Always far from perfect, yet always good enough

Farrell Ackerman[1], Ágnes Kalivoda[2], Robert Malouf[3]
[1] University of California, San Diego
[2] Hungarian Research Centre for Linguistics
[3] San Diego State University
fackerman@ucsd.edu, kalivoda.agnes@nytud.hu, rmalouf@sdsu.edu

A goal of morphological inquiry is identifying patterns of formal similarity and difference within particular languages and distilling common organizing principles between languages viewed as a complex adaptive and discriminative system. An informing assumption has been the explanatory role of systemic and diachronic dimensions in accounting for the stunning variety of morphological systems attested in the world's languages. One rich area of instructive complexity is the domain of complex predicates.

Cross-linguistically, complex predicates often exhibit *tmesis* and *univerbation* (Lehmann 2018) with respect to preverbs and prefixes, where adpositions, adverbials, (non-finite) verbs and incorporated elements are the historical resources exploited for a new morphological category participating in lexeme-formation. Concerning the isolate Chitimacha, Hieber (2018) argues that the development of preverbs is an instance of *category genesis* whereby a previously extant category in a language is redeployed, eventually assuming a novel grammatical function. A language system originally devoid of preverbs which affect the syntactic role, valence and meaning of verbs can undergo systemic reorganization in which this new category and its effects become embedded in the language's morphology and clausal syntax (Nash, Craig and Hale 1988, Booij ed., Los, Craig and Imbert 2008, Arkadiev 2015, among others). Category genesis demonstrates the need for a construction theoretic perspective on grammatical exponence and "wordhood".

We argue that an important goal for morphologists, and construction theoretical grammarians more broadly, is to explain the observation that grammar systems are better viewed in terms of *satisficing* than in in terms of *optimization*: the contingencies of history, usage and learnability conspire/self-organize to yield systems of related constructions that provide a resource for the generation of novel forms of known lexemes. This insight into the essential imperfection and imperfectability of language was identified by Elizabeth Bates (1979) in an informing analogy between the contingent basis of evolution in the development for biological life-forms and the contingent basis of evolution and development for both Language and languages (more recently see research ecological evolutionary developmental systems (Gilbert and Sarkar 2000, Sultan 2015, Balari et al. 2020)):

> We want to demonstrate that Nature builds many new systems out of old parts, and selects for organisms that can carry out the same reconstruction process ontogenetically, jerrybuilding the same new machines from the same old parts in a highly reliable fashion. Human language may be just such a jerrybuilt system... (Bates, E. 1979. "On the evolution and development of symbols." *The emergence of symbols: Cognition and communication in infancy,* 1–32.)

To illustrate how this operates, we focus on interdependencies in the empirical domain of cross-linguistic complex predicate constructions consisting of preverbs/prefixes and verb stems. We taxonomize these constructions in terms of two intersecting dimensions, namely, synthetic versus periphrastic exponence and inflection as external versus internal to derivation. We focus on what appear to be two stable synchronic systems; several Souian languages (Helmbrecht and Lehmann 2009, Kasak 2019, Marsault 2021) exemplifying synthetic exponence with the inflection internal to and interleaving with derivation (also see Vajda, Carter on Ket), and

Hungarian exemplifying periphrastic exponence with inflection internal to derivation. We will pay particular attention to the Hungarian constructions where we provide diachronic motivation for the typologically unusual phenomenon of inflecting preverbs.

Example (1) from Hočąnk (Souian; Helmsbrecht and Lehmann 2017) shows the effects of the diachronic development of preverbs (glossed as ISC) from transparently meaningful prefixes and into semantically opaque prefixes as parts of discontinuous lexemic roots, where inflectional markers are interposed between these prefixes and associated verb roots. In Hočąnk, as in many other Siouan languages, there has been no evidence of a shift towards externalizing these so-called trapped inflectional markers (in the sense of Harris and Farlund 2006, Haspelmath 1993). In (1) the inflectional pronominal marker for agentive 2nd singular exhibits multiple exponence with one marker interposed between the derived lexemic root *gi…ruk'as* consisting of an "initial derivational affix or submorpheme (ISC)" and verb root. Note that the Hočąnk forms are all synthetic.

(1) *ra-**gi-**šu-ruk'as*
A.2s.SG-**ISC**-A.2SG-take.off
'you take off'

The Hungarian constructions in which inflection-internal-to-derivation occurs in **both synthetic and periphrastic** expressions:

(2) a. *A vita és a **belé-m**-köt-és nem ugyanaz!*
the argument and the PV-1SG-tie-NMLZ not same
'Arguing is not the same as provoking me.'

b. *Bocs, hogy megint **belé-d**-köt-ök.*
sorry that again PV-2SG-tie-1SG
'I'm sorry to provoke you again.'

While there appears to be a cross-linguistically common process of complex predicate formation this eventuates in diverse language particular encodings, indicating the use of recurrent old parts to create new and viable systems of organization for phenomenally similar types of constructions. In each of the examples discussed, we demonstrate that the verbal derivations participate in a family of constructions: it is the systemic organization within each language family that facilitates a native speaker's acquisition and use of forms that they have never encountered before. Despite evident departures from any defensible notion of 'perfection', the existing systems appear to be organized as 'good enough' for learnability and creative use.

Morphological overabundance: Systemic and idiosyncratic factors in the use of parallel forms

Mari Aigro & Virve Vihman
University of Tartu
virve.vihman@ut.ee

Morphological overabundance may be expected to be more productively used by speakers if it shows systematicity through some subsection of the paradigm, such as plural noun forms. Estonian nominal morphology shows overabundance in various parts of the system, both as part of particular declension classes and for lexemes which can be inflected according to several classes. In addition to lemmas which exhibit parallel forms through the plural paradigm, various lemmas exhibit overabundance in singular cells, but this is assumed to be more varied. For instance, in the most overabundant singular cell, the illative, the use of parallel forms is believed to depend on derivational morphology, lexical semantics and frequency (Siiman, 2019: 24).

While usage-based accounts have been proposed for the parallel illative singular (Hasselblatt 2000, Kio 2006) and partitive plural forms (Siiman 2013, 2015, Kaalep 2010), we lack a general view connecting systemic overabundance with the question of speakers' online choices between forms. For example, the available parallel plural forms are known not to be all in productive use, but there is little research into the factors affecting the choice between them. Our corpus data suggest that the choice between forms may be affected as much by individual lexemes and their semantics as morphological paradigms and inflectional classes. In this paper, we examine the form and function of particular overabundant cells and the lexical semantics of lexemes exhibiting overabundance, as well as the corpus frequencies of parallel forms, in order to address the question of usage in the context of overabundance.

References

Hasselblatt, C. 2000. Estonian between German and Russian: Facts and fiction about language interference. *Languages in Contact*, 135-144.

Kaalep, H.-J. 2010. Mitmuse osastav eesti keele käändesüsteemis [Partitive plural in the Estonian case system]. *Keel Ja Kirjandus*, *2*, 94–111.

Kio, K. 2006. *Sisseütleva käände kasutus eesti kirjakeeles* [The use of the illative in Standard Estonian]. Master's thesis, University of Tartu. Available at: http://dspace.ut.ee/bitstream/handle/10062/865/kio.pdf

Siiman [Metslang], A. 2013. Mitmuse osastava vormidest toimetamata kirjakeeles [Partitive plural forms in unedited written Estonian]. *Oma Keel*, 21−29.

Siiman [Metslang], A. 2015. Mitmuse osastava sid- ja si-lõpu varieerumise kasutuspõhine analüüs [A usage-based analysis of partitive plural -*sid* and -*si* affix variation]. *Keel ja Kirjandus*, 11, 792−803.

Siiman, A. 2019. *Vormikasutuse varieerumine ning selle põhjused osastava ja sisseütleva käände näitel* [Form usage variation of the partitive and illative case and reasons for it.] PhD thesis, University of Tartu. https://dspace.ut.ee/handle/10062/64787

Hubs, dearth and inter-individual differences, analogically – the case of English *–ic* and *-ical*

Sabine Arndt-Lappe, Universität Trier

arndtlappe@uni-trier.de

Analogy-based morphological theories (and their computational implementations) assume that language users create novel word forms on the basis of similar existing forms in their Mental Lexicons (e.g. Skousen 1989; Daelemans & van den Bosch 2005). One important question about such theories is which of all forms in the Mental Lexicon count as 'similar', and why. Skousen's (2002 et seq.) computational model AML is particularly interesting in this respect because, unlike e.g. many k-NN models, decisions on what counts as 'similar' are made for each lexical item individually, on the basis of what is essentially a measure ensuring that decisions are made maximising the certainty of predictions.

