What is learned from exposure

Divjak, Dagmar; Milin, Petar; Ez-zizi, Adnane; Jozefowski, Jaroslaw; Adam, Christian

DOI:
10.1080/23273798.2020.1815813

Document Version
Peer reviewed version

Citation for published version (Harvard):
Divjak, D, Milin, P, Ez-zizi, A, Jozefowski, J & Adam, C 2021, 'What is learned from exposure: an error-driven approach to productivity in language ', *Language, Cognition and Neuroscience*, vol. 36, no. 1, pp. 60-83. https://doi.org/10.1080/23273798.2020.1815813

Title

What is learned from exposure: an error-driven approach to productivity in language

Authors, Affiliations & Addresses

Dagmar Divjak[1 †], Petar Milin[2 †], Adnane Ez-zizi[2], Jarosław Józefowski[3] and Christian Adam[2]

[1]University of Birmingham

Department of Modern Languages & Department of English Language and Linguistics

Edgbaston

B15 2TT Birmingham

UK

[2]University of Birmingham

Department of Modern Languages

Edgbaston

B15 2TT Birmingham

UK

[3]Samsung Research & Development Institute Plac Europejski 1

00-844 Warszawa

Poland

[†] Joint first authors

Dagmar Divjak – corresponding author <d.divjak@bham.ac.uk>

Abstract

How language users become able to process forms they have never encountered in input is central to our understanding of language cognition. A range of models, including rule-based models, stochastic models, and analogy-based models have been proposed to account for this ability. Despite the fact that all three models are reasonably successful, we argue that productivity in language is more insightfully captured through learnability than by rules or probabilities. Using a combination of computational modelling and behavioural experimentation we show that the basic principle of error-driven learning allows language users to detect relevant patterns of any degree of systematicity. In case of allomorphy, these patterns are found at a level that cuts across phonology and morphology and is not considered by mainstream approaches to language. Our findings thus highlight how a learning-based approach applies to phenomena on the continuum from rule-based over probabilistic to "unruly" and constrains our inferences about the types of structures that should be targeted on a cognitively realistic account of allomorphic representation.

Keywords

productivity; error-driven learning; emergentism; inflectional morphology; allomorphy

1. Introduction

Natural languages are rich in structure and, as such, harbour an impressive range of patterns waiting to be discovered. Yet, arguably, language patterns are not a fully regular and easily predictable domain. Instead, alongside many regular patterns, languages feature a wide range of probabilistic tendencies that co-exist with (usually highly frequent) exceptions and irregularities (cf., McClelland & Patterson, 2002; Mirković, Vinals, & Gaskell, 2018; Seidenberg & Plaut, 2014). How language patterns are learned has been the topic of much debate. This debate has largely played out in the area of inflectional morphology, where novel forms frequently need to be generated. Of particular interest are allomorphic patterns, that is variation in form that does not imply any change in meaning (cf., Lieber, 1982; Lyons, 1968; Spencer, 2001). The English past tense or the English noun plurals are among the most well-known cases in point, but the phenomenon appears in many natural languages, including but not limited to Arabic (Ratcliffe, 1998), varieties of Australian languages (Patz, 1991), Czech (Bermel & Knittl, 2012), Dutch (Daelemans, Berck, & Gillis, 1997; Keuleers et al., 2007), German (Köpcke, 1993), English (cf., Berko, 1958; for a recent overview see Plag, 2018), Hungarian (Hayes & Londe, 2006; Kertész, 2003), Italian (Eddington, 2002a), Russian (Blevins, 2004), Serbian (Ivić, 1990; Zec, 2006), and Spanish (Eddington, 2002b).

Four main models have been proposed to explain the principles underlying the human ability to generate novel forms of words.[1] While a purely rule-based mechanism was long the standard, in particular in generative approaches to language cognition, emergentist or usage-based approaches to language favour a view in which productivity is exclusively based on analogy (Bybee, 1985; Rumelhart

[1] Another line of work takes an Information Theoretical view on morphological complexity. Languages that appear to differ greatly in their number of exponents and inflectional classes can be quite similar in terms of the challenge they pose for a language user who already knows how the system works: low conditional entropy among related word forms in paradigms makes it possible for speakers to make accurate guesses about unknown forms of words based on information from known words, regardless of the complexity of the system. This approach makes it possible to quantify how likely it is for a speaker of a language to produce a previously unknown form and explains what makes complex systems learnable (Ackerman & Malouf, 2013).

& McClelland, 1985; Skousen, 1989). Eventually, these two approaches were united in the dual mechanism model (Pinker & Prince, 1988) that relies on rules to cover regular processes and analogy to deal with all other cases. Albright and Hayes (2003) proposed another solution that relies on multiple stochastic rules. In this paper we side-step these four accounts and approach productivity from the point of view of learnability. We focus specifically on the following questions. How does the human capacity for learning handle this formal variational extravagance? Might a possible explanation for the learnability and retention of rich variation relate to the type of patterns that naïve language users detect in the input; that is, are there any reliable patterns present in the data that are not typically considered in linguistic analyses? And if there are, do these patterns reside at the level of the language as a system that is internalized by individuals or are they a by-product of aggregating over individual preferences for one ending or the other?

Addressing these questions will provide new insights into how speakers learn and use language and what they know about their language, which may affect the way in which linguists and psychologists approach the language system.

1.1. Modelling allomorphy: rules or analogies?

Allomorphy has been the focus of much attention across the language sciences. It is typically defined as "a situation in which a single lexical item, meaning, function, or morphosyntactic category has two or more different phonological realizations depending on context" (Paster, 2014, p. 219). That is, allomorphy refers to a situation where there are two (or more) endings per paradigmatic cell.[2] Due to their combination of regularity with a limited number of exceptions, English past tense and plural noun allomorphy have served as strategic examples in the battle for and against rule-based language cognition: it seems that, if a single (context-free) rule is insufficient, we can either (i) augment it to

[2] Allomorphy thus refers to two (or more) endings per paradigmatic cell and differs from doubletism (Bermel & Knittl, 2012) which refers to two (or more) endings per lexical cell, and from overabundance (Thornton, 2011) where there are two (or more) endings per lexical cell but no lexical differentiation.

accommodate exceptions, (ii) recast it in stochastic terms or (iii) account for it using analogical principle(s).

The English past tense, in particular, has been the topic of many computational simulation efforts over the past three decades. The Rumelhart and McClelland (1985) parallel distributed processing (PDP) connectionist model used a very simplified phonological representation (i.e., Wickelphones or Wickelfeatures; cf., Pinker & Prince, 1989) to show that rule-like behaviour could emerge from interactions among network units that encode how root forms are mapped to past tense forms. As the models have been given access to more extensive phonological representations (distinctive features combined with information on the position of the unit in the syllabic template in B. MacWhinney & Leinbach, 1991) which have also become more sophisticated (recall the phonological constraints in Albright & Hayes, 2003), their performance has improved considerably. Connectionist models have been shown to work well and mimic the behaviour of speakers of the language for less regular phenomena across languages too, e.g., German noun plurals (Marcus, Brinkmann, Vlahsen, Wiese, & Pinker, 1995; Hahn & Charles, 2000; Nakisa & Hahn, 1996) and Serbian case endings (Mirković, Seidenberg, & Joanisse, 2011), given a suitable specification of the input and sufficient training examples.

While connectionist models have concentrated on correlational representations that are sensitive to the statistical structure of the input, another fruitful computational approach to the previously mentioned set of phenomena relies on memory-based analogical modelling. Analogy is a structure-sensitive process that involves the comparison of systems of relations between items in a domain. It has been known to account for challenging phenomena such as Dutch diminutives (Daelemans et al., 1997), German and Dutch noun plurals (Keuleers et al., 2007; Wulf, 2002), Italian verb conjugations (Eddington, 2002a), Spanish gender assignment (Eddington, 2002b) and linking elements in Dutch and German (Krott, Baayen, & Schreuder, 2001; Krott, Schreuder, Baayen, & Dressler, 2007). Nevertheless, Albright and Hayes (2003) found that their stochastic rule-based model for the English past tense outperformed their implementation of Nosofsky's (1986; 1990) analogical

model in correlations to the experimental data because of the analogical model's overreliance on variegated similarity. The authors concluded that analogy, in its most basic form, is too powerful a mechanism to account for how morphological systems in human languages work (Albright & Hayes, 2003, p. 153). This power may well be the reason why analogy has been the mechanism of choice for modelling instances of allomorphy that defy any type of rules, absolute and probabilistic. TiMBL (Daelemans, Zavrel, Van der Sloot, & Van den Bosch, 2007) exploits mechanisms of analogical inference based on a cohort of the most similar entries and has been used successfully for the computational treatment of rule-defying inflectional allomorphy that is observed with masculine nouns in the Croatian and Serbian instrumental case (Lečić, 2016; Milin, Keuleers, & Filipović Đurđević, 2011), for example. Applying TiMBL to the English past tense, Keuleers (2008) showed that, after correcting for an inconsistent scaling in Albright and Hayes (2003), performance is on par with the rule-based model. He further showed that the iterative rule discovery that the Albright and Hayes (2003) model relies on is equivalent to a particular type of exemplar model. TiMBL also outperformed Analogical Modelling (Skousen, 1989) in handling allomorphy in the Croatian instrumental singular and genitive plural (Lečić, 2016).

More generally, while some researchers go as far as to describe models based on analogy as "the only game in town" (Ambridge, 2019), these models do not all perform identically. The architecture of Analogical Modelling (Skousen, 1989) is computationally expensive and can only handle a few features to distinguish exemplars, while TiMBL comes with several parameters that are crucial to achieving adequate performance (cf., the simulations in Milin, Keuleers, & Filipović-Đurđević, 2011). In fact, these models rely on extensive (hand)coding of input features (Milin, Divjak, Dimitrijević, & Baayen, 2016), which distinguishes them from the learning-based approach we advocate.

1.2. Modelling allomorphy: memory or learning?

In line with the storage metaphors that dominated research on memory until relatively recently, memory-based models, of which analogical models are a particular instantiation, are in essence large stores of tokens of multidimensional experience, or exemplars. Their engine relies on two cognitive processes to explain the linguistic knowledge that drives comprehension and production: the categorization of and analogical extension from stored exemplars. Because of their focus on the storage of tokens of experience on which certain cognitive operations are performed, they fit well with usage-based and emergentist theories of language (cf., Bybee, 2013; Brian MacWhinney, 1999). These theories assume that cognitive representations are built up as language users encode utterances and categorize them on the basis of phonetic form, meaning, and context. Utterances come in a variety of sizes, ranging from a single segment, such as a vowel, to whole texts, such as song lyrics or poems. As incoming utterances are sorted and matched by similarity to existing representations, units such as syllable, word, and construction emerge through abstraction. Thus, grammar can be viewed as the cognitive organization of one's experience with language (Bybee, 2006). It is, however, recognized that languages are constantly changing, and usage depends on the individual and the situation (Hopper 1987). The cognitive organization of language knowledge is, therefore, not a static structure, expressible as a fixed set of representations.

In other words, usage-based linguists recognize that both the experience and the organization of the experience are in a perpetual state of flux. This aligns with the starting point for theories of learning. In its most basic form learning is often presented, not as memory-based, but as error-driven and discriminative (cf., Ramscar, Yarlett, Dye, Denny, & Thorpe, 2010). Learning is sensitive to the non-randomness in the environment: "exposure to relations among events in the environment" helps an organism acquire knowledge about "the structure of its world" (Rescorla, 1988, p. 152). Learning models could provide a valuable alternative approach to exemplar-based models, especially from a Cognitive Linguistic stance (cf., Divjak, 2015; Dąbrowska, 2016; Milin et al., 2016). In fact, an approach

couched in learning is ideally suited for testing the emergentist perspective on language knowledge, which takes a holistic approach to language knowledge and recognizes gradience and continuity in input and representation. As in usage-based approaches to language, in learning theory, initially, all is one and events are discerned or contrasted (discriminated) only under adaptive pressure: "nothing separates except what must" (James, 1890, p. 488; for references in contemporary research see Ramscar, 2010; Ramscar et al., 2010).

Learning models provide a principled account of dealing with the patterns not (yet) present in experience by generalizing from input (Widrow & Lehr, 1990). In that respect, they empower linguists to build a consistent and consequential account of the effects of (partial and varying) systematicity in input and to model spoken or written language in a way that respects predictions that are realistic and can be tested experimentally (Milin et al., 2016): given a sample of naturalistic language use, are the abstractions (of various degree of schematicity and/or conventionality) learnable? Practically speaking, cues in usage events are allowed to compete for associability with (or discriminability for) outcomes. Based on the strengths of the associations between such cues and outcomes that emerge during learning, possible dimensions along which the experience may vary – and these can be linguistic abstractions (see Milin, Divjak, & Baayen, 2017) – can be proposed.

For analogical models, the features that distinguish experiences, along with the data these models receive, are the biggest factor determining the performance (Skousen, 2002). Milin et al. (2011), for example, coded exemplars in terms of orthography for modelling Serbian/Croatian case, yet also noted that "a clear improvement in predictions is to be expected if additional similarities were included". Krott et al. (2001), on the other hand, report a decrease in performance when phonological properties are included in the representation to model linking elements in Dutch. More generally, as Albright & Hayes (2003) point out, how an exemplar is encoded is typically defined by the researcher, yet a learner is faced with the task of discovering their distinguishing properties.

Another key difference between exemplar and learning models lies in how they generalize to unseen instances. Exemplar-based models organize their memory by forming neighbourhoods of units

on the basis of similarity between units (which can involve different criteria and computational rules). A new item is compared to available neighbourhoods, and sorted with the group to which it is most similar. Discrimination learning, on the other hand, gradually (iteratively) tunes the weights between cues and outcomes available in the environment and, hence, learns which cues are informative (or discriminative). These differences are precisely what the learning approach described above is concerned with[3].

Our approach bridges the gap between linguistic theory and learning theory: while the latter is interested in the process of learning, the former focuses on the outcomes of that process. Adopting a, what we will call, usage-based learning stance allows empirical evidence to accrue and alter the way in which we think about linguistic knowledge and language cognition. In Section 4, we will propose a learning-based computational model that relies on raw input only and does not require (hand)coded input features. To be recognized as general explanatory principle for the accumulation of language knowledge, however, learning should be able to deal with the whole gamut of distributions, ranging from (a) fully rule-governed cases, over (b) cases where histories of usage-events (or 'idiographic' experiences) can be summarized and rephrased as reliable probabilities (rules of thumb, in laymen's terms), to (c) those phenomena that are learned (as much as possible, but far from perfectly) even though a relatively stable experiential equilibrium cannot be reached (Danks, 2003 proposes a systematic account of such an equilibrium for a simple learning rule). An instructive case in point is the allomorphy exhibited by the Polish genitive singular of masculine inanimate nouns, which can be marked by -a or -u.

