Modeling morphology with Linear Discriminative Learning: considerations and design choices

Maria Heitmeier, Yu-Ying Chuang, and R. Harald Baayen
Eberhard-Karls Universität Tübingen

May 2021

Abstract

This study addresses a series of methodological questions that arise when modeling inflectional morphology with Linear Discriminative Learning. Taking the semi-productive German noun system as example, we illustrate how decisions made about the representation of form and meaning influence model performance. We clarify that for modeling frequency effects in learning, it is essential to make use of incremental learning rather than the endstate of learning. We also discuss how the model can be set up to approximate the learning of inflected words in context. In addition, we illustrate how in this approach the wug task can be modeled in considerable detail. In general, the model provides an excellent memory for known words, but appropriately shows more limited performance for unseen data, in line with the semi-productivity of German noun inflection and generalization performance of native German speakers.

Keywords: German nouns, Linear discriminative learning, semi-productivity, multivariate multiple regression, Widrow-Hoff learning, frequency of occurrence, semantic roles, wug task

1 Introduction

Computational models of morphology fall into two broad classes. The first class, which comprises the largest number of models, addresses the question of how to produce a morphologically complex word given a morphologically related form (often a stem, or an identifier of a stem or lexeme) and a set of inflectional or derivational features. We will refer to these models as form-oriented models. The second, much smaller class, covers models that seek to understand the relation between words' forms and their meanings. We will refer to these models as meaning-oriented models. We first consider some important form-oriented models.

Analogical Modeling of Language (Skousen, 1989, 2002) and Memory Based Learning (Daelemans and Van den Bosch, 2005) are nearest-neighbor classifiers. Input to these models are tables with observations (words) in rows, and factorial predictors and a factorial response in columns. The response variable specifies, for each observation, a particular outcome class (e.g., an allomorph), and the model is given the task to predict the outcome classes from the other predictor variables (for allomorphy prediction, typically position-specific specifications of words' phonological make-up). Predictions are based on sets of nearest neighbors, serving as constrained exemplar sets for generalization. These models have proved very useful for understanding a range of morphological phenomena, ranging from the allomorphy of the Dutch diminutive (Daelemans et al., 1995) to stress assignment in English (Arndt-Lappe, 2011).

Within the tradition of generative grammar, Minimum Generalization Learning (Albright and Hayes, 2003) offers an algorithm for rule induction (for comparison with nearest neighbor methods, see Keuleers et al., 2007). The model finds rules by an iterative process of minimal generalization that combines specific rules into ever more general rules. Each rule comes with a measure of prediction accuracy, and the rule with the highest accuracy is selected for predicting a word's form. The model as laid out in Albright and Hayes (2003) works with fine-grained phonological features.

Another model coming from the generative tradition is that of Belth et al. (2021), which makes use of a particular implementation of recursive partitioning. Their study illustrates the algorithm for a dataset with words as observations, with as predictors a word's stem, some stem-final segments, and its inflectional features, and as response a categorical variable specifying the morphological change that produces the inflection from the stem.

Ernestus and Baayen (2003) compared the performance of the MBL, AML, and GLM models, as well as a logistic regression model and a recursive partitioning tree (Breiman et al., 1984), on the task of predicting whether word-final obstruents in Dutch alternate with respect to their voicing. They observed similar performance across all models, with the best performance, surprisingly, for the only parameter-free model, AML. Their results suggest that the quantitative structure of morphological data sets may be straightforward to discover for any reasonably decent classifier.

All models discussed thus far are exemplar-based, in the sense that the input to any of these models consists of a table with exemplars, exemplar features selected on the basis of domain knowledge, and a categorical response variable specifying targeted morphological form changes. In other words, all these models are classifiers that absolve the analyst from hand-engineering lexical entries, rules or constraints operating on these lexical entries, and theoretical constructs such as inflectional classes. In this respect, they differ fundamentally from the following three computational methods.

Evans and Gazdar (1996) introduced the DATR language for defining non-monotonic inheritance networks for lexical knowledge representation. This language is optimized for removing any redundancy from lexical descriptions. A DATR model requires the analyst to set up lexical entries that specify information about, for instance, inflectional class, gender, the forms of exponents, and

various kinds of phonological information. The challenge for the analyst is to set up the lexicon in such a way that the number of lexical entries is kept as small as possible, while still allowing the model, through its mechanism of inheritance, to correctly predict all inflected variants. The theory of realizational morphology (RM) of Stump (2001), which sets up rules for realizing bundles of inflectional and lexical features in phonological form, can also be seen as a formal language (a finite-state transducer) that provides mappings from underlying representations onto their corresponding surface forms and vice versa (Karttunen, 2003). Finally, the Gradual Learning Algorithm (Boersma, 1998; Boersma and Hayes, 2001, GLA) works within the framework of optimality theory (Prince and Smolensky, 2008). The algorithm is initialized with a set of constraints and gradually learns an optimal constraint ranking by incrementally moving through the training data, and upgrading or downgrading constraints according to the algorithm's current predictions.

A third group of form-oriented computational models comprises connectionist models. The famous past-tense model developed by Rumelhart and McClelland (1986) used as its core engine a simple network, mapping input form features to output form features. This model was trained to produce English past-tense forms given the corresponding present-tense form. An early enhancement of this model was proposed by MacWhinney and Leinbach (1991), for an overview of the many follow-up models, see Kirov and Cotterell (2018). Kirov and Cotterell proposed a sequence-to-sequence deep learning network, the ED learner, that they argue does not suffer from the drawbacks noted by Pinker and Prince (1988) for the original model of Rumelhart and McClelland (1986). Malouf (2017) introduced a recurrent deep learning model trained to predict upcoming segments, and showed that this model has high accuracy for predicting paradigm forms given the lexeme and the inflectional specifications of the desired paradigm cell. An independent line of research focuses on incremental topological learning using temporal self-organizing maps (TSOMs, Ferro et al., 2011; Chersi et al., 2014; Marzi et al., 2012, 2018).

In summary, the class of form-oriented models comprises three subsets of models: statistical classifiers (AML, MBL, GLM, recursive partitioning), generators based on linguistic knowledge engineering (DATR, RM, GLA), and connectionist models (paste-tense model, ED learner). The models just referenced presuppose that when speakers use a morphologically complex form, this form is derived on the fly from its underlying form. The sole exception is the model of Malouf (2017), which takes the lexeme and its inflectional features as point of departure. As pointed out by Blevins (2016), the focus on how to create one form from another has its origin in pedagogical grammars, which face the task of clarifying to a second language learner how to create inflected variants. Unsurprisingly, applications within natural language processing also have need of systems that can generate inflected and derived words.

However, it is far from self-evident that native speakers of English would create past-tense forms from present-tense stems, or that speakers of Estonian would inflect nouns on the basis of criteria such as inflectional class and a set of stem allomorphs. The class of meaning-oriented models for morphological processing, which is more sparsely populated than the class of form-oriented models, comprises models proposing that in comprehension, the listener or reader can go straight from the auditory or visual input to the intended meaning, without having to go through a pipeline requiring identification of underlying forms and exponents. Likewise, speakers are argued to start from a meaning, and realize this meaning in written or spoken form.

The class of meaning-oriented models comprises both symbolic and subsymbolic models. The symbolic models of Levelt et al. (1999) and Dell (1986) implement, albeit in different ways, the general approach of realizational morphology. Concepts activate morphemes, which in turn activate

stems and exponents. Both models hold that the production of morphologically complex words is a compositional process in which at various hierarchically ordered levels, units are assembled together and ordered for articulation. It is worth noting that these psycholinguistic models have been worked out only for English, and to our knowledge have not been applied to languages with non-trivial morphological systems.

Unlike these symbolic models, the subsymbolic triangle model of Harm and Seidenberg (2004) sets up multilayer networks between orthographic, phonological, and semantic units. No attempt is made to define morphemes, stems, or exponents. To the extent that such units have any reality, they are assumed to arise, statistically, at the hidden layers. Likewise, the model for auditory comprehension of Gaskell and Marslen-Wilson (1997) uses a three-layer recurrent network to map speech input onto distributed semantic representations, without any attempt to isolate units such as phonemes or morphemes. The triangle model is applied by Mirković et al. (2005) to a language with a rich morphological system, Serbian. Instead of taking gender to be a theoretical primitive (serving, for instance, as input to a classifier), this study argues that gender is an emergent property of the network that arises from statistical regularities governing both words' forms and their meanings (see Corbett, 1991, for discussion of semantic motivations for gender systems).

The naive discrimination learning (NDL) model proposed by Baayen et al. (2011) represents words' forms subsymbolically, but words' meanings symbolically. It thus is a hybrid model. The modeling set-up that we discuss in the remainder of this study, that of linear discriminative learning (LDL, Baayen et al., 2019), replaces the symbolic representation of word meaning in NDL by subsymbolic representations that build on distributional semantics (Landauer and Dumais, 1997; Mikolov et al., 2013b).

LDL is an implementation of Word and Paradigm Morphology (Matthews, 1974; Blevins, 2016), and as such explicitly eschews sublexical units such as stems and exponents. However, semantic representations in LDL are analytical, in the sense that the semantic vector (word embedding) of an inflected word is constructed (by means of vector addition) from the semantic vector of the content lexeme of that word and the semantic vectors of the inflectional functions that are to be expressed. Below, we introduce this concept in more detail. Here, we note that both NDL and LDL make use of the simplest possible networks, networks with only input and output layers, without any intervening hidden layers. Mathematically, NDL implements multiple label classification, whereas LDL implements multivariate multiple regression (see, e.g., Baayen and Smolka, 2020; Chuang and Baayen, 2021).

To place LDL in perspective, the distinction made by Breiman et al. (2001) between statistical models and machine learning is useful. The goal of statistical models is to provide insight into the mechanisms that are likely to have generated the data. The goal of machine learning, on the other hand, is to optimize prediction accuracy, and if the system that best optimizes prediction accuracy is a black box, this is no reason for concern. LDL is much closer to statistical modeling than to machine learning. All representations at input and output levels can be set up to be transparently interpretable (Baayen et al., 2019). Furthermore, because the model is a multivariate multiple regression model, the mathematical properties of which are well-understood, modeling results do not depend on architectural hyper-parameters (such as how many LSTM layers with how many LSTM units to build into the model), and are completely determined by the representations chosen by the analyst.

The goals of this study are, first, to clarify how choices of representation affect LDL model performance; second, to illustrate how much can be achieved simply with multivariate multiple

regression; and third, to call attention to the kind of problems that are encountered when the modeling of word meaning is taken seriously. We do so by addressing the comprehension and production of German nouns. In what follows, we first introduce some basic properties of the German noun system, and review some of the models that have been proposed for German nouns. We then introduce the framework of LDL, after which we proceed to the heart of this study, a systematic overview of modeling choices with respect to the representation of form, the representation of meaning, and learning algorithm (incremental learning versus the regression 'endstate of learning' solution).

2 German noun morphology

The German noun system is characterized by three different genders. As can be seen in Table 1, plural forms are marked with one of four suffixes (*-(e)n*, *-er*, *-e*, *-s*) or without adding a suffix (−0; a "zero" morpheme (Köpcke, 1988, p. 306)), three of which can pair with stem vowel fronting (e.g. *a* (/a/) → *ä* (/ɛ/)) (e.g. Köpcke, 1988). There are some additional suffixes which usually apply to words with foreign origin, such as *-i* (e.g. *Cello* → *Celli*, 'cellos') (Cahill and Gazdar, 1999). These eight classes can be further subdivided according to various sub-regularities in nouns' phonology and gender. For example, Cahill and Gazdar (1999) subcategorise the nouns into 11 classes, based on whether singular forms have a different suffix than plural forms (e.g. *Album* → *Alben*, 'albums'). On the other hand, Nakisa and Hahn (1996) distinguished 60 different classes. None of the plural classes is prevalent overall (Köpcke, 1988), and it is impossible to fully predict plural class from gender, syntax, phonology or semantics (Köpcke, 1988; Cahill and Gazdar, 1999; Trommer, 2021). To illustrate, consider the neuter nouns *Fett, Brett* and *Bett* with their nominativ plurals *Fette, Bretter* and *Betten* or the masculine nouns *Schmerz* → *Schmerzen* and *Scherz* → *Scherze*.

The five broad classes of German nouns that can be set up by considering just the plural exponents have to be further subdivided into more fine-grained declension classes once case is taken into account. German has four cases: *nominative, genitive, dative*, and *accusative*. There are only two additional endings available to mark case: *-(e)n* and *-(e)s* (Schulz and Griesbach, 1981). Since many forms do not receive a separate marker for case in plural forms, the system has been described as "degenerate" (Bierwisch, 2018, p. 245) (see Table 2). Just as plural forms, case forms are not fully predictable from gender, phonology or meaning.

Plural class	Example	Type frequency
-(e)n	Tasse → Tassen 'cup(s)'	56.5%
(uml+)-e	Tag → Tage 'day(s)'	
	Topf → Töpfe 'pot(s)'	23.9%
(uml+)-er	Brett → Bretter 'board(s)'	
	Glas → Gläser 'glass(es)'	2.3%
(uml+)-0	Daumen → Daumen 'thumb(s)'	
	Apfel → Äpfel 'apple(s)'	13.3%
-s	Kamera → Kameras 'camera(s)'	2.6%

Table 1: Plural classes of German nouns (relative frequencies from Gaeta (2008)). Most of the classes can appear with both masculine and neuter nouns. Feminine nouns belong mostly to the *-(e)n* class (97%).

The German noun system is in many ways irregular and unpredictable. Unsurprisingly, it has been the subject of a long-standing debate whether a distinction between regular and irregular nouns is useful for German. It is also unsurprising that the system shows limited productivity. Several so-called 'wug' studies, where participants are asked to provide inflected forms of presented nonce words, clarified that German native speakers struggle with generalizing the system to new plural forms. Köpcke (1988); Zaretsky et al. (2013); McCurdy et al. (2020) reported a high variability across speakers with respect to the plural forms that they produced. Köpcke (1988) took this as evidence for a "modified schema model" of German noun inflection. According to Köpcke, plural forms are generated based not only on a speaker's experience with the German noun system, but also based on the "cue validity" of the different plural markers. For example, *-(e)n* is a very valid cue for plural, as it does not occur with many singular forms, and therefore is informative for plurality. By contrast, *-er* has low cue validity for plurality, as it occurs with many singular forms. According to Köpcke, additional factors such as grammatical gender can also modify cue validity.