The present paper will use the AML algorithm to shed light on how inter-individual differences in Mental Lexicon structure impinge on differences in productivity profiles among rival affixation processes, something that has remained largely unexplored so far. Empirically, we will be concerned with the rivalry of adjectival *–ic* and *-ical* in English. Examples of existing *-ic* and *-ical* words are provided in (1).

(1)	historic,	historical	(< history)
		biological	(< biology)
	magnetic		(< magnet)

This rivalry is particularly interesting because the distribution of *-ic* and *-ical* reflects the whole range of possibilities in the continuum between 'productive', 'regular' patterns on the one hand, and 'unproductive', 'irregular' patterns on the other (esp. Lindsay & Aronoff 2013; Aronoff & Lindsay 2014; Bauer, Lieber & Plag 2013; Hamawand 2007), according to conventional terminologies. Thus, *-ic* is generally more frequent than *-ical*, in terms of both types and tokens, and hence usually held to be more productive. At the same time, *-ical* is more productive than *-ic* in some formally defined 'niches' (Lindsay & Aronoff 2013); this particularly holds for stems ending in <olog> (like *biolog-ical* in (1)), where *-ic* is rare, and hence can be described as irregular. Finally, there is a strikingly high proportion of *-ic* and *-ical* doublets in the lexicon, suggesting that the variation is stable.

In an analogical model, different degrees of productivity may emerge when distributions provide different degrees of support for a pattern. In such a system, 'regularity' and 'irregularity' have no independent theoretical status. Instead, 'irregularity' is the extreme end on a scale of dwindling support.

This paper will present a series of simulation studies of *–ic* and *–ical* rivalry based on different lexicons. The statistical analysis of model performance shows that patterns of productivity are clearly predicted to differ for different lexicons. For example, on the basis of a lexicon that is restricted to high-frequency words like *practical* and *tactical* AML predicts *-ical* for creating the adjective *syntactical* from the base *syntax*. However, on the basis of a lexicon that also contains low-frequency, academic adjectives like *phonotactic*, *climactic* and *chiropractic*, the algorithm predicts *syntactic*. This is because in a big lexicon that also contains low-frequency words, model classification is based on a large set of highly similar items most of which end in [æktɪk] (a 'hub'). In a smaller lexicon that only has low-frequency items, however, the hub is much smaller.

In other cases, classification is based on sets of lexical items that are much less similar to each other, on the basis of a much more sparsely populated similarity space ('dearth'). This is the case, for example, for the classification of *cleric/clerical* on the basis of a lexicon that is restricted to high-frequency exemplars that are common in spoken language. In this case, AML makes its decision on the basis of words like *chemical* (sharing only the nuclear vowel [e]), and the pair *classic* and *classical* (sharing only onset consonants). By contrast, when given a lexicon that also includes rare words, classification is based on much more similar items like *spherical* and *anti-clerical*.

On a general level, we see that high productivity may be associated with both strong hubs and with 'dearth' configurations. Among the models presented, big lexicons raise chances of having strong hubs; predictions from small lexicons, by contrast, are more likely to be based on less specific, more sparse similarity configurations. Crucially, variation is predicted to be rather stable, with the exception of a few homogenous hubs, challenging the assumption commonly made that morphological systems tend to not sustain variability of this type.

Overall, accuracy of model predictions is in line with global descriptions of the productivity of *-ic* and *-ical* in the literature. Greater overall productivity of *-ic* is reflected in *-ic* being predicted more often and with greater certainty than *-ical*, and greater 'niche' productivity of *-ical* is predicted for bases ending in <olog>. Going beyond previous accounts, 'niches' in the AML models are neither restricted to <olog> nor to *-ical*.

Model predictions for inter-individual differences will be tested on novel *–ic* and *–ical* words that were experimentally elicited from speakers with different lexicons. So far, we have analysed data from a pilot study comparing formations by native and non-native (German L1) speakers of English. Our findings are promising so far. For example, native speakers choose *-ic* more often than non-native speakers. This is in line with the fact that support for *–ical* is greater than support for *–ic* when AML is run on a lexicon that is restricted to high-frequency items. Collection of more data is under way.

References

Aronoff, Mark & Mark Lindsay. 2014. Productivity, blocking and lexicalization. In Rochelle Lieber & Pavol Štekauer (eds.), *The Oxford handbook of derivational morphology* (Oxford Handbooks in Linguistics), 67–83. First edition. Oxford: Oxford University Press.

Baayen, Harald R., Petar Milin, Dusica Filipović Durdević, Peter Hendrix & Marco Marelli. 2011. An amorphous model for morphological processing in visual comprehension based on naive discriminative learning. *Psychological Review* 118. 438–482.

Bauer, Laurie, Rochelle Lieber & Ingo Plag. 2013. *The Oxford Reference Guide to English Morphology*. Oxford: Oxford University Press.

Daelemans, Walter & Antal van den Bosch. 2005. *Memory-Based Language Processing*. Cambridge: CUP.

Hamawand, Zeki. 2007. *Suffixal Rivalry in Adjective Formation*. London: Equinox Publishing

Lindsay, Mark & Mark Aronoff. 2013. Natural selection in self-organizing morphological systems. In Fabio Montermini, Gilles Boyé & Jesse Tseng (eds.), *Selected Proceedings of the 7th Décembrettes*, 133–153. München: LINCOM Europa.

Skousen, Royal. 1989. *Analogical Modeling of Language*. Dordrecht: Kluwer.

Structuralist Tradition Meets Empirical Data: Corpus Data Enhancing the Czech Internet Language Reference Book

Martin Beneš, Czech Language Institute, Czech Academy of Sciences

The paper is focused on empirically-based improvements of the Internet Language Reference Book (Internetová jazyková příručka in Czech). It applies the data from the Czech National Corpus processed by the corpus interface KonText to the traditional representation of nominal paradigms. Its overall goal is to discuss empirically-based changes in the paradigm representation so that the ILRB can offer up-to-date as well as more detailed information to the users. The paper focuses on cell overabundance. It deals with the question of how many (empirically based) variants are to be presented in an individual cell and under what circumstances. In doing so, it uses, as an illustration, the sample of 5,516 Npl cells of animate masculine nouns present at this moment both in the ILRB and in the SYN v8 corpus. It shows that Npl forms of some lemmas are not attested enough in SYN v8 to be said to be empirically based and deals with the principal question, where the real variation starts – what level of less frequent variant's relative frequency makes it attested enough to make it presentable as empirically based. The paper presents how different limits for attestedness in corpus and for relative frequency of the less frequent variant changes the number of: 1) Npl cells whose forms are not attested enough in the SYN v8 to be safely presented in the ILRB as empirically based, 2) Npl cells whose forms are attested enough and are not overabundant and 3) Npl cells whose forms are attested enough and are overabundant. In doing so, the paper tries to find optimal and safe limits for presenting empirical-based data concerning overabundance in the Czech declension system.

Ideal and real paradigms: confronting evidence from grammars and corpora

Neil Bermel (University of Sheffield, n.bermel@sheffield.ac.uk)
Martin Alldrick (University of Oxford, martin.alldrick@seh.ox.ac.uk)
Luděk Knittl (University of Sheffield, l.knittl@sheffield.ac.uk)

Working definitions of defectivity such as Sims (2015:2) and overabundance (Thornton 2012:183) assume the ability to generalize from descriptions (in a normative handbook) or found data (in a corpus) that a paradigm cell is functionally unoccupied or overoccupied, but as both Sims and Thornton note, the data and descriptions do not always match. How far, then, can linguists working on morphologically complex languages trust the descriptions they have at their behest? In our current project, which has involved experimental work with Czech native speakers to look at defective and overabundant paradigm cells, we found we first needed to operationalize the terms 'defective' and 'overabundant'. The current paper attempts to reconcile our experimental findings with the conflicting data sources used to choose our target lexemes, and to evaluate the usefulness of these sources in identifying defective and overabundant paradigm cells.

Czech poses an interesting problem for scholars interested in the canonicity (as per Corbett 2005 and subsequent works) of paradigms. Standard Czech conserves many features from the archaic sixteenth-century written variety on which it was based. It is not used in informal spoken settings and does not constitute anyone's 'native' variety. Differences between the morphology of the standard and that of current spoken varieties are profound and result in numerous well-documented examples of multiple exponents serving a single function in standard and spoken usage. Some of these differences assume high salience for native speakers and serve as markers of informal speech vs. formal speech or writing, forming what Thornton (2011: 362) terms *diaphasic overabundance,* whose variants in this instance distinguish 'high' vs. 'low' styles. Other examples of variation, however, seem to have low salience, and it is difficult to detect any principled motivation for them. (Similar unmotivated variation has been noticed e.g. for Croatian by Lečić (2015).)

While overabundance in Czech has been treated in a number of publications (*inter alia* Bermel, Knittl & Russell 2015, 2017), defective behaviour has gone largely unremarked as a scholarly topic and is not reflected consistently in any published sources; dictionaries, for example, use a variety of terms to mark forms that are suspect. To construct a study testing overabundant and defective items, we therefore attempted to uncover 'hidden' examples of defectivity in Czech.