[3] Notably, TiMBL (Daelemans et al., 2007) weights features by how well they discriminate classes (formalized as information gain), which can be seen as an engineering short-cut (an approximation) of what learning as described here is sensitive to.

1.3. Cellmates: the Polish genitive of inanimate masculine nouns

For all but a few animate masculine Polish nouns the genitive desinence is -a. The situation is rather different for inanimate masculine nouns, such as *telefon* (phone), *komputer* (computer) or *tablet* (tablet) which become *telefonu*, *komputera* and *tabletu* respectively although all three are loanwords naming small electronic devices. It has been argued that, in Polish, the genitive singular of inanimate masculine nouns lacks a default ending (Dąbrowska, 2008): both -a and -u are used and although the majority of nouns take only one ending (or prefer one ending), there are no reliable principles to determine which of the two endings ought to be chosen. This situation is not unusual as "the formal variation associated with lexically conditioned allomorphy typically shows only a loose correlation with systematic phonological or semantic conditions (synchronically, at least) and often seems to serve no apparent communicative function" (Ackerman & Malouf, 2015, p. 1). The main culprit for the chaos in this area is historical change (Janda 1996, p. 329). To explain current use, criteria have been proposed that are phonological, morphological or semantic in nature (Bodnarowska, 1962; Mańczak, 1953; Westfal, 1956; Zwoliński, 1948), but most of these are rather unreliable. Even when considered in conjunction, the three dimensions do not yield unequivocal solutions; in fact they are often in conflict (cf., Dąbrowska, 2005; Kottum, 1981).

Do these considerations imply that a hierarchical system of constraints is waiting to be uncovered or, on the contrary, that there is no system at all and desinences are learned on a case-by-case basis? Some evidence for the latter stance would come from the observation that, in some cases, different genitive endings distinguish homonyms, as when the genitive of *zamek* is *zamka* when it means 'lock', but *zamku* when it means 'castle' and *przypadek* has *przypadka* when it means 'grammatical case' but *przypadku* when it means 'chance, coincidence' (see Zasina, 2017 for an overview). A case-by-case scenario of acquisition also fits with results from an analysis of spontaneous speech by young children: children appear to have little trouble in restricting each ending to a particular subset of nouns, even though this subset is essentially arbitrary (Dąbrowska, 2008, p. 571).

In a series of experiments, Dąbrowska (2004, 2005, 2008) established that, although productivity with the genitive inflection begins to develop before the age of 2;0, it does not reach adult-like levels until the age of 10;0. From the age of 10;0, speakers begin to show sensitivity to lexico-constructional patterns (such as those that mark the count vs mass distinction), but explicit influence of the referential properties of the noun (such as object vs substance) is not observed until adulthood, if at all. Dąbrowska (2008) therefore concluded that learners may be more sensitive to observable probabilistic distributional cues than to abstract semantic cues.

The question we address in this paper is, hence, which learnable cues, if any, are informative for allomorphic realisation. Can such cues be detected by a linguistically naïve algorithm that mimics human learning? And does that which is learned by the algorithm help explain human performance on a language task?

1.4. This study

We approach the issue of productivity in language with a study on seemingly unmotivated allomorphy that has nevertheless been passed on from generation to generation for many centuries. Instead of relegating such phenomena to the corners of the system, at the mercy of item-by-item memorization, we take a learning-based approach and re-examine the input, unconstrained by linguistic tradition. Is there a system, not under the purview of linguistic convention, that we 'patternovores' discern and learn? Is the apparent systematicity merely an epiphenomenon of aggregation and, hence, masks individual preferences in allomorphic choice?

In Section 2, we present a computational simulation study that sheds light on why allomorphy is learned and transmitted across generations, rather than being levelled analogically. In Section 3, we present a corpus-linguistic model that tests the predictive value of a range of variables identified in the literature on their ability to account for the -a/-u alternation. In Section 4, we build on previous work (e.g., Milin et al., 2016) and rely on a biologically and cognitively plausible discriminative learning

algorithm to discern qualitatively different patterns in the input; like linguistically naïve language users, the algorithm feeds on what words sound like and how they are used in context. The findings are compared against performance by native speakers on an experimental task in Section 5.

2. Learning and maintaining an 'unruly' distribution

We mentioned earlier that, to be recognized as general explanatory principle for the accumulation of language knowledge, learning should be able to deal with the whole gamut of distributions observed in usage. These range from fully rule-governed cases over cases where stochastic tendencies are apparent to those phenomena that appear 'unruly', because multiple rules compete and usage is not definitively established for all items (cf. Swan, 2002). The 'unruly' distributions are surprising on at least two counts. First, these distributions appear to be a mix of several rules that each hold to some extent. Second, this complex situation is transmitted across generations rather than abandoned in favour of the 'default' ending, even though there does not appear to be a clear functional motivation for the form variation.

To shed light on the mechanisms that drive the learning and maintenance of rule-governed, probabilistic and 'unruly' form variation we ran three computational simulation studies using an error-driven learning algorithm, the Widrow-Hoff rule (WH: Widrow & Hoff, 1960; this rule is in essence identical to the better-known Rescorla-Wagner rule, RW: Rescorla & Wagner, 1972; Rescorla, 2008). In a nutshell, the rule defines how an organism (an animal, human or a computer device) learns from its own errors in order to adapt to the task at hand. More specifically, the rule learns to associate, on an event-by-event basis, the presence of a given outcome with the cues that are informative about its occurrence by estimating connection strengths or weights from each cue to each outcome: if a given cue is consistently present when an outcome is present, their connection is strengthened, but if a given cue is repeatedly present when the outcome is absent, the weight on the connection between them is weakened. This dynamic ensures minimal error given all prior experience. Error minimization

depends heavily on cue competition: the more cues, the higher the competition, which increases the importance of misses (weights weakening) and decreases the importance of hits (weights strengthening). These weights are updated as experience accumulates and over time, some cues become discriminative for an outcome, while many become irrelevant. The overall support that an outcome gets from the cues, its activation, is the sum of the weights on the connection between those cues and the outcome. If there is a systematicity that can be established between cues and outcomes, experience accrues and learning evolves driven solely by the discrepancy between current 'best guess' and true outcome. This forms, somewhat simplified, the essence of error-driven learning (cf., Baayen, Milin, Filipović Đurđević, Hendrix, & Marelli, 2011; Chen, Haykin, Eggermont, & Becker, 2008; Milin, Feldman, Ramscar, Hendrix, & Baayen, 2017; Ramscar et al., 2010).

Our three learning simulations covered the three typical distributions (given in Table A1 in Appendix 1): one for a rule-governed distribution, one for a probabilistic distribution and one for an allomorphic distribution of the type seen in the Polish genitive singular. In each case, there were 10 items that could take one of two endings, here -x or -y, across 1000 training iterations for each item. The training regimes are summarized in Table A1 in Appendix A. For the current learning setup there are two specifics that need to be mentioned. First, in addition to the 10 cues (items) we have assumed a constant learning background; this essentially means that everything else in the context of learning is taken to be equally informative (or uninformative). A constant background is informative about the relative frequency of an outcome as well as about the number trials on which cue and outcome are not coupled (see background rate: Rescorla, 1968). Inclusion of such a background appears more feasible in small-scale computational simulations such as the ones presented below, where it is rather straightforward to draw inferences about how informative 'everything else' (i.e., the background) is for an outcome (compare here, for example, the use of a constant background in Spellman, 1996 and

the simulation of this study in Danks, 2003; a more detailed discussion about the background or context cue can be found in Milin, Nenadić, & Ramscar, Under revision).[4]

First, in the present study the, the background was implemented as an additional cue that was present on all trials. Second, the training assumed the learning rate parameter $\gamma = 0.01$, which is a commonly used value, small enough to guarantee incremental learning (cf., Rescorla & Wagner, 1972; Enquist, Lind, & Ghirlanda, 2016). The learning process is visualized in Figures 1, 2 and 3. Figure 1 displays how a rule-based phenomenon with an exception is learned.

Comparing the left panel to the right panel in Figure 1 reveals an interesting contrast: the background rate (the black dotted line), or the rate with which the outcome (-x) occurs, stands at 0.8 for the default ending -x, which is much higher than for the 'exceptional' ending -y (the former occurs in 9 out of 10 cases and the latter in 1 out of 10 cases, with an equal base frequency of 1000 for each of the 10 cues). Interestingly, in the left panel, the background rate seems to overpower the individual exemplars that appear only weakly predictive. (This is, indeed, an interesting side-effect of cue competition where, as it appears, the context itself becomes the most informative or predictive, relative to the exemplars which are focal cues). The one exception ends up being strongly negatively associated with the default ending -x, at -0.8. The right panel shows that this exception is a strong cue for -y, at 0.9, while the bulk of the exemplars is weakly negatively associated with -y, at -0.1.

[4] Note, however, that the inclusion of a background in large-scale computational training may become problematic because the question of "how constant may a background be" (to paraphrase Tweedie & Baayen, 1998) becomes next to impossible to answer. In language corpora, one may argue for a variety of backgrounds, which all change more or less often.

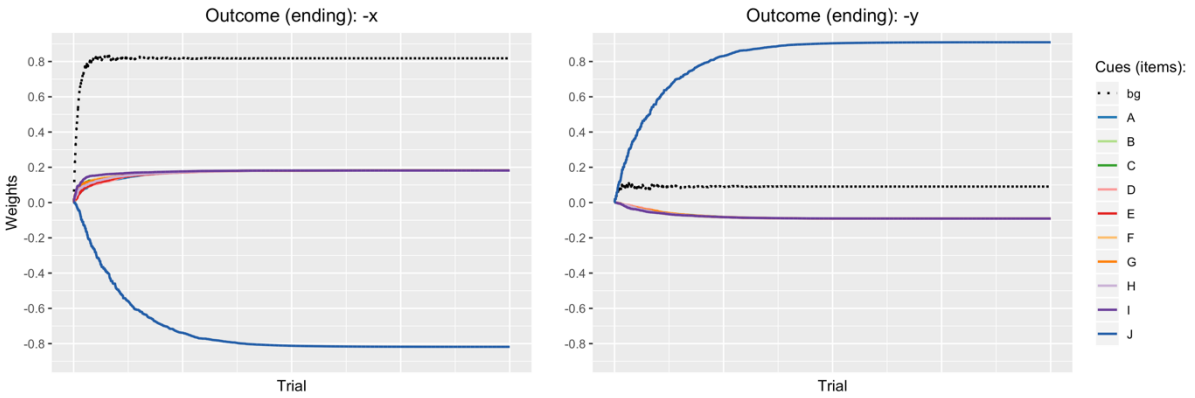

Figure 1. Simulation of learning a rule-based phenomenon with an exception.

The probabilistic training regime contained the very same 10 items and the same, constant background. In this case, usage is not definitively established for any of the 10 items but, instead, each item is attested with both endings, in different proportions (see column 4 in Table A1 in Appendix A). Figure 2 shows that in such cases, overall, the background rate of the more frequently occurring ending -x is higher (0.6) than that of -y (0.4). In all cases, there is an initial period of high uncertainty, but over time a relatively stable pattern of instabilities (fluctuations) emerges. The association strengths between the items and the endings in the left and right panels are mirror images of each other. This constellation seems to be ideal for statistical learning to demonstrate its strengths.

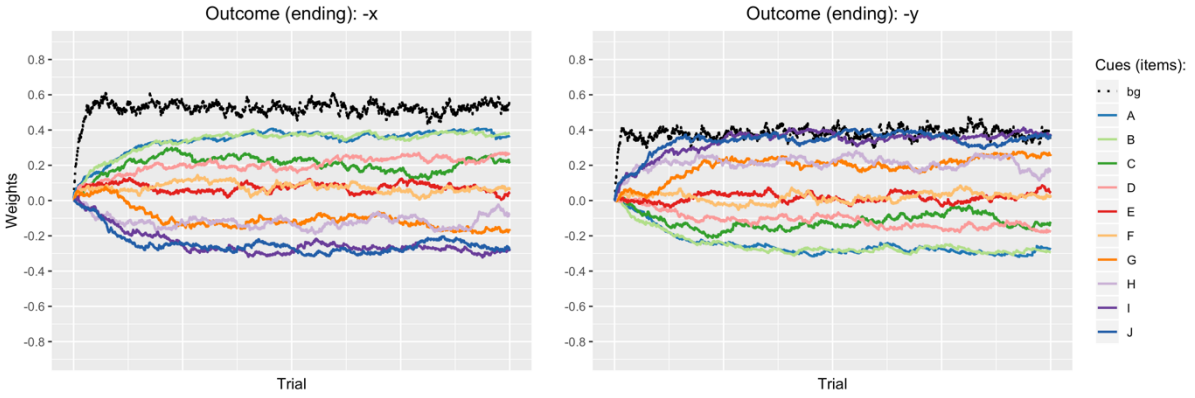

Figure 2. Simulation of learning a probabilistic phenomenon.

The training regime for allomorphy consists of a combination of rules and probabilities: six items always take -x, three items always take -y and one item that takes either ending, 50% of the time. In this particular case, -x represents -u, while -y represents -a, and frequencies for each ending are taken from our corpus data. The left panel in Figure 3 shows that the majority that always takes -x is reasonably well associated with -x (0.4) while the minority that that never takes -x is reasonably well dissociated from -x (-0.6). The one item that takes both endings does not become associated with -x – it hovers around 0 initially, then becomes inhibitory, remaining around -0.1. The right panel shows that against an overall lower background rate, associations are stronger: the minority of 3 items that always takes -y is more strongly associated with -y (0.6) than the x-taking items are associated with -x (0.4). The majority of 6 items that never takes -y is less well dissociated from -y (-0.3) than the y-taking minority is dissociated from -x (-0.6). The one item that takes both endings is associated with -y to some extent: it hovers around 0 initially, then goes up to a 0.2 association. In other words, not only are both endings learned to a relatively stable level (0.4 or 0.6), but the proportion of 6 -x taking and 3 -y taking items appears to create such a learning environment that even an alternation can thrive: a chance level item will evolve as weakly negative for the outcome that is more frequent (given the background rate) and has more competition among supportive cues (6 out of 10), but it will become positive and stabilise as such for the less frequent outcome and when the cohort of cues competing for positive evidence is smaller (3 out of 10).

