defectivity in this way. While cognitively this approach makes a great deal of sense in that it shows how people negotiate the common absence of forms in their experience (Janda and Tyers, 2018), it can lead in practice to counterintuitive results in which an accidental corpus lacuna deriving from a pragmatic situation is interpreted as an inherent paradigmatic gap. The logical results are documented in Kováříková et al., 2020: the form *zavraždila* 'murder-PAST-FEM' is a lacuna, and yet fails our 'inherent defectivity' test, as native speakers have no difficulty in providing it or accepting it. Its near-absence in the corpus turns out to be a function of the fact that far more men are murderers than women.

The description of absence also poses issues for language planning, and our investigations have shown that different traditions handle paradigmatic gaps differently. Some languages' authoritative codificatory bodies and works, influenced perhaps by the traditional acceptance of paradigmatic gaps in classical Greek and Latin based on which forms were or were not attested in texts, have internalized descriptions of defectivity readily and easily: French, Spanish and Russian fall into this camp. In other languages, such as Czech, Finnish, Croatian and Estonian, there is little tradition of describing such gaps, and instead, a bias towards provision and systematicity prevails, even in the face of empirical evidence that a cell is defective. The role of prescriptivism and descriptivism in language culture is thus thrown into sharp relief by considering the fate of these cells.

Overabundance, on the other hand, has been studied primarily as a corpus phenomenon (see e.g. Thornton, 2011, 2012; Brown 2007), and there has been a strong presumption that variation in form must be somehow conditioned, whether that conditioning be stylistic, diachronic, syntactic or regional.

However, even if we accept the premise described above in section 2, by which overabundance is an accidental and, in the grand scheme of things, transient by-product of historical processes, we are left with the question of how such a system can emerge and be maintained in the language of speakers at any particular time. The Principle of Contrast (Clark, 1987) proposes a strong hypothesis in support of this: that this tendency is linked to language acquisition and is a fundamental part of how we perceive and operationalize variation. Our noticing of these differences is what allows us to define and distinguish both easily distinguished and defined categories, as well as those that are less systematic and idiosyncratic to particular lexemes or groups of lexemes.

Numerous studies have documented situations in which this received view demonstrably obtains, but a number of recent studies on usage in less-commonly-studied languages have found the situation to be less clear-cut than this. These studies examine how native speakers react to such forms (Lečić 2015, Bermel and Knittl 2012a, 2012b; Bermel, Knittl and Russell 2015, 2018) and a curious finding across them is that, while conditioned change is undoubtedly possible and even common, there are also examples of non-conditioned or weakly-conditioned variation that are apparently stable over time, with the only salient difference between them being "frequency of appearance". This brings us back to our original question: in places where variation is unconditioned or weakly conditioned, speakers must still learn these systems as children; they must manage them throughout their adult lives; and we should be able to find ways to model these systems using computational methods without defaulting to the notion of a single "target" form. We need an explanatory framework that can

manage unmotivated as well as motivated variation.

Here too, local interpretations of how language should be managed and presented to its users come into play. In Croatian, for example, it has been shown that there is a strong bias in handbooks towards presenting users with clear "rules" for usage of one or another doublet form, which, as it turns out, have no basis in actual usage as determined by corpus searches (Lečić, 2015: 378, 380; Polančec 2017: 203-205; Bošnjak Botica, Polančec and Sviben 2022: 45-46). The conclusion is that handbook authors have attempted to implement the Principle of Contrast using individual intuitions, in the hope of giving their users more security and confidence. Once we identify and distance ourselves from this approach, however, the question remains of what we will replace it with.

An important goal of this project and the contributions here has been to increase the methodological diversity of the ways in which we examine these phenomena. The articles in this issue encompass a number of approaches, utilising corpus data; experimental data; and evidence from handbooks and other normative literature.

Corpus methods play a major role in our investigations, partially for reasons discussed below, and partly because they have in part replaced the reliance on handbooks and manual excerption characteristic of previous eras as foundational data for research, especially but not exclusively in computational modelling of language (see *inter alia* Chuang et al. 2022). As the contributions in this issue show, however, the question of how to use corpora to investigate defectivity is not straightforward to answer. Even overabundance, which can be straightforwardly researched and reported in descriptive terms, is problematic in terms of what those corpus numbers mean beyond the recitation of percentage figures: what does a 60:40 or 70:30 split in frequency mean in terms of how users perceive and use these forms? Approaches to corpus data on defectivity are considered in Nikolaev and Bermel's contribution below, while Aigro and Vihman's article considers the interpretation of corpus data on overabundance.