Köpcke (1988) also observed that *-s* is used slightly more in his wug experiments than would be expected from corpus data. Marcus et al. (1995) and Clahsen (1999) took this as a starting point for a dual-route model of German noun inflection. They argued that *-s* serves as the regular default plural marker in German, in contrast to all other plural markers that are supposed to be irregular and rote-learned. Others, however, have argued that an *-s* default rule does not provide any additional explanatory value in a theory of German plurals (Nakisa and Hahn, 1996; Zaretsky and Lange, 2015; Behrens and Tomasello, 1999; Indefrey, 1999). Furthermore, Baayen et al. (2002) showed that the kind of arguments used by Clahsen (1999) to support the default status of the German *-s* exponent don't generalize to Dutch.

Subregularities within the German noun system have also been pointed out (Wunderlich, 1999; Wiese, 1999). For instance, Wunderlich (1999, p.7f.) reports a set of rules that German nouns adhere to, which can be overridden on an item-by-item basis through 'lexical storage'. For example, he notes that

a. Masculines ending in schwa are weakly inflected (and thus also have n-plurals).
b. Non-umlauting feminines have an n-plural.
c. Non-feminines ending in a consonant have a ə-plural. [...]
e. All untypical nouns have an s-plural. [...]

He also allows for semantics to co-determine class membership. For instance, masculine animate

case & number	masculin I	masculin II	neutral	feminin
Nom. sg.	der Freund	der Mensch	das Kind	die Mutter
Gen. sg.	des Freund**es**	des Mensch**en**	des Kind**es**	der Mutter
Dat. sg.	dem Freund	dem Mensch**en**	dem Kind	der Mutter
Acc. sg.	den Freund	den Mensch**en**	das Kind	die Mutter
Nom. pl.	die Freund**e**	die Mensch**en**	die Kind**er**	die Mütter
Gen. pl.	der Freund**e**	der Mensch**en**	der Kind**er**	der Mütter
Dat. pl.	den Freund**en**	den Mensch**en**	den Kind**ern**	den Mütter**n**
Acc. pl.	die Freund**e**	die Mensch**en**	die Kind**er**	die Mütter

Table 2: German noun declension. Plural endings vary with declension class. Table adapted from Schulz and Griesbach (1981, p. 105).

nouns show a tendency to belong to the -*n* plural class (see also Gaeta, 2008). A further remarkable aspect of the German noun system, especially for second language learners, is that whereas it is remarkably difficult to learn to produce the proper case-inflected forms, understanding these forms in context is straightforward.

In the light of these considerations, the challenges for computational modeling of German noun inflection, specifically from a cognitive perspective, are the following:

1. to construct a memory for a highly irregular, degenerate, semi-productive system,

2. to ensure that this memory shows some moderate productivity for novel forms, but with all the uncertainties that characterize the generalization capacities of German native speakers, and

3. to furthermore ensure that the performance of the mappings from form to meaning, and from meaning to form, within the framework of the discriminative lexicon (Baayen et al., 2019), are properly asymmetric with respect to comprehension and production accuracy (see also Chuang et al., 2020a).

2.1 Computational models for German nouns

Unsurprisingly, the complexity of the German declension system has inspired many researchers to come to grips with this system with the help of computational modeling. Currently, a wide range of models is available. The DATR model of Cahill and Gazdar (1999) belongs to the class of generating models based on linguistic knowledge engineering. It divides German noun lexemes into carefully designed hierarchies of declension classes. Each class inherits the properties from classes further up in the hierarchy, but will override some of these properties. This model provides a successful and succint formal model for German noun declension. The downside of the model is that for new nouns, the correct declension class has to be assigned manually. The model of Trommer (2021), which draws on Optimality Theory (OT), falls into the same class of models. This model requires carefully hand-crafting and ranking a set of constraints. Again, for novel words, proper diacritics have to be assigned to the underlying forms in the lexicon before the model can be made to work.

The model of Belth et al. (2021) is an instance of a statistical classifier. It makes use of recursive partitioning, with as response variable the set of morphological changes required to transform a singular into a plural, and as predictors the final segments of the lexeme, number, and case. At each node, nouns are divided by their features, with one branch including the most frequent plural ending with those features (which will inevitably include some nouns with a different plural ending, which are labelled as exceptions), the other branch including the remainder of the nouns. Each leaf node of the resulting tree is said to be productive if a criterion for node homogeneity is met. Node homogeneity is determined by applying a tolerance principle, such that leaf nodes with a smaller number of noun types can tolerate a higher number of minority plural endings compared to leaf nodes with larger numbers of types. An older model that also is a classifier, was developed 20 years earlier by Hahn and Nakisa (2000).

Connectionist models for the German noun system include a model using a simple recurrent network (Goebel and Indefrey, 2000), and a deep learning model implementing a sequence-to-sequence encoder-decoder (McCurdy et al., 2020). The latter model takes letter-based representations of German nouns in their singular form as input, together with information on the grammatical gender of the noun. The model is given the task to produce the corresponding plural form. The model

learned the task with high accuracy on held out data (close to 90%), but was more locked in on the 'correct' forms compared to native speakers, who in a wug task showed substantially more variability in their choices.

This short overview of the computational models for the German noun system clearly illustrates the marked difference between linguistically insightful models, such as the DATR model of Cahill and Gazdar (1999) that require careful hand-crafting, and black boxes such as sequence-to-sequence deep learning (McCurdy et al., 2020). The deep learning models show generalization to novel nouns, which is not possible with the DATR model without further complementary algorithms that assign inflectional class probabilities to novel forms. In fact, especially for paradigms much richer than those of German, a speaker needs to have encountered all principal parts (the minimal subset of forms one needs to know in order to predict all other forms in a paradigm) for successful generalization across the paradigm (Finkel and Stump, 2007). For German, for instance, given a dative plural with the exponent *-en*, it is impossible to decide whether a word belongs to the masculin II class (*Menschen*) or the masculin I class (*Freunden*). Thus, evaluating performance on held-out data is not straightforward, but can in principle be implemented also for models based on the DATR language.

Interestingly, both DATR-based models and deep learning models may perform better than native speakers. The deep learning model of McCurdy et al. (2020) is an example of a morphological artificial intelligence that provides more focused predictions than those available to human learners.

It is against this background that the LDL model comes into its own. This model is mathematically highly constrained: it implements multivariate multiple linear regression, and hence it cannot handle non-linearities that even shallow connectionist models (Goldsmith and O'Brien, 2006) can take in their stride. Although it is widely believed that nonlinearities are ubiquitous, our hypothesis is that morphological systems are by and large linear in nature, given appropriate representations for form and meaning. We do not commit ourselves to the position that morphological systems are completely linear, and hence cases where model predictions are less precise under linearity can be seen as indicative of learning bottlenecks. In short, LDL is developed as a model of human lexical processing, with all its limitations and constraints, rather than as an optimized computational system for generating (or understanding) morphologically complex words.

By applying LDL to the modeling of the German noun system (including its case forms), we also address a question that has thus far not been addressed computationally, namely the incorporation of semantics. Semantic subregularities in the German noun system have been noted by several authors (e.g. Wunderlich, 1999; Gaeta, 2008), and although deep learning models can be set up that incorporate semantics (see, e.g., Malouf, 2017), LDL by design must take semantics into account.

In what follows, we first introduce the LDL model in more detail, and then proceed with an overview of the many modeling decisions that have to be made, even for this model that implements the most simple network mathematically possible. An important part of this overview is devoted to moving beyond the modeling of isolated words, as words come into their own only in context (Elman, 2009), and case labels do not correspond to contentful semantics, but instead are summary devices for syntactic distribution classes (Blevins, 2016; Baayen et al., 2019).

3 Linear Discriminative Learning

Linear Discriminative Learning (LDL) is the computational engine of the discriminative lexicon model (DLM) proposed by Baayen et al. (2019). The DLM implements mappings between form and meaning for both reading and listening, and mappings from meaning to form for production. It also allows for multiple routes operating in parallel. For reading in English, for instance, it sets up a direct route from form to meaning, in combination with an indirect route from visual input to a phonological representation that in turn is mapped onto the semantics (cf. Coltheart et al., 1993). In what follows, we restrict ourselves to the mappings from form onto meaning (comprehension) and from meaning onto form (production). Both mappings are set up with Linear Discriminative Learning. Mappings can be obtained either with trial-to-trial learning, or by estimating the endstate of learning. In the former case, the model implements incremental regression using the learning rule of Widrow and Hoff (1960), in the latter case, it implements multivariate multiple linear regression, which is mathematically equivalent to a simple network with input units, output units, no hidden layers, and simple summation of incoming activation without using thresholding or squashing functions.

Each word form of interest is represented by a set of cues. For example, *wordform1* might feature the cues `cue1`, `cue2` and `cue3`, while *wordform2* could be marked by `cue1`, `cue4` and `cue5`. We can thus express a word form as a binary vector, where 1 denotes the presence and 0 the absence of a particular cue. This information is coded in the cue matrix **C**:

$$\mathbf{C} = \begin{array}{c} \\ \text{wordform1} \\ \text{wordform2} \end{array} \begin{array}{c} \begin{array}{ccccc} \texttt{cue1} & \texttt{cue2} & \texttt{cue3} & \texttt{cue4} & \texttt{cue5} \end{array} \\ \begin{pmatrix} 1 & 1 & 1 & 0 & 0 \\ 1 & 0 & 0 & 1 & 1 \end{pmatrix} \end{array}$$

Next, we need to decide on how to represent words' meanings. Here, we have to choose between discrete semantic outcomes, as in Naive Discriminative Learning (NDL) (Baayen et al., 2011), and continuous outcomes (LDL). Focussing on LDL, the semantic outcomes can again be represented by a vector, where each entry denotes the strength of a certain semantic feature. Semantic features can either have a concrete meaning or they can be 'latent', abstract, dimensions (see Section 4.2 below). In the following example, *wordform1* has strong negative support for semantic features `S3` and `S5`, while *wordform2* has strong positive support for `S4` and `S5`. This information is brought together in a semantic matrix **S**:

$$\mathbf{S} = \begin{array}{c} \\ \text{wordform1} \\ \text{wordform2} \end{array} \begin{array}{c} \begin{array}{ccccc} \texttt{S1} & \texttt{S2} & \texttt{S3} & \texttt{S4} & \texttt{S5} \end{array} \\ \begin{pmatrix} 0.1 & 0.004 & -1.95 & 0.03 & -0.54 \\ -0.49 & -0.32 & 0.03 & 1.06 & 0.98 \end{pmatrix} \end{array}$$

Comprehension and production in LDL are modelled by means of simple linear mappings from the form matrix $\boldsymbol{C}$ to the semantic matrix **S**, and vice versa. The mappings specify how strongly input nodes are associated with output nodes. The weight matrix for a given mapping can be obtained in two ways. First, using the mathematics of multivariate multiple regression, a compre-

hension weight matrix $\mathbf{F}$ is obtained by solving

$$\mathbf{S} = \mathbf{C} \cdot \mathbf{F},$$

and a production weight matrix $\mathbf{G}$ is obtained by solving

$$\mathbf{C} = \mathbf{S} \cdot \mathbf{G}.$$

As for linear regression modeling, the predicted row vectors are approximate, and borrowing notation from statistics, we write

$$\hat{\mathbf{S}} = \mathbf{C} \cdot \mathbf{F}$$

for predicted semantic vectors (row vectors of $\hat{\boldsymbol{S}}$), and

$$\hat{\mathbf{C}} = \mathbf{S} \cdot \mathbf{G}$$

for predicted form vectors (row vectors of $\hat{\mathbf{C}}$).

Estimating the mappings $\mathbf{F}$ and $\mathbf{G}$ using the matrix algebra of multivariate multiple regression provides optimal estimates, in the least squares sense, of the connection weights (or equivalently, beta coefficients) for datasets that are type-based, in the sense that each pair of row vectors $\boldsymbol{c}$ of $\boldsymbol{C}$ and $\boldsymbol{s}$ of $\boldsymbol{S}$ is unique. Having multiple instances of the same pair of row vectors in the dataset does not make sense, as it renders the input completely singular and does not add any further information. Thus, models based on the regression estimates of $\mathbf{F}$ and $\mathbf{G}$ are comparable to type-based models such as AML, MBL, MGL, and models using recursive partitioning.

In order to make the estimates of the mappings sensitive to frequency of use, the weight matrices have to be estimated using incremental learning, which updates weights after each word token that is presented for learning. Incremental learning is implemented using the learning rule of Widrow and Hoff (1960), which defines the matrix $\boldsymbol{W}^{t+1}$ with updated weights at time $t+1$ as the weight matrix $\boldsymbol{W}^{t}$ at time t, modified as follows:

$$\boldsymbol{W}^{t+1} = \boldsymbol{W}^{t} + \boldsymbol{c} \cdot (\boldsymbol{o}^{T} - \boldsymbol{c}^{T} \cdot \boldsymbol{W}^{t}) \cdot \eta,$$

where $\boldsymbol{c}$ is the current cue (vector), $\boldsymbol{o}$ the current outcome vector, and η the learning rate. Conceptually, this means that after each newly encountered word token, the weight matrix is changed such that the next time that the same cue vector has to be mapped onto its associated outcome vector, it will be slightly closer to the target outcome vector than it was before. Details on the Widrow-Hoff formula and its applications in language sciences can be found in Milin et al. (2020), an example of its use in the context of the DLM is given in Chuang et al. (2020a). The learning rule of Widrow-Hoff implements incremental regression. As the number of times that a model is trained again and again on a training set increases (training epochs), the network's weights will converge to the matrix of beta coefficients obtained by approaching the estimation problem with multivariate multiple regression (see, e.g. Shafaei-Bajestan et al., 2021). As a consequence, the regression-based estimates pertain to the 'endstate of learning', at which the data have been worked through infinitely many times. Unsurprisingly, effects of frequency and order of learning are not reflected in model predictions based on the regression estimates. Such effects do emerge with incremental learning, as we will demonstrate in Section 4.5.