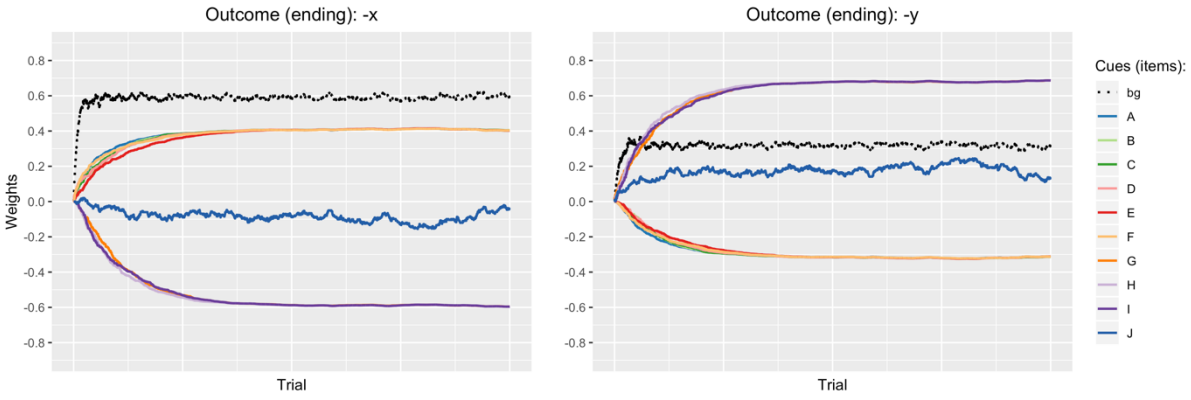

Figure 3. Simulation of learning a phenomenon that combines rules and probabilities.

Across the three probability distributions (as summarized in Table A1 in Appendix A), we observe different learning dynamics. First, the rule-favouring distribution allows for the canonical case to be associated with the most global cue, i.e., the background or the language itself, while associating the exception with the more specific cue. This shows how tuning to a more or less 'precise' cue emerges naturally and favours parsimony. Next, the probabilistic distribution, again, favours learning of statistical regularities. No cue is strongly singled out, but they are all evaluated by their respective probabilistic merits that reflect both local instabilities as well as general trends. Finally, the distribution with one chance-level cue that exists alongside a larger (the default or canonical) and a smaller (the exceptional) cohort of cues, reveals how allomorphy can persist in the long run due to an imbalance in pressures that stems from competing with two cue cohorts of different sizes. Overall, this shows that our implementation of the Widrow-Hoff rule (or of the Rescorla-Wager rule) can perform very rule-like when the co-occurrences of cues and outcomes are quite regular, but it can also detect co-occurrences of cues and outcomes when these are much more diffuse.

In the next sections we will explore two different ways of accounting for the dimensions of experience that govern the -a/-u distribution in Polish. Firstly, we will evaluate the predictive potential of a wide range of linguistic variables identified in the literature on this particular instance of allomorphy on their ability to account for the -a/-u alternation (Section 3). Secondly, we present the results of statistical modelling with learning-based measures, to draw inferences about their importance for predicting the allomorphic alternations in Polish (Section 4). Thirdly, we complete the series of empirical studies with an experiment designed to test whether native speakers are sensitive to the same predictors that fell out from our computational analyses as the most important (Section 5).

3. A corpus-based model of -a/-u variation

In this section we present a corpus-linguistic model that tests the predictive power of the 'usual suspects' that can be found in linguistic studies. After an introduction to the data annotation (Section 3.1), we present the results of our statistical models (Section 3.2) and discuss the results in the light of existing descriptive and experimental findings (Section 3.3).

3.1. Data and annotation

The 250-million-token downloadable version of the morpho-syntactically annotated NKJP corpus (the National Corpus of Polish) was queried (on 8 June 2015) for masculine inanimate nouns in the genitive singular case via the poliqarp search tool (Janus, 2006-2012). For a comparison of frequencies of occurrence across a number of datasets, we refer to SupMat1.

In a bid to evaluate existing accounts of the allomorphy affecting Polish masculine inanimate nouns against actual usage data and to reveal what drives -a/-u allomorphy, we annotated all 5,393 retrieved nouns[5] for those phonological, morphological and semantic properties that have been proposed in those accounts. On the one hand, some properties of the nominative form have been found to determine whether a noun obtains an -a or -u ending in the genitive. These phonologically and morphologically (a)typical endings are listed in Table 1 below; the nouns in our sample were automatically annotated for these properties.

[5] Note that in the analyses that follow we use those 4872 nouns that occur with one of the two endings only; as explained above, different genitive endings may be pressed into use to distinguish homonyms, and it is not possible to control for this in a large corpus.

Table 1. Endings typical for the -a and -u suffixes.

category	**suffix**	**typical endings**
phonetically typical	-a	-k, -g, -h
phonetically typical	-u	-m, -st, -ft, -zg, -szt
morphologically typical	-a	-ak, -er, -ec, -yk, -ik, -acz, -erz, -arz
morphologically typical	-u	-zm, -ot

On the other hand, semantic properties have been identified that would make a noun a candidate for the -a or -u ending. We avoided 'hypernomial' classifications such as the one proposed by Westfal (1956) that are too detailed for any of the categories to generalise beyond a handful of lexical items. Instead we opted for two composite semantic variables with several variable levels, i.e. EntitySize (immovable, manipulable, other) and Type (abstract, collective, count, mass, brand name, place). Small and easily manipulable objects, such as *papieros* 'cigarette', tend to prefer the -a genitive suffix, while large and immovable objects such as *basen* 'pool', on the other hand, tend to prefer the -u suffix; this distinction applies to one third of all nouns in the sample. Mass nouns (which subsume substances) are thought to prefer the -u suffix (e.g. *azot* 'hydrogen'), while count nouns (which subsume objects) prefer the -a suffix (e.g. *kolec* 'spike'). The count/mass distinction applies to half of all nouns in the sample. All nouns in our sample were manually annotated for these properties.[6] Table 2 summarizes the frequency of occurrence of each variable level.

[6] Our sample was manually annotated by the fourth author, a linguist who is also a native speaker of Polish.

Table 2. Frequency of occurrence of each variable level.

		Frequency
Phonology	**Phon.typical**	
	typical-a	1290
	typical-u	487
	atypical	3095
Morphology	**Morph.typical**	
	typical-a	748
	typical-u	361
	atypical	3763
Semantics	**Entity size**	
	immovable	435
	manipulable	777
	other	3660
	Type	
	abstract	2177
	collective	97
	count	1670
	mass	632
	brand name	78
	place	218

Interestingly, three out of the four variables apply to one third or less of all cases. For the variable PhonologicallyTypical, 64% of the data is labelled ‘atypical’; for MorphologicallyTypical, 77% of the data gets the label ‘atypical’; and with EntitySize, 75% of the data is labelled ‘other’. Moreover,

the label EntitySize only applies to the subset of nouns that qualify as 'count' nouns. Finally, even though all levels of the variable Type are directly meaningful, almost half of the sampled nouns (45%) belong to the category of 'abstract' nouns. In other words, many preceding studies have focused on a subset of the problem space only, contrasting e.g., place names with brand names, and count with mass nouns.

3.2. Statistical models

Despite the size of our sample (total of $N = 4872$ noun types), a substantial number of variable combinations remained unattested. Crossing the four target predictors – MorphologicallyTypical, PhonologicallyTypical, EntitySize and Type – revealed 66% zero cells. Many of these zero cells are so-called structural rather than sampling zeros. This means that we do not expect instances of a particular combination to occur. For example, MorphologicallyTypical -a is not expected to co-occur with PhonologicallyTypical -u. Similarly, abstract nouns (Type) like 'absurdity' or 'affection', are not classifiable as levels of EntitySize. The dependent variable shows a similar bias too, with 66% of nouns preferring -u and the remaining 34% preferring -a. For these reasons we will rely on Log-Linear modelling (implemented in the core of the R software environment by R Core Team, 2019), as this particular type of statistical models is not constrained by any specific distributional assumptions. Next, we will focus our analyses on meaningful variable levels only, leaving aside labels such as 'atypical' and 'other'. This affects the sample sizes to different degrees, and we will therefore test each variable separately as predictor of a particular GenitiveEnding (-a versus -u).

3.2.1. Phonological typicality

A subset of $N = 1777$ data points (36.5% of the total sample size) was categorized as PhonologicallyTypical for -a or -u and, thus, qualified for further analysis. Log-Linear Modelling showed

a significant contribution of the targeted predictor (decrease in $\text{Deviance} = 376.35$; $p < 0.0001$; decrease in $\text{AIC} = 374.35$). However, the Residual Deviance of the model remained significant ($\text{Residual Deviance} = 373.6$; $p < 0.0001$), which indicates a considerable lack of fit.

Details of the lack of fit between observed (bars) and predicted (black lines) frequencies are presented in Figure 4. We see a correctly predicted trend for PhonologicallyTypical -u, but the difference in predicted frequencies for -u versus -a is less prominent than the actually observed difference. This same trend of a predicted preference for -u is also given for PhonologicallyTypical -a where the observed difference between -a and -u is, however, negligible. In other words, the observed values for PhonologicallyTypical -a split almost equally into an -a and -u ending for a noun, while the predicted values incorrectly show a difference in favour of the -u ending.

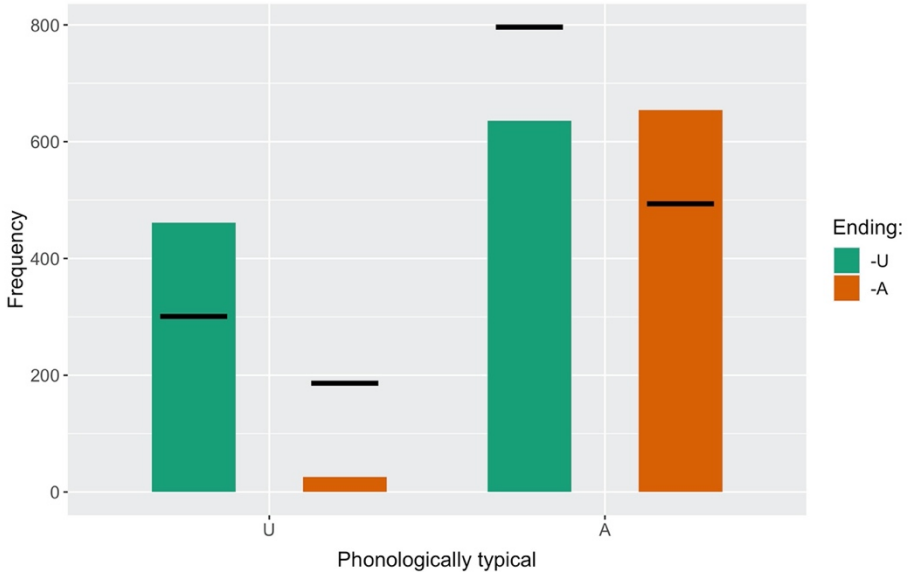

Figure 4. Plot of observed versus predicted frequencies for PhonologicallyTypical.

3.2.2. Morphological typicality

Only a fraction of the data (23% of the sample) was categorised as MorphologicallyTypical -a or -u. The subset ($N = 1109$) showed that the targeted predictor did contribute to a reduction of Deviance

(decrease in Deviance $= 137.93$; $p < 0.0001$; decrease in AIC $= 135.93$), yet as with the analysis of PhonologicallyTypical, the Residual Deviance of the model remained high (Residual Deviance $= 570.1$; $p < 0.0001$), indicating a significant lack of fit.

Figure 5 visualizes the lack of fit between observed and predicted frequencies. Predictions (black lines) show only the preference for the -a ending, but observed frequencies (bars) clearly indicate the interaction. Essentially, MorphologicallyTypical -u is a very good predictor of an actual -u ending, with almost no cases of -a (only 14 out of 1109 cases). MorphologicallyTypical -a, however, is much less indicative of the ending being -a in reality, with many more cases of actual -u (190 out of 1109 cases).

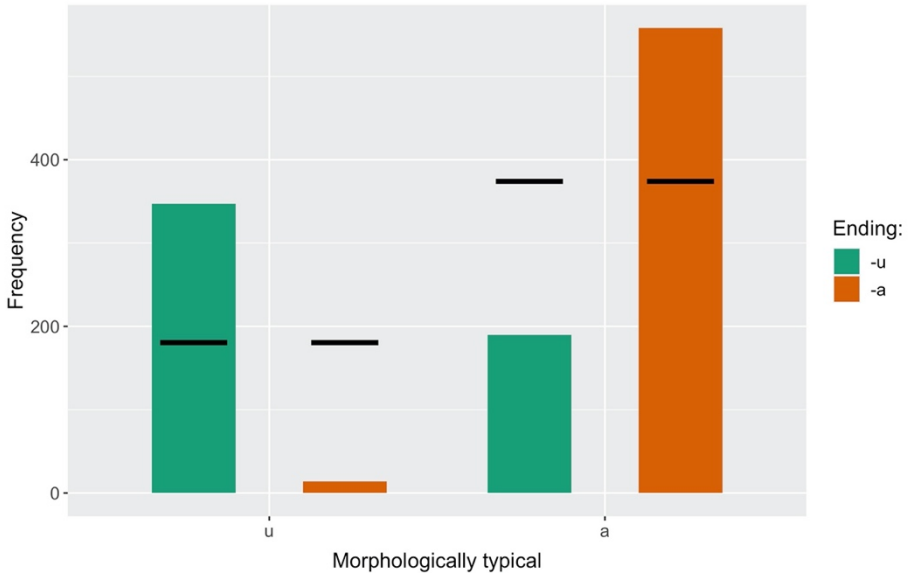

Figure 5. Plot of observed versus predicted frequencies for MorphologicallyTypical.

3.2.3. Type

Log-Linear Modelling showed that Type significantly contributed to explaining the frequencies of -a versus -u (decrease in Deviance $= 5034.5$; $p < 0.0001$; decrease in AIC $= 5024.53$). Overall, however, the model did not perform very well as the unaccounted Deviance remained significant (Residual Deviance $= 923.4$; $p < 0.0001$).

Figure 6 illustrates the specifics of the lack of fit between observed (bars) and predicted (black dashes) frequencies. First, low-frequency categories like 'collective' and 'name' show a good match between observed and predicted frequencies, with a correct preference for the -u ending. Next, some frequent labels, like 'mass' and 'abstract' show a predicted trend that goes in the right direction – i.e., towards -u. However, the differences between the predicted frequencies of the -a and -u endings are less extreme than in the observed frequencies (recall that this was also the case for predictions by PhonologicallyTypical or MorphologicallyTypical; see Figure 4 and 5). Finally, the categories of 'place' and 'count', in particular, reveal a complete mismatch between observed and predicted frequencies, with -a being observed more frequently than -u, but -u being predicted as more frequent.