The interplay between corpus data, handbook data and experimental data from language use is another evolving area of research. One team in Zagreb (see below) and another in Tartu are looking at how overabundance is handled by young children of different ages, while a team in Sheffield and Joensuu looks at this interaction as concerns adult language. Other project investigations are examining ways of viewing second language acquisition through the prism of overabundance and defectivity. While investigations in these areas are at an early stage, the goal is to show that defective and overabundant slots can be profitably viewed as a kind of managed uncertainty that users cope with in daily usage, with predictably differing results.

Finally, handbooks and other 'official' interpretations of language structure continue to play a major role in our ability to confidently label an inflectional slot as 'defective' or 'overabundant'. Overtly empirical and experimental investigations (see this issue, and Chuang et al. 2022) include at least in part a labelling scheme for individual lexemes that relies in part on handbook data. The questions of how handbooks come by their interpretations, how they convey them and how they can incorporate new findings is a critical one for languages with well-developed mechanisms for language regulation and advisory services. A curious and oft-overlooked contrast is with non-standardised languages, of which there are several in the Slavonic and Finno-Ugric families; in the absence of prescriptive or descriptive handbooks, without an educational system and lacking the prestige of established national languages, the dynamics of competition between potential forms may prove to be significantly different. All these areas are currently under investigation in the scope of the project.

4. Outline of the volume

All five articles in the volume derive from the AHRC-funded 'Feast and Famine project', and consider various aspects of how speakers of Slavonic and Finno-Ugric languages manage

defective and overabundant data. Our focus in the project's initial phases, which took place during the pandemic, was of necessity on exploiting virtual resources such as corpora and conducting surveys that could be administered remotely. The results presented here form the basis on which further studies in the above areas are being conducted.

Mari Aigro and **Virve Vihman's** contribution opens the issue with a look at the distribution of overabundant noun forms based on evidence from corpora. Their study of the use of overabundant forms in Estonian nouns contrasts with the focus of earlier work in this area on the availability of multiple forms for a given paradigm cell. They found that syntactic context did not play a significant role in choosing between overabundant forms in use. This goes against claims about the use of one of the forms (the short form) being less acceptable in certain contexts. Furthermore, overabundance in the Estonian noun system as it is actually observed in use appears to be restricted and more canonical than what is potentially available.

Another perspective on overabundance involving competing processes (overgeneralisation) - from the point of view of child language development - is provided by **Gordana Hržica, Tomislava Bošnjak Botica and Sara Košutar**. They use a parental reporting questionnaire to investigate verbs with stem changes in the longitudinal Croatian corpus of child language, providing us with evidence in relation to potential and actual forms used in child directed speech and children's productions. Their study shows that overgeneralized forms are reported in all classes. Verb frequency and class size correlate the proportion of overgeneralizations. They argue that this provides evidence for the gradual abstraction of morphological rules based on the input.

Returning to corpus linguistics, **Alexandre Nikolaev and Neil Bermel** tackle the question of how to interpret negative evidence in corpora of Finnish: are data of this sort conclusive evidence of defectivity, or not? They examine instances of the Finnish instructive case, a marginal category represented at low frequencies in corpora, whose forms are susceptible to adverbialisation. They contrast evidence from corpora of different sizes and consider the sorts of statistical tests that have been and could be applied to such data, asking finally to what extent corpus data can provide definitive evidence of defectivity.

The final two papers in the issue compare and contrast defective and overabundant data.

Dominika Kováříková, Oleg Kovářík, Kamila Smejkalová and Martin Beneš's article looks at how data from the corpus tool GramatiKat can be used to develop more accurate and useful reference works for Czech, identifying and reflecting overabundant and defective slots more clearly than has been the case to date. They examine the data on which the Czech Internet Language Reference Book, a source of advisory information for users of Czech, has been created, and identify places where that information could be enhanced and made more accurate through the use of corpus data. Using the GramatiKat tool created as part of this project, they describe a method for investigating potentially overabundant and defective cells on the basis of the animate nominative plural, a frequent source of uncertainty for Czech speakers; this method, which is contrasted to others previously used in other studies, is presented as a model for future use. They propose and discuss potential manipulations of the Reference Book's presentation of data that would improve the information made available to users.

Neil Bermel, Luděk Knittl and Alexandre Nikolaev's study contrasts native-speaker reactions to a gap-filling exercise containing overabundant and defective lexemes. By asking native speakers to fill guided gaps in sentence-long prompts, they gathered evidence of multiple potential and actual forms, and were able to compare how speakers reacted to three proposed conditions (defective cells, overabundant cells, and a control condition). They identify types of reactions clearly separating defective cells from others both in terms of the number of responses produced, the character of the responses, and the time taken to produce

them. They further find that overabundant reactions pattern largely with the control condition, but pattern with the defective condition when it comes to the identification of a single 'default' response. They propose that these distinctions capture different aspects of speaker uncertainty faced with multiple potential choices.