Potential defective lexemes were identified using:

1. *Normative handbooks and manuals, primarily dictionaries* (grammars do not mention the phenomenon except in the context of *singularia/pluralia tantum* nouns);
2. *Identification of analogous paradigmatic slots* to those where Russian, a language well-known for defective paradigms, sometimes has absent forms; and adduction of analogous structural examples to those found in Rosen's (2003) unpublished sounding;
3. *Corpus searches*: Trawls in balanced 100m-word corpora and two nonbalanced larger corpora of 5.5bn-10bn word forms for lexemes where forms from slots found above were absent; and outputs of the GramatiKat tool (Kovářiková 2021) to speculatively identify forms in theorized slots that had much lower-than-average occurrence.

A streamlined process was repeated on overabundant items drawn from the same grammatical categories as our defective ones:

1. *Normative handbooks and manuals*: The focus was primarily on grammar handbooks, where the phenomenon is well-described;
2. *Corpus searches*: Trawls of items found in handbooks were made in two balanced 100m-word corpora (SYN2015, SYN2020) of contemporary written language and larger corpora (SYN and CzechTenTen) for extremely low-frequency items.

Our findings suggested that handbooks were an adequate starting point for research into defective and overabundant slots, but not a reliable source of data compared to corpora:

- Czech handbooks may avoid implying that a slot is ‘empty’, and instead describe such slots as occupied by a form that may be labelled as ‘rare’ or ‘not typically used’. There were isolated examples from dictionaries to support Baerman’s (2008: 83) point that lexemes can become defective over time.
- Explicit designations of defectivity in handbooks pointed to a lower-than-expected frequency of the form(s) in question in corpus data, as expected. However, the reverse was not true. Forms lacking or unexpectedly infrequent in a corpus were sometimes nonetheless described in handbooks as canonical. The experimental data from our survey confirmed that respondents treated both groups of lexemes as defectives, suggesting that corpus data were more reliable in this instance than handbooks.
- Overabundance is comprehensively described in Czech handbooks, although they differ in their descriptions of how this variation is distributed and in its application to specific lexemes. Non-corpus-based reference works apply a partly asynchronic approach, presenting some marginal forms with historical provenance that may have been carried over from earlier works. Corpora give a broadly stable picture of variation, and native-speaker responses to our questionnaire tend to confirm corpus patterns, although there are some notable differences in the forms preferred.

References

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

Bermel, N., Knittl, L. & Russell, J. 2015. “Morphological variation and sensitivity to frequency of forms among native speakers of Czech.” *Russian Linguistics* 39, 283–308.

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, 197–231.

Corbett, G. 2005. The canonical approach in typology. In Frajzyngier, Z. et al. (eds.), *Linguistic Diversity and Language Theories*. Amsterdam & Philadelphia: John Benjamins, 25–49.

Kováříková, D. 2021. Sharing data through specialized corpus-based tools: The case of GramatiKat. *Jazykovedný časopis* 72, 531–544.

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

Rosen, A. 2003. Funkce bez formy aneb o nevyslovitelných věcech. Unpublished seminar presentation.

Sims, A. 2015. *Inflectional Defectiveness*. Cambridge: CUP.

Thornton, A. 2011. Overabundance (multiple forms realizing the same cell): A non-canonical phenomenon in Italian verb morphology. In Maiden, M. et al. (eds.), *Morphological Autonomy*. Oxford: OUP, 358–381.

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

An uncertain future? On ‘problematic’ paradigmatic cells in Czech

Neil Bermel (University of Sheffield, n.bermel@sheffield.ac.uk)
Alexandre Nikolaev (University of Eastern Finland, alexandre.nikolaev@helsinki.fi)
Luděk Knittl (University of Sheffield, l.knittl@sheffield.ac.uk)

A goal of the Imperfectability of Morphology workshop is to ‘arrive at tentative explanations, functional or otherwise, for recurrent patterns of adaptation’, i.e. the mechanisms by which we handle ‘noisy deviations… intrinsic to the system’ (Workshop Proposal, p. 4). The current paper starts with two phenomena presented in the linguistic literature as contrasting examples of gaps and superfluous forms (WP, p. 3). *Overabundant* cells occur in paradigms such as *prove,* which has two perfect participles *(I have proved/proven);* they represent unmotivated choice beyond what seems to be ‘needed’ in the grammatical system of a language. *Defective* cells lack a specified form and exemplify the hesitation of users in producing the past tense or participle of verbs such as *troubleshoot* or *output,* where expected forms like *troubleshot/troubleshooted* or *output/outputted* are reported to be unsuitable, leaving periphrastic (*did some troubleshooting; created outputs)* or synonymic substitution as the most satisfactory options. Following the idea that language is a complex adaptive system (workshop proposal; Beckner et al. 2009; Divjak et al. 2021), we look at ‘defective’ and ‘overabundant’ as shorthand labels for speakers’ prior exposure and experience.

We bring forward examples from Czech verbal paradigms where the lexeme has multiple plausible stems, introducing high uncertainty along the paradigmatic axis (WP, pp. 1-2). Typically, in some forms of the word, only one stem is expected, whereas in some others, a second or even a third stem is potentially available. These are:

- The non-past tense of some verb paradigms
- The passive participle of some transitive verbs (irregulars and some paradigms)

In such instances, three resolutions are possible: (a) the uncertainty resolves consistently in favour of one form, i.e. non-variance (**N**), a ‘canonical’ cell as per Corbett (2005); (b) the uncertainty persists, and two or more forms are found and accepted, i.e. overabundance (**OA**); (c) the use of any of the forms is avoided, i.e. defectivity (**D**).

To examine respondents’ reactions to multiple-stem nouns and verbs, we designed an experiment testing the following hypotheses:

- **H1**: Czech native speakers produce a greater variety of forms for D and OA cells than for N cells.
- **H2**: Czech NSs produce the ‘most regular’ choices less often for D and OA cells than for N cells.
- **H3**: Czech NSs are slower at producing forms to fill D and OA cells than they are for N cells.

We predicted in each instance that the values for D, OA and N cells would differ in predictable ways. For each, D < OA < N, where ‘<’ can mean ‘less uniformity’, ‘less often’ or ‘less decisiveness (longer reaction time)’, whereas for the null hypotheses, there would be no distinguishable differences between D, OA and N cells.

The lexemes in our experiment were identified by examining handbooks for typical contexts in which variation and gaps occurred, and checking these against corpus data (Czech National Corpus [SYN2015] and Czech TenTen to validate defectivity). Exploratory methods for identifying lexemes with paradigm gaps were also used (Kovaříková et al. 2020).

In our experiment, the high paradigmatic uncertainty referenced above was paired with low syntagmatic uncertainty: it consisted of a gap-filling exercise, in which respondents encountered paired sentences containing target lexemes, some of which were posited in advance to have defective or overabundant cell(s). In the first sentence, respondents were presented with a form where deviation was highly improbable (e.g., the nominative singular); in the second, there was a syntactic context requiring a form where variation was conceivable, and respondents were instructed to fill that gap. (There were a matched number of filler items, using the same syntactic contexts, where no variation is typically found.)

Our results were in some respects expected and in others surprising.

We found discrepancies between expected and actual answers, with a greater variety of items populating each slot than anticipated, even where no variation had been foreseen. Work with nouns had led us to predict that corpus frequencies of forms, rather than of lemmas, would be most predictive, but with verbs, frequency effects were absent. Between-speaker variation was visible, with educational level and age affecting the answers given and the speed at which they were produced. Even allowing for the artificiality of the task in eliciting responses in unusual contexts, it seems that the prescriptive works used as the basis for our classifications were not always accurate in predicting what sorts of answers might appear, and even corpus data underestimated the amount of variation found.

Respondents reacted to defective cells with a significantly greater variety of answers, and they were significantly slower to initiate and complete their responses to the triggers. Differences between overabundant and non-variant cells were less pronounced, differing significantly only in that in overabundant cells, respondents frequently failed to converge on a single variant, while in non-variant cells they always did. Overabundance can thus be seen operationally as an intermediate step between non-variance and defectivity, in which uncertainty is present but does not typically cause unusual behaviour for language users.

References

Czech National Corpus [Český národní korpus]. 2015. SYN2015: A representative corpus of written Czech. Prague: Faculty of Arts and Philosophy, Charles University. (http://www.korpus.cz)

csTenTen: Corpus of the Czech Web. 2018. (https://www.sketchengine.eu/cstenten-czech-corpus)

Beckner, Clay, Richard Blythe, Joan Bybee, Morten H. Christiansen, William Croft, Nick C. Ellis, John Holland, Jinyun Ke, Dian Larsen-Freeman & Tom Schoenemann. 2009. Language is a complex adaptive system: Position paper. *Language Learning* 59, 1–26.