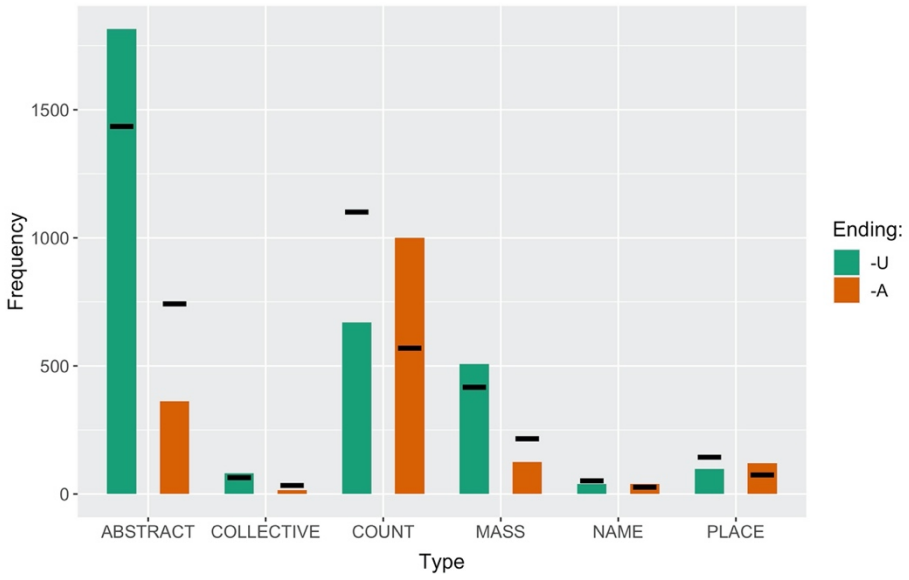

Figure 6. Plot of observed versus predicted frequencies for Type.

3.2.4. Entity size

A sample of $N = 1212$ data points was used (25% of the full sample size), that was categorized as either immovable or manipulable (note that all nouns for which EntitySize plays a role are classified as 'count' nouns under Type). Statistical modelling showed a significant contribution of the predictor

EntitySize (decrease in $\text{Deviance} = 97.828$; $p < 0.0001$; decrease in $\text{AIC} = 95.83$). Again, the unaccounted Deviance remained significant ($\text{Residual Deviance} = 201$; $p < 0.0001$). Figure 7 presents the lack of fit between observed (bars) and predicted (black dashes) frequencies. Similar to other predictors, we see both a match and mismatch in predicted trends where the differences are, again, less pronounced than in the observed frequencies: manipulable objects show a tendency in the right direction, i.e., towards -a, while the predicted frequencies for immovable objects show an incorrect predicted tendency.

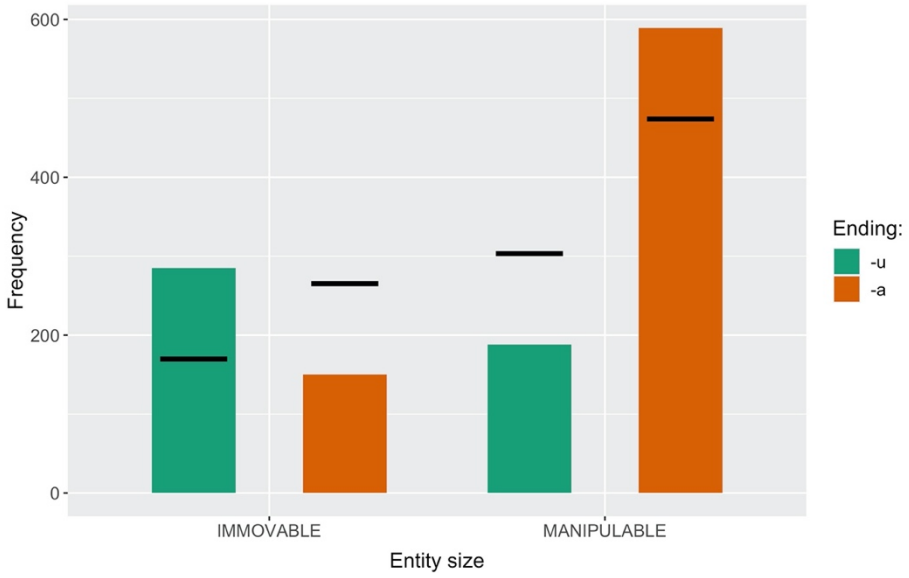

Figure 7. Plot of observed versus predicted frequencies for EntitySize.

3.3. Comparison of predictor strength

In sum, all predictors show a significant contribution to reducing the Deviance but in each particular case the fit remains unsatisfactory. Moreover, for PhonologicallyTypical, MorphologicallyTypical and EntitySize this fit is obtained for a rather small fraction of the data, as in the majority of cases these three criteria default to 'atypical' and 'other'. Overall, given the significant mismatches between observed and predicted endings, to achieve a satisfactory fit, we would need theoretically unattractive

models, without restriction on the respective 2-way tables (i.e., a maximal model, with $\mathrm{df} = 0$ and $\mathrm{Deviance} = 0$).

As none of the corpus-based variables contributed to a sufficient reduction in Deviance, comparing their effectiveness as predictors can only be descriptive and strictly provisory. Since the datasets for each of the four critical predictors differ in size, once restricted to 'meaningful' categories, we made use of the Odds Ratios from the respective 2-way tables as this measure is not sensitive to actual frequencies and, thus, allows for straightforward comparisons (Rudas, 1998).

Since all tables, except for the variable Type, were 2x2 tables, comparisons were straightforward. For the variable Type, we used the harmonic mean (H) of the 2x2 Local Odds Ratios (as explained in Rudas, 1998), in one or the other direction.[7] The Odds Ratios present the four predictors in decreasing order: MorphologicallyTypical ($\mathrm{OddsRatio} = 72.79$), PhonologicallyTypical ($\mathrm{OddsRatio} = 18.23$), Type ($\mathrm{H}_{\mathrm{OddsRatio}>1} = 6.36$; $\mathrm{H}_{\mathrm{OddsRatio}<1} = 0.17$; $1/0.17 = 5.88$), and finally EntitySize ($\mathrm{OddsRatio} = 5.95$).

MorphologicallyTypical and PhonologicallyTypical come out as the most significant variables. Crudely put, the main difference between the phonological and morphological variables in our analysis boils down to the number of letters that is considered, with phonology being limited to any 1 or 2 letters from the end of the word, and morphology stretching to 3 that allegedly form a minimally meaningful unit. Type and EntitySize show a much weaker association with the inflectional ending (-a or -u). As the Log-Linear model with the Type showed (Section 3.2.3, Figure 6), the levels abstract noun, mass noun, collective noun and brand name facilitate reliable predictions, but place name and count noun do not. Finally, EntitySize, which applies to count nouns only and subdivides them further

[7] Note that the Type variable is nominal and any ordering of the local 2x2 tables is strictly arbitrary. We opted for ordering its categories by decreasing (marginal) frequencies, i.e., 'abstract', 'count', 'mass', 'place', 'collective', and 'name', but other orderings would be equally valid. In general, the Odds Ratio of 1 indicates no association, and the farther its value is from 1, the stronger the association becomes, with, for example, 0.5 and 2 being equally distant from 1, as $1/2 = 0.5$. The exact value of the Odds Ratio (e.g., 0.5 or 2) reveals the direction of association (see Rudas, 1998).

into manipulable or moveable and immovable objects, provides reliable guidance for nouns that qualify as manipulable only (Section 3.2.4, Figure 7).

Dąbrowska (2008) presented experimental evidence that some users show sensitivity to lexico-constructional pattern, such as 'count' versus 'mass' or 'object' versus 'substance', that are subsumed under the variable Type. However, sensitivity to the 'count' versus 'mass' distinction is attested only from the age of 10;0 onwards, and 'count' was insignificant in our model. Further to Dąbrowska (2008), sensitivity to the contrast 'object' versus 'substance' appears to be rare in linguistically naïve speakers. The combination of these corpus-based and experimental findings makes it unlikely that the variable Type would be a true reflection of what underlies the usage of these endings.

In other words, the predictor Type appears unconvincing from a user's perspective: the categories that this variable relies on are time-consuming to develop or assume sophisticated linguistic insights, which makes it unlikely that Type-information would guide the development of the relevant morphological knowledge in language users. Those variables that appear to be the most readily 'available' to naïve users, such as phonology, appear to be only weakly related to the ending of an inflected word form. In the next section we will present a computational model that aims to capture how language users arrive at making reliable -a versus -u decisions without assuming a sophisticated knowledge base and instead staying close to how the input sounds and what it means.

4. A discriminative computational model

As shown in Section 3, some parts of the semantic system capture the dimensions that govern allomorphic choices to a significant degree. This fit, however, is far from perfect (or even from good) and, more importantly, concerns have been raised about the availability of these concepts during the acquisition of the system. To model how accessible raw phonological, morphological and semantic information is to a naïve speaker and indeed how utilizable this information might be in day-to-day

language use, we rely on the principle of error-driven discrimination learning postulated by Rescorla and Wagner (1972). Rescorla and Wagner proposed a very simple learning mechanism that can detect and learn patterns present in experience to better navigate the environment (cf., Rescorla, 1988). We rely on the NDL implementation of the Rescorla-Wagner learning rule (for examples of use see Baayen et al., 2011; Milin, Feldman, et al., 2017; while for implementations in **R** and **Python** see Arppe et al., 2018; Sering, Weitz, Künstle, & Schneider, 2017)[8]. After an introduction to the data and training regime for our computational simulations (Section 4.1) we present and discuss our findings in the light of existing literature (Section 4.2).

4.1. Data and training

We trained two Naive Discriminative Learning models, an orthographic model and a lexical model. For both models, the outcomes are fully inflected word forms. For the orthographic model (G2F; grapheme-to-form), each sentence corresponds to one learning event. The cues are all overlapping trigraphs of that sentence and the outcomes are all the word forms that the sentence contains. For example, the event corresponding to the sentence "Peter loves Mary" would consist of three outcomes (peter, loves, mary) and 16 trigraph cues (#pe, pet, ete, ter, er#, r#l, #lo, lov, ove, ves, es#, s#m, #ma, mar, ary, ry# - the hashtag indicates word boundaries). Like for the orthographic G2F model, for the lexical model (C2F; context-to-form), each sentence corresponded to one learning event, with the cues and outcomes being all the words in that given sentence (excluding self-cueing). Importantly, the C2F network bears strong resemblance to so-called Distributed Semantic Models (DSMs: e.g., Griffiths, Steyvers, & Tenenbaum, 2007; Landauer & Dumais, 1997; Lund & Burgess, 1996; Marelli & Baroni, 2015; Mikolov, Sutskever, Chen, Corrado, & Dean, 2013; Shaoul & Westbury, 2010)

[8] NDL has given rise to LDL, a computational model for linear discriminative learning that implements Word and Paradigm Morphology (Baayen, Chuang, Shafaei-Bajestan, & Blevins, 2019). Different from NDL, LDL takes a semantic representation as input, and maps this onto a phone sequence as output (production) but can also take a phone sequence as input and map that onto a semantic representation (comprehension). LDL has been used to model inflectional allomorphy and case inflection (Baayen, Chuang, & Blevins, 2018).

The learning events were generated from the Araneum Polonicum web corpus (Benko, 2014). All sentences were extracted and filtered to retain those that only contain Polish characters. Punctuation and numeric characters were removed, and letters were lowercased. This resulted in a total of 67,876,701 G2F and 67,155,758 C2F learning events (i.e., sentences; the slight difference is due to a number of single-word sentences). Duplicate cues and outcomes were removed from all events. The number of outcomes for both training sessions amounted to 13,929 word forms; there were 35,629 unique cues (attested Polish trigraphs) for G2F, and 46,236 word forms as cues for C2F, and they were also included as outcomes. The set of outcomes was supplemented with an additional 32,307 randomly sampled words (using stratified sampling over frequency-bands with a cut-off of ≥ 5, following the idea of Ellis et al., 2004; see also Ellis, Simpson-Vlach, & Maynard, 2008 for a similar approach).[9]

All pre-processing was done using the **Python** programming language, and the models were then trained using the NDL implementation from the **pyndl** package (Sering et al., 2017). All learning parameters were set to their default values (the maximum learnability or associability of outcomes, λ, was set to 1.0 for all outcomes, which normalises them and makes them comparable; the learning rate – $\alpha \times \beta$ equalled 0.0001, a conveniently small value to allow learning to be gradual; see Milin, Divjak, et al., 2017; Milin, Feldman, et al., 2017 and references therein for details).

4.2. Computational simulations with discriminative measures

In this section, we present the results of the statistical analyses that use NDL-generated measures as derived from the two learning models introduced above. The aim of these analyses is to understand how likely the genitive form in -a or -u is for a given lexical item (for example, the genitive *antraktu* from the nominative *antrakt*). More specifically, this analysis aims to shed light on the knowledge a

[9] Note that by adding more outcomes we indirectly allowed cues to compete more as it increases the cues' chances of being individually or jointly present or absent for many outcomes.

language user might rely on when selecting one of two genitive endings. It also parallels the corpus-linguistic analysis, and thus offers a comparison of the linguistic and discriminative approaches. To make this comparison possible we intersected the data used for the corpus analysis and NDL training that for technical reasons had to be carried out using different Polish language corpora, i.e. the NKJP and Araneum Polonicum, respectively.[10] There were $\mathrm{N} = 3570$ word forms attested in both corpora which comprised the final dataset used in this part of the analyses.

With the genitive ending as the response variable, we considered a range of discriminative predictors, derived from the G2F and the C2F networks (details of these measures are described in Milin, Feldman, et al., 2017). These derived measures are important for understanding what exactly is learned. In authentic language data many cues are present simultaneously, which may obscure why the weights between cues and outcomes develop in a certain way. Derived measures such as priors and activation diversity measures are crucial for pinpointing what the network has actually learned.

The predictors from the G2F network used in the present study include G2F Activation (how strongly our word is supported by the trigraphs present in visual input), G2F Prior activation (how well could the orthographic system know our word; this is similar to frequency, whereby more frequently encountered words are likely to be known better and more widely) and G2F a-Diversity (as a measure of competition that shows the diversity of forms that are activated by the same set of cues in the visual input). From the latter measure we derived two more specific ones: G2F Word End Diversity (the diversity of the last two letters of the word and the hashtag that signals the word end, henceforth G2F Word End Diversity) and G2F Final Trigram Diversity (the diversity of the last three letters of the word, henceforth G2F Tri-Diversity). From the C2F network we derived C2F Prior (a measure of how well the semantic system could know the word; cf. frequency as discussed above), and C2F Typicality (how typical is our word, i.e. is our word unusual with respect to the contexts it appears in).

[10] The NKJP does not allow access to the raw data, which makes computational modelling impossible. The Araneum Polonicum, on the other hand, is uncleaned and untagged, which makes automatic morphological annotation (which is required to retrieve genitive singular forms of nouns) unreliable.

Prior to statistical modelling we checked our set of predictors for multicollinearity using the **usdm** package (Naimi, Hamm, Groen, Skidmore, & Toxopeus, 2014) in **R**. In the first step, we removed G2F Activation and G2F Prior because they were exceptionally highly correlated among themselves ($\mathrm{r} = 0.89$) and with several other NDL variables (e.g., with C2F Prior, the two variables' correlations were, respectively, $\mathrm{r} = 0.86$ and $\mathrm{r} = 0.99$). Subsequent variance inflation analysis, utilising the **usdm** package, indicated that no further variable removal was required, and only one correlation, between C2F Prior and C2F Typicality, stood out as higher ($\mathrm{r} = 0.73$). All remaining NDL measures (both G2F and C2F) were rank-normalized because of the spiky density distribution of their NDL learning weights (cf., Milin, Feldman, et al., 2017). For rank-normalization we implemented the procedure suggested by Karssen, van Duijn, and Aulchenko (2016).