This completes the model specification for comprehension. Model accuracy for a given word ω is assessed by comparing its predicted semantic vector $\hat{s}_{\omega}$ with all gold standard semantic vectors in

$\boldsymbol{S}$, using either the cosine similarity measure or the Pearson correlation measure. In what follows, we use the correlation measure, and select as the meaning that is recognized that gold standard row vector $\boldsymbol{s}_{\max}$ of $\boldsymbol{S}$ that shows the highest correlation with $\hat{s}_\omega$. If $\boldsymbol{s}_{\max}$ is the targeted semantic vector, the model's prediction is classified as correct, otherwise, it is taken to be incorrect.

For the modeling of production, a supplementary algorithm is required for constructing actual word forms. The predicted vectors $\hat{\boldsymbol{c}}$ provide information about the amount of support that cues receive from the semantics. However, information about the amount of support received by the full set of cues does not provide any information about the order in which a small subset of these cues have to be woven together into actual words. The problem can be conceptualized using graph theory, by taking cues to be the vertices of a graph. The question then amounts to finding a proper path in the graph that represents a word's form. The algorithms that are available for setting up such paths all build on the insight that when form cues are defined as n-grams ($n > 1$), the cues contain implicit information about order. For instance, for digraph cues, cues `ab` and `bc` can be combined into the string `abc`, but cues `ab` and `cd` cannot be merged. Therefore, when n-grams are used as cues, directed edges can be set up in the graph for all vertices with the proper partial overlap. By distinguishing between initial n-grams (starting with an initial word edge symbol, typically a # is used) and final n-grams (ending with #), a word is uniquely defined by a path in the graph from an initial to a final n-gram. This raises the question of how to find a word's path. The core idea is straightforward: first discard n-grams with low support from the semantics below a threshold θ, then calculate all possible remaining paths, and select for articulation that path for which the corresponding predicted semantic vector (obtained by mapping its corresponding cue vector $\boldsymbol{c}$ onto $\boldsymbol{s}$ using comprehension matrix $\boldsymbol{F}$) best matches the semantic vector that is the target for articulation. This approach is described as 'synthesis by analysis', see Baayen et al. (2019) and Baayen et al. (2018) for further details and theoretical motivation.

The first algorithm that was used to enumerate possible paths made use of a shortest-paths algorithm from graph theory. This works well for small datasets, but becomes prohibitively expensive for large datasets. The **JudiLing** package (Luo et al., 2021) offers a new algorithm that scales up much better. This algorithm is first trained to predict, from either the $\hat{\boldsymbol{C}}$ or the $\boldsymbol{S}$ matrix, for each possible position in the word, which cues are best supported at that position. All possible paths with the top k best-supported cues are then calculated, and subjected to synthesis by analysis. Details about this algorithm, implemented in julia in the **JudiLing** package as the function `learn_paths` can be found in (Luo, 2021). The `learn_paths` function is used throughout the remainder of the present study. A word form is judged to be produced correctly when it exactly matches the targeted word form.

4 Modelling considerations

When modelling a language's morphology within the framework of the DLM, the analyst is faced with a range of considerations and choices. Figure 1 provides an overview of the most important choice points. From left to right, choices are listed for representing form, for the unit of analysis, for the representation of semantics, for the handling of context, and for the learning regime. With respect to form representations, we need to decide on what kind of n-grams to use (setting n, defining the kind of grams to use, and deciding on how to model stress or lexical tone). With respect to the unit of analysis, the analyst has to decide whether to model isolated words, or words in phrasal contexts. A third set of choices concerns what semantic representations to use: simulated

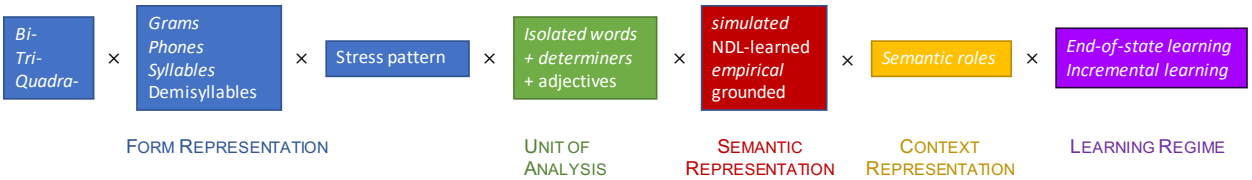

Figure 1: Options when modelling a language's morphology with LDL. Examples with options in italics are discussed in the present study.

word form	pronunciation	lemma	case	number	frequency	gender
Aal	`al`	Aal	nominative	singular	29	m
Aal	`al`	Aal	dative	singular	29	m
Aal	`al`	Aal	accusative	singular	29	m
Aale	`al@`	Aal	nominative	plural	34	m
Aale	`al@`	Aal	genitive	plural	34	m
Aalen	`al@n`	Aal	dative	plural	17	m
Aalen	`al@n`	Aal	accusative	plural	17	m

Table 3: Representation of the paradigm for *Aal* 'eel' in our dataset. Genitive singular (*Aals*) is not included as it does not appear in CELEX.

representations, or word embeddings such as *word2vec* (Mikolov et al., 2013b), or grounded vectors (Shahmohammadi et al., 2021). A further set of choices for languages with case concerns how to handle case labels, as these typically refer to syntactic distribution classes rather than contentful inflectional features (Blevins, 2016). Finally, a selection needs to be made with respect to whether incremental learning is used, or instead the endstate of learning using regression-based estimation. In what follows, we consider several of these choice points using examples addressing the German noun system, and discuss their advantages and drawbacks.

The dataset on German noun inflection that we use for our worked examples was compiled as follows. First, we extracted about 6,000 word forms from German CELEX (Baayen et al., 1995). Of these we retained the 5,486 word forms for which we could retrieve grammatical gender from Wiktionary, thus including word forms of 2,732 different lemmas. The resulting data was expanded such that each attested word form was listed once for each possible paradigm cell it could belong to. For instance, *Aal* ('eel') would be listed once as singular nominative, once as dative and once as accusative, see Table 3. This resulted in a dataframe with 18,147 entries, with word form frequencies ranging from 1 to 5,828 (M log frequency 2.56, SD 1.77). Word forms are represented in their DISC notation, which represents German phones with single characters[1].

From Table 3 we can immediately notice that there are many homophones, words sharing the same form but differing in meaning. In German, because many of the word forms are not marked for case and number, even though we have a relatively large dataset, the actual number of distinct word forms is only 5,486, which amounts to on average about two word forms per lemma.

There are many ways in which model performance can be evaluated. First, we may be interested in how well the model performs as a memory. How well does the model learn to understand and produce words it has encountered before? Note that because the model is not a list of forms, this

[1]Data and code for this study are available in the supplementary materials at `https://osf.io/zrw2v/`.

is not a trivial question. For evaluation of the model as a memory, we can consider its performance on the training data (henceforth `train`). Second, we may be interested in the extent to which the memory is productive. Does it generalize so that new forms can be understood or produced? Above, we observed that the German noun system is semi-regular, and that German native speakers are unsure about what the proper plural is of words they have not encountered before (McCurdy et al., 2020). If our modeling approach mirrors the human limitations on generalization from data with only partial regularities, evaluation on unseen data should not be perfect. In the light of these considerations, it is important to assess model performance on held-out data. At this point, however, several issues arise that require careful thought.

For one, from the perspective of the linguistic system, it seems unreasonable to assume that any held-out form can be properly produced (or understood) if some of the principal parts (Finkel and Stump, 2007) of the lexeme are missing in the training data. In what follows, we will make the simplifying assumption that under cross-validation with sufficient training data, this situation will not arise.

A further question that arises is how to evaluate held-out words that have homophones in the training data. On the one hand, these homophones present novel combinations of a form vector (shared with another data point in the training data) and a semantic vector (not attested for this form in the training data). We may therefore evaluate comprehension performance under the strict criterion that it should get the semantic vector exactly right. But then, when presented with a homophone, a human listener cannot predict which of a potentially large set of paradigm cells is the targeted one. We may therefore want to use a lenient evaluation criterion for comprehension according to which comprehension is judged to be accurate when the predicted semantic vector $\hat{\boldsymbol{s}}$ is associated with one of a homophonic word's possible semantic interpretations. Yet a further possible evaluation metric is to see how well the model performs on words that have forms that have not been encountered in the training data. These possibilities are summarized in Table 4. Below, in section 4.3.1, we will consider further complications that can arise in the context of testing the model on unseen forms.

Table 4: Types of model evaluation

evaluation type			
simple	blind evaluation of all held-out data		`val_all`
nuanced	evaluation on novel forms only		`val_newform`
	evaluation on homophones	strict	`val_strict`
		lenient	`val_lenient`

For evaluating the productivity of the model, we split the full dataset into 80% training data and 20% validation data, with 14,518 and 3,629 word forms respectively. In the validation data, 3309 forms are also present in the training data, and 320 are new forms. Among the 320 new forms, 8 have novel lemmas that are absent in the training data. Since it is unrealistic to expect the model to understand or produce inflected forms of completely new words, these 8 words are excluded from the validation dataset for new forms, although they are taken into consideration when calculating the overall accuracy for the validation data. The same training and validation data are used for all the simulations reported below, unless indicated otherwise.

4.1 Representing words' forms

Decisions about how to represent words' forms depend on the modality that is to be modelled. For auditory comprehension, Arnold et al. (2017) and Shafaei-Bajestan et al. (2021) explore ways in which features can be derived from the audio signal. Instead of using low-level audio features, one can also use more abstract symbolic representations such as phone n-grams. For visual word recognition, one may use letter n-grams, or, as lower-level visual cues, for instance, features derived from histograms of oriented gradients (Dalal and Triggs, 2005; Linke et al., 2017). In what follows, we use vectors with combinations of phonological units to represent the forms of German nouns. We first consider form representations with n-phones as cues. Next, we will present results for when n-syllables are used as cues.

4.1.1 Phone-based representations

Sublexical phone cues can be of different granularity, such as biphones and triphones. For the word *Aale* (pronunciation `al@`), the biphone cues are `#a`, `al`, `l@`, and `@#`, and the triphone cues are `#al`, `al@`, and `l@#`. The number of unique cues (and hence the dimensionality of the form vectors) increases as granularity decreases. For the present dataset for example, there are 931 biphone cues, but 4,656 triphone cues. For quadraphones, there are no less than 9,068 unique cues. Although model performance tends to become better with more unique cues, we also run the risk of overfitting. That is, the model does not generalize and thus performs worse on validation data. The choice of granularity therefore determines the balance of having a precise memory on the one hand and a productive memory on the other hand. In the simulation examples with n-phones that follow, we made use of simulated semantic vectors. Details on the many different kinds of semantic vectors that can be used are presented in Section 4.2.1.

	comprehension				production			
	train	val_all	val_lenient	val_newform	train	val_all	val_lenient	val_newform
biphone	22%	16%	17%	8%	48%	31%	33%	12%
triphone	93%	88%	92%	51%	84%	64%	68%	21%
quadraphone	97%	93%	97%	53%	91%	67%	73%	11%
bisyllable	99%	93%	99%	20%	95%	63%	69%	0.3%
word2vec	87%	72%	80%	0.3%	97%	88%	94%	25%

Table 5: Comprehension and production accuracy for train and validation datasets, with biphones, triphones, quadraphones, and bisyllables as cues. For the first four rows, we used simulated semantic vectors. For the last row, cues are triphones, and semantic vectors are *word2vec* embeddings (discussed in Section 4.2.2). For the `learn_paths` algorithm, the threshold θ was set to 0.05, 0.008, 0.005, 0.005, and 0.008 respectively.

Model accuracy for n-phones is presented in the first three rows of Table 5. For the training data, comprehension accuracy is high with both triphones and quadraphones. For biphones, the small number of unique cues clearly does not offer sufficient discriminatory power to distinguish word meanings. Under strict evaluation, unsurprisingly given the large number of homophones in German noun paradigms, comprehension accuracy plummets substantially to 8%, 33%, and 35% for biphone, triphone, and quadraphone models respectively. Given that there is no way to tell the meanings of homophones apart without further contextual information, we do not provide further

details of strict evaluation. However, in Section 4.4.1 we will address the problem of homophony by incorporating further contextual information into the model.

With regards to model accuracy for validation data, we see that overall accuracy (`val_all`) is quite low for biphones, while it remains high for both triphones and quadraphones. Closer inspection reveals that this high accuracy is mainly contributed by homophones (`val_lenient`). Since these forms are already present in the training data, a high comprehension accuracy under lenient evaluation is unsurprising. As for unseen forms (i.e., `val_newform`), quadraphones perform slightly better than triphones.

Production accuracy, presented in the right half of Table 5, is highly sensitive to the threshold θ used by the `learn_paths` algorithm. Given that usually only a relatively small number of cues receive strong support from a given meaning, we therefore set the threshold such that the algorithm does not need to take into account large numbers of irrelevant cues. Depending on the form and meaning representations selected, some fine-tuning is generally required to obtain a threshold value that optimally balances both accuracy and computation time. That is, we aim for the best accuracy that the algorithm can achieve within a reasonable time span. Once the threshold is fine-tuned for the training data, the same threshold is used for the validation data.

Production accuracy is similar to comprehension accuracy, albeit systematically slightly lower. Triphones and quadraphones again outperform biphones by a large margin. For the training data, triphones are somewhat less accurate than quadraphones. Interestingly, in order to predict new forms in the validation data, triphones outperform quadraphones. Clearly, triphones offer better generalizability compared to quadraphones, suggesting that we are overfitting when modeling with quadraphones as cues. Accuracy under the `val_newform` criterion is quite low, which is perhaps not unexpected given the uncertainty that characterizes native speakers' intuitions about the forms of novel words (McCurdy et al., 2020). In Section 4.3.2 we return to this low accuracy, and consider in further detail the best supported top candidates.

4.1.2 Syllable-based representations

Instead of using n-phones, the unit of analysis can be a combination of n syllables. The motivation for using syllables is that some suprasegmental features, such as lexical stress in German, are bound to syllables. Although stress information is not considered in the current simulation experiments, suprasegmental cues can incorporated (see Chuang et al., 2020a, for an implementation).