Corbett, Greville. 2005. The canonical approach in typology. In Zygmunt Frajzyngier, Adam Hodges & David S. Rood (eds.). Linguistic Diversity and Language Theories. Amsterdam/Philadelphia: John Benjamins, 25–49

Divjak, Dagmar, Petar Milin, Adnane Ez-zizi, Jarosław Józefowski & Christian Adam. 2021. What is learned from exposure: an error-driven approach to productivity in language. *Language, Cognition and Neuroscience* 36, 60–83.

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

Form predictability and cell frequency: a behavioural study

Maria Copot Olivier Bonami
maria.copot@etu.u-paris.fr olivier.bonami@linguist.univ-paris-diderot.fr

Background

Much recent work in morphology moving away from the traditional view of exceptionality as a binary property (Prasada & Pinker, 1993; O'Donnell, 2015; Yang, 2016) and takes it to be a continuum (Bybee & Slobin, 1982; Rumelhart and McClelland, 1986; Smolensky, 1995; Albright, 2002; Blevins, 2016; Herce, 2019). The renewed interest in paradigmatic structure and information theory has provided a useful framework for thinking about word form exceptionality in a quantitative fashion: a word form's exceptionality can be operationalised in terms of surprisal or entropy involved in predicting it from another member of its paradigm (the intuition behind Albright, 2002; Albright & Hayes, 2003; made more explicit in Ackerman & Malouf, 2013; Bonami & Beniamine, 2016), which in turn can be derived from the type frequency of the patterns that exist between the two cells.

It is also increasingly a matter of interest that paradigmatic form predictability interacts with various frequency measures, for reasons to do with linguistic processing and learnability (Milin et al., 2009; Divjak, 2019). The more high-frequency a word is, the more it can afford to have an unpredictable form, because its frequency ensures that its phonological form is highly active in memory and thus easily accessible. On the flip side, low frequency words are more likely to be easily predictable from other members of the paradigm: if a word is already syntagmatically uncertain (low-frequency words are tautologically an unexpected way to continue the average utterance), it's unlikely to tolerate additional uncertainty on the paradigmatic axis (Filipović Đurđević & Milin, 2018).

Building on this, Copot & Bonami (2021) show in a corpus study that the frequency in use of a word is negatively correlated to its paradigmatic predictability (at parity of lexeme frequency, the word with the target meaning that is most easily accessible will be employed by speakers), but this relationship is moderated by the frequency of all members of the lexeme's paradigm (high-frequency lexemes and word forms will have representations in memory that are more independent of the pattern they instantiate, and so can be accessed through more direct retrieval, rather than have to be produced as the result of analogy), and by the frequency of the cell (if a cell is very frequent, it will rarely need to be predicted, but rather it will form the basis of prediction).

Motivation

Following Copot & Bonami (2021)'s findings, we perform a behavioural experiment on the interaction between word frequency, cell frequency (the summed frequency of all words filling a particular paradigm cell across lexemes), and paradigm predictability (how expected the form filling one cell of a paradigm is given the rest of the makeup of that paradigm).

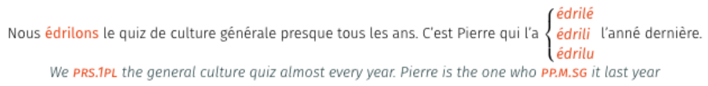
The corpus study employed average paradigmatic predictability (the average of a form's predictability based on each of its other paradigm members) as a predictor, so while it appears that form predictability matters on average, more work is necessary to establish the impact that paradigmatic predictability has on language processing. On this matter, we can ask 1) Are speakers sensitive to individual relationships of paradigmatic predictability between two cells/word forms in a larger paradigmatic system? Previous research on this topic has looked at small two-cell subsystems (the English past tense, the English plural) - claims concerning the paradigmatic nature of predictability would be stronger if evidence for them could be found in more complex systems. 2) Is the effect of paradigmatic predictability bidirectional? Or, as per Jun & Albright (2017), are predictability relationships only exploited when predicting from the base form?

Moreover, the corpus study used token frequency of a word as the variable to be predicted as a function of paradigmatic word form predictability, lexeme frequency and cell frequency, but token frequency interacts in complex ways with lexeme and cell frequency. To disentangle such interactions, we employ pseudoword stimuli - this enables us to isolate the effect of paradigmatic predictability and cell frequency.

Methods

To tackle these questions, we implement a modified version of Jun and Albright (2017)'s methodology with French data. Experimental items are sentences containing the same pseudolexeme twice, in two different inflected forms. Participants are asked to use a continuous, unmarked slider to express a well-formedness judgement on the second inflected form, under the assumption that the first form belonged to the same lexeme. The hypothesis is that the more predictable the second form is from the first, the higher the score it will receive. Furthermore, form predictability is expected to matter more when predicting towards more frequent cells (for which speakers have a better grip on pattern distribution), and that judgements towards less frequent cells will on average be higher (speakers are more willing to be accepting of forms in cells which they've been exposed to fewer examples of).

Experimental items varied in two dimensions: the identity of the paradigmatic cells involved, and the degree of predictability of the second form from the first. Two cell pairs were chosen based on the range of predictability values of the possible patterns of alternation to be found between them: INF⇆IND.PRS.2PL, IND.PRS.1PL ⇆PP.M.SG. To test if the effect of predictability is bidirectional, items for each cell pair varied which cell was first vs second in the sentence (2 cell pairs * 2 directions of prediction = 4 cell conditions). To test the effect of form predictability on judgements, each experimental item had three possible versions of the second word form, differing in the degree to which the second inflected form was predictable based on the first inflected form.



Nous édrilons le quiz de culture générale presque tous les ans. C'est Pierre qui l'a {édrilé
édrili
édrilu} l'anné dernière.

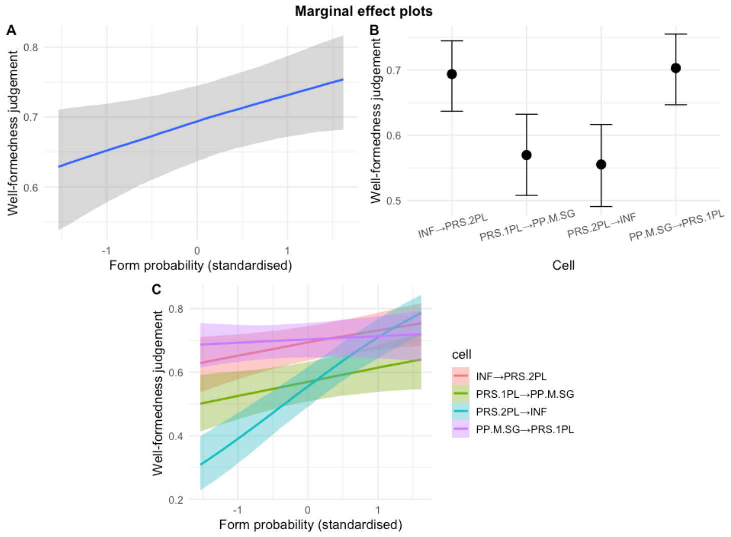
We PRS.1PL the general culture quiz almost every year. Pierre is the one who PP.M.SG it last year

An example IND.PRS.1PL → PAST.PART.M.SG item in its three variations according to the predictability of the second form from the first.

A maximal bayesian zero- and one-inflated beta regression with by-participant and by-item random effects was fitted to the experimental data. Form probability was obtained using Calderone, Hathout & Bonami (2021)'s methodology.

Results & Discussion

Form predictability has a positive effect on well-formedness judgements (fig. A): on the margin, an increase of one standard deviation in form probability will lead to a 6.45% increase in well-formedness score. The result corroborates previous empirical findings on the cognitive relevance of paradigmatic predictability, and confirms the importance of treating anomaly as a matter of degree.



Participants are more generous when scoring forms in less frequent cells (fig. B). The mean scores for infinitive and past participle forms (the two most frequent verbal cells in French) are 15% lower than those for the two present indicative cells (of middling frequency for French verbs). Speakers are willing to accept rarer patterns more readily when they are less familiar with the distribution of possible patterns in a cell. Cell frequency also dictates how confident speakers are about the distribution of patterns: the importance of form predictability for well-formedness judgements is proportional to the frequency of the cell (fig. C).

Contrary to Jun & Albright (2016), speakers exploit paradigmatic predictability relationships between two cells bidirectionally. In fact, once other factors have been controlled for, form predictability matters most when predicting the INF, which is both the citation form and the best overall predictor of the rest of the paradigm. Despite their opposite conclusion, Jun and Albright's results are compatible with ours: we predict that once cell frequency is taken into account, their findings can be given the same interpretation as ours.