Next, we compared the performance of the learning-based measures (NDL) using the **gbm** package (Greenwell, Boehmke, Cunningham, & GBM Developers, 2019) in the **R** software environment (R Core Team, 2019). We tested the importance of all NDL variables in predicting the noun ending, -a versus -u, utilising the Bernoulli distribution and applying 5-fold cross-validation. G2F Word End Diversity and G2F Final Trigram Diversity came out as the strongest and second strongest predictors, respectively. Jointly, these two variables accounted for approximately 83% of relative informativity of the full set of NDL predictors (individually: 46.5% and 36.1%).

With these two predictors we constructed a Generalised Additive Model (GAM) to predict the variation in the choice of ending using the **mgcv** and **itsadug** packages in **R** (Wood, 2006; 2011; van Rij, Wieling, Baayen, & van Rijn, 2016). We considered only one of these two G2F diversity measures at a time, since they are both informative about overlapping final trigraphs. The difference between the two diversity measures lies in the number of characters considered, and in the position these characters can occupy in other words: while G2F Word End Diversity focuses on the last two characters of a word, G2F Tri- Diversity takes the last three into account. Because G2F Tri-Diversity does not include the hashtag that marks the end of the word, this very character combination could in principle occur in any position in words other than the ones considered here. We found a model including G2F

Word End Diversity to be the best. This is consistent with our GBM-based analyses, where the same predictor also came out as the most informative learning-based predictor.

To facilitate model fitting we applied conservative trimming of the somewhat heavy and discontinuous tails of G2F Word End Diversity, which reduced the dataset by 83 data points; thus, the final dataset had $\mathrm{N} = 3487$ data points. The final model is summarised in Table 3. The model includes a smooth for G2F Word End Diversity and a tensor product of C2F Prior and C2F Typicality.

Table 3. Outputs of the generalised additive model of the association between the genitive ending and discrimination-based predictors using only nouns allowing a single ending: AIC = 3430.976; ML = -4897.5; R-sq (adj) = 0.146.

A. Parametric coefficients	**Estimate**	**Std. Error**	**t-value**	**p-value**
Intercept	1.384	0.0534	25.72	< 0.0001
B. smooth terms	**edf**	**Rel.df**	**F-value**	**p-value**
s(G2F Word End Diversity)	7.959	8.674	291.5	< 0.0001
te(C2F Prior, C2F Typicality)	3.002	3.003	100.6	< 0.0001

Figure 8 below depicts both model terms. In the left-hand panel, the Y-axis is mapped onto a 0 to 1 interval, where 1 assumes a tendency towards -u, and 0 towards -a. When G2F Word End Diversity is below average (i.e., 0.0, as the variable was rank-transformed and scaled), we are looking at trigraphs that activate comparatively fewer words, and this appears to coincide with a slight preference for the ending -a. Conversely, when G2F Word End Diversity is above 0.5, we are looking at trigraphs that activate comparatively more words, and the -u ending appears to thrive. Simply put, the sparser the neighbourhood, the greater the uncertainty, and the more likely the speaker is to choose -a.

The right panel shows the interaction between C2F Prior and C2F Typicality (recall that the C2F Prior captures how well the semantic system could know the word, while C2F Typicality indicates

whether our word is unusual with respect to the contexts it appears in); low values (blue) signal -a, whereas high values (ochre) signal -u. The isolines seem to indicate a rather regular quadratic inter-relationship – i.e., in inverse U-shaped trend across the main diagonal (upper left to lower right) and a regular U-shaped trend on the minor diagonal (lower left to upper right). In other words, there is preference for the -u ending when both C2F Prior and C2F Typicality are high and there is a somewhat less pronounced preference for the same ending when both predictors are at their lowest values. By extension, the -a ending appears to be (slightly) preferred when C2F Prior is high but C2F Typicality is low, and vice versa, i.e., when C2F Typicality is high while C2F Prior is low. Another way to say this is that -u words seem to require a good deal of experience (to be well entrenched), while -a words combine less experience (they are poorly entrenched) with contextual peculiarity.

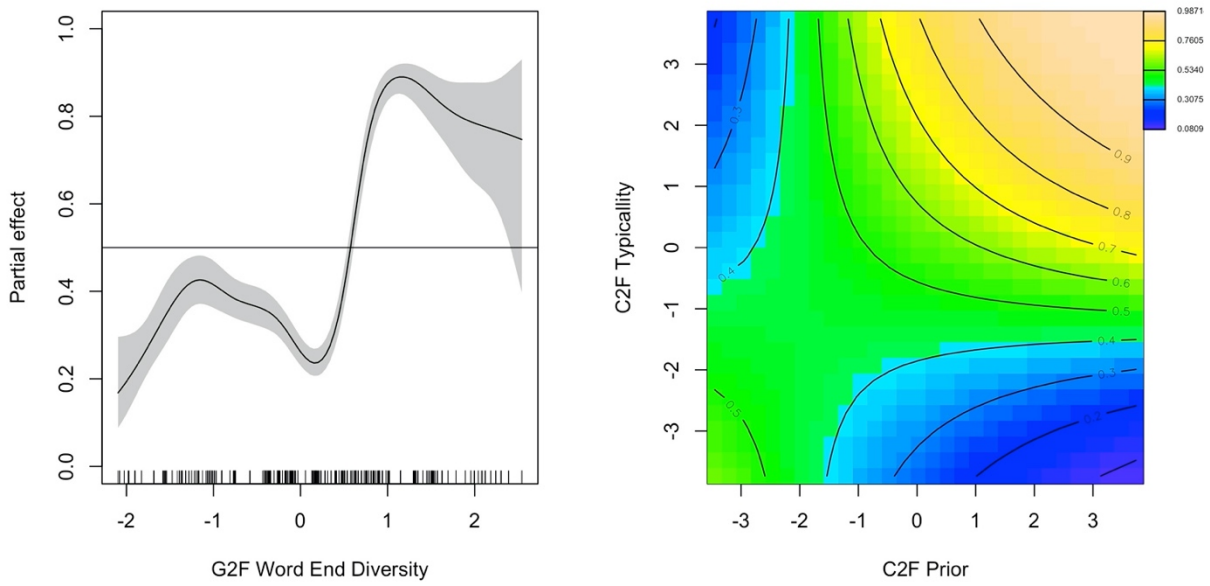

Figure 8. Model terms of the generalised additive model of the association between the genitive ending and discrimination-based predictors using only nouns allowing a single ending. In the left-hand panel, the Y-axis is mapped onto a 0 to 1 interval, where 1 assumes a tendency towards -u, and 0 towards -a. In the right-hand panel, low values (in blue) signal a tendency towards -a, whereas high values (in ochre) signal a tendency towards -u.

The results of both the G2F and C2F computational networks together teach us the following: words that end in -a tend to contain trigraphs that activate comparatively fewer words in the language. Semantically speaking (recall that the C2F network is similar to distributional semantic models), these words tend to be poorly entrenched or rather atypical. Words that end in -u, on the other hand, contain trigraphs that activate many other words in Polish. Semantically speaking, these words are well entrenched and contextually typical. Intriguingly, this finding provides a possible explanation for Westfal's (1956) conclusion that the -u ending would be the elegant one while the -a ending would come with a tinge of vulgarity (or roughness). Words that take -u are made up of trigraphs that are distributed over many other words and are hence typical for the Polish language; words that take -a, on the other hand, contain trigraphs that are distributed over fewer words, i.e. trigraphs that would be less typical for Polish and, hence, such words appear less desirable.

Words that are typical for Polish are naturally better known than words that are atypical. This finding may explain why the minority ending -a is the one that attracts proportionally more foreign words. Westfal (1956) found that -u is the most frequently used genitive ending, except for borrowed words, which are more likely to take -a. That is, in his sample, 30% of -a words are foreign (259/880), versus 18% of -u words (602/3322). This creates a situation in which overall, the raw number of -u words that are foreign is higher, although the proportion of -a words that are foreign is larger (Westfal, 1956, pp. 363-364).

5. Experimental verification of learning and learnability of NDL measures at the individual and group levels and comparison with the corpus-based measures

The error-driven learning model presented in Section 4 suggested that language users may be sensitive to two dimensions of their experience with language: are the last two letters of the word distributed over many other words, and is the word well known and does it behave like many other words. In this section, we report on the results of a forced-choice task, designed to test whether the

top discriminative measure, G2F Word End Diversity or the diversity of the last trigraph only, which includes the last two letters of the word and the hashtag that signals the word end, can explain the preferences of native speakers for one of the two genitive forms (-a vs. -u). In keeping with previous work on morphological productivity, we make use of genitive forms derived from Polish pseudo-words as pseudo-words do not link an extra-linguistic experience to linguistic (semantic) space and have an occurrence frequency of zero, which simplifies statistical analysis and model interpretation. This allows us to isolate the connection between letter clusters and genitive forms ending in -a or -u.

5.1. Experimental validation of the G2F Word End Diversity effect

Using **Wuggy** for macOS (Keuleers & Brysbaert, 2010), 5000 potential nonce words were created, based on our list of 5,500 words attested in the Araneum Polonicum corpus (Benko, 2014). NDL weights for the trigraphs in the nonce word list were extracted from the matrix of NDL weights and a sample of 750 forms with the same mean and standard deviation for G2F Word End Diversity that described the initial list of 5000 nonce words was selected. After removing illegal nonce words (i.e., those with illegal syllable combinations), 563 Polish pseudo-words remained; for these, genitive forms in both -u and -a were created.

The experiment was administered online using **Qualtrics** software (Qualtrics, Provo, UT). We recruited participants via emails and social media (Twitter and Facebook). Our sample of volunteers consisted of 223 native Polish speakers (164 females, 66 males; mean age 31.1 years; age range 18–65 years).[11] We asked participants to complete a demographic and a reading habits questionnaire as well as the main forced-choice task. The demographic questionnaire prompted participants to provide

[11] To encourage participation, we introduced a prize draw after we had collected data from around 70 participants. The effect of the introduction of the prize draw was tested and controlled for in the statistical analyses presented below by running the analyses with and without a variable encoding whether the prize was used in the model. Since the inclusion of this variable did not change the direction of the effects (the variable was not significant and did not significantly interact with the main predictor in the main model), we only present the results of the statistical analyses on all data, but without the prize draw variable.

information about their gender, age, education and the foreign languages they speak. The reading habits questionnaire included questions such as how often the respondents read for enjoyment and how many hours they spend each week in reading different types of materials (e.g., emails, newspapers and fiction books). In the main task, participants were instructed to choose, on each trial, the appropriate -a or -u form for the pseudo-words. The two forms were presented below the clause "Nie ma …" (There is no …), which triggers the genitive on nouns, and participants had to select one of the two forms. For each participant, we generated 50 trials based on the original list of 563 pseudo-words.

Prior to running statistical analyses, we discarded data from participants who produced fewer than five responses in the main task (0.2%). As a result, we excluded 7 participants, which left us with a set of 10,287 data points from 216 participants. The available data were then entered into a Generalised Additive Mixed Modelling (GAMM), with random intercepts for both participants and items (nonce words). As fixed effects, we considered G2F a-Diversity, G2F Word End Diversity and G2F Trigram Diversity, but only retained one of them in the final model as we did in the NDL-based analysis (see Section 3), using an AIC-based selection criterion. As our experimental items were nonce words, they do not actually occur in the Polish language and, hence, cannot be used as the outcomes for an NDL learning network; for this reason, C2F predictors were not considered in this part of the analyses.

To select the diversity predictor to use, we compared all GAMMs that contained one diversity predictor using the likelihood ratio test. The best model was based on the G2F Word End Diversity measure, which reached an overall training accuracy rate of 73.2%. The model outputs are summarised in Table 4.

Table 4. Outputs of the generalised additive model of form choices in the behavioural experiment. AIC = 12284; fREML = 14616; R-sq (adj) = 0.232; n = 10,287.

A. Parametric coefficients	**Estimate**	**Std. Error**	**t-value**	**p-value**

Intercept	.10	.06	1.56	.119
B. smooth terms	**edf**	**Rel.df**	**F-value**	**p-value**
s(G2F Word End Diversity)	1.00	4.84	12.34	<.001
by-Participant random intercept	170.17	215.00	925.43	<.001
by-Word random intercept	426.92	561.00	1809.49	<.001

The smooth term for the G2F Word End Diversity predictor was highly significant ($p < .001$), revealing a negative relationship between the orthographic diversity of the final trigraph, attested in many Polish words, and the preference for the -a ending: the more diverse the trigraph, the less likely it is that the -a will be chosen as ending (Figure 9). The other G2F predictors – G2F a-Diversity and G2F Trigram Diversity – do not only fit the experimental data less well, as is shown by the likelihood ratio tests, but they were also non-significant in their respective models ($p = .863$ for G2F a-Diversity and $p = .166$ for G2F Trigram Diversity).

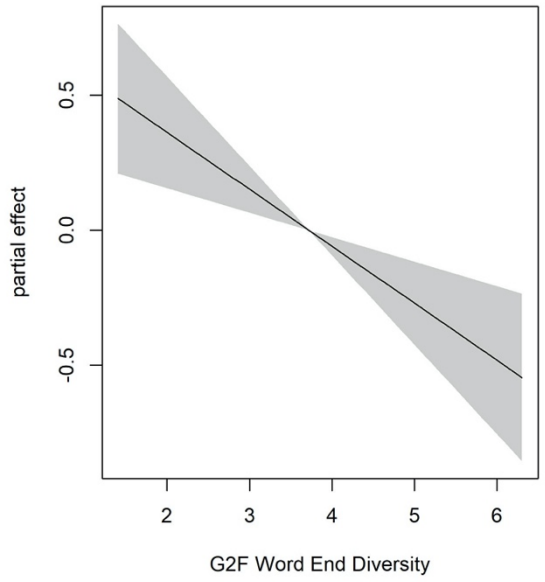

Figure 9. Partial effect of G2F Word End Diversity on the probability of choosing the -a genitive form.

The findings from a forced choice task straddle the boundaries between production and comprehension. This conflation appears to be the case more generally in language processing, not only in production (see Hickok, 2014 for an overview on the need for feedback control), but also in comprehension: even "strict" comprehension tasks are known to involve the other component (e.g., for the effect of phonology on (silent) reading see Perrone-Bertolotti et al., 2012; Newman, Jared, & Haigh, 2012). NDL and its sibling LDL have been shown to be able to accommodate this tension gracefully, both in practice and in theory (Hendrix, Bolger, & Baayen, 2016; Baayen et al., 2018; Chuang et al., 2020), and the results reported here are no exception.