As for n-phones, when using n-syllables, we have to choose a value for the unit size n. For the word *Aale*, the bi-syllable cues are `#-a`, `a-l@`, and `l@-#`, with "-" indicating syllable boundary. When unit size equals two, there are in total 8,401 unique bi-syllable cues. For tri-syllables, the total number of unique cues triples increases to 10,482. Above, we observed that the model was already overfitting with 9,068 unique quadraphone cues. We therefore do not consider tri-syllable cues, and only present modeling results for bi-syllable cues.

As shown in the fourth row of Table 5, comprehension accuracy for the training data is almost error-free, 99%, the highest among all the cue representations. For the validation data, the overall accuracy is also high, 93%. This is again due to the high accuracy for the seen forms (`val_lenient` = 99%). Less than a quarter of the unseen forms, however, is recognized successfully (`val_newform` = 23%). As for production, accuracies for the training and validation data are 95% and 63% respectively. The model again performs well for homophones (`val_lenient` = 69%) but fails to produce unseen forms (`val_newform` = 0.3%). This extremely low accuracy is in part due to the large number of cues that appear only in the validation dataset (325 for bisyllables, but only 23 for

triphones). Since such novel cues do not receive any training, words with such cues are less likely to be produced correctly. We will come back to the issue of novel cues in Section 4.3. For now, we conclude that triphone-based form vectors are a good choice.

4.2 Semantic representation

There are many ways in which words' meanings can be represented numerically. The simplest method is to use one-hot encoding, as implemented in NDL (Baayen et al., 2011). One-hot encoding, however, misses out on the semantic similarities between lemmas: under one-hot encoding, all lemmas have meaning representations that are orthogonal. Instead of using one-hot encoding, binary vectors with multiple bits on can be derived from WordNet (Chuang et al., 2020a). In what follows, however, we will work with real-valued vectors, known as 'word embeddings' in Natural Language Processing. In the present study, we refer to word embeddings as semantic vectors. Semantic vectors can either be simulated, or derived from corpora using methods from distributional semantics (see, e.g. Landauer and Dumais, 1997; Mikolov et al., 2013b).

4.2.1 Simulated semantic vectors

When corpus-based semantic vectors are unavailable, semantic vectors can be simulated. The **JudiLing** package enables the user to simulate such vectors using normally distributed random numbers for content lexemes and inflectional functions. By default, the dimension of the semantic vectors is set to be identical to the dimension of the form vectors. Thus, the dimension of the semantic vectors was smallest for the simulation using biphones (931), followed by that using triphones (4,656), and largest for that using quadraphones (9,068).

The semantic vector for an inflected word is obtained by summing the vector of its lexeme and the vectors of all the pertinent inflectional functions. As a consequence, all vectors sharing a certain inflectional feature are shifted in the same direction in semantic space. By way of example, consider the German plural genitive of *Aal* 'eel', *Aale*. We compute its semantic vectors by adding the semantic vector for PLURAL and GENITIVE to the lemma vector $\overrightarrow{Aal}$:

$$\overrightarrow{Aale} = \overrightarrow{Aal} + \overrightarrow{\text{PLURAL}} + \overrightarrow{\text{GENITIVE}}$$

The corresponding singular can be coded as:

$$\overrightarrow{Aals} = \overrightarrow{Aal} + \overrightarrow{\text{SINGULAR}} + \overrightarrow{\text{GENITIVE}}$$

Alternatively, the singular form could be coded as unmarked, following a privative opposition approach:

$$\overrightarrow{Aals} = \overrightarrow{Aal} + \overrightarrow{\text{GENITIVE}}$$

For the remainder of the paper, we treat *number* as equipollent opposition (the former approach). Finally, a small amount of random noise is added to each semantic vector, as an approximation of further semantic differences in word use other than number and case (see Sinclair, 1991; Tognini-Bonelli, 2001, and further discussion below). The results reported above in Table 5 were all obtained with simulated vectors.

It is worth noting that when working with simulated semantic vectors, the meanings of lexemes will still be orthogonal, and that as a consequence, all similarities between semantic vectors originate exclusively from the semantic structure that comes from the inflectional system.

4.2.2 Empirical semantic vectors

A second possibility for obtaining semantic vectors is to derive them from corpora. Baayen et al. (2019) constructed semantic vectors from the TASA corpus, in such a way that semantic vectors were obtained not only for lexemes but also for inflectional functions. With their semantic vectors, the semantic vector of *Aale* can be straightforwardly constructed from the semantic vectors of *Aal*, PLURAL, and GENITIVE.

However, semantic vectors that are created with standard methods from machine learning, such as *word2vec* (Mikolov et al., 2013a), *fasttext* (Bojanowski et al., 2017) or *GloVe* (Pennington et al., 2014), can also be used. In what follows, we illustrate this for 300-dimensional vectors generated with *word2vec*, trained on the German Wikipedia (Yamada et al., 2020). For representing words' forms, we used triphones.

Results are presented in the last row of Table 5. The model in general performs well for the training data. For the validation data, while the homophones are easy to recognize and produce, the unseen forms are again prohibitively difficult. Interestingly, if we compare the current results with the results of simulated vectors (cf. second row, Table 5), we observe that while the `train` and `val_all` accuracies are fairly comparable for the two vector types, their `val_newform` accuracies nonetheless differ. Specifically, understanding new forms is substantially more accurate with simulated vectors (51% *vs.* 0.3%), whereas *word2vec* embeddings yield slightly better results for producing new forms (21% *vs.* 25%).

To understand why these differences arise, we note, first, that lexemes are more similar to each other than is the case for simulated vectors (in which case lexemes are orthogonal), and second, that *word2vec* semantic vectors are exactly the same for each set of homophones within a paradigm, so that inflectional structure is much less precisely represented. The lack of inflectional structure may underlie the inability of the model to understand novel inflected forms correctly. Furthermore, the lack of differentiation between homophones simplifies the mapping from meaning to form, leading to more support from the semantics for the relevant triphones, which in turn facilitates synthesis by analysis.

To better understand the difference between simulated vectors and *word2vec* semantic vectors, we took the *word2vec* vectors, and reconstructed from these vectors the vectors of the lexemes and of the inflectional functions. For a given lexeme, we created its lexeme vector by averaging over the vectors of its inflectional variants. For plurality, we averaged over all vectors of forms that can be plural forms. Using these new vectors, we constructed semantic vectors for a given paradigm cell by adding the semantic vector of the lexeme and the semantic vectors for its number and case values. The mean correlation between the new "analytical" *word2vec* vectors and the original empirical vectors was 0.79 ($sd = 0.076$). It follows that there is considerable variability in how German word forms are actually used in texts, a finding that has also emerged from corpus linguistics (Sinclair, 1991; Tognini-Bonelli, 2001). The idiosyncracies in the use of individual inflected forms renders the comprehension of a novel, but nevertheless idiosyncratic, word form difficult if not impossible. From this we conclude that the small amount of noise that we added to the simulated semantic vectors is likely to be unrealistically small compared to real language use.

Interestingly, semantic similarity may facilitate the production of unseen forms. A Linear Discriminant Analysis (LDA) predicting nine plural classes (the eight sub-classes presented in Table 1 plus one 'other' class) from the *word2vec* semantic vectors has a prediction accuracy of 62.7% (50.5% under leave-one-out cross validation). Conducting 10-fold cross-validation with Support Vector Machine (SVM) gives us an average accuracy of 57.2%, which is significantly higher than

the percentage of majority choice (35.6% for the -n plural class). This indicates that semantically similar words do tend to inflect in similar ways. When a novel meaning is encountered in the validation set, it is therefore possible to predict to some extent its general form class. Given the similarities between LDA and regression, it seems likely that the same kind of information is captured by the mapping from meaning to form in LDL.

4.3 Missing forms and missing semantics

Evaluation on held-out data is a means for assessing the productivity of the network. However, it often happens during testing that the model is confronted with novel, unseen cues, or with novel, unseen semantics. Here, linguistically and cognitively motivated choices are required.

4.3.1 Novel cues

For the cross-validation results presented thus far, the validation data comprise a random selection of words. As a consequence, there often are novel cues in the validation data that the model has never encountered during training. The presence of novel cues is especially harmful for production. As mentioned in Section 4.1.2, the model with bi-syllables as cues fails to produce unseen forms, due to the large number of novel cues in the validation data.

What is the theoretical status of novel cues? To answer this question, first consider that actual speakers rarely encounter new phones or new phone combinations in their native languages. Furthermore, novel sounds encountered in loan words are typically assimilated into the speaker's native phonology. Second, many cues that are novel for the model actually occur not only in the held-out nouns, but also in verbs, adjectives, and compounds. Thus, the presence of novel cues is in part a consequence of modeling only part of the German lexicon.

Since novel cues have zero weights on their efferrent connections (or, equivalently, zero beta coefficients), they are completely inert for prediction. One way to address this issue is to select the held-out data with care. That is, instead of randomly holding out words, we make sure that in the validation data all cues are already present in the training data. This is a linguistically more interesting, and statistically more sensible, alternative for evaluating a model's productivity.

As before, we split the dataset into 80% training and 20% validation data, now making sure that there are no novel triphone cues for the validation dataset. Among the 3629 validation words, 3331 are homophones, and 298 are unseen forms. Changing the kind of cues used typically has consequences for how many datapoints can be held out for validation. For instance, when bisyllables are used instead of triphones, due to the sparsity of bisyllable cues, we have to increase the percentage of validation data to include sufficient numbers of unseen forms. Even for 65% training data and 35% validation data, we still have that the majority of validation data are homophones (98.5%), and only 76 cases represent unseen forms (but with known cues).

	comprehension				production			
	train	val_all	val_lenient	val_newform	train	val_all	val_lenient	val_newform
triphone	92%	88%	91%	53%	85%	62%	67%	14%
bisyllable	99%	99%	99%	61%	95%	52%	52%	14%

Table 6: Comprehension and production accuracy for train and validation datasets, which are split in such a way that no novel cues are present in the validation set. Both the triphone and bisyllable models make use of simulated semantic vectors.

For the triphone model (top row, Table 6), for both comprehension and production, the `train`, `val_all` and `val_lenient` accuracies are similar to the results presented previously (Table 5). For the evaluation of unseen forms (`val_newform`), there is a slight improvement for comprehension (from 51% to 53%), for other datasets, the improvement can be larger. However, for production, `val_newform` becomes worse after we make sure that there are no novel cues in the validation data (from 21% to 14%). The reason is that even though all triphone cues of the validation words are present in the training data, they obtain insufficient support from the semantics. The solution is to allow a small number of triphone cues with weak support (below the threshold θ) to be taken into account by the algorithm that orders triphones into words. This requires turning on the `tolerance` mode on in the `learn_paths` function of the **JudiLing** package). By allowing at most two weakly supported triphones to be taken into account, production accuracy for unseen forms increases to 56%.

The bi-syllable model, on the other hand, benefits more from the removal of novel cues in the validation data. Especially for comprehension, the accuracy of unseen forms reaches 61% (compared to 20% with random selection). For production, we observe a non-negligible improvement as well (from 0.3% to 14%). Further improvements are expected when tolerance mode is used (but given the large number of bisyllables, this comes at considerable computation costs). In other words, bisyllables provide a model that is an excellent memory, but a memory with very limited productivity specifically for production.

4.3.2 Unseen semantics

In real language, speakers seldomly encounter words that are completely devoid of meaning: even novel words are typically encountered in contexts which narrow down their possible meanings. In the wug task, by contrast, participants are often confronted with novel words presented without any indication of their meaning, as, for instance, in the experiment on German nouns reported by McCurdy et al. (2020). Within the framework of the discriminative lexicon, this raises the question of how to model the semantics of nonwords, as without a semantic representation for a nonword, the model has no way to produce inflected variants.

In order to model the wug task, and compare our model's performance with that of German native speakers, we take as starting point the observation that the comprehension system generates meanings for nonwords. Chuang et al. (2020b) showed that measures derived from the semantic vectors of nonwords were predictive for both reaction times in an auditory lexical decision task and for nonwords' acoustic durations in a reading task. In order to model the wug task, we therefore proceeded as follows:

1. We first simulated a speaker's lexical knowledge prior to the experiment by training a comprehension matrix using all the words described in Section 4. In what follows, we made use of simulated semantic vectors.

2. We then used the resulting comprehension network to obtain semantic vectors $\boldsymbol{s}_{\text{nom.sg}}$ for the nominative singular forms of the nonwords by mapping their cue vectors into the semantic space, resulting in semantic vectors $\boldsymbol{s}_{\text{nom.sg}}$.

3. Next, we created the production mapping from meaning to form, using not only all real words but also the nonwords (known only in their nominative singular form).

4. Then, we created the semantic vectors for the plurals ($\boldsymbol{s}_{nom.pl}$) of the nonwords by adding the plural vector to their nominative singular vectors after subtracting the singular vector.

5. Finally, these plural semantic vectors were mapped onto form vectors ($\hat{\boldsymbol{c}}_{\text{nom.pl}}$) using the production matrix, in combination with the `learn_paths` algorithm that orders the triphones for articulation.

We applied these modeling steps to a subset of the experimental materials provided by Marcus et al. (1995) (reused by McCurdy et al., 2020), in order to compare the predictions of our model with those reported by McCurdy et al. (2020). The full materials of Marcus et al. (1995) contained nonwords that were set up such that only half of them had an existing rhyme in German. We restricted ourselves to the nonwords with existing rhymes, first, because non-rhyme words have many cues that are not in the training data; second, because, as noted by Zaretsky and Lange (2015), many of the non-rhyme words have unusual orthography and thus are strange even for German speakers, and third, because many of the non-rhyme nonwords share their endings and therefore do not provide strong data for testing model predictions.

McCurdy et al. (2020) presented nonwords visually and asked participants to provide the plural form in writing. In what follows, we therefore made use of letter trigrams rather than triphones. We represented words without their articles as the wug task implemented by McCurdy et al. (2020) presented the plural article as a prompt for the plural form, so that participants only produced the plural form without the article. In order to assess what forms are potential candidates for production, we examined the set of candidate forms, ranked by how well their internally projected meanings (obtained with the synthesis-by-analysis algorithm, see Section 3), correlated with the meaning $\boldsymbol{s}_{\text{nom.pl}}$ targeted for production. We then examined the top best candidates as possible alternative plural forms.

