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

R. Harald Baayen[1], Yu-Ying Chuang[1], Elnaz Shafaei-Bajestan[1], and James P. Blevins[2]

1: Seminar für Sprachwissenschaft, Eberhard-Karls University of Tübingen, Wilhelmstrasse 19, 72074 Tübingen, Germany
2: Homerton College, University of Cambridge, Hills Road, Cambridge CB2 8PH, UK.

November 16, 2018

Abstract

The discriminative lexicon is introduced as a mathematical and computational model of the mental lexicon. This novel theory is inspired by word and paradigm morphology, but operationalizes the concept of proportional analogy using the mathematics of linear algebra. It embraces the discriminative perspective on language, rejecting the idea that words' meanings are compositional in the sense of Frege and Russell, and arguing instead that the relation between form and meaning is fundamentally discriminative. The discriminative lexicon also incorporates the insight from machine learning that end-to-end modeling is much more effective than working with a cascade of models targeting individual subtasks. The computational engine at the heart of the discriminative lexicon is linear discriminative learning: Simple linear networks are used for mapping form onto meaning, and meaning onto form, without requiring the hierarchies of post-Bloomfieldian 'hidden' constructs such as phonemes, morphemes, and stems. We show that this novel model meets the criteria of accuracy (it properly recognizes words, and produces words correctly), productivity (the model is remarkably successful in understanding and producing novel complex words), and predictivity (it correctly predicts a wide array of experimental phenomena in lexical processing). The discriminative lexicon does not make use of static representations that are stored in memory and that have to be accessed in comprehension and production. It replaces static representations by states of the cognitive system that arise dynamically as a consequence of external or internal stimuli. The discriminative lexicon brings together visual and auditory comprehension as well as speech production into an integrated dynamic system of coupled linear networks.

keywords:

discriminative learning, morphology, linear mappings, mental lexicon, lexical processing

1 Introduction

Theories of language and language processing have a long history of taking inspiration from mathematics and computer science. For more than a century, formal logic has been influencing linguistics (Frege, 1879; Russell, 1942; Montague, 1973), and one of the most widely-known linguistic theories, generative grammar, has strong roots in the mathematics of formal languages (Chomsky, 1957;

Levelt, 2008). Similarly, Bayesian inference is currently seen as an attractive framework for understanding language processing (Kleinschmidt and Jaeger, 2015).

However, current advances in machine learning present linguistics with new challenges and opportunities. Methods across machine learning, such as random forests (Breiman, 2001; Strobl et al., 2007) and deep learning (Hannun et al., 2014; Tüske et al., 2014; Schmidhuber, 2015), offer unprecedented prediction accuracy. At the same time, these new approaches confront linguistics with deep fundamental questions, as these models typically are so-called end-to-end models that eschew representations for standard linguistic constructs such as phonemes, morphemes, syllables, and word forms.

There are mainly three possible responses to these new algorithms. A first response is to dismiss machine learning as irrelevant for understanding language and cognition. Given that machine learning algorithms currently outperform algorithms that make use of standard concepts from linguistics, this response is unlikely to be productive in the long run.

A second response is to interpret the units on the hidden layers of deep learning networks as capturing the representations and their hierarchical organization familiar from standard linguistic frameworks. The dangers inherent in this approach are well illustrated by the deep learning model proposed by Hannagan et al. (2014) for lexical learning in baboons (Grainger et al., 2012). The hidden layers of their deep learning network were claimed to correspond to areas in the ventral pathway of the primate brain. However, Scarf et al. (2016) reported that pigeons can also learn to discriminate between English words and nonword strings. Given that the avian brain is organized very differently from the primate brain, and yet accomplishes the same task, the claim that the layers of Hannagan et al.'s deep learning network correspond to areas in the ventral pathway must be premature. Furthermore, Linke et al. (2017) showed that baboon lexical learning can be modeled much more precisely by a two-layer wide learning network. It is also noteworthy that while for vision some form of hierarchical layering of increasingly specific receptive fields is well established (Hubel and Wiesel, 1962), it is unclear whether similar neural organization characterizes auditory comprehension and speech production.

A third response is to take the ground-breaking results from machine learning as a reason for rethinking, against the backdrop of linguistic domain knowledge, and at the functional level, the nature of the algorithms that underlie language learning and language processing.

The present study presents the results of an ongoing research program that exemplifies the third kind of response, narrowed down to the lexicon and focusing on those algorithms that play a central role in making possible the basics of visual and auditory comprehension as well as speech production. The model that we propose here brings together several strands of research across theoretical morphology, psychology, and machine learning. Our model can be seen as a computational implementation of paradigmatic analogy in word and paradigm morphology. From psychology, our model inherits the insight that classes and categories are constantly recalibrated as experience unfolds, and that this recalibration can be captured to a considerable extent by very simple principles of error-driven learning. We adopted end-to-end modeling from machine learning, but we kept our networks as simple as possible as the combination of smart features and linear mappings is surprisingly powerful (Baayen and Hendrix, 2017; Arnold et al., 2017). Anticipating discussion of technical details, we implement linear networks (mathematically, linear mappings) that are based entirely on discrimination as learning mechanism, and that work with large numbers of features at much lower levels of representation than in current and classical models.

Section 2 provides the theoretical background for this study, and introduces the central ideas of the discriminative lexicon that are implemented in subsequent sections. Section 3 introduces our operationalization of lexical semantics, section 4 discusses visual and auditory comprehension, Section 5 presents how we approach the modeling of speech production. This is followed by a brief

discussion of how time can be brought into the model (section 6). In the final section, we discuss the implications of our results.

2 Background

This section lays out some prerequisites that we will build on in the remainder of this study. We first introduce Word and Paradigm morphology, an approach to word structure developed in theoretical morphology within the broader context of linguistics. Word and Paradigm morphology provides the background for our highly critical evaluation of the morpheme as theoretical unit. The many problems associated with morphemes motivated the morpheme-free computational model developed below. Subsection 2.2 introduces the idea of vector representations for word meanings, as developed within computational linguistics and computer science. Semantic vectors lie at the heart of how we model word meaning. The next subsection, 2.3, explains how we calculated the semantic vectors that we used in this study. Subsection 2.4 discusses naive discriminative learning, and explains why, when learning the relation between form vectors and semantic vectors, it is advantageous to use linear discriminative learning instead of naive discriminative learning. The last subsection explains how linear discriminative learning provides a mathematical formalization of the notion of proportional analogy that is central to Word and Paradigm morphology.