5.2. Investigating individual differences

Do the patterns we observe reside at the level of the language as a whole, or at the level of the individual speaker? In other words, is any systematicity we have registered here an epiphenomenon of between-subject variability, which is covered up by using aggregated data such as language corpora?

To check whether participants, individually, are biased towards one of the genitive endings we plotted the histogram of the proportions of -u choices (see Figure 10). The obtained distribution of the -u choices made by participants looks close to a normal distribution that is symmetrical about the chance level (Shapiro-Wilk test: $p = .221$). This suggests that participants were mostly unbiased, since otherwise the distribution would have been skewed.

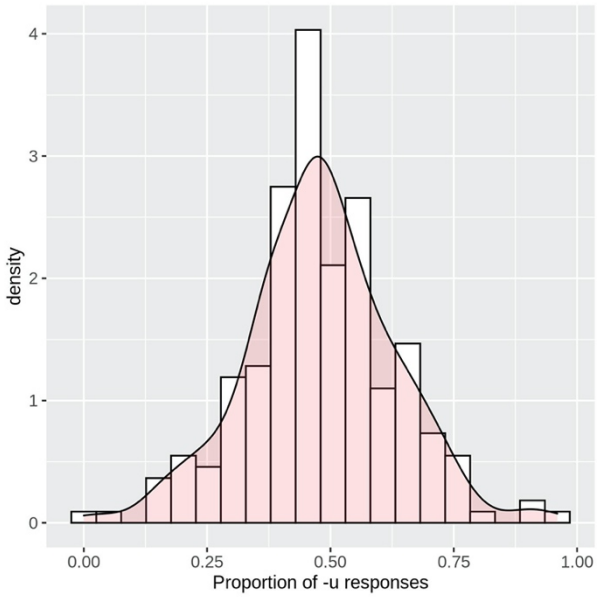

Figure 10. Histogram of the distribution of the -u response proportions from participants and its probability density function estimate.

To make further assertions about the (un)systematicity of participants' choices and, at the same time, to understand better the relationship between participant-relevant behaviour with the characteristics of language-relevant items (i.e., pseudo-words), we grouped participants into those that are biased towards one of the genitive endings and those that are not and classified the nonce words into those that have a preference towards one genitive ending and those that do not. This allowed for a simple test: if a pseudo-word's morphological characteristics are the main driver of our participants' behaviour, then participants' choices should vary depending on the category a (nonce) word belongs to.

Because each participant and each nonce word contributed or attracted a different number of responses (the number of responses belonging to a particular participant ranged from 5 to 50 while it ranged between 14 and 21 for nonce words), we constructed a 95% confidence interval for the proportion of -u choices (by assuming that the choice of ending is completely random). If the observed proportion of -u choices for a particular participant or a pseudo-word falls outside the confidence interval, then this provides direct evidence that the participant or pseudo-word is biased towards one

of the two endings; more precisely, biased towards -u if the observed proportion is above the upper bound of the confidence interval or biased towards -a if the observed proportion is below the lower bound. For example, assume that a certain participant responded to all the 50 questions she was presented with. Her 95% confidence interval is then [0.36, 0.64], and thus she would be considered biased towards the -a ending if her proportion of -u choices is below the lower bound 0.36; conversely, she would be considered '-u-biased individual' if her choices of -u endings would be above 0.64.

Finding the confidence interval for each participant and each nonce word allowed us to group participants and nonce words into three clusters: those biased towards -u, those biased towards -a and those with no evidence of bias. These three clusters are depicted in Figure 11. The left panel provides evidence for heterogeneous behaviour among participants, while the right panel shows heterogeneity of (lexical) properties of nonce words.

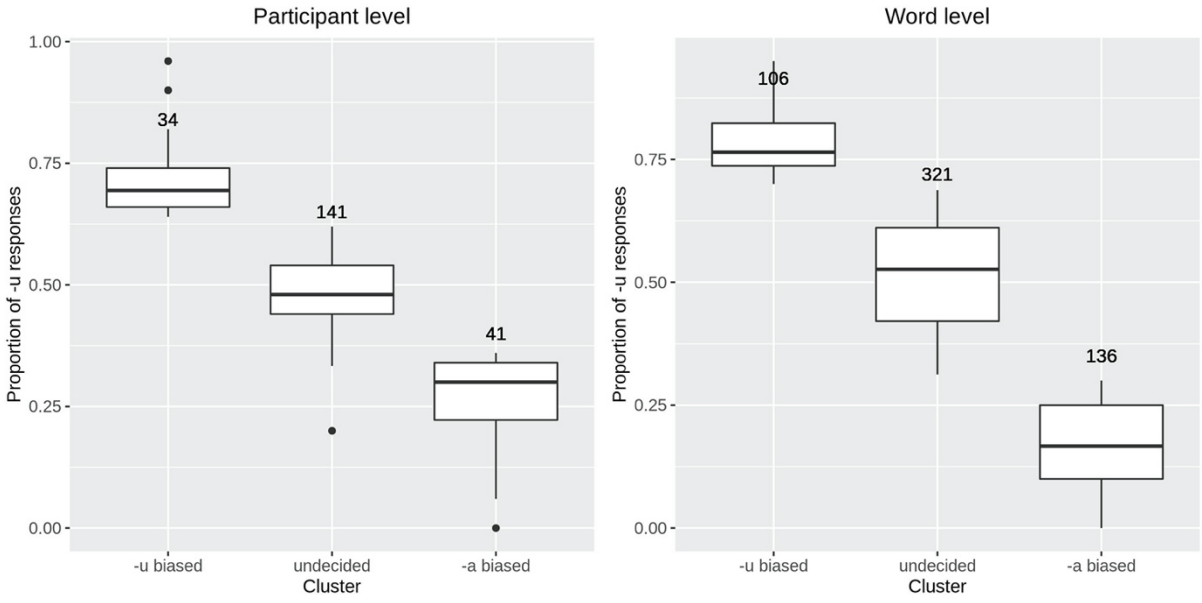

Figure 11. Distribution of the proportion of -u choices in each of the clusters at the participant level (left) and word level (right). The number above each box indicates the number of members in the cluster.

To compare our findings to the literature, we extracted the three most highly associated final bigrams with the clusters containing words that are -u biased, -a biased or non-biased. For that we

first estimated and then ranked the proportions of co-occurrences of all bigrams with each of the three clusters relative to the other clusters. To improve the robustness and generalisability of our results, we only considered bigrams that appeared at least five times in the data set. This was to prevent endings that appear rarely (say once) but co-occur often with one of the clusters from being ranked highly for that cluster. Interestingly, as Table 5 shows, only two endings out of six matched those that are considered as typical phonological or morphological endings in the literature. These are -st- for the suffix -u and -rz- for the suffix -a (compare here the endings listed in Table 1). We refer to Table B1 in Appendix B for a complete list of final bigrams that trigger -u or -a.

Table 5. The three most predictive last bigrams for each word cluster in descending order.

-u biased cluster	**-a biased cluster**	**uncategorised cluster**
st	eń	at
ar	**rz**	nd
et	ań	rf

Figure 12 depicts the proportions of -u responses by each cluster of participants and for each cluster of words. The figure shows that the three groups of participants reacted differently to different clusters of words, with the proportion of -u choices well spread between the two extreme ends, 0 and 1, on the Y-axis. More specifically, for all participant clusters, the proportion of -u choices decreases as one moves from the cluster that is biased towards -u to the cluster that is biased towards -a. The proportion of -u choices is lowest in the cluster combining -a biased words and -a biased respondents, and highest in the cluster combining -u biased words and -u biased respondents. This indicates that even biased participants did make use of the characteristics of the nonce words that they were facing when choosing between the two genitive endings. This conclusion is also supported by the results of the GAMM analyses above.

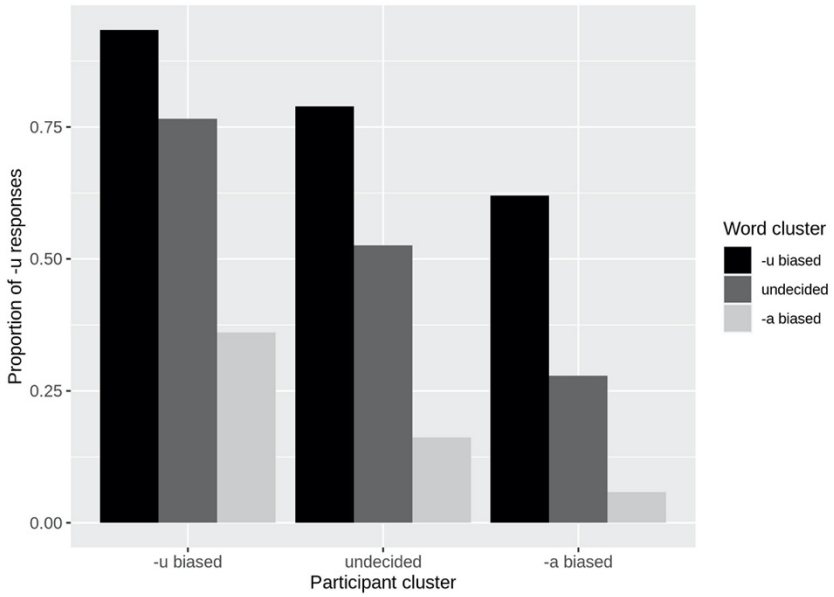

Figure 12. Interaction between the clusters of participants and the clusters of words in term of proportion of -u choices.

5.3. Comparison of the NDL-based measures with the corpus-based measures

To test the performance of the learning-based measures (NDL) against the corpus-based measures we ran a Gradient Boosting Machine (GBM) analysis.[12] We tested the importance of all variables in predicting the noun ending, -a versus -u, utilising the Bernoulli distribution and applying 5-fold cross-validation.

In order to have the largest number of available data points available, we first compared the variable importance of the corpus-based variable Type with the five retained learning-variables (G2F a-Diversity, G2F Word End Diversity, G2F Final Trigram Diversity, C2F Prior, and C2F Typicality). This reduced the full sample of $\mathrm{N} = 4872$ to 3570 data points as some nouns were not attested in both corpora (NKJP and Araneum Polonicum). The GBM analysis showed that the variables Type and G2F Word End Diversity were the most important ones (on a relative, percent scale): 31.6% and 31.1%.

[12] We note, however, that mixing categorical and numerical variables is, in principle, not recommended as variable importance procedures tend to favour categorical variables (cf., Strobl, Boulesteix, Zeileis, & Hothorn, 2007).

They are followed by G2F Final Trigram Diversity at 23.9%. The remaining three NDL variables appeared to be much less important.

In a second step, we added another corpus-based variable, PhonologicallyTypical, which our Log-Linear Models had identified as the second most predictive variable. Unfortunately, retaining the meaningful categories of PhonologicallyTypical -a or -u reduced the number of data points substantially from $\mathrm{N} = 4872$ to $\mathrm{N} = 822$. The results show that the importance of Type increased to 33.0% (from 31.6%) while the influence of both G2F Word End Diversity dropped (from 31.1% to 22.7%), as did the effect of G2F Final Trigram Diversity (from 23.9% to 16.3%). PhonologicallyTypicality came out as fourth variable by relative importance (13.1%). The importance of the remaining three NDL variables remained unaffected yet low.

In other words, our naïve variable G2F Word End Diversity that relies on final letter combinations only, not only gets strong behavioural support but is, furthermore, virtually indistinguishable in its performance from the semantically sophisticated variable, Type. Moreover, G2F Word End Diversity has the added advantage of being available for any word, right from the start of language learning, while Type relies on conceptual categories that develop gradually during childhood and adolescence.

To examine whether what experienced users observe as a feature can be linked to a more basic principle, we ran a Gradient Boosting Machines model to try and predict Type from the C2F predictors that we used in the discriminative approach. The predictors were C2F Prior and C2F Typicality. The model did not perform well, mainly predicting ‘abstract’, the most frequent category (‘abstract’ was predicted 95.6% of the times when ‘abstract’ was the true category and 87.1% of the times when it was not). The accuracy rate was about 52.6%, which is slightly better than always predicting the predominant category (with an accuracy rate of 49.7%). This leads to the conclusion that the concepts that make up Type are not directly derived from lexical co-occurrence information.

6. General discussion

In this study we approached productivity from a novel perspective that focuses on the learnability of the relation between cues and outcomes, which depends on its systematicity. Our case study targets allomorphy, a phenomenon that exhibits (in the majority of cases) systematicity at the lexical level yet creates the impression of low systematicity at the category level. Moreover, the variation in form is not reflected in concurrent variation in meaning, which is a common occurrence in morphology (Ackerman & Malouf, 2015, p. 1) but challenges frameworks with a strong requirement for form-function mappings. We asked how the human capacity for learning deals with such seemingly unmotivated variation in order to draw inferences about the types of structures that should be targeted on a cognitively realistic account of morphological representation. We showed that, for morphology, there is systematicity at a level not typically under the purview of linguistic approaches. Importantly, the tendencies we discovered reside at the level of the individual speaker and are not simply a by-product of data aggregation.

6.1. Modelling productivity: what are the relevant dimensions of experience with language?

To arrive at answers, we pitted a 'traditional' linguistic account of allomorphic variation against an account based on insights from error-driven learning. As representatives of the traditional linguistic account, we selected a manageable combination of phonological, morphological and semantic properties that have been proposed in the literature dealing with this phenomenon. While the linguistic account relies on sophisticated knowledge, our computational learning algorithm runs on approximations of what the word sounds like and how it positions itself in semantic space, as measured while occurring and co-occurring in a sample of attested language. Such knowledge simply falls out of exposure to and use of language. We will discuss both findings in turn.

We found that the semantic strand of the linguistic account, represented by the variables Type and EntitySize, performs rather poorly. EntitySize lacks coverage while Type is unlikely to be available before the allomorphy is acquired. Dąbrowska (2008) established experimentally that speakers begin to show sensitivity to the lexico-constructional patterns that make up the variable Type, such as 'count' versus 'mass', from the age of 10;0 only and that only a very small minority (10% in her sample) of adult language users shows sensitivity to the referential properties of the noun such as 'object' versus 'substance'. The variable Type is hardly what drives learning the -a versus -u distinction; instead, it appears to be a generalization that emerges gradually, as speakers build up knowledge of the -a/-u distinction. As it is not directly related to lexical co-occurrence information, it is quite possibly a post-hoc rationalization that is pressed into service to make sense of the variation.