The model provided a plausible plural form as the best candidate in 7 out of 12 cases. Five of these belonged to the *-en* class. A further plausible candidate was also only provided in 5 of the cases. The lack of diversity as well as the bias for *-en* plurals does not correspond to the responses given by German speakers in McCurdy et al. (2020).

Upon closer inspection, it turns out that a more variegated wug performance can be obtained by changing two parameters. First, we replaced letter trigrams by letter bigrams. This substantially reduces the number of n-grams that are present in the nonwords, but that do not occur in the training data. Second, we made a small but important change to how semantic vectors were simulated. The default parameter settings provided with the *JudiLing* package generate semantic vectors with the same standard deviation for both content words and inflectional features. Therefore, the magnitudes of the values in the semantic vectors is very similar for content words and inflectional features. Since words are inflected for case and number, their semantic vectors are numerically dominated by the inflectional meanings. To enhance the importance of the lexeme, and to reduce the dominance of the inflectional functions, we reduced the standard deviation when generating the semantic vectors for number and case. As a consequence, the mean of the absolute values in the plural vector decreased from 3.25 to 0.32. (Technical details are provided in the supplementary materials.)

With these two changes, the model generated a more diverse set of plural nonword candidates, as shown in Table 7. Model performance is now much closer to the performance of native speakers, as reported by (Zaretsky et al., 2013; McCurdy et al., 2020).

The model also produces some implausible plural candidates, all of which are phonotactically legal; these are marked with an asterisk in Table 7. Sometimes a plural marker is interfixed instead

Bral	Kach	Klot	Mur	Nuhl	Pind	Pisch	Pund	Raun	Spand	Spert	Vag
Bralen	Kachen	Klot	Muren	Nuhlen	Pinden	Pischen	Punden	Raunen	*Spanend	Sperten	Vag
Bral	Kach	*Klotten	Murn	Nuhl	Pind	Pisch	*Punend	Raun	Spand	Sperte	Vagen
*Bralenen	Kacher	*Klotte	Mur	Nuhle	Pinder	Pischer	Pund	*Raunern	*Spanende	Sperter	Vage
*Bralern	Kache	*Klotter	*Murnen	*Nuhlern	Pinde	Pische	Punde	Rauner	*Spanenden	*Spererten	Vager
Braler	*Kachern	*Klieloten	Murer	*Nuhlere	*Pindern	*Pischern	*Pundene	Raune	*Spatend	*Spererte	*Vagern

Table 7: First five candidates for the plural forms of nonwords. Forms that are implausible as plurals are marked with an asterisk.

of suffixed (e.g., Spand, Span-en-d; Pund, Pun-en-d), almost all words have a candidate which shows double plural marking (e.g. Bral, Bral-en-en; Nuhl, Nuhl-er-e; cf. Dutch kind-er-en; Pind, Pind-er-n), or a mixture of both (e.g. Span, Span-en-d-e; Spert, Sper-er-t-en). For *Klot*, doubling of the *-t* can be observed, as this form is presumably more plausible in German (e.g. Motte ('moth'), Gott ('god'), Schrott ('scrap, rubbish')) and one plural has been attracted to an existing singular (Spand, Spaten-d). Apparently, by downgrading the strength (or more precisely, the L1-norm) of the semantic vectors of inflectional functions, the model moves in the direction of interfixation-like changes.

Interestingly, the model does not produce a single plural form with an umlaut. This is surprising in the sense that in corpora, umlauted plurals are relatively frequent (see e.g. Gaeta, 2008). The model's performance may simply reflect that pertinent bigrams are 'still' missing, but then this processing limitation does seem to reflect the performance of German speakers: The German speakers in McCurdy (2019) also tended to avoid umlaut forms with as exception *Kach* → *Kächer*).

Finally, it is noteworthy to see that most nonwords have a plural in *-en* as one of the candidates (10 out of 12 cases), with as runners-up the *-e* plural (8 out of 12 cases), and the *-er* plural (8 out of 12). There is not a single instance of an *-s* plural, which fits well with the low prevalence (around 5%) of *-s* plurals in the experiment of McCurdy et al. (2020).

In summary, this simulation study shows that it is possible to make considerable headway with respect to modeling the wug task for German. The model is not perfect, unsurprisingly, given that we have worked with simulated semantic vectors and estimates of nonwords' meanings. Furthermore, the strong weight imposed on the stem shifts model performance in the direction of interfixation-like morphology. Last but not least, the model has no access to information about words' frequency of use, and hence is blind to an important factor shaping human learning (see Section 4.5 for further discussion). Nevertheless, the model does mirror the uncertainties of German speakers fairly well.

4.4 Words in context

Thus far, we have modeled words in isolation. However, in German, case and number information is to a large extent carried by preceding determiners. In addition, in actual language use, a given grammatical case denotes one of a wide range of different possible semantic roles. In other words, the simplifying assumption that an inflectional function can be represented by a single vector, which may be reasonable for grammatical number, is not at all justified for grammatical case. In this section, we therefore explore how context can be taken into account. In what follows, we first present modeling results of nouns learned together with their articles. Next, we break down grammatical cases into actual semantic functions, and show how we can begin to model the noun declension system with more informed semantic representations.

4.4.1 Articles

We first consider definite articles. Depending on gender and case, a noun can follow one of the six definite articles in German — der, die, das, dem, den, des. These articles, transcribed in DISC notation, are added before the nouns. Although in writing articles and nouns are separated by space (e.g. `der Aal`), to model auditory comprehension we remove the space (e.g., `deral`). By adding the articles to the noun forms, the number of homophones in our dataset is reduced to a substantial extent, and the number of unique word forms now more than doubles (from 5,427 to 12,798).

In the first set of simulations we used the same semantic vectors as we did previously for modeling isolated words. That is, the meanings of the definite articles are not taken into account in the semantic vectors, as all forms would be shifted in semantic space in the same way. After including articles, the validation data now only contained 3,982 homophones, but the number of unseen forms increased to 3,260. Using triphones as cues, we ran two models, one with simulated vectors and the other with *word2vec* semantic vectors. As shown in Table 8, for simulated vectors the results are generally similar to those obtained without articles (Table 5). However, if we look at the evaluation of comprehension with the strict criterion (according to which recognizing a homophonic meaning is considered incorrect), without articles `val_strict` is 0.6%, whereas it is 30% with articles. The generalizability of the model also improves as the number of homophones in the dataset decreases. Even though there are more unseen forms in the current dataset with articles than in the original one without articles, the `val_newform` for comprehension still increases by 12% from 51% to 63%.

With respect to *word2vec* embeddings, the addition of articles in form representations also benefitted the comprehension of unseen forms: the `val_newform` astonishingly increases from 0.3% to 58%. This is because previously homophones all shared the same form representations and exactly the same *word2vec* vectors. Many triphone cues are superfluous in the sense that they cannot serve as good predictors for lemma or inflectional meanings. Now, with the addition of articles, the form space is better discriminated. Given that the number of predictors (triphone cues) has increased, the model is now able to predict and generalize more accurately for comprehension. However, for production, model performance is generally worse when articles have to be produced. For the training data, for instance, production accuracy drops from 97% (without articles) to 48%. This is again unsurprising. In the simulation with articles, the semantic representations remain the same, but now these semantic vectors have to predict more variegated triphone vectors. The learning task has become more challenging, and inevitably results in less accurate performance. Replacing the contextually unaware *word2vec* vectors by contextually aware vectors obtained using language models such as BERT (Devlin et al., 2018; Miaschi and Dell'Orletta, 2020) should alleviate this problem.

We can test the model on more challenging data by including indefinite articles (ein, eine, einem, einen, einer, eines), and creating two additional semantic vectors, one for definiteness and one for indefiniteness. This doubles the size of our dataset: half of the words are preceded by definite articles, and the other half by indefinite articles. However, because German indefinite articles are restricted to singular forms, only indefinite singular forms are preceded by indefinite articles. On the meaning side, the $\overrightarrow{\text{DEFINITE}}$ vector is added to the semantic vectors of words preceded by definite articles, and the $\overrightarrow{\text{INDEFINITE}}$ vector is added to those of words preceded by either indefinite articles in the singular, or no article in the plural.

The validation data of this dataset is faced with in total 3,982 homophones and 3,260 unseen

	comprehension				production			
	train	val_all	val_lenient	val_newform	train	val_all	val_lenient	val_newform
simulated	94%	76%	92%	63%	81%	37%	57%	19%
word2vec	91%	69%	81%	58%	48%	14%	28%	0.1%
def + indef	94%	80%	93%	64%	82%	40%	61%	15%

Table 8: Comprehension and production accuracy for train and validation datasets with articles. All three simulations use triphones as cues. The first two rows present results with simulated vectors and *word2vec* embeddings as semantic representations. The simulation presented in the bottom row also makes use of simulated vectors, but includes both definite and indefinite articles.

forms. Homophones comprise slightly more words with indefinite articles (57%) whereas unseen forms consist of slightly more definite articles (59%). The results, presented in the bottom row of Table 8, are very similar to those with only definite articles (top row). Closer inspection of the results for the validation data shows that for comprehension, accuracies do not differ much across definite and indefinite forms. For production, however, especially for unseen forms, the accuracy for definite articles is twice higher than that for indefinite articles (20% and 9%, averaging out to 15%). This is a straightforward consequence of the much more diverse realizations of indefinite nouns. For definite nouns, the possible triphone cues at the first two positions in the word are always limited to the triphone cues of the six definite articles. For indefiniteness, however, in addition to the six indefinite articles, initial triphone cues also originate from words' stems, given that indefinite plural forms are realized without articles. The mappings for production are faced with a more complex task for indefinites, and the model is therefore more likely to fail on indefinite forms.

4.4.2 Semantic roles

The simulation studies thus far suggest it is not straightforward to correctly comprehend a novel German word form in isolation, even when articles are provided. This is perhaps not that surprising, as in natural language use, inflected words appear in context, and usually realize not some abstract case ending, but a specific semantic role (also called *thematic role*, see, e.g., Harley, 2010). For example, a word in the nominative singular might express a theme, as *der Apfel* in *Der Apfel fällt vom Baum.* ('The apple falls from the tree'), or it might express an agent as *der Junge* in *Der Junge isst den Apfel.* ('The boy eats the apple.'). Exactly the same lemma, used with exactly the same case and number, may still realize very different semantic roles. Consider the two sentences *Ich bin bei der Freundin* ('I'm at the friend's') and *Ich gebe der Freundin das Buch.* ('I give the book to the friend'). *der Freundin* is dative singular in both cases, but in the first sentence, it expresses a location while in the second it represents the beneficiary or receiver. Semantic roles can also be reflected in a word's form, independently of case markers. For example, German nouns ending in *-er* are so-called "Nomen Agentis" (Baeskow, 2011). As pointed out by Blevins (2016), case endings are no more (or less) than markers for the intersection of form variation and a distribution class of semantic roles. Since within the framework of the DLM, the aim is to provide mappings between form and meaning, a case label is not a proper representation of a word's actual meaning. All it does is specify a range of meanings that the form can have, depending on context. Therefore, even though we can get the mechanics of the model to work with case specifications, doing so clashes with the 'discriminative modeling approach'. In what follows, we therefore present an attempt to implement mappings with more realistic semantic representations of German inflected nouns.

Case	Semantic roles
Nominative	agent (50%), theme (40%), patient (10%)
Genitive	possessive (90%), partitive (10%)
Dative	beneficiary (50%), location (50%)
Accusative	patient (40%), motion (30%), experiencer (30%)

Table 9: Probabilities of semantic roles by cases in the German noun system. Semantic roles are informed by Schulz and Griesbach (1981). Percentages are chosen arbitrarily.

Our starting point is that in German, different cases can realize a wide range of semantic roles. For our simulations, we restrict ourselves to some of the most prominent semantic roles for each case (see Table 9). Even though these clearly do not reflect the full richness of the semantics of German cases, they suffice for a proof-of-concept simulation.

In order to obtain a data set with variegated semantic roles, we expanded the previous data set, with each word form (including its article) appearing with a specification of its semantic role, according to the probabilities presented in Table 9. The resulting dataset had 45,605 entries, which we randomly split into 80% training data and 20% validation data. For generating the semantic matrix, we again used *number*, but instead of a case label, we provided the *semantic role* as inflectional features. Comprehension accuracy on this data is comparable to the previous simulations: 89% for the training data `train`, and 85% `val_lenient`. Comprehension accuracy on the validation set drops dramatically when we use strict evaluation (4% accuracy). This is unsurprising given that it is impossible for the model to know which semantic role is indicated when only being exposed to the word form and its article in isolation, without syntactic context. Production accuracy is likewise comparable to previous simulations with `val_lenient` at 61% (`val_newform` = 25%). This simple result clarifies that in order to properly model German nouns, it is necessary to take the syntactic context in which a noun occurs into account. Future research will also have to face the challenge of integrating words' individual usage profiles into the model (see also Section 4.2.1 above).

4.5 Incremental learning versus the endstate of learning

In the simulation studies presented thus far, we made use of the regression method to estimate the mappings between form and meaning. The regression method is strictly type based: the data on which a model is trained and evaluated consists of all unique combinations of form vectors $\boldsymbol{c}$ and semantic vectors $\boldsymbol{s}$. In this respect, the regression method is very similar to models such as AML, MBL, MGL, and to statistical analyses with the GLM or recursive partioning methods. However, word types (understood as unique sets $\{\boldsymbol{c}, \boldsymbol{s}\}$) are not uniformly distributed in language, and there is ample evidence that the frequencies with which word types occur co-determines lexical processing (see, e.g., Baayen et al., 1997, 2016, 2007; Tomaschek et al., 2018). While some formal theorists flatly deny that word frequency effects exist for complex words (Yang, 2016), others have argued that there is no problem with integrating frequency of use into theories of the lexicon (Jackendoff, 1975; Jackendoff and Audring, 2019), and yet others have argued that it is absolutely essential to incorporate frequency into any meaningful account of language in action (Langacker, 1987; Bybee, 2010).