2.1 Word and Paradigm morphology

The dominant view of the mental lexicon in psychology and cognitive science is well represented by Zwitserlood (2018, p. 583):

> Parsing and composition — for which there is ample evidence from many languages — require morphemes to be stored, in addition to information as to how morphemes are combined, or to whole-word representations specifying the combination.

Words are believed to be built from morphemes, either by rules or by constructional schemata (Booij, 2010), and the meanings of complex words are assumed to be a compositional function of the meanings of their parts.

This perspective on word structure, which has found its way into almost all introductory textbooks on morphology, has its roots in post-Bloomfieldian American structuralism (Blevins, 2006). However, many subsequent studies have found this perspective to be inadequate (Stump, 2001). Beard (1977) pointed out that in language change, morphological form and morphological meaning follow their own trajectories, and the theoretical construct of the morpheme as a minimal sign combining form and meaning therefore stands in the way of understanding the temporal dynamics of language. Before him, Matthews (1974, 1991) had pointed out that the inflectional system of Latin is not well served by analyses positing that its fusional system is best analyzed as underlyingly agglutinative (i.e, as a morphological system in which words consist of sequences of morphemes, as (approximately) in Turkish; see also Hockett, 1960). Matthews argued that words are the basic units, and that proportional analogies between words within paradigms (e.g., *walk*:*walks* = *talk*:*talks*) make the lexicon as a system productive. By positing the word as basic unit, Word and Paradigm morphology avoids a central problem that confronts morpheme-based theories, namely, that systematicities in form can exist without corresponding systematicities in meaning. Minimal variation in form can serve to distinguish words or to predict patterns of variation elsewhere in a paradigm or class. One striking example is given by the locative cases in Estonian, which express meanings that in English would be realized with locative and directional prepositions. Interestingly, most plural case endings in Estonian are built on the form of the partitive singular (a grammatical case that

can mark for instance the direct object). However, the semantics of these plural case forms do not express in any way the semantics of the singular and the partitive. For instance, the form for the partitive singular of the noun for 'leg' is *jalga*. The form for expressing 'on the legs', *jalgadele*, takes the form of the partitive singular and adds the formatives for plural and the locative case for 'on', the so-called adessive (see Erelt, 2003; Blevins, 2006, for further details). Matthews (1993, p. 92) characterized the adoption by Chomsky and Halle (1968) of the morpheme as 'a remarkable tribute to the inertia of ideas', and the same characterization pertains to the adoption of the morpheme by mainstream psychology and cognitive science. In the present study, we adopt from Word and Paradigm morphology the insight that the word is the basic unit of analysis. Where we take Word and Paradigm morphology a step further is in how we operationalize proportional analogy. As will be laid out in more detail below, we construe analogies as being system-wide rather than constrained to paradigms, and our mathematical formalization better integrates semantic analogies with formal analogies.

2.2 Distributional semantics

The question that arises at this point is what word meanings are. In this study, we adopt the approach laid out by Landauer and Dumais (1997), and approximate word meanings by means of semantic vectors, referred to as embeddings in computational linguistics. Weaver (1955) and Firth (1968) noted that words with similar distributions tend to have similar meanings. This intuition can be formalized by counting how often words co-occur across documents or within some window of text. In this way, a word's meaning comes to be represented by a vector of reals, and the semantic similarity between two words is evaluated by means of the similarities of their vectors. One such measure is the cosine similarity of the vectors, a related measure is the Pearson correlation of the vectors. Many implementations of the same general idea are available, such as HAL (Lund and Burgess, 1996), HiDEx (Shaoul and Westbury, 2010), and WORD2VEC (Mikolov et al., 2013). Below, we provide further detail on how we estimated semantic vectors.

There are two ways in which semantic vectors can be conceptualized within the context of theories of the mental lexicon. First, semantic vectors could be fixed entities that are stored in memory in a way reminiscent of a standard printed or electronic dictionary, the entries of which consist of a search key (a word's form), and a meaning specification (the information accessed through the search key). This conceptualization is very close to the currently prevalent way of thinking in psycholinguistics, which has adopted a form of naive realism in which word meanings are typically associated with monadic concepts (see, e.g. Roelofs, 1997; Levelt et al., 1999; Taft, 1994; Schreuder and Baayen, 1995). It is worth noting that when one replaces these monadic concepts by semantic vectors, the general organization of (paper and electronic) dictionaries can still be retained, resulting in research questions addressing the nature of the access keys (morphemes, whole-words, perhaps both), and the process of accessing these keys.

However, there is good reason to believe that word meanings are not fixed, static representations. The literature on categorization indicates that the categories (or classes) that arise as we interact with our environment are constantly recalibrated (Love et al., 2004; Marsolek, 2008; Ramscar and Port, 2015). A particularly eloquent example is given by Marsolek (2008). In a picture naming study, he presented subjects with sequences of two pictures. He asked subjects to say aloud the name of the second picture. The critical manipulation concerned the similarity of the two pictures. When the first picture was very similar to the second (e.g., a grand piano and a table), responses were slower compared to the control condition in which visual similarity was reduced (orange and table). What we see at work here is error-driven learning: when understanding the picture of a grand piano as signifying a grand piano, features such as having a large flat surface are strengthened

to the grand piano, and weakened to the table. As a consequence, interpreting the picture of a table has become more difficult, which in turn slows the word naming response.

2.3 Discrimination learning of semantic vectors

What is required, therefore, is a conceptualization of word meanings that allows for the continuous recalibration of these meanings as we interact with the world. To this end, we construct semantic vectors using discrimination learning. We take the sentences from a text corpus, and, for each sentence, in the order in which these sentences occur in the corpus, train a linear network to predict the words in that sentence from the words in that sentence. The training of the network is accomplished with a simplified version of the learning rule of Rescorla and Wagner (1972), basically the learning rule of Widrow and Hoff (1960). We denote the connection strength from input word i to output word j by w_{ij}. Let δ_i denote an indicator variable that is 1 when input word i is present in sentence t, and zero otherwise. Likewise, let δ_j denote whether output word j is present in the sentence. Given a learning rate ρ ($\rho \ll 1$), the change in the connection strength from (input) word i to (output) word j for sentence (corpus time) t, Δ_{ij}, is given by

$$\Delta_{ij} = \delta_i \rho (\delta_j - \sum_k \delta_k w_{kj}), \tag{1}$$

where δ_i and δ_j vary from sentence to sentence, depending on which cues and outcomes are present in a sentence. Given n distinct words, an $n \times n$ network is updated incrementally in this way. The row vectors of the network's weight matrix $\boldsymbol{S}$ define words' semantic vectors. Below, we return in more detail to how we derived the semantic vectors used in the present study. Importantly, we now have semantic representations that are not fixed but dynamic: they are constantly recalibrated from sentence to sentence.[1] Thus, in what follows, a word's meaning is defined as a semantic vector that is a function of time: it is subject to continuous change. By itself, without context, a word's semantic vector at time t reflects its 'entanglement' with other words.