Morphological Typicality and Phonological Typicality are somewhat predictive for the -a versus -u alternation. Interestingly, the difference between the phonological and morphological variables in accounts of Polish allomorphy boils down to the number of letters that is considered, with phonology being limited to 1 or 2 letters from the end of the word, and morphology stretching to 3 that should form a minimally meaningful unit. Our computational algorithm was straightforwardly trained on two- and three-letter combinations. Due to the relatively shallow orthography of Polish, our grapheme-based approach approximates reasonably accurately 'what the word sounds like' to native learners of Polish. This was complemented by a context-based network that captures how words are used in (sentential) context and represents a distribution-based approach to meaning. The pattern that emerges here is one that signals an alignment of orthographic and distributional semantic information. Words ending in -a are characterized by a sound and semantic profile that are rather atypical; words that end in -u tend to be well entrenched and display a typical sound and semantic profile. In other words, words favouring a genitive ending in -a sound different from other words and are used in a semantically distinctive way, while words favouring a genitive in -u sound like many other words and are used in contexts that fit many other words. This finding is in line with the impression Westfal (1956) had when he said that -a is the rough and vulgar counterpart to elegant -u: the latter

occurs in many words, whereas the former is more limited. In fact, it is specialized for borrowings. On Westfal's count (Westfal, 1956, p. 363-364), the proportion of borrowings is larger among words that take -a than among words that take -u: although -u is, overall, the more common ending (68%) only 18% of -u words is foreign, versus 30% of -a words. If native speakers, like our algorithm, pick up on this correlation between the -a ending and the foreign origin of a word, the preference of foreign words taking -a would perpetuate itself by attracting ever more foreign words into that 'vulgar' -a category. This is quite likely as sensitivity to phonotactic regularities has been shown to influence word learning, with more common phonotactic patterns being learned more rapidly than rare sound sequences (Storkel, 2001; Storkel, 2003). Furthermore, there is evidence that, while part of a word's phonological form may be arbitrary, another part appears to be systematic and assists in acquiring lexical category information (Monaghan, Christiansen, & Fitneva, 2011). Our computational learning simulation did differ markedly from the corpus-based approach in one crucial respect: what it learned was not explicitly marked on the data. Nevertheless, the learning algorithm managed to nurture discrimination weights from naive input units (bi- or tri-graphs and words co-occurring in context) to our targeted allomorphic variants (the -a vs. -u endings). It did so without the explicit, intervening steps of classification and/or extension by analogy. In the process, it showed remarkable subtlety which captured Westfal's sense of elegance associated with the -u ending and the alleged vulgarity of the -a ending.

Our study thus shows that there may be systematicity at a different level than typically considered by linguists, and that naïve language users are sensitive to patterns at this particular level (see also Stevens, Harrington, & Schiel, 2019) even though there is no clear form-function relationship. Both rule-based and probabilistic grammars struggle to account for phenomena where the dimensions along which the items differ are difficult to identify. Our corpus study showed that linguistic intuitions, honed over years of exposure, are accurate but have three drawbacks: (1) the options offered by form-based rules or tendencies are too limited because of linguists' focus on or 'trust' in traditional concepts such as morphemes (but Baayen et al., 2011 and Blevins, 2016 offer amorphous accounts); (2) often,

coveted meaning-based labels apply only to part of the variation that needs explaining, thus forcing us to posit different, possibly hierarchical, tendencies or rules for different parts of the system (3) the meaning-based labels take language users years to build up because they require discovering commonalities between exemplars, as well as gradual generalizing and abstracting over these commonalities (cf., Dąbrowska, 2008). The traditional linguistic account is thus not only expensive in terms of up-front 'investment' in that it requires rather sophisticated formal or semantic insights; these insights themselves need to be built up and are therefore unlikely to guide the learning of the system they aim to explain.

6.2. Modelling productivity: memory or learning?

Our findings suggest that even seemingly unmotivated morphological form-variation can be learned without having to resort to item-by-item memorization, and this can be done using the same mechanism that can account for the development of rule-like and probabilistic linguistic behaviour. A forced choice task involving nonce words that were manipulated to incorporate relevant orthographic properties, confirmed the validity of our modelling results. Native speakers appear to be sensitive to the same elements that our simple learning principle picked up on, i.e. a variety of 2-letter combinations, most of which would not be considered legal morphemes. Further statistical analysis revealed that these results are not an epiphenomenon of data aggregation across speakers, but point to internalized tendencies in the usage as exhibited by individuals. And this brings us to the question we set out to answer: what exactly do we mean when we say 'internalized tendencies'? Form variation challenges rule-based and probabilistic approaches to language on a theoretical and methodological level. In this respect, two major points deserve further consideration.

First, memory-based models dominate efforts to conceptualize the accumulation and organization of linguistic knowledge within usage-based linguistics. These models are in essence large stores of exemplars upon which two cognitive processes, i.e. the categorization of and analogical

extension from exemplars, operate to explain the linguistic knowledge that drives comprehension and production. We have shown that we can identify a learning mechanism that allows us to be systematic where systematicity is apparent, but also where systematicity appears to be absent. The result is not necessarily parsimonious and for this reason it could be queried whether such an approach has any advantage over storing full forms. In fact, it has been suggested that "words (and other past experiences) are stored, and calculations of contextual similarity are only performed when needed, such as when one is asked to inflect nonce verbs" (Eddington, 2004: 102). Yet, how is something committed to and stored in memory? The image of storage highlights what is stored in memory, but obscures how it got there. It is often overlooked that learning plays a key role in committing experiences to memory, even though the experimental study of memory started with Ebbinghaus who measured memory by learning lists of words and testing how many had been forgotten versus could be recalled after experimentally manipulated periods of time had passed (Ebbinghaus, 1885). Furthermore, it is rarely made explicit that all formal learning traditions assume that what is learned are probabilistic regularities, i.e., the systematicity present in the environment (cf., Rescorla, 1988). How would we achieve 'mindless' memory storage of items upon which we can later perform operations? And when we need to perform operations on the items in this warehouse, how could these operations be parsimonious, i.e., fast and efficient, given that we are such 'memory hoarders'? For this, memory would need to be organized, and associative (discriminative) learning is an optimal candidate-mechanism for achieving this. As Sherlock Holmes famously put it, "when you have eliminated all which is impossible, then whatever remains, however improbable, must be the truth".

Usage-based linguists recognize that both the experience and the organization of the experience are in a perpetual state of flux; it allows, in principle, for units of different sizes to be identified in the input. This aligns with the starting point for learning in general: learning deals with the identification of predictive aspects of the environment. A holistic perspective adopts the position that, initially, all is one, and events will only be discriminated and wholes are only separated into their component parts (for original theorizing on these points see James, 1890; for a contemporary

synthesis see Ramscar, 2010; Ramscar et al., 2010) if this is indicated by adaptive pressures. A learning-based approach, which acknowledges gradience and continuity in input and representation, can handle less obviously systematic systematicity. In the linguistic hierarchy, rules trump probabilities: when describing phenomena, linguists typically aim for rules (and their exceptions), and resort to tendencies only in case of rule-defeat. Moreover, they like those tendencies to be reliably predicted from a few dimensions that can be meaningfully labelled. Many phenomena continue to defy adequate description under these terms. If a probabilistic approach falters, do we conclude that learning such phenomena requires item-by-item memorization? The results we have presented provide evidence against such a conclusion and for the plausibility of a more parsimonious mechanism. Rules work if there is certainty, while probabilities apply when uncertainty remains within bounds and the outcome is, hence, reasonably predictable. When the latter condition is not met (when there is no equilibrium to reach, viz. Danks, 2003), a system (be it an animal, human, or machine) does not simply 'give up'; instead, it engages with whatever degree of systematicity that is present in the usage-events. That residual systematicity will leave traces which will be learned over time to the extent possible. Learning these traces will eventually make prediction possible, even though it is unlikely to become error-free.

The naïve learning-based approach offers a bottom-up discovery procedure rather than a post-hoc interpretation or top-down labelling of categories. This way, we argue, it safe-guards the cognitive commitment (Lakoff, 1990; Divjak, 2015; Divjak, Levshina, & Klavan, 2016) in an organic manner, by gradually (iteratively) deriving all that a language user would need using only one simple principle of penalising current erroneous predictions about immediate future-state (which is the essence of the Rescorla & Wagner, 1972 rule). Our learning-based approach also comes with the added benefit that we can distinguish a different kind of learnable patterns, not constrained by any particular linguistic tradition, nor by the requirement for a direct form-function mapping. Bridging the representational (linguistic) and processing (psychological) perspectives with learning makes it possible to appreciate the complexity and adaptive dynamics of the system to their fullest.

Data and code

Script files: https://github.com/ooominds/TBC

Data files: https://doi.org/TBC

Author contributions

DD and PM are first authors, jointly responsible for the conception and design of the study, the analysis and interpretation of the data, and the writing and critical revision of the manuscript. AE was involved in the statistical analysis of the corpus and computational data and in drafting the corresponding parts of the manuscript; he also collected and analysed the behavioural data. JJ extracted and annotated the corpus data, drafted the description of the procedure and ran preliminary statistical analyses. CA contributed to the literature review, trained NDL and prepared the accompanying paper package.

Acknowledgements

We are greatly indebted to Tamas Rudas for extensive discussions regarding the analysis of categorical data. We are grateful to Neil Bermel, Jim Blevins, Paul O'Neill and the two journal reviewers for detailed comments on earlier versions of this manuscript. We would also like to thank the audiences at conferences and seminars where we presented parts of this study for comments and discussion.

Disclosure statement

The authors report no conflict of interest.

References

Ackerman, F., & Malouf, R. (2013). Morphological organization: The low conditional entropy conjecture. *Language, 89*(3), 429-464.

Ackerman, F., & Malouf, R. (2015). *The No Blur Principle effects as an emergent property of language systems.* Paper presented at the Proceedings of the annual meeting of the berkeley linguistics society.

Albright, A., & Hayes, B. (2003). Rules vs. analogy in English past tenses: A computational/experimental study. *Cognition, 90*, 119-161.

Ambridge, B. (2019). Against stored abstractions: A radical exemplar model of language acquisition. *First Language*. doi:https://doi/org/10.1177/0142723719869731

Arppe, A., Hendrix, P., Milin, P., Baayen, R. H., Sering, T., & Shaoul, C. (2018). ndl: Naive Discriminative Learning (Version 0.2.18). Retrieved from https://CRAN.R-project.org/package=ndl

Baayen, H., Chuang, Y.-Y., & Blevins, J. P. (2018). Inflectional morphology with linear mappings. *The mental lexicon, 13*(2), 230-268.

Baayen, H., Chuang, Y.-Y., Shafaei-Bajestan, E., & Blevins, J. P. (2019). The discriminative lexicon: A unified computational model for the lexicon and lexical processing in comprehension and production grounded not in (de) composition but in linear discriminative learning. *Complexity, 2019*.

Baayen, H., Milin, P., Filipović Đurđević, D., Hendrix, P., & Marelli, M. (2011). An amorphous model for morphological processing in visual comprehension based on naive discriminative learning. *Psychological Review, 118*(3), 438.

Benko, V. (2014). *Aranea: Yet another family of (comparable) web corpora.* Paper presented at the International Conference on Text, Speech, and Dialogue.

Berko, J. (1958). The child's learning of English morphology. *Word, 14*(2-3), 150-177.

Bermel, N., & Knittl, L. (2012). Corpus frequency and acceptability judgments: A study of morphosyntactic variants in Czech. *Corpus linguistics and linguistic theory, 8*(2), 241-275.

Blevins, J. P. (2004). Inflection classes and economy. *Explorations in Nominal Inflection, Berlin: Mouton de Gruyter*, 51-96.

Blevins, J. P. (2016). *Word and paradigm morphology*: Oxford University Press.

Bodnarowska, J. (1962). Problematyka doboru końcówek -a/-u w dopełniaczu 1. p. rzeczowników męskich. *Język Polski 42*(1), 29-49.

Bybee, J. L. (1985). Regular morphology and the lexicon. *Language and Cognitive Processes*(10), 425–455.

Bybee, J. L. (2006). From usage to grammar: the mind's response to repetition. *Language, 82*(4), 711-733.

Bybee, J. L. (2013). Usage-based theory and exemplar representation. In T. Hoffman & G. Trousdale (Eds.), *The Oxford Handbook of Construction Grammar* (pp. 49-69). Oxford: Oxford University Press.

Chen, Z., Haykin, S., Eggermont, J. J., & Becker, S. (2008). *Correlative learning: a basis for brain and adaptive systems* (Vol. 49): John Wiley & Sons.

Chuang, Y.-Y., Vollmer, M.-l., Shafaei-Bajestan, E., Gahl, S., Hendrix, P., & Baayen, R. H. (2020). The processing of pseudoword form and meaning in production and comprehension: A computational modeling approach using Linear Discriminative Learning. *Behavior Research Methods*.

Dąbrowska, E. (2004). Rules or schemas? Evidence from Polish. *Language and cognitive processes*(19), 225-271.

Dąbrowska, E. (2005). Productivity and beyond: mastering the Polish genitive inflection. *Journal of Child Language*(32), 191-205.

Dąbrowska, E. (2008). The later development of an early-emerging system: the curious case of the Polish genitive. *Linguistics, 46*(3), 629–650.

Dąbrowska, E. (2016). Cognitive linguistics' seven deadly sins. *Cognitive linguistics, 27*(4), 479-491.

Daelemans, W., Berck, P., & Gillis, S. (1997). Data mining as a method for linguistic analysis: Dutch diminutives. *Folia Linguistica, 31*(1-2), 57-76.

Daelemans, W., Zavrel, J., Van der Sloot, K., & Van den Bosch, A. (2007). Timbl: Tilburg memory-based learner. *Version, 6*, 07-03.

Danks, D. (2003). Equilibria of the Rescorla–Wagner model. *Journal of Mathematical Psychology, 47*(2), 109-121. doi:https://doi.org/10.1016/S0022-2496(02)00016-0

Divjak, D. (2015). Four challenges for usage-based linguistics. In J. Daems, E. Zenner, K. Heylen, D. Speelman, & H. Cuyckens (Eds.), *Change of Paradigms: New Paradoxes. Recontextualizing Language and Linguistics* (pp. 297-309). Berlin: de Gruyter.

Divjak, D., Levshina, N., & Klavan, J. (2016). Cognitive Linguistics: Looking back, looking forward. *Cognitive Linguistics, 27*(4), 447-463.

Ebbinghaus, H. (1885). *Über das Gedächtnis: Untersuchungen zur Experimentellen Psychologie (About Memory: Studies on Experimental Psychology)*. Leipzig: Duncker & Humblot.

Eddington, D. (2002a). Dissociation in Italian conjugations: A single-route account. *Brain and Language 81*, 291–302.

Eddington, D. (2002b). Spanish gender assignment in an analogical framework. *Journal of Quantitative Linguistics, 9*, 49-75.

Eddington, D. (2004). *Spanish Phonology and Morphology: Experimental and quantitative perspectives*. Amsterdam: John Benjamins.