References: Ackerman, F. & Malouf, R. 2013. 'Morphological organization: The low conditional entropy conjecture.' Language 89 – **Albright, A**. 2002. 'Islands of Reliability for Regular Morphology: Evidence from Italian'. Language 78 – **Albright, A., & Hayes, B.** 2003. 'Rules vs. analogy in English past tenses: a computational/experimental study'. Cognition 90 – **Bonami, O & Beniamine, S.** 2016. 'Joint predictiveness in inflectional paradigms'. Word Structure 9 – **Bybee, J. & Slobin, D.** 1982. 'Rules and Schemas in the Development and Use of the English past Tense'. Language. 58 – **Calderone B. & Hathout. N & Bonami, O.** 2021 'Not quite there yet: Combining analogical patterns and encoder-decoder networks for cognitively plausible inflection' 18th SIGMORPHON Workshop **– Copot, M. & Bonami, O.** 2021. 'Spare us the surprise: the interplay of paradigmatic predictability and frequency'. Talk given at ISMo 2021 – **Divjak, D.** 2019. 'Frequency in language: Memory, attention and learning'. Cambridge UP – **Filipović Đurđević, D. & Milin, P.** 2019. Information and learning in processing adjective inflection. Cortex 116 – **Herce, B. 2019.** Deconstructing (ir)regularity. Studies in Language. – **Jun, J., & Albright, A.** 2017. 'Speakers' knowledge of alternations is asymmetrical: Evidence from Seoul Korean verb paradigms'. Journal of Linguistics 53 – **O'Donnell, T.J.** 2015. Productivity and Reuse in Language: A Theory of Linguistic Computation and Storage**. – R- Rumelhart, D. E. & J. L. McClelland** 1986. `On learning past tenses of English verbs´. In Parallel Distributed Processing: Vol 2, MIT Press – **Smolensky, P.** 1995. `Constituent structure and explanation in an integrated connectionist/symbolic cognitive architecture´. In: Macdonald, C., Macdonald, G. (Eds.), Connectionism: Debates on Psychological Explanation. Blackwell.

Is all morphological marking remembered in the same way? A perspective from research on memory

Dagmar Divjak [1], Petar Milin [1], and Srdan Medimorec [2]
[1] University of Birmingham, UK; [2] Teesside University, UK
ooominds@ooominds.org

There is an interesting parallel between the linguistic dichotomy of lexicon versus grammar, and the psychological dichotomy of declarative versus procedural memory. Declarative memory and non-declarative or procedural memory are long-term memory systems. Declarative memory is flexible, relational memory that can be accessed consciously and intentionally to guide behaviour in new situations (Reber, Knowlton, & Squire, 1996). Non-declarative memory, although less well understood than declarative memory, is generally accepted to be acquired and accessed unconsciously. Broadly speaking, the two long-term memory systems differ in learning (i.e., acquisition of new information) and retrieval (access or recall). Learning by the declarative system is fast, but access to the information it harbours is slow and controlled (MacDonald, 2008), as well as fallible (Squire, Knowlton, & Musen, 1993). On the contrary, learning in the non-declarative mode is slow, but access to the information is fast and automatic as well as reliable (Squire et al., 1993; MacDonald, 2008). In addition to difference in learning and retrieval, it is generally accepted that the formation of procedural memories can be attested only indirectly: through a change in behaviour.

It has been claimed that lexical knowledge would fall under the purview of declarative memory while procedural memory would harbour grammatical knowledge (Ullman, 2004). The exclusive focus on syntax and the lexicon is at least in part due to the central position that syntax and the lexicon occupy in theories of language and language cognition. Whereas generative, dual-route models are heavily invested in a lexicon-grammar split, for single-route models such as usage-based linguistics these are extremes of a continuum. In other words, for the former, the nature of the problem-space is *taxonomic*, while for the latter, it is *graded*. In this study, we turn the tables: rather than selecting stimuli of types that fit theoretical assumptions about memory and language, we select stimuli of types that represent language in order to detect their memory signatures. The implications are clear: if the involvement of either of the two memory domains is binary (present/absent) then the taxonomic account is justified; otherwise, if both domains are engaged to a certain extent, then the single-route graded account is correct.

Starting from the observation that, during retrieval, only declarative memory is disrupted by divided attention we implemented a dual-task paradigm to test how nominal morphology (case) and verbal morphology (aspect) behave compared to syntax (subordination) and the lexicon under increased working memory load. Thus, we contrast knowledge of language, ranging from morphology (case and aspect) and syntax (subordination) to lexical semantics (collocations), and explore whether different categories of grammatical knowledge, and different types of morphological knowledge, behave in the hypothesised fashion of a clear divide. Our experimental paradigm contrasts a single-task, full-attention condition, in which a main task only is executed, with a divided attention condition in which execution of the main task is paired with a concurrent task. Participants were thus asked to judge correct and incorrect sentences as sole task, or in conjunction with a 3-digit span task, i.e., while they held 3 random digits in memory.

Considering the typical sample size in studies on memory for language, we recruited 48 participants (9 self-identified as male; mean age = 24.5 years, range 18-62) at the University of Warsaw, Poland. All participants were native Polish speakers and 91.7% (n=44) of participants had learned English as their first foreign language. Our participants were asked to carry out an auditory grammaticality judgment task. All participants heard 192 Polish sentences in total (see https://tinyurl.com/bdfz7eba). The stimuli were divided into two sets of 96 sentences each. Each set of 96 sentences contained 48 incorrect and 48 correct items, half of which were experimental items representing the phenomena of interest (aspect, case, subordination and lexicon) and half filler items. The audio files were generated using Google cloud text-to-speech services: we used WaveNet, pl-PL-WaveNet-B voice (for more details see https://cloud.google.com/text-to-speech/). The sound duration range was 2050-5550 ms.

We analysed three dimensions of participants' performance: speed of judgment (using a Generalized Additive Mixed-Effects Model, Wood, 2006), accuracy of judgment (using Log-Linear Modelling, Rudas, 2018) and consistency of judgment (using a 3-point rolling standard deviation of time taken to reach a decision, and analysed with a Quantile Generalized Additive Mixed-Effects Model, Fasiolo, Wood, Zaffran, Nedellec, & Goude, 2021). For judgment speed, there was a significant interaction of single versus dual task condition and linguistic type whereby all linguistic types were affected by the memory load, albeit to different extents: in order of magnitude, collocations were followed closely by subordination, which was followed by case and aspect. In the single task condition, case and subordination group together, as do aspect and collocations. This pattern also shows under concurrent task conditions, but it is a consequence of collocations being affected most and aspect least by memory load. For accuracy of judgment, there was no effect of single versus dual task condition with case and subordination consistently causing significantly fewer mismatches between participant and experimenter judgment than aspect and collocation. Analysis of the rolling standard deviation on the time taken to reach a decision showed the same pattern: there is no interaction between Type and single versus dual task Condition and instead, case and subordination consistently show less variation in time to decision.

Our findings confirm the existence of a distinction between lexicon and grammar as a generative, dual-route model would predict, but the distinction is graded, as usage-based models assume. The hypothesized grammar–lexicon opposition appears as a continuum on which grammatical phenomena can be placed as being more or less 'ruly' or 'idiosyncratic'. However, usage-based models, too, need adjusting as not all types of linguistic knowledge are proceduralised to the same extent. The dual-task paradigm revealed that one of our four linguistic types is, indeed, mainly declarative in nature (lexical collocations) while the other types show traces of procedural memory to different extents. Crucially, syntax differs least from the lexicon in terms of memory load effect, and the real "opposition" is one between lexicon and morphology (aspect). Within conditions, however, morphology (case) and syntax (subordination) pair together and differ from morphology (aspect) and the lexicon (collocations), in terms of judgment speed, accuracy and stability.

This move-away from a simple dichotomy fundamentally changes how we think about memory for language, and hence how we design and interpret behavioural and neuro-imaging studies that probe into the nature of language cognition.

Fasiolo, M., Wood, S. N., Zaffran, M., Nedellec, R., & Goude, Y. (2021). Fast calibrated additive quantile regression. *Journal of the American Statistical Association, 116*(535), 1402-1412.

MacDonald, K. B. (2008). Effortful Control, Explicit Processing, and the Regulation of Human Evolved Predispositions. *Psychological Review, 115*(4), 1012-1031.

Reber, P. J., Knowlton, B. J., & Squire, L. R. (1996). Dissociable properties of memory systems: differences in the flexibility of declarative and nondeclarative knowledge. *Behavioral neuroscience, 110*(5), 861-871.

Rudas, T. (2018). *Lectures on categorical data analysis*. New York: Springer.

Squire, L. R., Knowlton, B. J., & Musen, G. (1993). The structure and organization of memory. *Annual Review of Psychology, 44*, 453-495.