Obviously, it is extremely unlikely that our understanding of the classes and categories that we discriminate between are well approximated just by lexical co-occurrence statistics. However, we will show that text-based semantic vectors are good enough as a first approximation for implementing a computationally tractable discriminative lexicon.

Returning to the learning rule (1), we note that it is well-known to capture important aspects of the dynamics of discrimination learning (Trimmer et al., 2012; Miller et al., 1995). For instance, it captures the blocking effect (Kamin, 1969; Rescorla and Wagner, 1972; Rescorla, 1988) as well as order effects in learning (Ramscar et al., 2010). The blocking effect is the finding that when one feature has been learned to perfection, adding a novel feature does not improve learning. This follows from (1): As w_{ij} tends towards 1, Δ_{ij} tends towards 0. A novel feature, even though by itself it is perfectly predictive, is blocked from developing a strong connection weight of its own. The order effect has been observed for, e.g., the learning of words for colors. Given objects of fixed shape and size but varying color, and the corresponding color words, it is essential that the objects are presented to the learner before the color word. This ensures that the object's properties become the features for discriminating between the color words. In this case, application of the learning rule (1) will result in a strong connection between the color feature of the object to the appropriate

[1] The property of semantic vectors changing over time as experience unfolds is not unique to the present approach, but is shared with other algorithms for constructing semantic vectors such as word2vec. However, time-variant semantic vectors are in stark contrast to the monadic units representing meanings in models such as proposed by Baayen et al. (2000), Taft (1988), Levelt et al. (1999). The distributed representation for words' meanings in the triangle model (Harm and Seidenberg, 2004) were derived from WordNet and hence are also time-invariant.

color word. If the order is reversed, the color words become features predicting object properties. In this case, application of (1) will result in weights from a color word to object features that are proportional to the probabilities of the object's features in the training data.

Given the dynamics of discriminative learning, static lexical entries are not useful. Anticipating more detailed discussion below, we argue that actually the whole dictionary metaphor of lexical access is misplaced, and that in comprehension meanings are dynamically created from form, and that in production forms are dynamically created from meanings. Importantly, what forms and meanings are created will vary from case to case, as all learning is fundamentally incremental and subject to continuous recalibration. In order to make the case that this is actually possible and computationally feasible, we first introduce the framework of naive discriminative learning, discuss its limitations, and then lay out how this approach can be enhanced to meet the goals of this study.

2.4 Naive discriminative learning

Naive discriminative learning (NDL) is a computational framework, grounded in learning rule (1), that was inspired by prior work on discrimination learning (e.g., Ramscar et al., 2010; Ramscar and Yarlett, 2007). A model for reading complex words was introduced by Baayen et al. (2011). This study implemented a two-layer linear network with letter pairs as input features (henceforth cues), and units representing meanings as output classes (henceforth outcomes). Following Word and Paradigm morphology, this model does not include form representations for morphemes or word forms. It is an end-to-end model for the relation between form and meaning, that set itself the task to predict word meanings from sublexical orthographic features. Baayen et al. (2011) were able to show that the extent to which cues in the visual input support word meanings, as gauged by the activation of the corresponding outcomes in the network, reflects a wide range of phenomena reported in the lexical processing literature, including the effects of frequency of occurrence, family size (Moscoso del Prado Martín et al., 2004), relative entropy (Milin et al., 2009), phonaesthemes (non-morphemic sounds with a consistent semantic contribution such as *gl* in *glimmer*, *glow*, *glisten*) (Bergen, 2004), and morpho-orthographic segmentation (Milin et al., 2017b).

However, the model of Baayen et al. (2011) adopted a stark form of naive realism, albeit just for reasons of computational convenience: A word's meaning was defined in terms of the presence or absence of an outcome (δ_j in equation 1). Subsequent work sought to address this shortcoming by developing semantic vectors, using learning rule (1) to predict words from words, as explained above (Baayen et al., 2016a; Milin et al., 2017b). The term "lexome" was introduced as a technical term for the outcomes in a form-to-meaning network, conceptualized as pointers to semantic vectors.

In this approach, lexomes do double duty. In a first network, lexomes are the outcomes that the network is trained to discriminate between given visual input. In a second network, the lexomes are the 'atomic' units in a corpus that serve as the input for building a distributional model with semantic vectors. Within the theory unfolded in this study, the term lexome refers only to the elements from which a semantic vector space is constructed. The dimension of this vector space is equal to the number of lexomes, and each lexome is associated with its own semantic vector.

It turns out, however, that mathematically a network discriminating between lexomes given orthographic cues, as in naive discriminative learning models, is suboptimal. In order to explain why the set-up of NDL is sub-optimal, we need some basic concepts and notation from linear algebra, such as matrix multiplication and matrix inversion. Readers unfamiliar with these concepts are referred to Appendix A, which provides an informal introduction.