Ellis, N. C., Natsume, M., Stavropoulou, K., Hoxhallari, L., Van Daal, V. H., Polyzoe, N., . . . Petalas, M. (2004). The effects of orthographic depth on learning to read alphabetic, syllabic, and logographic scripts. *Reading Research Quarterly, 39*(4), 438-468.

Ellis, N. C., Simpson-Vlach, R., & Maynard, C. (2008). Formulaic language in native and second language speakers: Psycholinguistics, corpus linguistics, and TESOL. *Tesol Quarterly, 42*(3), 375-396.

Enquist, M., Lind, J., & Ghirlanda, S. (2016). The power of associative learning and the ontogeny of optimal behaviour. *Royal Society open science, 3*(11), 160734.

Griffiths, T. L., Steyvers, M., & Tenenbaum, J. B. (2007). Topics in semantic representation. *Psychological Review, 114*(2), 211.

Hahn, U., & Charles, N. R. (2000). German inflection: Single Route or Dual Route. *Cognitive Psychology, 41*, 313-360.

Hayes, B., & Londe, Z. C. (2006). Stochastic phonological knowledge: The case of Hungarian vowel harmony. *Phonology, 23*(1), 59-104.

Hendrix, P., Bolger, P., & Baayen, R. H. (2016). Distinct ERP signatures of word frequency, phrase frequency, and prototypicality in speech production. *Journal of Experimental Psychology: Language, Memory and Cognition, 43*(1), 128 - 149.

Hickok, G. (2014). The architecture of speech production and the role of the phoneme in speech processing. *Language, Cognition and Neuroscience, 29*(1), 2-20.

Ivić, P. (1990). *O jeziku nekadašnjem i sadašnjem [On past and contemporary language]*. Belgrade: Bigz.

James, W. (1890). *The principles of psychology (Vol. 1)* (Vol. 474). New York: Henry Holt and Co.

Janda, L. A. (1996). Figure, ground, and animacy in Slavic declension. *Slavic and East European Journal, 40*(2), 325-355.

Janus, D. (2006-2012). Poliqarp. Retrieved from http://poliqarp.sourceforge.net/index.html

Karssen, L. C., van Duijn, C. M., & Aulchenko, Y. S. (2016). The GenABEL Project for statistical genomics. *F1000Research, 5*(914). doi:https://doi.org/10.12688/f1000research.8733.1

Kertész, Z. (2003). Vowel harmony and the stratified lexicon of Hungarian. *The Odd Yearbook, 7*, 62-77.

Keuleers, E. (2008). *Memory-based learning of inflectional morphology.* (PhD), University of Antwerp, Antwerp.

Keuleers, E., & Brysbaert, M. (2010). Wuggy: A multilingual pseudoword generator. *Behavior Research Methods, 42*(3), 627-633.

Keuleers, E., Sandra, D., Daelemans, W., Gillis, S., Durieux, G., & Martens, E. (2007). Dutch plural inflection: The exception that proves the analogy. *Cognitive Psychology, 54*, 283–318.

Köpcke, K.-M. (1993). *Schemata bei der Pluralbildung im Deutschen*. Tübingen: Narr.

Kottum, S. E. (1981). The genitive singular form of masculine nouns in Polish. *Scando-Slavica, 27*(1), 179-186.

Krott, A., Baayen, R. H., & Schreuder, R. (2001). Analogy in morphology: modeling the choice of linking morphemes in Dutch. *Linguistics, 39*(1; ISSU 371), 51-94.

Krott, A., Schreuder, R., Baayen, R. H., & Dressler, W. U. (2007). Analogical effects on linking elements in German compound words. *Language and cognitive processes, 22*(1), 25-57.

Lakoff, G. (1990). The Invariance Hypothesis: Is abstract reason based on image-schemas? . *Cognitive Linguistics 1*(1), 39-74.

Landauer, T. K., & Dumais, S. T. (1997). A solution to Plato's problem: The latent semantic analysis theory of acquisition, induction, and representation of knowledge. *Psychological Review, 104*(2), 211.

Lečić, D. (2016). *Morphological Doublets in Croatian: A multi-methodological analysis.* (Doctor of Philosophy), The University of Sheffield,

Lieber, R. (1982). Allomorphy. *Linguistic Analysis Seattle, Wash., 10*(1), 27-52.

Lund, K., & Burgess, C. (1996). Producing high-dimensional semantic spaces from lexical co-occurrence. *Behavior research methods, instruments, & computers, 28*(2), 203-208.

Lyons, J. (1968). *Introduction to theoretical linguistics*: Cambridge university press.

MacWhinney, B. (Ed.) (1999). *The emergence of language*. Mahwah, NJ: Lawrence Erlbaum Associates.

MacWhinney, B., & Leinbach, J. (1991). Implementations are not conceptualizations: Revising the verb learning model. *Cognition, 40*(1/2), 121-157.

Mańczak, W. (1953). O repartycji końcówek dopelniacza -a:-u. *Język Polski 33*(2), 71-72.

Marcus, G. F., Brinkmann, U., Vlahsen, H., Wiese, R., & Pinker, S. (1995). German Inflection: The Exception That Proves the Rule. *Cognitive Psychology, 29*, 189-256.

Marelli, M., & Baroni, M. (2015). Affixation in semantic space: Modeling morpheme meanings with compositional distributional semantics. *Psychological Review, 122*(3), 485.

McClelland, J. L., & Patterson, K. (2002). 'Words or Rules' cannot exploit the regularity in exceptions. *Trends in Cognitive Sciences, 6*(11), 464-465.

Mikolov, T., Sutskever, I., Chen, K., Corrado, G. S., & Dean, J. (2013). *Distributed representations of words and phrases and their compositionality.* Paper presented at the Advances in neural information processing systems.

Milin, P., Divjak, D., & Baayen, R. H. (2017). A learning perspective on individual differences in skilled reading: Exploring and exploiting orthographic and semantic discrimination cues. *Journal of Experimental Psychology: Learning, Memory, and Cognition, 43*(11), 1730.

Milin, P., Divjak, D., Dimitrijević, S., & Baayen, R. H. (2016). Towards cognitively plausible data science in language research. *Cognitive Linguistics, 27*(4), 507-526. doi:https://doi.org/10.1515/cog-2016-0055

Milin, P., Feldman, L. B., Ramscar, M., Hendrix, P., & Baayen, R. H. (2017). Discrimination in lexical decision. *PloS one, 12*(2), e0171935.

Milin, P., Keuleers, E., & Filipović Đurđević, D. (2011). Allomorphic Responses in Serbian Pseudo-Nouns as a Result of Analogical Learning. . *Acta Linguistica Hungarica*(58), 65-84.

Milin, P., Keuleers, E., & Filipović-Đurđević, D. (2011). Allomorphic responses in Serbian pseudo-nouns as a result of analogical learning. *Acta Linguistica Hungarica, 58*(1), 65-84.

Milin, P., Nenadić, F., & Ramscar, M. (Under revision). Approaching text genre: How contextualized experience shapes task-specific performance. *Scientific Study of Literature*.

Mirković, J., Seidenberg, M., & Joanisse, M. (2011). Rules Versus Statistics: Insights From a Highly Inflected Language. *Cognitive Science, 35*, 638-681.

Mirković, J., Vinals, L., & Gaskell, M. G. (2018). The role of Complementary Learning Systems in learning and consolidation in a quasi-regular domain. *Cortex*.

Monaghan, P., Christiansen, M. H., & Fitneva, S. A. (2011). The arbitrariness of the Sign: Learning Advantages From the Structure of the Vocabulary. *Journal of Experimental Psychology: General., 140*(3), 325-347.

Naimi, B., Hamm, N. A., Groen, T. A., Skidmore, A. K., & Toxopeus, A. G. (2014). Where is positional uncertainty a problem for species distribution modelling? *Ecography, 37*(2), 191-203.

Nakisa, R., & Hahn, U. (1996). *Where defaults don't help: The case of the German plural system.* Paper presented at the 18th annual conference of the Cognitive Science Society

Newman, R. L., Jared, D., & Haigh, C. A. (2012). Does phonology play a role when skilled readers read high-frequency words? Evidence from ERPs. *Language and cognitive processes, 27*(9), 1361– 1384.

Nosofsky, R. M. (1986). Attention, similarity, and the identification–categorization relationship. *Journal of experimental psychology: General, 115*(1), 39-57.

Nosofsky, R. M. (1990). Relations between exemplar-similarity and likelihood models of classification. *Journal of Mathematical psychology, 34*(4), 393-418.

Paster, M. (2014). Allomorphy. In R. Lieber & P. Štekauer (Eds.), *The Oxford Handbook of Derivational Morphology* (pp. 219-234). Oxford: Oxford University Press.

Patz, E. (1991). Djabugay. The handbook of Australian languages 4, ed. by RM W. Dixon and BJ Blake, 245–347. In: Melbourne: Oxford University Press.

Perrone-Bertolotti, M., Kujala, J., Vidal, J. R., Hamame, C. M., Ossandon, T., Bertrand, O., . . . Lachaux, J.-P. (2012). How silent is silent reading? intracerebral evidence for top-down activation of temporal voice areas during reading. *Journal of Neuroscience, 32*(49), 17554–17562.

Pinker, S., & Prince, A. (1988). On language and connectionism: analysis of a parallel distributed processing model of language acquisition *Cognition*(28), 73–193.

Plag, I. (2018). *Word-formation in English*. Cambridge: Cambridge University Press.

Qualtrics. Provo, Utah, USA. Retrieved from https://www.qualtrics.com

Ramscar, M. (2010). Computing machinery and understanding. *Cognitive science, 34*(6), 966-971.

Ramscar, M., Yarlett, D., Dye, M., Denny, K., & Thorpe, K. (2010). The Effects of Feature- Label- Order and Their Implications for Symbolic Learning. *Cognitive Science, 34*(6), 909-957. doi:https://doi.org/10.1111/j.1551-6709.2009.01092.x

Ratcliffe, R. R. (1998). *The broken plural problem in Arabic and comparative semitic: allomorphy and analogy in non-concatenative morphology* (Vol. 168): John Benjamins Publishing.

Rescorla, R. A. (1968). Probability of shock in the presence and absence of CS in fear conditioning. *Journal of comparative and physiological psychology, 66*(1), 1-5.

Rescorla, R. A. (1988). Pavlovian conditioning: It's not what you think it is. *American psychologist, 43*(3), 151-160.

Rescorla, R. A. (2008). Rescorla-Wagner model. *Scholarpedia, 3*(3), 2237. http://www.scholarpedia.org/article/Rescorla-Wagner_model

Rescorla, R. A., & Wagner, A. R. (1972). A theory of Pavlovian conditioning: Variations in the effectiveness of reinforcement and nonreinforcement. In A. H. Black & W. F. Prokasy (Eds.), *Classical conditioning II: Current research and theory* (pp. 64-99). New York: Appleton-Century-Crofts.

Rudas, T. (1998). *Odds ratios in the analysis of contingency tables*. Thousand Oaks: Sage.

Rumelhart, D. E., & McClelland, J. L. (1985). *On learning the past tenses of English verbs*. Retrieved from La Jolla, CA:

Seidenberg, M. S., & Plaut, D. C. (2014). Quasiregularity and its discontents: the legacy of the past tense debate. *Cognitive science, 38*(6), 1190-1228.

Sering, K., Weitz, M., Künstle, D.-E., & Schneider, L. (2017). Pyndl: Naive discriminative learning in python. Retrieved from https://doi.org/10.5281/zenodo.597964

Shaoul, C., & Westbury, C. (2010). Exploring lexical co-occurrence space using HiDEx. *Behavior Research Methods, 42*(2), 393-413.

Skousen, R. (1989). *Analogical modeling of language*. Berlin: Springer Science & Business Media.

Spellman, B. A. (1996). Conditionalizing causality. In D. R. Shanks, K. J. Holyoak, & D. L. Medin (Eds.), *The psychology of learning and motivation: Advances in research and theory* (pp. 167-206). New York: Academic Press.

Spencer, A. (2001). Morphology. In M. Aronoff & J. Rees-Miller (Eds.), *The handbook of linguistics* (pp. 213-238). Oxford MA: Blackwell.

Stevens, M., Harrington, J., & Schiel, F. (2019). Associating the origin and spread of sound change using agent-based modelling applied to /s/-retraction in English. *Glossa: a journal of general linguistics 4*(1), 1-30. doi:https://doi.org/10.5334/gjgl.620

Storkel, H. L. (2001). Learning new words: Phonotactic probability in language development. *Journal of Speech, Language, and Hearing Research : JSLHR, 44*(6), 1321-1337.

Storkel, H. L. (2003). Learning new words II: Phonotactic probability in verb learning. *Journal of Speech, Language, and Hearing Research : JSLHR, 46*(6), 1312-1323.

Strobl, C., Boulesteix, A. L., Zeileis, A., & Hothorn, T. (2007). Bias in random forest variable importance measures: illustrations, sources and a solution. *BMC bioinformatics, 8*(25). doi:https://doi.org/10.1186/1471-2105-8-25

Swan, O. (2002). *A grammar of contemporary Polish*. Bloomington, IN: Slavica.

Thornton, A. M. (2011). Overabundance (Multiple Forms Realizing the Same Cell): A Non-canonical Phenomenon in Italian Verb Morphology. In M. Maiden, J. C. Smith, M. Goldbach, & M.-O. Hinzelin (Eds.), *Morphological Autonomy*. Oxford: Oxford University Press.

Tweedie, F. J., & Baayen, R. H. (1998). How variable may a constant be? Measures of lexical richness in perspective. *Computers and the Humanities, 32*(5), 323-352.

Westfal, S. (1956). *A study in Polish morphology: The genitive singular masculine* (Vol. 8). 's-Gravenhage Mouton & Co.

Widrow, B., & Hoff, M. E. (1960). *Adaptive switching circuits.* Paper presented at the WESCON Convention Record Part IV.

Widrow, B., & Lehr, M. A. (1990). 30 Years of Adaptive Neural Networks: Perceptron, Madaline, and Backpropagation. *Proceedings of the IEEE, 78*(9), 1415-1442.

Wulf, D. J. (2002). pplying analogical modeling to the German plural. In R. Skousen (Ed.), *Analogical modeling: An exemplar-based approach to language* (pp. 109-122). Amsterdam: John Benjamins.

Zasina, A. J. (2017). Konkurence koncovek -a a -u v genitivu singuláru neživotných maskulin v polštině. In M. Stluka & M. Škrabal (Eds.), *Liſka a czban – Sborník příspěvků k 70. narozeninám prof. Karla Kučery* (pp. 90-98). Praha: NLN.

Zec, D. (2006). Phonology within morphology in South Slavic: the case of OV augmentation. *Handout. Nova Gorica, University of Nova Gorica, 2*.

Zwoliński, P. (1948). Przyczynki do repartycji polskich końcówek -a/-u w dopełniaczu l. poj. rzeczowników męskich. *Język Polski 28*(6), 174-177.