Within the present approach, effects of frequency of occurrence can be incorporated seamlessly

word form	lemma	case	number	semantic role	form frequency	form+role frequency
Adresse	Adresse	nominative	singular	agent	137	20
Adresse	Adresse	nominative	singular	theme	137	16
Adresse	Adresse	nominative	singular	patient	137	0
Adresse	Adresse	genitive	singular	possessive	137	0
Adresse	Adresse	genitive	singular	partitive	137	35
Adresse	Adresse	dative	singular	beneficiary	137	18
Adresse	Adresse	dative	singular	location	137	18
Adresse	Adresse	accusative	singular	patient	137	0
Adresse	Adresse	accusative	singular	motion	137	0
Adresse	Adresse	accusative	singular	experiencer	137	35

Table 10: Example of simulated frequencies for combinations of case and semantic role for the word form "Adresse".

by using incremental learning instead of the endstate of learning as defined by the regression equations (see Danks, 2003; Evert and Arppe, 2015; Shafaei-Bajestan et al., 2021, for the convergence over learning time of incremental learing to the regression endstate of learning). We illustrate this for our German nouns dataset with number and semantic role as crucial constructors of simulated semantic vectors.

We begin with noting that word forms usually do not instantiate all possible semantic roles equally frequently. For instance, a word such as *der Doktor* ('doctor') will presumably occur mostly as *agent* in the nominative singular form, rather than as *theme* or *patient*. If the model is informed about the probability distributions of semantic roles in actual language use, it may be expected to make more informed decisions when coming across new forms, for instance, by opting for the best match given its past experience.

Incremental learning with the learning rule of Widrow-Hoff makes it possible to start approximating human word-to-word learning as a function of experience. As a consequence, the more frequent a word type occurs in language use, the better it can be learned: practice makes perfect. This sets the following simulation study apart from models such as proposed by Belth et al. (2021) or McCurdy et al. (2020), who base their training regimes on types rather than tokens.

In the absence of empirical frequencies with which combinations of semantic roles and German nouns co-occur, we simulated frequencies of use. To do so, we proceeded as follows. First, we collected token frequencies for all our word forms from CELEX. Next, we assigned an equal part $freq_p$ of this frequency count to each case/number cell realising this word form. Third, for each paradigm cell, we randomly set to zero some semantic roles, drawing from a Binomial distribution with $n = 1$, $p = \frac{1}{K}$, with K the number of semantic roles for the paradigm cell. In this way, on average, one semantic role was omitted per paradigm cell. Finally, given a proportional frequency count $freq_p$, the semantic roles associated with a paradigm cell received frequencies proportional to the percentages given in Table 9. Further details on this procedure are available in the supplementary materials, a full example can be found in Table 10.

Having obtained the simulated frequencies, we proceeded by randomly selecting 274 different lemmas (1,289 distinct word forms with definite articles included), in order to keep the size of the simulation down — simulating with the Widrow-Hoff rule is computationally expensive. The total

number of tokens in this study was 4,470. For the form vectors, we used triphones. Semantic vectors were simulated. The dimension of the semantic vectors was identical to that of the cue vectors. As before, the data was split into 80% training and 20% validation data. We followed the same procedure as in the previous experiments, but instead of computing the mapping matrices in their closed form solution, we used incremental learning.

While for comprehension, the implementation of the learning algorithm is relatively straightforward, this is not the case for production. The `learn_paths` algorithm calculates the support for each of the n-grams, for each possible position in a word. In the current implementation of **JudiLing**, the calculation of positional support is not implemented for incremental learning. Therefore, we do not consider incremental learning of production here.

Comprehension accuracy was similar to that observed for previous experiments. Training accuracy when taking into account homophones was 85%, validation accuracy on the full data was 79% (`val_lenient`). Without considering homophones, validation accuracy drops substantially (`val_strict` = 7%). This is unsurprising given that from the form alone it is once again impossible to predict the proper semantic role.

The accuracy of the model's predictions is also closely linked to the frequencies with which words' form+role combinations are encountered in the training data. If a word's form+role combination is very frequent, it is learned better. Figure 2 presents the correlations of words' predicted and targeted semantic vector against their frequency of occurrence. The left panel presents the results for the incrementally learned model, the right panel for the endstate of learning. Clearly, after incremental learning the model predicts the semantics of more frequent form+role combinations more accurately than for less frequent ones. For the endstate of learning on the other hand, no such effect can be observed. These results clearly illustrate the difference between a token-based model and a typed-based model.

The effect of frequency of use on the kind of errors made by the model is also of interest. We zoom in on those cases where the model was able to correctly identify the lemma and paradigm cell of the word form, but did not get the semantic role correct. Figure 3 provides scatterplots graphing the number of times a semantic role was (incorrectly) understood against the frequency of the form's semantic role, cross-classified by training method (incremental, left panels; endstate of learning, right panels) and by evaluation set (top panels: training data, bottom panels: validation data).

For incremental learning, there is a positive correlation between the number of times a semantic role was (incorrectly) identified and the frequency of the semantic role in the training data. Note that the relation is not linear, but curvilinear. A linear relation would have implied that a fixed proportion of word forms would be incorrectly recognized, across all semantic roles. What we see, by contrast, is that greater exposure in language use has an increasingly detrimental effect on learning, with more probable semantic roles being over-identified. Importantly, for the endstate of learning, this curvilinear effect of frequency on learning is absent, with the PATIENT role representing an outlier. This outlier status may be due to the *patient* semantic role being realized by two cases: *nominative* and *accusative*. As a consequence, it is not only frequent, but it is also predicted by many more different cues (especially cues from the articles) than is the case for other semantic roles.

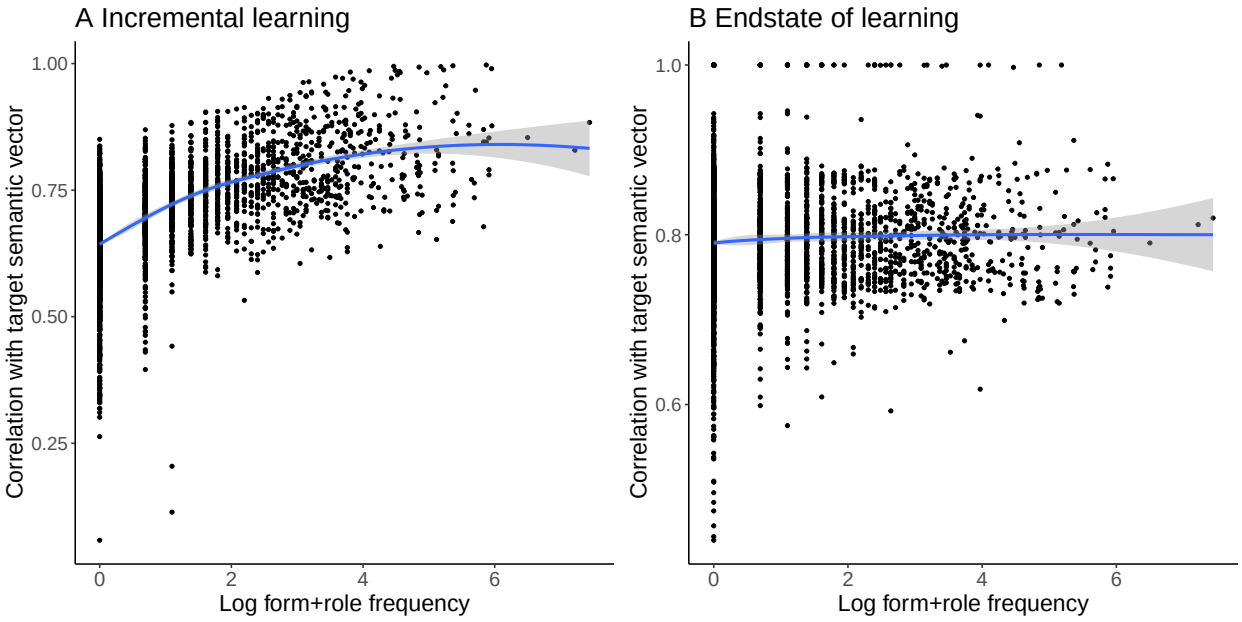

Figure 2: Correlation between the simulated frequency and correlation of the predicted semantic vector with its target. Generally, the more frequent a word form is, the more accurate its semantic vector is predicted. The blue line indicates a loess smooth with a .95 confidence interval.

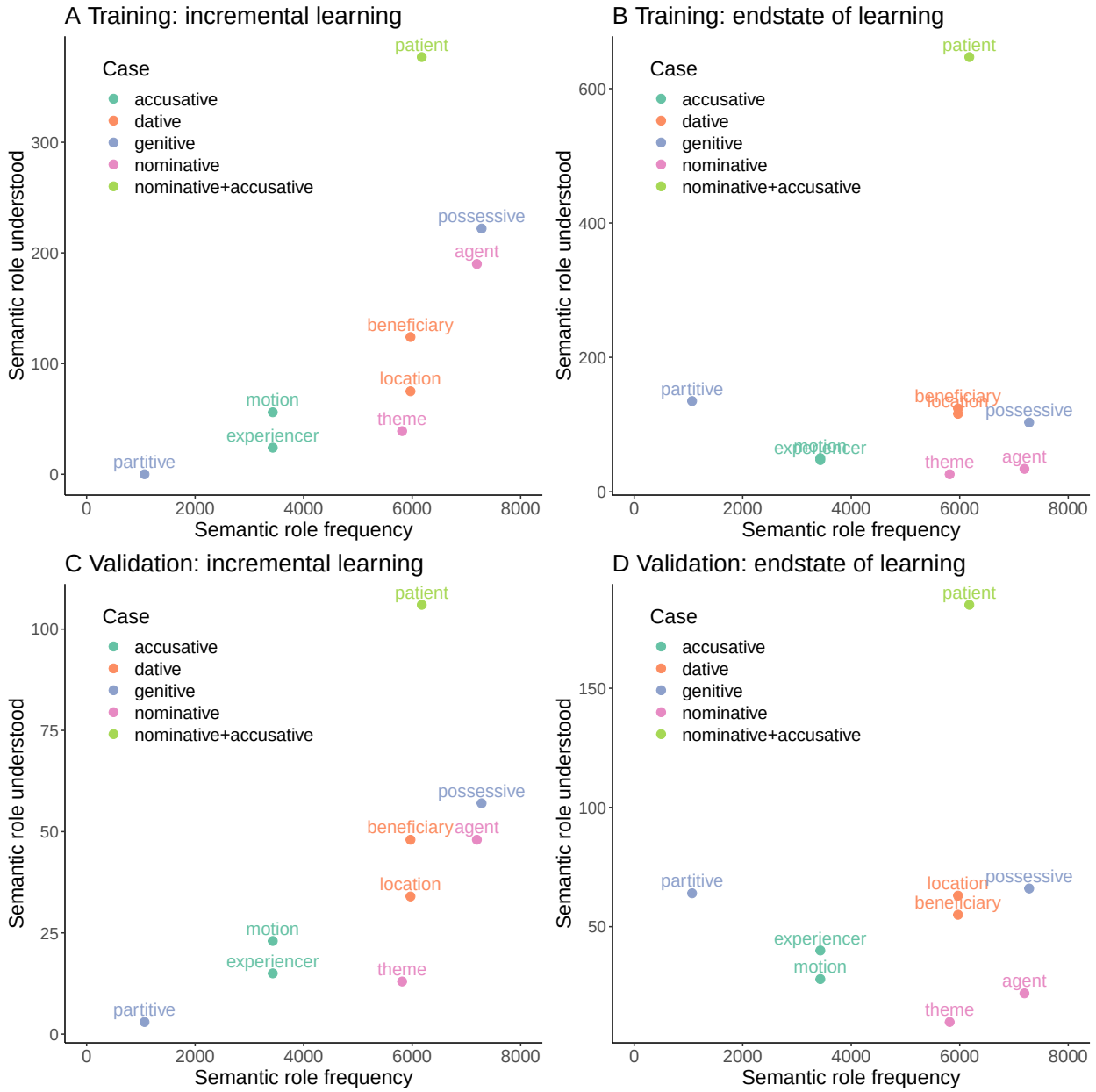

Figure 3: Counts of overgeneralization errors of semantic roles for training (top) and test data (bottom), for incremental learning (left) and the endstate of learning (right), conditional on the model having understood lexeme, number, and case correctly.

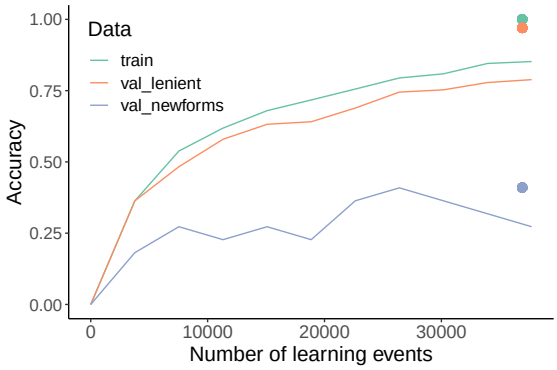

Figure 4: Comprehension accuracy over the course of learning. After a very fast increase in accuracy over the first 15,000 learning events, the amount of learning levels off. Points indicate the accuracy at the endstate of learning which the incremental model would reach eventually after an infinite number of learning events.

In other words, with incremental learning, strong frequency effects emerge, hand in hand with overgeneralization of semantic roles. (The study by Ramscar et al. (2013) makes the same point for irregular English noun plurals.) By contrast, for the endstate of learning, such effects are absent. Mathematically, this makes sense: as experience (i.e., volume of training data) goes to infinity, all forms are learned an infinite number of times, and frequency is no longer distinctive.

With incremental learning, it is also possible to follow the learning trajectory of the model. Figure 4 presents this trajectory at 10 evaluation points. Learning proceeds rapidly during the first 15,000 learning events and slows down afterwards. Validation accuracy `val_lenient` closely follows training accuracy, which is a straightforward consequence of the large numbers of homophones. `val_newforms` on the other hand stays relatively low, in accordance with the semi-productivity of the German declension system.