2.4.1 Limitations of naive discriminative learning

Mathematically, naive discrimination learning works with two matrices, a cue matrix $\boldsymbol{C}$ that specifies words' form features, and a matrix $\boldsymbol{S}$ that specifies the targeted lexomes (Sering et al., 2018). We illustrate the cue matrix for four words, *aaa, aab, abb* and *abb* and their letter bigrams *aa, ab, bb*. A cell c_{ij} is 1 if word i has bigram j, and zero otherwise.

$$\boldsymbol{C} = \begin{array}{c} \\ \text{aaa} \\ \text{aab} \\ \text{abb} \\ \text{abb} \end{array} \begin{array}{c} \begin{array}{ccc} \text{aa} & \text{ab} & \text{bb} \end{array} \\ \begin{pmatrix} 1 & 0 & 0 \\ 1 & 1 & 0 \\ 0 & 1 & 1 \\ 0 & 1 & 1 \end{pmatrix} \end{array}. \tag{2}$$

The third and fourth words are homographs: they share exactly the same bigrams. We next define a target matrix $\boldsymbol{S}$ that defines for each of the four words the corresponding lexome λ. This is done by setting one bit to 1 and all other bits to zero in the row vectors of this matrix:

$$\boldsymbol{S} = \begin{array}{c} \\ \text{aaa} \\ \text{aab} \\ \text{abb} \\ \text{abb} \end{array} \begin{array}{c} \begin{array}{cccc} \lambda_1 & \lambda_2 & \lambda_3 & \lambda_4 \end{array} \\ \begin{pmatrix} 1 & 0 & 0 & 0 \\ 0 & 1 & 0 & 0 \\ 0 & 0 & 1 & 0 \\ 0 & 0 & 0 & 1 \end{pmatrix} \end{array}. \tag{3}$$

The problem that arises at this point is that the word forms jointly set up a space with three dimensions, whereas the targeted lexomes set up a four-dimensional space. This is problematic for the following reason. A linear mapping of a lower-dimensional space $\mathcal{A}$ onto a higher-dimensional space $\mathcal{B}$ will result in a subspace of $\mathcal{B}$ with a dimensionality that cannot be greater than that of $\mathcal{A}$ (Kaye and Wilson, 1998). If $\mathcal{A}$ is a space in $\mathbb{R}^2$ and $\mathcal{B}$ is a space in $\mathbb{R}^3$, a linear mapping of $\mathcal{A}$ onto $\mathcal{B}$ will result in a plane in $\mathcal{B}$. All the points in $\mathcal{B}$ that are not on this plane cannot be reached from $\mathcal{A}$ with the linear mapping.

As a consequence, it is impossible for NDL to perfectly discriminate between the four lexomes (which set up a four-dimensional space) given their bigram cues (which jointly define a three-dimensional space). For the input `bb`, the network therefore splits its support equally over the lexomes λ_3 and λ_4. The transformation matrix $\boldsymbol{F}$ is

$$\boldsymbol{F} = \boldsymbol{C}'\boldsymbol{S} = \begin{array}{c} \\ \text{aa} \\ \text{ab} \\ \text{bb} \end{array} \begin{array}{c} \begin{array}{cccc} \lambda_1 & \lambda_2 & \lambda_3 & \lambda_4 \end{array} \\ \begin{pmatrix} 1 & 0 & 0 & 0 \\ -1 & 1 & 0 & 0 \\ 1 & -1 & 0.5 & 0.5 \end{pmatrix} \end{array},$$

and the matrix with predicted vectors $\hat{\boldsymbol{S}}$ is

$$\hat{\boldsymbol{S}} = \boldsymbol{C}\boldsymbol{F} = \begin{array}{c} \\ \text{aaa} \\ \text{aab} \\ \text{abb} \\ \text{abb} \end{array} \begin{array}{c} \begin{array}{cccc} \lambda_1 & \lambda_2 & \lambda_3 & \lambda_4 \end{array} \\ \begin{pmatrix} 1 & 0 & 0 & 0 \\ 0 & 1 & 0 & 0 \\ 0 & 0 & 0.5 & 0.5 \\ 0 & 0 & 0.5 & 0.5 \end{pmatrix} \end{array}.$$

When a cue matrix $\boldsymbol{C}$ and a target matrix $\boldsymbol{S}$ are set up for large numbers of words, the dimension of the NDL cue space will be substantially smaller than the space of $\boldsymbol{S}$, the dimension of which is

equal to the number of lexomes. Thus, in the set-up of naive discriminative learning, we do take into account that words tend to have similar forms, but we do *not* take into account that words are also similar in meaning. By setting up $\boldsymbol{S}$ as completely orthogonal, we are therefore making it unnecessarily hard to relate form to meaning.

This study therefore implements discrimination learning with semantic vectors of reals replacing the one-bit-on row vectors of the target matrix $\boldsymbol{S}$. By doing so, we properly reduce the dimensionality of the target space, which in turn makes discriminative learning more accurate. We will refer to this new implementation of discrimination learning as *linear* discriminative learning (LDL) instead of *naive* discriminative learning, as the outcomes are no longer assumed to be independent and the networks are mathematically equivalent to linear mappings onto continuous target matrices.

2.5 Linear transformations and proportional analogy

A central concept in word and paradigm morphology is that the form of a regular inflected form stands in a relation of proportional analogy to other words in the inflectional system. As explained by Matthews (1991, p. 192f), :

> In effect, we are predicting the inflections of servus by analogy with those of dominus. As Genitive Singular domini is to Nominative Singular dominus, so x (unknown) must be to Nominative Singular servus. What then is x? Answer: it must be servi. In notation, dominus domini = servus servi. (Matthews 1991: 192f)

Here, form variation is associated with 'morphosyntactic features'. Such features are often naively assumed to function as proxies for a notion of 'grammatical meaning'. However, in word and paradigm morphology, these features actually represent something more like distribution classes. For example, the accusative singular forms of nouns belonging to different declensions will typically differ in form and be identified as the 'same' case in different declensions by virtue of distributional parallels. The similarity in meaning then follows from the distributional hypothesis, which proposes that linguistic items with similar distributions have similar meanings (Weaver, 1955; Firth, 1968). Thus, the analogy of forms

$$\text{dominus} : \text{domini} = \text{servus} : \text{servi}.$$

is paralleled by an analogy of distributions d:

$$d(\text{dominus}) : d(\text{domini}) = d(\text{servus}) : d(\text{servi}).$$

Borrowing notational conventions of matrix algebra, we can write

$$\begin{pmatrix} \text{dominus} \\ \text{domini} \\ \text{servus} \\ \text{servi} \end{pmatrix} \sim \begin{pmatrix} d(\text{dominus}) \\ d(\text{domini}) \\ d(\text{servus}) \\ d(\text{servi}) \end{pmatrix}. \tag{4}$$