Ullman, M. T. (2004). Contributions of memory circuits to language: the declarative/procedural model. *Cognition, 92*(1), 231-270.

Wood, S. N. (2006). *Generalized additive models*. Boca Raton: Chapman & Hall.

More than just ‘adding the -ed’: Can we predict verb overgeneralizations in morphologically rich languages?

Gordana Hržica
University of Zagreb
gordana.hrzica@erf.unizg.hr

Tomislava Bošnjak Botica
Institute of Croatian Language and Linguistics
tbosnjak@ihjj.hr

Sara Košutar
University of Zagreb
sara.kosutar@erf.unizg.hr

The acquisition of inflectional morphology can be particularly challenging for children and often leads to the production of incorrect forms, including overgeneralization of a given rule to irregular word forms (e.g., *bring* to *bringed* instead of *brought* in English). Overgeneralized forms reflect the complexity of the morphological system and reveal the strategies children use when confronted with it. Highly inflected languages exhibit some features related to early language acquisition that are not entirely consistent with findings in less inflected languages (Dressler 2005), especially English, which is usually taken as a model to draw inferences about this phenomenon. The Croatian conjugation system displays different degrees of complexity based not only on the number of inflectional morphemes but also on an elaborate system of stem changes. For most verbs it is not sufficient to attach the corresponding affix to the infinitive stem because the infinitive stem and the simple present stem differ in phonological features (cf. *šet-a-ti.*INF ‘to walk’ and *šeć-e-m.*PRS.1SG ‘I walk’). During early language development, children are likely to use overgeneralized forms to overcome this complexity. These forms are often used interchangeably with ‘adult-like’ forms (e.g., *plesati.*INF ‘to dance’ > *plesam* / *plešem.*PRS.1SG ‘I dance’), that is, children use more than one form per slot in the morphological paradigm. Both the correct and overgeneralized forms are attested simultaneously, but not with the same frequency. In the research on the acquisition of inflectional morphology, two main factors influencing overgeneralization can be identified – *token frequency* (how often a child is exposed to a particular verb form) and *phonological neighbourhood density* or *class size* (the number of verbs with phonologically similar word stems bearing the same corresponding inflectional morpheme). However, not many studies investigated the influence of these factors in morphologically rich languages, especially on the acquisition of verbal morphology (cf. Kirjavainen, Nikolaev & Kidd 2012; Engelmann et al. 2019).

The aim of this study was to investigate the production of overgeneralized verb forms in preschool Croatian-speaking children aged 2;6 to 5;11 using a questionnaire in which parents report overgeneralizations in the language of their children. We hypothesized that parents will report overgeneralized forms in all verb classes in which the stem changes, but that the frequency of overgeneralizations will depend on the features of the input, i.e. it will correlate with the frequency of verbs (higher rate of overgeneralizations for infrequent verbs) and class size in lemmas, tokens and tokens of selected verb types (higher rate of overgeneralizations for verbs with smaller class size). To date, studies have used a corpus-based method to retrieve overgeneralizations in child language, which has had limited success in capturing this phenomenon due to the low density of language sampling and low-frequency phenomena such as overgeneralizations tend to be underrepresented. A parental questionnaire could provide more precise information about the overgeneralization of verb forms. The verbs for the questionnaire were selected from the longitudinal *Croatian corpus of child language* (Kovačević 2002) to ensure that they occur in *child-directed speech* (CDS). To refine the selection, we used two lexical databases that contain information on the estimated age of acquisition and subjective word frequency retrieved from native speakers. A total of 36 verbs were included in the study. For each verb, two items were created: the 'adult-like' form and the overgeneralized form. Frequency of verbs was calculated from the longitudinal child language corpus (CDS) and corpus of written adult language (*hrWaC*, Ljubešić & Klubička 2014). We obtained two types of frequency information for each of the preselected verbs: the overall frequency of all morphological types of a verb and the frequency of specific morphological types of a verb that were central to our study. The class size was calculated from the same sources in lemmas, tokens and tokens of selected types. Parents were asked to indicate how often their child produces a particular verb form using a 5-point Likert scale. Altogether, 87 parents completed the online questionnaire, therefore, we obtained data for 87 children.

The results showed that parents reported overgeneralized forms in child language for all verbs included in the questionnaire. We found a negative relationship between the proportion of overgeneralized forms and the overall verb frequency in both CDS and *hrWaC*. Verbs with a lower frequency have a higher proportion of overgeneralized forms. We also found a negative relationship between the class size in tokens (CDS) and both the frequency of overgeneralized forms and the proportion of overgeneralized forms. The same results were found for the class size in specific morphological types of a verb (*hrWaC*). The frequency of overgeneralized forms and the proportion of overgeneralized forms are lower in larger classes, that is, in classes that are more frequent in the language surrounding children. The frequency and class size showed better correlations with the measures in the questionnaire when calculated on tokens of specific morphological forms, i.e. those where overgeneralization is expected. Our results are consistent with previous studies that have confirmed a facilitating effect of frequency and class size on the acquisition of inflectional morphology (e.g., Kirjavainen, Nikolaev & Kidd 2012; Engelmann et al. 2019). The present study reveals that that preschool children still resort to the mechanism of overgeneralizations to overcome the complexity of verbal morphology.

References

1. Dressler, W. U. 2005. Morphological Typology and First Language Acquisition: Some Mutual Challenges. *Morphology and Linguistic Typology, On-line Proceedings of the Fourth Mediterranean Morphology Meeting* (MMM4), eds. Geert Booij et al., University of Bologna.
2. Kirjavainen, M., Nikolaev, A. & Kidd, E. 2012. The effect of frequency and phonological neighbourhood density on the acquisition of past tense verbs by Finnish children. *Cognitive Linguistics 23*(2), 273–315.
3. Engelmann, F., Granlund, S., Kolak, J., Szreder, M., Ambridge, B., Pine, J. M., Theakston, A., & Lieven, E. 2019. How the input shapes the acquisition of verb morphology: Elicited production and computational modelling in two highly inflected languages. *Cognitive Psychology* 110, 30–69.
4. Kovačević, M. 2002. *Hrvatski korpus dječjeg jezika*. CHILDES project.
5. Ljubešić, N. & Klubička, F. 2014. {bs,hr,sr}WaC - Web Corpora of Bosnian, Croatian and Serbian. In Felix Bildhauer & Roland Schäfer (eds.), *Proceedings of the 9th Web as Corpus Workshop (WaC-9) @ EACL 2014*, pp. 29–35.

Studying negative evidence in Finnish language corpora

Alexandre Nikolaev (University of Eastern Finland, alexander.nikolaev@uef.fi)

In morphologically complex languages such as Finnish, the inflectional paradigm of one noun can include as many as two thousand members (inflected word forms). When inflectional paradigms are that large, it could be a matter of chance that some inflected word forms are absent even from a large corpus. However, if we use another corpus of a similar size or increase the corpus size, e.g., by a factor of ten, then what are the chances that these particular word forms will still be absent (showing zero frequency)?

The aim of the present study is to explore a boundary between paradigmatic gaps that are purely accidental and those that are missing for reasons beyond simple chance. We compare the frequentist and Bayesian approaches that different statistical tests offer. In the Bayesian approach we are allowed to use background information while setting our priors (expectations), and this is the area where subjective grammaticality judgments could be used to set priors, which then could be updated by the corpus data.

We use a corpus of modern Finnish (84,308,641 tokens) which is based on written conversations of thousands of users in a Reddit-like internet community (Aller Media ltd., 2014). We analyze corpus frequencies of all inflectional forms of the two thousand most frequent Finnish nouns (selected from the list of the 10,000 most frequent Finnish words; CSC, 2004). We compare how frequencies of the word forms appearing in one of the 14 possible Finnish cases – the instructive – are distributed among these nouns, and then discuss factors that may contribute to overrepresentation or underrepresentation of some nouns in the instructive case.

In the next step, we discuss the syntactic and semantic functions that make some nouns more suitable for use in the instructive case (resulting in situations when some of these inflected word forms can function as a different part of speech; namely, adverbials whose clear origin as instructive case forms of nouns is reflected in their corpus tagging), and make some other nouns less suitable candidates for the instructive case. The latter group might be considered ‘accidental zeros’, where an instructive form is missing due to a lack of suitable context, rather than being missing despite suitable contexts being available. The question is then how to distinguish these two qualitatively different zeros from each other.