Note that in this simulation we only pass through the data once, in the sense that if a word form has a form+role frequency of 1, it is only seen a single time during training. As such, it is not possible for the model to reach accuracies as high as at the endstate of learning (indicated as dots in Figure 4), which would be reached eventually after an infinite number of passes through the data (Danks, 2003; Evert and Arppe, 2015; Shafaei-Bajestan et al., 2021). This sets our approach apart from deep learning, where models are trained on many epochs of the data set until the loss function reaches a local minimum. Whereas such a procedure makes sense for language engineering, it does not make sense for human learning: we don't relive the same exposure to data multiple times, and for healthy people, there is no point in learning after which performance degrades. For instance, vocabulary learning is a continuous process straight into old age (Keuleers et al., 2015).

In summary, what this simulation clarifies is that the present modeling framework offers the possibility to approximate incremental human learning and the consequences of frequency of exposure for learning (see also Chuang et al., 2020a, for learning in a multilingual setting).

4.6 Model complexity

LDL is costly in the number of connection weights, or equivalently, the number of beta coefficients. For example, the mapping matrix **F** for the dataset discussed in Section 4.4.2 has 35 million weights

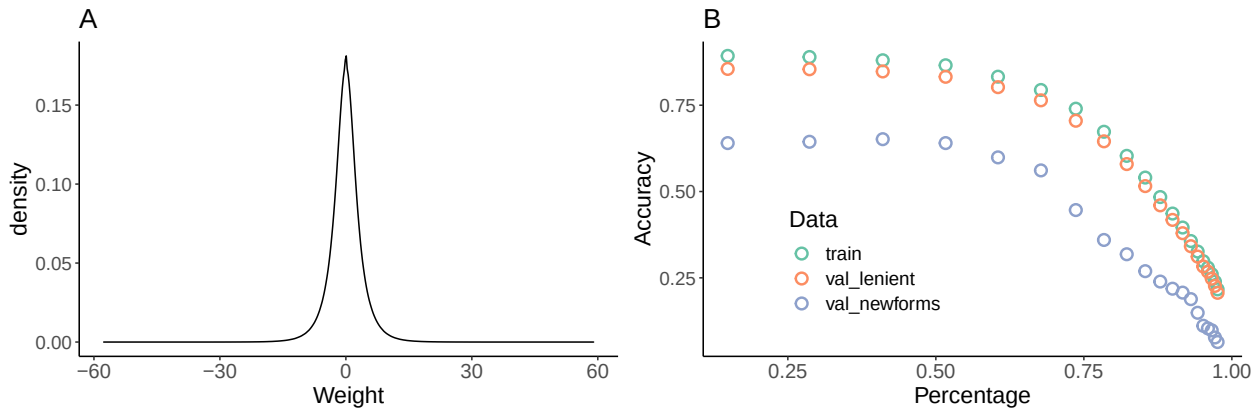

Figure 5: (A) Distribution of weights in the mapping matrix from form to meaning for the dataset with semantic roles. (B) Accuracy of the endstate model as a function of the proportion of connection weights close to zero are pruned. About 40% of the weights can be set to zero without seriously affecting the performance of the model.

(5913 × 5913 dimensions), rendering it much more costly in terms of the number of weights than deep-learning models, or models such as AML, MBL, and recursive partitioning methods.

Inspection of the distribution of weights, however, clarifies that most weights are very close to zero. In other words, most cues have low discriminative value. This suggests they can be pruned without seriously affecting model performance. This can be tested by selecting a threshold ϑ and setting all absolute values in the mapping matrix that fall below this threshold to zero. Figure 5 shows, for varying ϑ, that up to 40% of the small weights can be pruned without substantially impacting the performance of the model. As neural pruning is part and parcel of human cortical development (see, e.g. Gogtay et al., 2004), an interesting topic for further research is to integrate incremental learning with neural pruning of uninformative connections.

5 Discussion

In this study, we illustrated the methodological consequences of the many different choices that have to be made when modelling morphological systems within the discriminative lexicon framework, using LDL as modeling engine. We illustrated these choices for the German noun system. In one way, this system is 'degenerate', as many of its paradigm cells share the same word forms (homophones). This system is also in many ways irregular: a noun's declension class can often not be fully predicted by its phonology, gender, or semantics (Köpcke, 1988). The results we obtained with LDL reflect this complexity. The model can learn word forms very well, achieving accuracies of more than 90% on both comprehension and production when evaluated on training data. It can also generalize very well to new paradigm cells when it comes to word forms it has already seen, thanks to the ubiquitous homophony that characterizes German noun paradigms. However, it also mirrors the unpredictability of German inflections when it comes to word forms it hasn't seen before. Accuracies for both comprehension and production suffer, but nevertheless the model shows some semi-productivity and succeeds in generalizing to many of the subregularities found in the German noun system (Wunderlich, 1999), reaching accuracies of 50% on comprehension and

20% on production. Since German speakers encounter similar problems with new German word forms, as has been demonstrated in various wug studies (Zaretsky et al., 2013; McCurdy et al., 2020), our model properly exhibits the limitations that are also encountered by native speakers.

In this study, we also probed the modeling of German nouns in context. The rampant homophony that characterizes German noun paradigms is a straightforward consequence of considering words in isolation. The amount of homophony can be substantially reduced by including articles, in which case the model still performs well. In context, case-inflected words typically do not realize a specific case meaning, but rather a specific semantic role. As case endings typically do not stand in a one-to-one relation with semantic roles, we also examined to what extent we can make the model more realistic by replacing semantic vectors for cases with semantic vectors for a variety of semantic roles. For the simulated dataset that we constructed, the model again performed well. For this dataset, we also demonstrated how the consequences of frequency of occurrence can be brought into the model, namely, by moving from the endstate of learning (estimated with regression) to incremental learning using the Widrow-Hoff learning rule. One limitation of our model is that in most of the implementations, we have been using very high-level abstract representations. The phone-based representation, for example, involves tremendous simplifications compared to real speech, as variability in pronunciations is enormous (Johnson, 2004; Ernestus et al., 2002). On the meaning side, traditional case labels have no intrinsic semantic content, and although we can replace cases with semantic roles, these too are still too simplistic to be able to capture the full complexity of the semantics of words in context. However, we note that even with the present high-level representations, the model can still generate useful predictions, and various studies carried out within this framework have successfully modeled a range of aspects of human lexical processing (see Chuang and Baayen, 2021, for further details). In summary, even though the current framework undoubtedly misses out on a great number of nuanced but potentially informative features of forms and meanings in real language use, it can still serve as a useful linguistic tool to explore the strengths and weaknesses of morphological systems.

A question that inevitably arises in the context of computational modeling is how cognitively plausible a model is. In the introduction, we called attention to the distinction made by Breiman et al. (2001) between statistical models and machine learning models. We view LDL primarily as a statistical model that enables us to clarify quantitative structure in the lexicon. However, since the matrix of beta coefficients of multivariate multiple regression model can be conceptualized as the weight matrix characterizing connection strengths in a simple network, and given that such a network can be trained incrementally, it is worth noting that the principle of error-driven learning with the very simple learning rules of Widrow-Hoff and Rescorla-Wagner has excellent credentials across a wide range of domains of inquiry (see, e.g., Rescorla, 1988; Schultz, 1998; Marsolek, 2008; Oppenheim et al., 2010; Trimmer et al., 2012).

It is possible to take the model as point of departure for addressing questions at the level of neural organization in the brain. For instance, Heitmeier and Baayen (2021) were interested in clarifying whether the framework of the discriminative lexicon properly predicts the dissociations of form and meaning observed for aphasic speakers producing English regular and irregular past-tense forms, following Joanisse and Seidenberg (1999). They took the unordered banks of units of form and meaning (the column dimensions of the $\boldsymbol{C}$ and $\boldsymbol{S}$ matrices) and projected them onto two-dimensional surfaces approximating, however crudely, cortical maps. This made it possible to lesion the network in a topologically cohesive way, rather than by randomly taking out connections across the whole network. They made use of an algorithm from physics (`http://www.schmuhl.`

org/graphopt/) used to display graphs, but temporal self-organizing maps (TSOMs, Ferro et al., 2011; Chersi et al., 2014; Marzi et al., 2012, 2018) offer a much more fine-grained and principled way for modeling morphological organization building on principles of error-driven learning.

Deep learning algorithms provide the analyst with powerful modeling tools, but it seems that current architectures are too powerful (see, e.g., McCurdy et al., 2020) for understanding not only the strengths but also the weaknesses and the frailties of human lexical memory and lexical processing. However, linguistic models are in a different way also too powerful on the one hand, and too underspecified on the other hand. Paradigms are typically constructed to accommodate any contrast between forms and inflectional functions, even when a contrast is attested only for a few forms. The result is an overabundance of homophones, which are severely underspecified with respect to their actual meanings in real language use (such as their semantic roles). Furthermore, in actual language use, many paradigm cells remain empty (Karlsson, 1986), which in turn has clear consequences for lexical processing (Loo et al., 2018).

In this study, we have provided an overview of the many choice points that arise in psycho-computational modeling, each of which requires knowledge of morphology and morphological theory. The implications of our approach to psycho-computational modeling for morphological theory depends on the specifics of a given (often rival) theory of morphology. Our approach is broadly consistent with usage-based approaches to morphology (Bybee, 1985, 2010), and with Word and Paradigm Morphology (Blevins, 2016). It is less clear whether our modeling approach is informative for theories that are only interested in defining possible words. With this methodological study, we have shed some light on the many questions and issues that do not arise in formal theories of morphology, but that have to be addressed in a linguistically informed way when the goal of one's theory is to better understand, and predict, in all its complexity, human lexical processing across comprehension and production.

References

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

Arndt-Lappe, S. (2011). Towards an exemplar-based model of stress in english noun-noun compounds. *Journal of Linguistics*, pages 549–585.

Arnold, D., Tomaschek, F., Lopez, F., Sering, T., and Baayen, R. H. (2017). Words from spontaneous conversational speech can be recognized with human-like accuracy by an error-driven learning algorithm that discriminates between meanings straight from smart acoustic features, bypassing the phoneme as recognition unit. *PLOS ONE*, 12(4):e0174623.

Baayen, R. H., Chuang, Y.-Y., and Blevins, J. P. (2018). Inflectional morphology with linear mappings. *The Mental Lexicon*, 13(2):232–270.

Baayen, R. H., Chuang, Y.-Y., Shafaei-Bajestan, E., and Blevins, J. (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*.

Baayen, R. H., Dijkstra, T., and Schreuder, R. (1997). Singulars and plurals in Dutch: Evidence for a parallel dual route model. *Journal of Memory and Language*, 36:94–117.

Baayen, R. H., Milin, P., Filipović Durdević, D., Hendrix, P., and Marelli, M. (2011). An amorphous model for morphological processing in visual comprehension based on naive discriminative learning. *Psychological Review*, 118:438–482.

Baayen, R. H., Milin, P., and Ramscar, M. (2016). Frequency in lexical processing. *Aphasiology*, 30(11):1174–1220.

Baayen, R. H., Piepenbrock, R., and Gulikers, L. (1995). The CELEX lexical database [cd rom]. Philadelphia: Linguistic Data Consortium, University of Pennsylvania.

Baayen, R. H., Schreuder, R., De Jong, N. H., and Krott, A. (2002). Dutch inflection: the rules that prove the exception. In Nooteboom, S., Weerman, F., and Wijnen, F., editors, *Storage and Computation in the Language Faculty*, pages 61–92. Kluwer Academic Publishers, Dordrecht.

Baayen, R. H. and Smolka, E. (2020). Modelling morphological priming in German with naive discriminative learning. *Frontiers in Communication, section Language Sciences*. preprint on PsyArXiv, doi:10.31234/osf.io/nj39v.

Baayen, R. H., Wurm, L. H., and Aycock, J. (2007). Lexical dynamics for low-frequency complex words. a regression study across tasks and modalities. *The Mental Lexicon*, 2:419–463.

Baeskow, H. (2011). *Abgeleitete Personenbezeichnungen im Deutschen und Englischen: kontrastive Wortbildungsanalysen im Rahmen des minimalistischen Programms und unter Berücksichtigung sprachhistorischer Aspekte*, volume 62. Walter de Gruyter.

Behrens, H. and Tomasello, M. (1999). And what about the chinese? *Behavioral and Brain Sciences*, 22(6):1014–1014.

Belth, C., Payne, S., Beser, D., Kodner, J., and Yang, C. (2021). The greedy and recursive search for morphological productivity. *arXiv preprint arXiv:2105.05790*.

Bierwisch, M. (2018). *Syntactic features in morphology: General problems of so-called pronominal inflection in German*. De Gruyter Mouton.

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

Boersma, P. (1998). *Functional Phonology*. Holland Academic Graphics, The Hague.

Boersma, P. and Hayes, B. (2001). Empirical tests of the gradual learning algorithm. *Linguistic Inquiry*, 32:45–86.

Bojanowski, P., Grave, E., Joulin, A., and Mikolov, T. (2017). Enriching word vectors with subword information. *Transactions of the Association for Computational Linguistics*, 5:135–146.

Breiman, L. et al. (2001). Statistical modeling: The two cultures (with comments and a rejoinder by the author). *Statistical science*, 16(3):199–231.

Breiman, L., Friedman, J. H., Olshen, R., and Stone, C. J. (1984). *Classification and Regression Trees*. Wadsworth International Group, Belmont, California.

Bybee, J. (2010). *Language, usage and cognition*. Cambridge University Press, Cambridge.

Bybee, J. L. (1985). *Morphology: A study of the relation between meaning and form*. Benjamins, Amsterdam.

Cahill, L. and Gazdar, G. (1999). German noun inflection. *Journal of Linguistics*, pages 1–42.

Chersi, F., Ferro, M., Pezzulo, G., and Pirrelli, V. (2014). Topological self-organization and prediction learning support both action and lexical chains in the brain. *Topics in cognitive science*, 6(3):476–491.

Chuang, Y.-Y. and Baayen, R. H. (2021). Discriminative learning and the lexicon: Ndl and ldl. In *Oxford Research Encyclopedia of Linguistics*, page accepted. Oxford University Press.