In this study, we operationalize the proportional analogy of word and paradigm morphology by replacing the word forms in (4) by vectors of features that are present or absent in a word, and by replacing words' distributions by semantic vectors. Linear mappings between form and meaning formalize two distinct proportional analogies, one analogy for going from form to meaning, and a second analogy for going from meaning to form. Consider, for instance, a semantic matrix $\boldsymbol{S}$ and a

form matrix $\boldsymbol{C}$, with rows representing words:

$$\begin{aligned}\boldsymbol{S} &= \begin{pmatrix} a_1 & a_2 \\ b_1 & b_2 \end{pmatrix} \\ \boldsymbol{C} &= \begin{pmatrix} p_1 & p_2 \\ q_1 & q_2 \end{pmatrix}\end{aligned}$$

The transformation matrix $\boldsymbol{G}$ that maps $\boldsymbol{S}$ onto $\boldsymbol{C}$,

$$\underbrace{\frac{1}{a_1b_2 - a_2b_1}\begin{pmatrix} b_2 & -a_2 \\ -b_1 & a_1 \end{pmatrix}}_{\boldsymbol{S}^{-1}}\underbrace{\begin{pmatrix} p_1 & p_2 \\ q_2 & q_2 \end{pmatrix}}_{\boldsymbol{C}} = \underbrace{\frac{1}{a_1b_2 - a_2b_1}\left[\begin{pmatrix} b_2p_1 & b_2p_2 \\ -b_1p_1 & -b_1p_2 \end{pmatrix} + \begin{pmatrix} -a_2q_1 & -a_2q_2 \\ a_1q_1 & a_1q_2 \end{pmatrix}\right]}_{\boldsymbol{G}},$$

can be written as the sum of two matrices (both of which are scaled by $a_1b_2 - a_2b_1$). The first matrix takes the first form vector of $\boldsymbol{C}$ and weights it by the elements of the second semantic vector. The second matrix takes the second form vector of $\boldsymbol{C}$, and weights this by the elements of the first semantic vector of $\boldsymbol{S}$. Thus, with this linear mapping, a predicted form vector is a semantically weighted mixture of the form vectors of $\boldsymbol{C}$.

The remainder of this paper is structured as follows. We first introduce the semantic vector space $\boldsymbol{S}$ that we will be using, which we derived from the TASA corpus (Ivens and Koslin, 1991; Landauer et al., 1998). Next, we show that we can model word comprehension by means of a matrix (or network) $\boldsymbol{F}$ that transforms the cue row vectors of $\boldsymbol{C}$ into the semantic row vectors of $\boldsymbol{S}$, i.e., $\boldsymbol{CF} = \boldsymbol{S}$. We then show that we can model the production of word forms given their semantics by a transformation matrix $\boldsymbol{G}$, i.e., $\boldsymbol{SG} = \boldsymbol{C}$, where $\boldsymbol{C}$ is a matrix specifying for each word which triphones it contains. As the network is informative only about which triphones are at issue for a word, but remains silent about their order, an algorithm building on graph theory is used to properly order the triphones. All models are evaluated on their accuracy, and are also tested against experimental data.

In the general discussion, we will reflect on the consequences for linguistic theory of our finding that it is indeed possible to model the lexicon and lexical processing using systemic discriminative learning, without requiring 'hidden units' such as morphemes and stems.

3 A semantic vector space derived from the TASA corpus

The first part of this section describes how we constructed semantic vectors. Specific to our approach is that we constructed semantic vectors not only for content words, but also for inflectional and derivational functions. The semantic vectors for these functions are of especial importance for the speech production component of the model. The second part of this section addresses the validation of the semantic vectors, for which we used paired associate learning scores, semantic relatedness ratings, as well as semantic plausibility and semantic transparency ratings for derived words.

3.1 Constructing the semantic vector space

The semantic vector space that we use in this study is derived from the TASA corpus (Ivens and Koslin, 1991; Landauer et al., 1998). We worked with 752,130 sentences from this corpus, to a total of 10,719,386 word tokens. We used the `treetagger` software (Schmid, 1995) to obtain for each word token its stem and a part of speech tag. Compounds written with a space, such as *apple pie* and *jigsaw puzzle* were, when listed in the CELEX lexical database (Baayen et al., 1995), joined into

one onomasiological unit. Information about words' morphological structure was retrieved from CELEX.

In computational morphology, several options have been explored for representing the meaning of affixes. Mitchell and Lapata (2008) proposed to model the meaning of an affix as a vector that, when added to the vector $\boldsymbol{b}$ of a base word, gives the vector $\boldsymbol{d}$ of the derived word. They estimated this vector by calculating $\boldsymbol{d}_i - \boldsymbol{b}_i$ for all available pairs i of base and derived words, and taking the average of the resulting vectors. Lazaridou et al. (2013) and Marelli and Baroni (2015) modeled the semantics of affixes by means of matrix operations that take $\boldsymbol{b}$ as input and produce $\boldsymbol{d}$ as output, i.e.,

$$\boldsymbol{d} = \boldsymbol{M}\boldsymbol{b}, \tag{5}$$

whereas Cotterell et al. (2016) constructed semantic vectors as latent variables in a Gaussian graphical model. What these approaches have in common is that they derive or impute semantic vectors given the semantic vectors of words produced by algorithms such as `word2vec`, algorithms which work with (stemmed) words as input units.

From the perspective of discriminative linguistics, however, it does not make sense to derive one meaning from another. Furthermore, rather than letting writing conventions dictate what units are accepted as input to one's favorite tool for constructing semantic vectors, including *not* and *again* but excluding the prefixes *un-*, *in-* and *re-*, we included not only lexomes for content words but also lexomes for prefixes and suffixes as units for constructing a semantic vector space. As a result, semantic vectors for prefixes and suffixes are obtained straightforwardly together with semantic vectors for content words.

To illustrate this procedure, consider the sentence *the boys' happiness was great to observe*. Standard methods will apply stemming and remove stop words, resulting in the lexomes BOY, HAPPINESS, GREAT, OBSERVE being the input to the algorithm constructing semantic vectors from lexical co-occurrences. In our approach, by contrast, the lexomes considered are THE, BOY, HAPPINESS, BE, GREAT, TO, OBSERVE, and in addition PLURAL, NESS, PAST. Stop words are retained,[2] inflectional endings are replaced by the inflectional functions they subserve, and for derived words, the semantic function of the derivational suffix is identified as well. In what follows, we outline in more detail what considerations motivate this way of constructing the input for our algorithm constructing semantic vectors. We note here that our method for handling morphologically complex words can be combined with any of the currently available algorithms for building semantic vectors. We also note here that we are not concerned with the question of how lexomes for plurality or tense might be induced from word forms, from world knowledge, or any combination of the two. All we assume is, first, that anyone understanding the sentence *the boys' happiness was great to observe* knows that more than one boy is involved, and that the narrator situates the event in the past. Second, we take this understanding to drive learning.