Applied to corpus studies, frequentist tests (such as Fisher’s exact test, e.g., Stefanowitsch 2020: 274) seem to ignore a zero’s ability to be a placeholder and hence to be something greater than just the number preceding one. This approach seems not very informative when one analyzes inflectional morphology in languages that are inflectionally as complex as, e.g., the Finnish language. When a noun has on average 52 word forms out of 2,000 possible in a corpus of 84.3 million tokens, there is still plenty of room for many word forms to be unattested even in a much larger corpus. To be able to correctly decode some word forms one has never encountered before, one needs to match them with other stored/familiar word forms based on formational, referential and constructional similarities between them (see, e.g., Blevins 2006 for the Word and Paradigm model of language, and Janda & Tyers, 2018 for the notion of accidental paradigmatic gaps). The Bayesian approach gives us more room for analyzing zero frequency than the frequentist one because we can assign prior probability to an event of occurrence (or non-occurrence) of a particular word form. This can allow us to analyze corpus frequencies in conjunction with subjective acceptability judgments. And these acceptability judgments in turn can reflect the concept of the varying semantic need for particular word forms (grammaticality is not a feature that is either on or off, but rather a gradient phenomenon that is subject to change). We argue that this hybrid approach has the potential to outperform a simple approach that would rely solely on either corpus frequencies or on subjective acceptability judgments.

Aller Media ltd. (2014). The Suomi 24 Sentences Corpus (2016H2) [text corpus]. Kielipankki. Retrieved from http://urn.fi/urn:nbn:fi:lb-2017021505

CSC – Tieteen tietotekniikan keskus (2004). Frequency Lexicon of the Finnish Newspaper Language. Retrieved from http://urn.fi/urn:nbn:fi:lb-201405272

Blevins, J. P. (2006). Word-based morphology. *Journal of Linguistics, 42*, 531–573.

Janda, A. L., & Tyers, M. F. (2018). Less is more: why all paradigms are defective, and why that is a good thing. *Corpus Linguistics and Linguistic Theory, (published online ahead of print 2018), doi: https://doi.org/10.1515/cllt-2018-0031*

Stefanowitsch, A. (2020). *Corpus linguistics: A guide to the methodology*. Language Science Press.

Dictators and oligarchies: a morphological perspective.

Paul O'Neill
University of Sheffield
paul.oneill@sheffield.ac.uk

Historical changes can be extremely informative as to how speakers organise and process morphological material and can serve as external evidence by which to evaluate morphological models and theories. A case in point is the wealth of evidence from developments in the Romance verb which show how abstract distributional patterns of allomorphy, which can be realised in phonologically diverse ways, can condition morphological change, suggesting that such patterns are cognitively real for speakers (Maiden 2018). In this talk, I suggest that these patterns also challenge assumptions about the distinction between regular and irregular forms, the latter often considered as lexically autonomous and entrenched (Bybee 2001:124). I present historical evidence to suggest that it is important to consider irregularity in the context of the interplay between type and token frequency. Individual frequent lexemes with irregular morphology in paradigmatic cells of high token frequency are unlikely to influence other verbs, however, a number of frequent verbs with the same 'irregular' patterns of alternation over cells of both relatively high and low token frequency can be extremely influential for the organisation of the entire verbal morphology. And, in line with findings in cognitive linguistics (Strack and Mussweiler 1997, Goldberg et al. 2004), act as a prototype and model for less frequent tokens. In sum, morphological systems are much more sensitive to oligarchies than dictatorships. I suggest that this tendency is a result of the acquisition processes which largely takes place on the basis of a relatively few number of high frequency lexemes. I also note that the historical data is compatible with computational models based on temporal self-organizing maps (Pirrelli, Ferro, and Marzi 2015, Marzi and Pirrelli 2014) since simulations predict that 'irregular forms should act as "models" of the morphological organization of the speaker's mental lexicon'(Pirrelli, Herreros, and Calderone 2007, 285).

References

Bybee, Joan L. 2001. Phonology and language use. Cambridge: Cambridge University Press.

Goldberg, A, D Casenhiser, and N Sethuraman. 2004. "Learning Argument Structure Generalizations." Cognitive Linguistics 14 (3):289-316.

Maiden, Martin. 2018. The Romance Verb: Morphomic Structure and Diachrony: Oxford University Press.

Marzi, Claudia, and Victor Pirrelli. 2014. "A neuro-Computational Approach to Understanding the Mental Lexicon." Journal of Cognitive Science 16 (4):491–533.

Pirrelli, Victor, Marcello Ferro, and Claudia Marzi. 2015. "Computational complexity of abstractive morphology." In Understanding and Measuring Morphological Complexity, edited by Matthew Baerman, Dunstan Brown and Greville Corbett, 141–166. Oxford: OUP.

Pirrelli, Victor, Ivan Herreros, and Basilio Calderone. 2007. "Learning inflection: the importance of starting big." Lingue e Linguaggio 6:175-199.

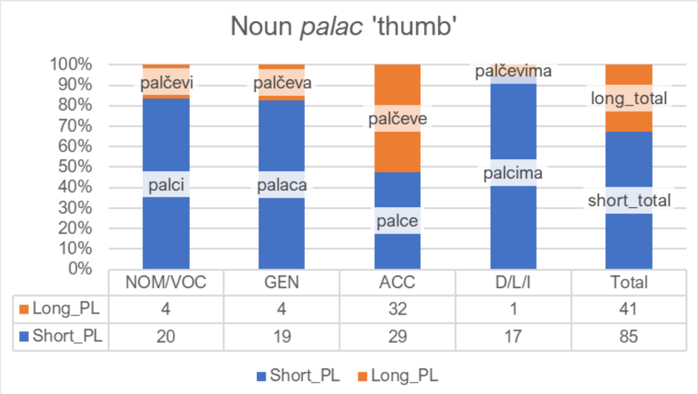
Strack, F, and T Mussweiler. 1997. "Explaining the enigmatic anchoring effect: mechanisms of selective accessibility." Journal of Personality and Social Psychology 73:437 - 446.

Mixed plural paradigms of masculine nouns in Croatian

Jurica Polančec, University of Zagreb, jpolance@ffzg.hr

Tomislava Bošnjak Botica, Institute of Croatian Language and Linguistics, tbosnjak@ihjj.hr

This talk deals with the subset of Croatian masculine nouns that form their plural paradigms by incorporating elements of two different plural formation strategies (hence the term *mixed paradigm*; cf. Štichauer 2018). The mixing refers to the situation when some paradigm cells employ one strategy while others use another. For instance, the noun *palac* 'thumb' typically employs the so-called short plural (*palc-i* 'thumbs, NOM. pl.'). In contrast, in the accusative, there is a slight preference for the so-called long plural, where the extension morph *-ev-* is added before the case ending *-e* (*palč-ev-e* 'thumbs, ACC. pl.'). The counts from the *Riznica* corpus of Standard Croatian can be seen in the chart to the right. Note that masculine nouns are usually uniform in this respect, as will be made clear by the general rules for the distribution of the two strategies, which are discussed next.



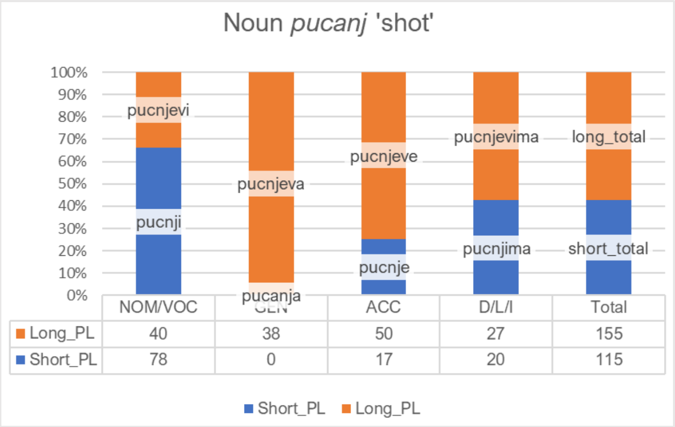
	NOM/VOC	GEN	ACC	D/L/I	Total
Long_PL	4	4	32	1	41
Short_PL	20	19	29	17	85

Croatian is unlike most other Slavic languages in that it employs two distinct plural strategies (Orlandi 1963). First, the long or extended plural paradigm is used with monosyllabic nouns (with very few exceptions). The plural extension morph has two forms (*-ov-* and *-ev-*) and is added to the noun before the plural ending (*stol* 'table, NOM. sg.' > *stol-ov-i* 'tables, NOM. pl.'). In contrast, the short plural paradigm, where no extension morph is used, occurs with polysyllabic nouns (*prozor* 'window, NOM. sg.' > *prozor-i* 'windows, NOM. pl.'). The plural paradigms for *stol* 'table' and *prozor* 'window' are summarized in the table. As can be seen, Croatian nouns (and other nominals) distinguish seven cases, but in the plural, only four forms (cells) are distinguished: the nominative/vocative, genitive, accusative, and dative/locative/instrumental cell.