Chuang, Y.-Y., Bell, M., Banke, I., and Baayen, R. H. (2020a). Bilingual and multilingual mental lexicon: a modeling study with Linear Discriminative Learning. *Language Learning*.

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

Clahsen, H. (1999). Lexical entries and rules of language: A multidisciplinary study of german inflection. *Behavioral and brain sciences*, 22(6):991–1013.

Coltheart, M., Curtis, B., Atkins, P., and Haller, M. (1993). Models of reading aloud: Dual-route and parallel-distributed-processing approaches. *Psychological review*, 100(4):589.

Corbett, G. G. (1991). Gender. introduction.

Daelemans, W., Berck, P., and Gillis, S. (1995). Linguistics as data mining: Dutch diminutives. In Andernach, T., Moll, M., and Nijholt, A., editors, *CLIN V, Papers from the 5th CLIN meeting*, pages 59–71. Parlevink, Enschede.

Daelemans, W. and Van den Bosch, A. (2005). *Memory-based language processing*. Cambridge University Press, Cambridge.

Dalal, N. and Triggs, B. (2005). Histograms of oriented gradients for human detection. In *2005 IEEE Computer Society Conference on Computer Vision and Pattern Recognition (CVPR'05)*, volume 1, pages 886–893.

Danks, D. (2003). Equilibria of the rescorla–wagner model. *Journal of Mathematical Psychology*, 47(2):109–121.

Dell, G. S. (1986). A spreading-activation theory of retrieval in sentence production. *Psychological review*, 93(3):283.

Devlin, J., Chang, M.-W., Lee, K., and Toutanova, K. (2018). Bert: Pre-training of deep bidirectional transformers for language understanding. *arXiv preprint arXiv:1810.04805*.

Elman, J. L. (2009). On the meaning of words and dinosaur bones: Lexical knowledge without a lexicon. *Cognitive science*, 33(4):547–582.

Ernestus, M. and Baayen, R. H. (2003). Predicting the unpredictable: Interpreting neutralized segments in Dutch. *Language*, 79:5–38.

Ernestus, M., Baayen, R. H., and Schreuder, R. (2002). The recognition of reduced word forms. *Brain and Language*, 81:162–173.

Evans, R. and Gazdar, G. (1996). DATR: A language for lexical knowledge. *Computational Linguistics*, 22:167–216.

Evert, S. and Arppe, A. (2015). Some theoretical and experimental observations on naive discriminative learning. *Proceedings* QITL *6, Tübingen*.

Ferro, M., Marzi, C., and Pirrelli, V. (2011). A self-organizing model of word storage and processing: implications for morphology learning. *Lingue e linguaggio*, 10(2):209–226.

Finkel, R. and Stump, G. (2007). Principal parts and morphological typology. *Morphology*, 17(1):39–75.

Gaeta, L. (2008). Die deutsche Pluralbildung zwischen deskriptiver Angemessenheit und Sprachtheorie. *Zeitschrift für germanistische Linguistik*, 36(1):74–108.

Gaskell, M. G. and Marslen-Wilson, W. (1997). Integrating form and meaning: A distributed model of speech perception. *Language and Cognitive Processes*, 12:613–656.

Goebel, R. and Indefrey, P. (2000). A recurrent network with short-term memory capacity learning the german-s plural. In *Models of language acquisition: Inductive and deductive approaches*, pages 177–200.

Gogtay, N., Giedd, J. N., Lusk, L., Hayashi, K. M., Greenstein, D., Vaituzis, A. C., Nugent, T. F., Herman, D. H., Clasen, L. S., Toga, A. W., et al. (2004). Dynamic mapping of human cortical development during childhood through early adulthood. *Proceedings of the National Academy of Sciences*, 101(21):8174–8179.

Goldsmith, J. and O'Brien, J. (2006). Learning inflectional classes. *Language Learning and Development*, 2(4):219–250.

Hahn, U. and Nakisa, R. (2000). German inflection: Single route or dual route? *Cognitive Psychology*, 41:313–360.

Harley, H. (2010). Thematic roles. In Hogan, P., editor, *The Cambridge Encyclopedia of the Language Sciences*, pages 861–862. Cambridge University Press.

Harm, M. W. and Seidenberg, M. S. (2004). Computing the meanings of words in reading: Cooperative division of labor between visual and phonological processes. *Psychological Review*, 111:662–720.

Heitmeier, M. and Baayen, R. H. (2021). Simulating phonological and semantic impairment of English tense inflection with Linear Discriminative Learning. *The Mental Lexicon*, 15:385–421.

Indefrey, P. (1999). Some problems with the lexical status of nondefault inflection. *Behavioral and Brain Sciences*, 22(6):1025.

Jackendoff, R. and Audring, J. (2019). *The texture of the lexicon: relational morphology and the parallel architecture*. Oxford University Press.

Jackendoff, R. S. (1975). Morphological and semantic regularities in the lexicon. *Language*, 51:639–671.

Joanisse, M. F. and Seidenberg, M. S. (1999). Impairments in verb morphology after brain injury: a connectionist model. *Proceedings of the National Academy of Sciences*, 96:7592–7597.

Johnson, K. (2004). Massive reduction in conversational American English. In *Spontaneous speech: data and analysis. Proceedings of the 1st session of the 10th international symposium*, pages 29–54, Tokyo, Japan. The National International Institute for Japanese Language.

Karlsson, F. (1986). Frequency considerations in morphology. *STUF-Language Typology and Universals*, 39(1-4):19–28.

Karttunen, L. (2003). Computing with realizational morphology. In *International Conference on Intelligent Text Processing and Computational Linguistics*, pages 203–214. Springer.

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

Keuleers, E., Stevens, M., Mandera, P., and Brysbaert, M. (2015). Word knowledge in the crowd: Measuring vocabulary size and word prevalence in a massive online experiment. *The Quarterly Journal of Experimental Psychology*, (8):1665–1692.

Kirov, C. and Cotterell, R. (2018). Recurrent neural networks in linguistic theory: Revisiting Pinker and Prince (1988) and the past tense debate. *Transactions of the Association for Computational Linguistics*, 6:651–665.

Köpcke, K.-M. (1988). Schemas in german plural formation. *Lingua*, 74(4):303–335.

Landauer, T. and Dumais, S. (1997). A solution to Plato's problem: The latent semantic analysis theory of acquisition, induction and representation of knowledge. *Psychological Review*, 104(2):211–240.

Langacker, R. W. (1987). *Foundations of cognitive grammar: Theoretical prerequisites*, volume 1. Stanford university press.

Levelt, W., Roelofs, A., and Meyer, A. S. (1999). A theory of lexical access in speech production. *Behavioral and Brain Sciences*, 22:1–38.

Linke, M., Broeker, F., Ramscar, M., and Baayen, R. H. (2017). Are baboons learning "orthographic" representations? probably not. *PLOS-ONE*, 12(8):e0183876.

Loo, K., Jaervikivi, J., Tomaschek, F., Tucker, B., and Baayen, R. (2018). Production of Estonian case-inflected nouns shows whole-word frequency and paradigmatic effects. *Morphology*, 1(28):71–97.

Luo, X. (2021). JudiLing: An implementation for Linear Discriminative Learning in JudiLing (unpublished Master's thesis).

Luo, X., Chuang, Y.-Y., and Baayen, R. H. (2021). Judiling: an implementation in Julia of Linear Discriminative Learning algorithms for language modeling.

MacWhinney, B. and Leinbach, J. (1991). Implementations are not conceptualizations: revising the verb learning model. *Cognition*, 40:121–157.

Malouf, R. (2017). Abstractive morphological learning with a recurrent neural network. *Morphology*, 27(4):431–458.

Marcus, G. F., Brinkmann, U., Clahsen, H., Wiese, R., and Pinker, S. (1995). German inflection: The exception that proves the rule. *Cognitive psychology*, 29(3):189–256.

Marsolek, C. J. (2008). What antipriming reveals about priming. *Trends in Cognitive Science*, 12(5):176–181.

Marzi, C., Ferro, M., Nahli, O., Belik, P., Bompolas, S., and Pirrelli, V. (2018). Evaluating inflectional complexity crosslinguistically: A processing perspective. In *Proceedings of the Eleventh International Conference on Language Resources and Evaluation (LREC 2018)*.

Marzi, C., Ferro, M., and Pirrelli, V. (2012). Word alignment and paradigm induction. *Lingue e Linguaggio*, 11(2):251–0.

Matthews, P. H. (1974). *Morphology. An Introduction to the Theory of Word Structure*. Cambridge University Press, Cambridge.

McCurdy, K. (2019). Neural Networks Don't Learn Default Rules for German Plurals, But That's Okay, Neither Do Germans. *Master's Thesis, University of Edinburgh*.

McCurdy, K., Goldwater, S., and Lopez, A. (2020). Inflecting when there's no majority: Limitations of encoder-decoder neural networks as cognitive models for German plurals. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 1745–1756. Association for Computational Linguistics.

Miaschi, A. and Dell'Orletta, F. (2020). Contextual and non-contextual word embeddings: an in-depth linguistic investigation. In *Proceedings of the 5th Workshop on Representation Learning for NLP*, pages 110–119.

Mikolov, T., Chen, K., Corrado, G., and Dean, J. (2013a). Efficient estimation of word representations in vector space. *arXiv preprint arXiv:1301.3781*.

Mikolov, T., Sutskever, I., Chen, K., Corrado, G. S., and Dean, J. (2013b). Distributed representations of words and phrases and their compositionality. In *Advances in neural information processing systems*, pages 3111–3119.

Milin, P., Madabushi, H. T., Croucher, M., and Divjak, D. (2020). Keeping it simple: Implementation and performance of the proto-principle of adaptation and learning in the language sciences. *arXiv preprint arXiv:2003.03813*.

Mirković, J., MacDonald, M. C., and Seidenberg, M. S. (2005). Where does gender come from? evidence from a complex inflectional system. *Language and cognitive processes*, 20:139–167.

Nakisa, R. C. and Hahn, U. (1996). Where defaults don't help: the case of the german plural system. In *Proc. 18th Annu. Conf. Cogn. Sci. Soc*, pages 177–182.

Oppenheim, G. M., Dell, G. S., and Schwartz, M. F. (2010). The dark side of incremental learning: A model of cumulative semantic interference during lexical access in speech production. *Cognition*, 114(2):227–252.

Pennington, J., Socher, R., and Manning, C. D. (2014). Glove: Global vectors for word representation. In *Empirical Methods in Natural Language Processing (EMNLP)*, pages 1532–1543.

Pinker, S. and Prince, A. (1988). On language and connectionism. *Cognition*, 28:73–193.

Prince, A. and Smolensky, P. (2008). *Optimality Theory: Constraint interaction in generative grammar*. John Wiley & Sons.

Ramscar, M., Dye, M., and McCauley, S. M. (2013). Error and expectation in language learning: The curious absence of mouses in adult speech. *Language*, 89(4):760–793.

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

Rumelhart, D. E. and McClelland, J. L. (1986). On learning the past tenses of English verbs. In McClelland, J. L. and Rumelhart, D. E., editors, *Parallel Distributed Processing. Explorations in the Microstructure of Cognition. Vol. 2: Psychological and Biological Models*, pages 216–271. The MIT Press, Cambridge, Mass.

Schultz, W. (1998). Predictive reward signal of dopamine neurons. *Journal of Neurophysiology*, 80:1–27.

Schulz, D. and Griesbach, H. (1981). *Grammatik der deutschen Sprache*. Max Hueber Verlag, München, 11 edition.

Shafaei-Bajestan, E., Tari, M. M., and Baayen, R. H. (2021). LDL-AURIS: Error-driven learning in modeling spoken word recognition. *Language, Cognition and Neuroscience*.

Shahmohammadi, H., Lensch, H., and Baayen, R. H. (2021). Learning zero-shot multifaceted visually grounded word embeddings via multi-task training. *arXiv preprint arXiv:2104.07500*.

Sinclair, J. (1991). *Corpus, concordance, collocation*. Oxford University Press.

Skousen, R. (1989). *Analogical Modeling of Language*. Kluwer, Dordrecht.

Skousen, R. (2002). *Analogical modeling*. Benjamins, Amsterdam.

Stump, G. (2001). *Inflectional Morphology: A Theory of Paradigm Structure*. Cambridge University Press.

Tognini-Bonelli, E. (2001). *Corpus linguistics at work*, volume 6. John Benjamins Publishing.

Tomaschek, F., Tucker, B. V., Fasiolo, M., and Baayen, R. H. (2018). Practice makes perfect: The consequences of lexical proficiency for articulation. *Linguistics Vanguard*, 4(s2).

Trimmer, P. C., McNamara, J. M., Houston, A. I., and Marshall, J. A. R. (2012). Does natural selection favour the Rescorla-Wagner rule? *Journal of Theoretical Biology*, 302:39–52.

Trommer, J. (2021). The subsegmental structure of german plural allomorphy. *Natural Language & Linguistic Theory*, 39(2):601–656.

Widrow, B. and Hoff, M. E. (1960). Adaptive switching circuits. *1960 WESCON Convention Record Part IV*, pages 96–104.

Wiese, R. (1999). On default rules and other rules. *Behavioral and brain sciences*, 22(6):1043–1044.

Wunderlich, D. (1999). German noun plural reconsidered. *Behavioral and Brain Sciences*, 22(6):1044–1045.

Yamada, I., Asai, A., Sakuma, J., Shindo, H., Takeda, H., Takefuji, Y., and Matsumoto, Y. (2020). Wikipedia2Vec: An efficient toolkit for learning and visualizing the embeddings of words and entities from Wikipedia. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing: System Demonstrations*, pages 23–30. Association for Computational Linguistics.

Yang, C. (2016). *The Price of Linguistic Productivity.* The MIT Press, Cambridge, MA.

Zaretsky, E. and Lange, B. P. (2015). No matter how hard we try: Still no default plural marker in nonce nouns in modern high german. In *A blend of MaLT: Selected contributions from the Methods and Linguistic Theories Symposium*, pages 153–178.

Zaretsky, E., Lange, B. P., Euler, H. A., and Neumann, K. (2013). Acquisition of german pluralization rules in monolingual and multilingual children. *Studies in Second Language Learning and Teaching*, 3(4):551–580.