We distinguished the following seven inflectional functions: COMPARATIVE and SUPERLATIVE for adjectives, SINGULAR and PLURAL for nouns, and PAST, PERFECTIVE, CONTINUOUS, PERSISTENCE[3], and PERSON3 (third person singular) for verbs.

The semantics of derived words tend to be characterized by idiosyncracies (Zeller et al., 2014) that can be accompanied by formal idiosyncracies (e.g., *business* and *resound*) but often are form-

[2] Although including function words may seem counterproductive from the perspective of semantic vectors in natural language processing applications, they are retained in the present approach for two reasons. First, since in our model, semantic vectors drive speech production, in order to model the production of function words, semantic vectors for function words are required. Second, although the semantic vectors of function words typically are less informative (they typically have small association strengths to very large numbers of words), they still show structure, as illustrated in Appendix C for pronouns and prepositions.

[3] PERSISTENCE denotes the persistent relevance of the predicate in sentences such as *London is a big city*.

ally unremarkable (e.g., *heady*, with meanings as diverse as intoxicating, exhilerating, impetuous, and prudent). We therefore paired derived words with their own content lexomes, as well as with a lexome for the derivational function expressed in the derived word. We implemented the following derivational lexomes: ORDINAL for ordinal numbers, NOT for negative *un* and *in*, UNDO for reversative *un*, OTHER for non-negative *in* and its allomorphs, ION for *ation, ution, ition, ion*, EE for *ee*, AGENT for agents with *er*, INSTRUMENT for instruments with *er*, IMPAGENT for impersonal agents with *er*, and CAUSER for words with *er* expressing causers (the differentiation of different semantic functions for *er* was based on manual annotation of the list of forms with *er*; the assignment of a given form to a category was not informed by sentence context and hence can be incorrect), AGAIN for *re*, and NESS, ITY, ISM, IST, IC, ABLE, IVE, OUS, IZE, ENCE, FUL, ISH, UNDER, SUB, SELF, OVER, OUT, MIS, DIS for the corresponding prefixes and suffixes. This set-up of lexomes is informed by the literature on the semantics of English word formation (Marchand, 1969; Bauer, 1983; Plag, 2003), and targets affixal semantics irrespective of affixal forms (following Beard, 1995).

This way of setting up the lexomes for the semantic vector space model illustrates an important aspect of the present approach, namely, that lexomes target what is understood, and not particular linguistic forms. Lexomes are not units of form. For example, in the study of Geeraert et al. (2017), the forms *die, pass away*, and *kick the bucket* are all linked to the same lexome DIE. In the present study, we have not attempted to identify idioms, and likewise, no attempt was made to disambiguate homographs. These are targets for further research.

In the present study, we did implement the classical distinction between inflection and word formation by representing inflected words with a content lexome for their stem, but derived words with a content lexome for the derived word itself. In this way, the sentence *scientists believe that exposure to shortwave ultraviolet rays can cause blindness* is associated with the following lexomes: SCIENTIST, PL, BELIEVE, PERSISTENCE, THAT, EXPOSURE, SG, TO, SHORTWAVE, ULTRAVIOLET, RAY, CAN, PRESENT, CAUSE, BLINDNESS, NESS. In our model, sentences constitute the learning events, i.e., for each sentence, we train the model to predict the lexomes in that sentence from the very same lexomes in that sentence. Sentences, or for spoken corpora, utterances, are more ecologically valid than windows of say two words preceeding and two words following a target word — the interpretation of a word may depend on the company that it keeps at long distances in a sentence. We also did not remove any function words, contrary to standard practice in distributional semantics (see Appendix C for a heatmap illustrating the kind of clustering observable for pronouns and prepositions). The inclusion of semantic vectors for both affixes and function words is necessitated by the general framework of our model, which addresses not only comprehension but also production, and which in the case of comprehension needs to move beyond lexical identification, as inflectional functions play an important role during the integration of words in sentence and discourse.

Only words with a frequency exceeding 8 occurrences in the TASA corpus were assigned lexomes. This threshold was imposed in order to avoid excessive data sparseness when constructing the distributional vector space. The number of different lexomes that met the criterion for inclusion was 23,562.

We constructed a semantic vector space by training an NDL network on the TASA corpus, using the **ndl2** package for R (Shaoul et al., 2015). Weights on lexome-to-lexome connections were recalibrated sentence by sentence, in the order in which they appear in the TASA corpus, using the learning rule of NDL (i.e., a simplified version of the learning rule of Rescorla and Wagner (1972), that has only two free parameters, the maximum amount of learning λ (set to 1) and a learning rate ρ (set to 0.001). This resulted in a 23,562 $\times$ 23,562 matrix. Sentence-based training keeps the carbon footprint of the model down, as the number of learning events is restricted to the number of utterances (approximately 750,000) rather than the number of word tokens (approximately 10,000,000). The row vectors of the resulting matrix, henceforth $\boldsymbol{S}$, are the semantic vectors that

we will use in the remainder of this study.[4]

The weights on the main diagonal of the matrix tend to be high, unsurprisingly, as in each sentence on which the model is trained, each of the words in that sentence is an excellent predictor of that same word occurring in that sentence. When the focus is on semantic similarity, it is useful to set the main diagonal of this matrix to zero. However, for more general association strengths between words, the diagonal elements are informative and should be retained.

Unlike standard models of distributional semantics, we did not carry out any dimension reduction, using, for instance, singular value decomposition. Because we do not work with latent variables, each dimension of our semantic space is given by a column vector of $\boldsymbol{S}$, and hence each dimension is linked to a specific lexome. Thus, the row vectors of $\boldsymbol{S}$ specify, for each of the lexomes, how well this lexome discriminates between all the lexomes known to the model.

Although semantic vectors of length 23,562 can provide good results, vectors of this length can be challenging for statistical evaluation. Fortunately, it turns out that many of the column vectors of $\boldsymbol{S}$ are characterized by extremely small variance. Such columns can be removed from the matrix without loss of accuracy. In practice, we have found it suffices to work with approximately 4 to 5 thousand columns, selected calculating column variances and using only those columns with a variance exceeding a threshold value.