Case (plural)	***stol*** **(long pl.)**	***prozor*** **(short pl.)**
NOM. / VOC.	*stol-ov-i*	*prozor-i*
GEN.	*stol-ov-a*	*prozor-a*
ACC.	*stol-ov-e*	*prozor-e*
DAT. / LOC. / INSTR.	*stol-ov-ima*	*prozor-ima*

However, contrary to the general rule, a substantial group of polysyllabic nouns occurs with the extended plural (*vitez* 'knight, NOM. sg.' > *vitez-ov-i* 'knights NOM. pl.'). Such nouns, in most cases, also allow for the plural without the extension morph, but the exact preferences are individual for each noun. With most nouns of the *vitez* type, preference for either strategy is uniform across the paradigm. For instance, with the noun *vitez,* the long plural is preferred across the plural paradigm (*vitezovi*, *vitezova*, *vitezove*, *vitezovima*), even though instances of short plural forms are also attested.

The nouns with mixed plural paradigms (e.g., *palac* 'thumb') are a subset of this group of polysyllabic nouns. However, unlike the nouns of the *vitez* type, preferences for either pattern is not uniform across the paradigm. Our earlier preliminary survey of plural paradigms of polysyllabic nouns, which was more narrowly focused on a sample of 18 nouns (Bošnjak Botica et al. 2022), showed that typically there is one cell that stands out in contrast to the other three, as with *palac* above. However, more complex patterns can also occur, as can be seen in the chart to the right. Here, the nom./voc. cell goes against the rest of the paradigm in its preference for the short plural, whereas in the gen. cell, the short plural is ruled out.



	NOM/VOC	GEN	ACC	D/L/I	Total
■ Long_PL	40	38	50	27	155
■ Short_PL	78	0	17	20	115

In this talk, we present a detailed, corpus-based description of nouns with mixed plural paradigms based on the sample of about a dozen instances of mixing that were identified in a larger sample of approximately 250 polysyllabic nouns. More specifically, we aim to make sense of this phenomenon by determining whether the mixing of plural paradigms can be seen as unmotivated, i.e., as a matter of lexical choice and historical accident, or, instead, as motivated, and therefore explainable by more general tendencies and rules of Croatian grammar (see Corbett 2016 for a distinction between motivated and unmotivated paradigm splits). Our preliminary analysis suggests that the answer may be different with different nouns.

REFERENCES

Bošnjak Botica, T., Polančec, J. & Sviben, R. 2022. Korpusno istraživanje hrvatskih imenica s dugom i kratkom množinom. *Jezikoslovlje* 23(1). 35–74.

Corbett, G. 2016. Morphomic splits. In A. Luís & R. Bermúdez-Otero (eds.), *The Morphome Debate*, 64–88. Oxford: Oxford University Press.

Orlandi, R. 1963. II plurale breve e lungo in serbo-croato. *Ricerche slavistiche* XI. 3–33.

Štichauer, P. 2018. Lexical splits within periphrasis: mixed perfective auxiliation systems in Italo-Romance. *Morphology* 28(1). 1–23.

What about meaning in inflection?

Elnaz Shafaei-Bajestan, Yu-Ying Chuang, Dunstan Brown, Roger Evans, Harald Baayen
University of Tübingen, {elnaz.shafaei-bajestan, yu-ying.chuang, harald.baayen}@uni-tuebingen.de;
University of York, dunstan.brown@york.ac.uk;
University of Brighton, roger.evans@nltg.org.uk

When it comes to inflection, introductory textbooks typically focus on words' forms, with scant attention to the details of words' meanings beyond the inflectional feature that they realize. The inflectional rule for pluralization in English, for instance, comes with the assumption that the exponent realizes the same semantics of 'plurality', irrespective of the semantics of the noun. However, an investigation of the word embeddings of English nouns using t-SNE (Van der Maaten and Hinton, 2008) with word2vec (Mikolov et al., 2013) in Figure 1 shows that the shift vectors between singulars and plurals vary widely, and are not well-captured by a single average shift vector. Of theoretical importance is that the orientation of shift vectors turns out to vary systematically with small sets of semantically similar nouns, resulting in some degree of semantic clustering. Since the constellations in which multiple objects of the same kind appear in the world varies greatly (compare apples, cars and skyscrapers), in hindsight, this is unsurprising. However, we are forced to conclude that the semantics of plurality is not constant, but highly variegated, challenging the current implementation of the Discriminative Lexicon model (Baayen et al., 2019). The variegated semantics of pluralization may also further advance our understanding of differences in the acoustic durations of the English plural (Tomaschek et al., 2019).

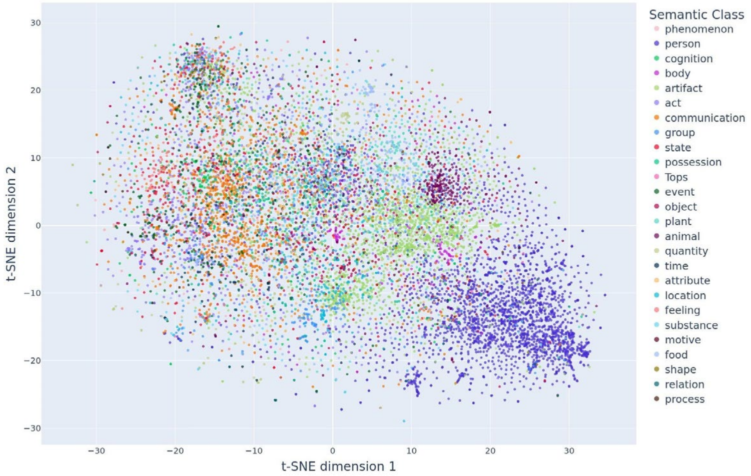


Fig 1: Projection of shift vectors onto a 2D plane using t-SNE shows semantic clustering.

An investigation of Russian using t-SNE for dimension reduction and visualisation on fasttext (Bojanowski et al., 2017) and word2vec shows an interesting effect based on paradigm size in which we can observe a flip in the t-SNE visualisation from a point where case and number forms are clustered around an abstract lexeme to one where forms are clustered according to case and number. On the left-hand side of Figure 2 word forms are included only for lexemes with form paradigms of size ≥ 12. Word forms cluster by lexeme with number (top left) and case (bottom left) distributed as we might expect across the clusters. When paradigms of size ≤ 10 are included, forms cluster according to morphosyntax; specifically by case (bottom right) and within cases by number (top right). Inclusion of lexemes with smaller paradigm sizes leads to a large increase in the number of tokens and consequent domination of case and number in determining clusters. This quantitative effect on the t-SNE reduction may have interesting implications for the structure of Russian speakers' mental noun lexicon. The right hand plots also show that the orientation of the shift vectors for number (singular → plural) varies by case, suggesting an interdependence between case and number in the distributional representations.

Our tentative conclusions are: (i) the singular-plural contrast is not semantically unitary and independent in English or Russian; (ii) in (case-free) English, it is mediated by lexeme semantic similarity; (iii) in Russian it is mediated primarily by case; (iv) case and number are not independent, but interacting unequal partners in the distributional behaviour of forms.

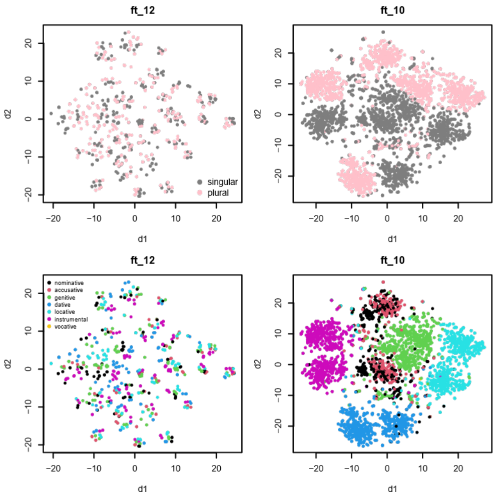


Fig 2: t-SNE clusters of Russian nouns classified by paradigm size and morphosyntactic feature.

These findings indicate that it may not make sense to discuss paradigm regularity and 'imperfectability' without a better understanding of the variegated semantics that characterize the system. For irregular forms there may be specific aspects of their distributional semantics that differ from the corresponding regularized forms (e.g., the semantics of irregular verbs in English, Baayen and Moscoso del Prado Mart´ın, 2005). For instance, for Russian nouns that are defective we have good evidence that there are semantic generalizations to be made in addition to the special form-based ones. They tend to have smaller paradigms, lower semantic transparency, as measured according to the method of Shen and Baayen (2021), and measurably reduced distributional semantic cohesion for case and number, as indicated by the fact that for case-number combinations the correlation (angle) with the average vector is less for defectives than non-defectives. Using the Russian WordNet thesaurus (Loukachevitch et al., 2016) we also find that defective nouns have fewer senses than non-defectives.

These results raise far-reaching questions about the extent to which the invariant phonemic representations and abstract semantic features with which paradigms are analysed block a more precise view on how language enables communication about the world.