The articles in this special issue should, we hope, provide more concrete evidence on what happens when these questions are addressed directly through corpus and experimental data. They also establish further questions and hypotheses about common mechanisms and processes within non-canonical paradigm cells that form the basis for future research in this area.

5. Acknowledgments

The authors of the articles in this special issue gratefully acknowledge the support of the UK Arts and Humanities Research Council (grant no. AH/T002859/1).

References

Baerman, M. (2008). Historical observations on defectiveness: The first singular non-past. *Russian Linguistics* 32(1). 81–97.

Baerman, M., & Corbett, G. G. (2010). Introduction: Defectiveness: Typology and Diachrony. In M. Baerman, G. G. Corbett, & D. Brown (Eds.), *Defective Paradigms: Missing forms and what they tell us* (pp. 1–18). Cambridge University Press.

Bermel, N. & Knittl, L. (2012a). Corpus frequency and acceptability judgements: A study of morphosyntactic variants in Czech. *Corpus Linguistics and Linguistic Theory* 8. 241275.

Bermel, N. & Knittl, L. (2012b). Morphosyntactic variation and syntactic constructions in Czech nominal declension: corpus frequency and native-speaker judgements. *Russian Linguistics* 36. 91119.

Bermel, N., Knittl, L., & Russell, J. (2015). Morphological variation and sensitivity to frequency of forms among native speakers of Czech. *Russian Linguistics*, *39*(3), 283–308. https://doi.org/10.1007/s11185-015-9149-2

Bermel, N., Knittl, L., & Russell, J. (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. https://doi.org/10.1515/cllt-2016-0032

Bošnjak Botica, T. & Polančec, J. & Sviben, R. (2022). Korpusno istraživanje hrvatskih imenica s dugom i kratkom množinom. *Jezikoslovlje* 23. 35-74.

Brown, D. (2007). Peripheral functions and overdifferentiation: the Russian second locative. *Russian Linguistics* 31. 61–76.

Chuang, Y-Y., Brown, D., Baayen, H., & Evans, R. (Accepted/In press). Paradigm gaps are associated with weird "distributional semantics" properties: Russian defective nouns and their case and number paradigms. To appear in *The Mental Lexicon*.

Clark, Eve V. 1987. The Principle of Contrast: a constraint on language acquisition. In MacWhinney, B. (Ed.), *Mechanisms of language acquisition* (pp. 1–33). Lawrence Erlbaum Associates, Inc.

Denison, D. (2003). Log(ist)ic and simplistic S-curves. In Hickey, R. (Ed.), *Motives for Language Change* (pp. 5470). Cambridge, Cambridge University Press.

Janda, A. L., & Tyers, M. F. (2021). Less is more: Why all paradigms are defective, and why that is a good thing. *Corpus Linguistics and Linguistic Theory*, *17*(1), 109–141. https://doi.org/10.1515/cllt-2018-0031

Kováříková, D., Škrabal, M., Cvrček, V., Lukešová, L., & Milička, J. (2020). Lexicographer's Lacunas or How to Deal with Missing Representative Dictionary Forms on the

Example of Czech. *International Journal of Lexicography*, *33*(1), 90–103. https://doi.org/10.1093/ijl/ecz027

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

Polančec, J. (2017). Naknadni prijeglas iza glasova *s t z* u hrvatskom jeziku. *Rasprave* 43. 197-225.

Sims, A. D. (2015). *Inflectional Defectiveness* (Vol. 148). Cambridge University Press.

Thornton, A. M. (2011). Overabundance (multiple forms realizing the same cell): A noncanonical phenomenon in Italian verb morphology. In Maiden, M., Smith,J. C., , Goldbach, M. & Hinzelin, M.-O. (Eds.), *Morphological autonomy: Perspectives from Romance inflectional morphology*, 358–381. Oxford: Oxford University Press.

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

Thornton, A. M. (2019a). *Oxford Research Encyclopedia of Linguistics* (M. Aronoff, Ed.). Oxford University Press. https://doi.org/10.1093/acrefore/9780199384655.013.554

Thornton, A. M. (2019b). Overabundance: a canonical typology. In Rainer, F., F. Gardani, W. U. Dressler & H. C. Luschützky (Eds.), *Competition in Inflection and Word-Formation* (pp. 223258). Dordrecht: Springer.