3.2 Validation of the semantic vector space

As shown by Baayen et al. (2016a); Milin et al. (2017b,a), measures based on matrices such as $\boldsymbol{S}$ are predictive for behavioral measures such as reaction times in the visual lexical decision task, as well as for self-paced reading latencies. In what follows, we first validate the semantic vectors of $\boldsymbol{S}$ on two data sets, one data set with accuracies in paired associate learning, and one dataset with semantic similarity ratings. We then considered specifically the validity of semantic vectors for inflectional and derivational functions, by focusing first on the correlational structure of the pertinent semantic vectors, followed by an examination of the predictivity of the semantic vectors for semantic plausibility and semantic transparency ratings for derived words.

3.2.1 Paired associate learning

The paired associate learning (PAL) task is a widely used test in psychology for evaluating learning and memory. Participants are given a list of word pairs to memorize. Subsequently, at testing, they are given the first word and have to recall the second word. The proportion of correctly recalled words is the accuracy measure on the PAL task. Accuracy on the PAL test decreases with age, which has been attributed to cognitive decline over the lifetime. However, Ramscar et al. (2014) and Ramscar et al. (2017) provide evidence that the test actually measures the accumulation of lexical knowledge. In what follows, we use the data on PAL performance reported by desRosiers and Ivison (1988). We fitted a linear mixed model to accuracy in the PAL task as a function of the Pearson correlation r of paired words' semantic vectors in $\boldsymbol{S}$ (but with weights on the main diagonal included, and using the 4275 columns with highest variance), with random intercepts for word pairs, sex and age as control variables, and crucially, an interaction of r by age. Given the findings of Ramscar and colleagues, we expect to find that the slope of r (which is always negative) as a predictor of PAL accuracy increases with age (indicating decreasing accuracy). Table 1 shows that this prediction is born out. For age group 20–29 (the reference level of age), the slope for r is

[4] Work is in progress to further enhance the $\boldsymbol{S}$ matrix by including word sense disambiguation and named entity recognition when setting up lexomes. The resulting lexomic version of the TASA corpus, and scripts (in python, as well as in R) for deriving the $\boldsymbol{S}$ matrix will be made available at http://www.sfs.uni-tuebingen.de/~hbaayen/.

estimated at 0.31. For the next age level, this slope is adjusted upward by 0.12, and as age increases these upward adjustments likewise increase to 0.21, 0.24, and 0.32 for age groups 40–49, 50-59, and 60-69 respectively. Older participants know their language better, and hence are more sensitive to the semantic similarity (or lack thereof) of the words that make up PAL test pairs. For the purposes of the present study, the solid support for r as a predictor for PAL accuracy contributes to validating the row vectors of $\boldsymbol{S}$ as semantic vectors.

A. parametric coefficients	Estimate	Std. Error	t-value	p-value
intercept	3.6220	0.6822	5.3096	< 0.0001
r	0.3058	0.1672	1.8293	0.0691
age=39	0.2297	0.1869	1.2292	0.2207
age=49	0.4665	0.1869	2.4964	0.0135
age=59	0.6078	0.1869	3.2528	0.0014
age=69	0.8029	0.1869	4.2970	< 0.0001
sex=male	-0.1074	0.0230	-4.6638	< 0.0001
r:age=39	0.1167	0.0458	2.5490	0.0117
r:age=49	0.2090	0.0458	4.5640	< 0.0001
r:age=59	0.2463	0.0458	5.3787	< 0.0001
r:age=69	0.3239	0.0458	7.0735	< 0.0001
B. smooth terms	edf	Ref.df	F-value	p-value
random intercepts word pair	17.8607	18.0000	128.2283	< 0.0001

Table 1: Linear mixed model fit to paired associate learning scores with the correlation r of the row vectors in $\boldsymbol{S}$ of the paired words as predictor. Treatment coding was used for factors. For the youngest age group, r is not predictive, but all other age groups show increasingly large slopes compared to the slope of the youngest age group. The range of r is $[-4.88, -2.53]$, hence larger values for the coefficients of r and its interactions imply worse performance on the PAL task.

3.2.2 Semantic relatedness ratings

We also examined performance of $\boldsymbol{S}$ on the MEN test collection (Bruni et al., 2014), that provides for 3000 word pairs crowd sourced ratings of semantic similarity. For 2267 word pairs, semantic vectors are available in $\boldsymbol{S}$ for both words. Figure 1 shows that there is a nonlinear relation between the correlation of words' semantic vectors in $\boldsymbol{S}$ and the ratings. The plot shows as well that for low correlations, the variability in the MEN ratings is larger. We fitted a Gaussian location scale additive model to this data set, summarized in Table 2, which supported r as a predictor for both mean MEN rating and the variability in the MEN ratings.

To put this performance in perspective, we collected the latent semantic analysis (LSA) similarity scores for the MEN word pairs using the website at `http://lsa.colorado.edu/`. The Spearman correlation for LSA scores and MEN ratings was 0.697, and that for r was 0.704 (both $p < 0.0001$). Thus, our semantic vectors perform on a par with those of LSA, a well established older technique that still enjoys wide use in psycholinguistics. Undoubtedly, optimized techniques from computational linguistics such as `word2vec` will outperform our model. The important point here is that even with training on full sentences rather than using small windows, and even when including function words, performance of our linear network with linguistically motivated lexomes is sufficiently high to serve as a basis for further analysis.

A. parametric coefficients	Estimate	Std. Error	t-value	p-value
Intercept [location]	25.2990	0.1799	140.6127	< 0.0001
Intercept [scale]	2.1297	0.0149	143.0299	< 0.0001
B. smooth terms	edf	Ref.df	F-value	p-value
TPRS r [location]	8.3749	8.8869	2766.2399	< 0.0001
TPRS r [scale]	5.9333	7.0476	87.5384	< 0.0001

Table 2: Summary of a Gaussian location scale additive model fitted to the similarity ratings in the MEN dataset, with as predictor the correlation r of word's semantic vectors in the $\boldsymbol{S}$ matrix. TPRS: thin plate regression spline. b: the minimum standard deviation for the logb link function. Location: parameter estimating the mean; scale: parameter estimating the variance through the logb link function ($\eta = \log(\sigma - b)$).

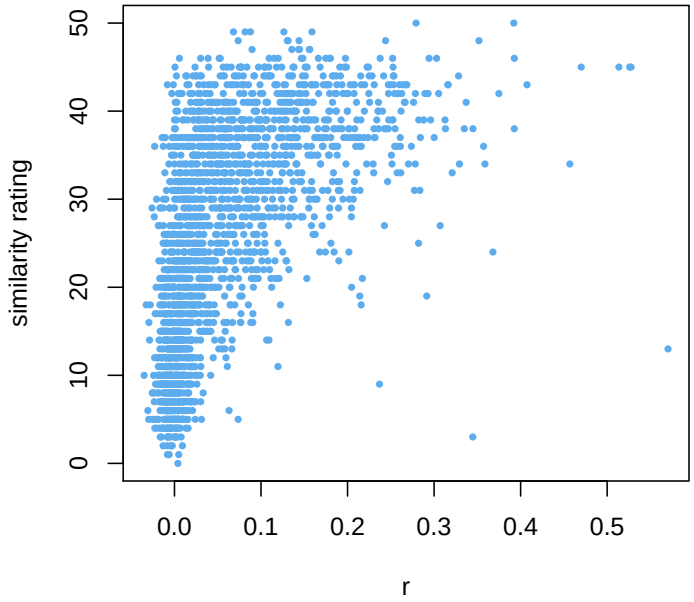

Figure 1: Similarity rating in the MEN dataset as a function of the correlation r between the row vectors in $\boldsymbol{S}$ of the words in the test pairs.

3.2.3 Correlational structure of inflectional and derivational vectors

Now that we have established that the semantic vector space $\boldsymbol{S}$, obtained with perhaps the simplest possible error-driven learning algorithm discriminating between outcomes given multiple cues (equation 1), indeed captures semantic similarity, we next consider how the semantic vectors of inflectional and derivational lexomes cluster in this space. Figure 2 presents a heatmap for the correlation matrix of the functional lexomes that we implemented as described above.

Nominal and verbal inflectional lexomes, as well as adverbial LY, cluster in the lower left corner, and tend to be either not correlated with derivational lexomes (indicated by light yellow), or to be negatively correlated (more reddish colors). Some inflectional lexomes, however, are interspersed

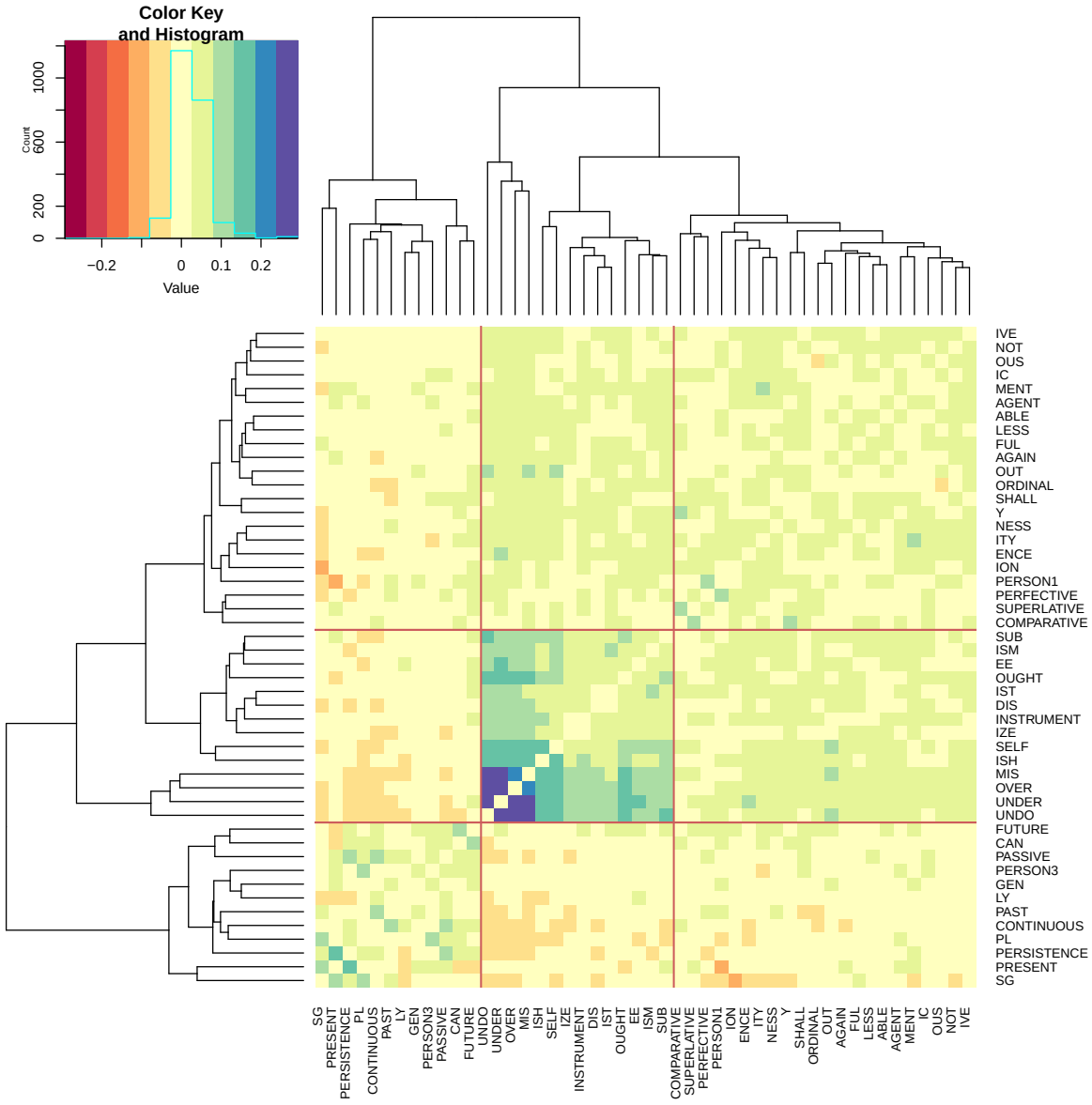

Figure 2: Heatmap for the correlation matrix of the row vectors in $\boldsymbol{S}$ of derivational and inflectional function lexomes. (Individual words will become legible by zooming in on the figure at maximal magnification.)

with derivational lexomes. For instance, COMPARATIVE, SUPERLATIVE and PERFECTIVE form a small subcluster together within the group of derivational lexomes.

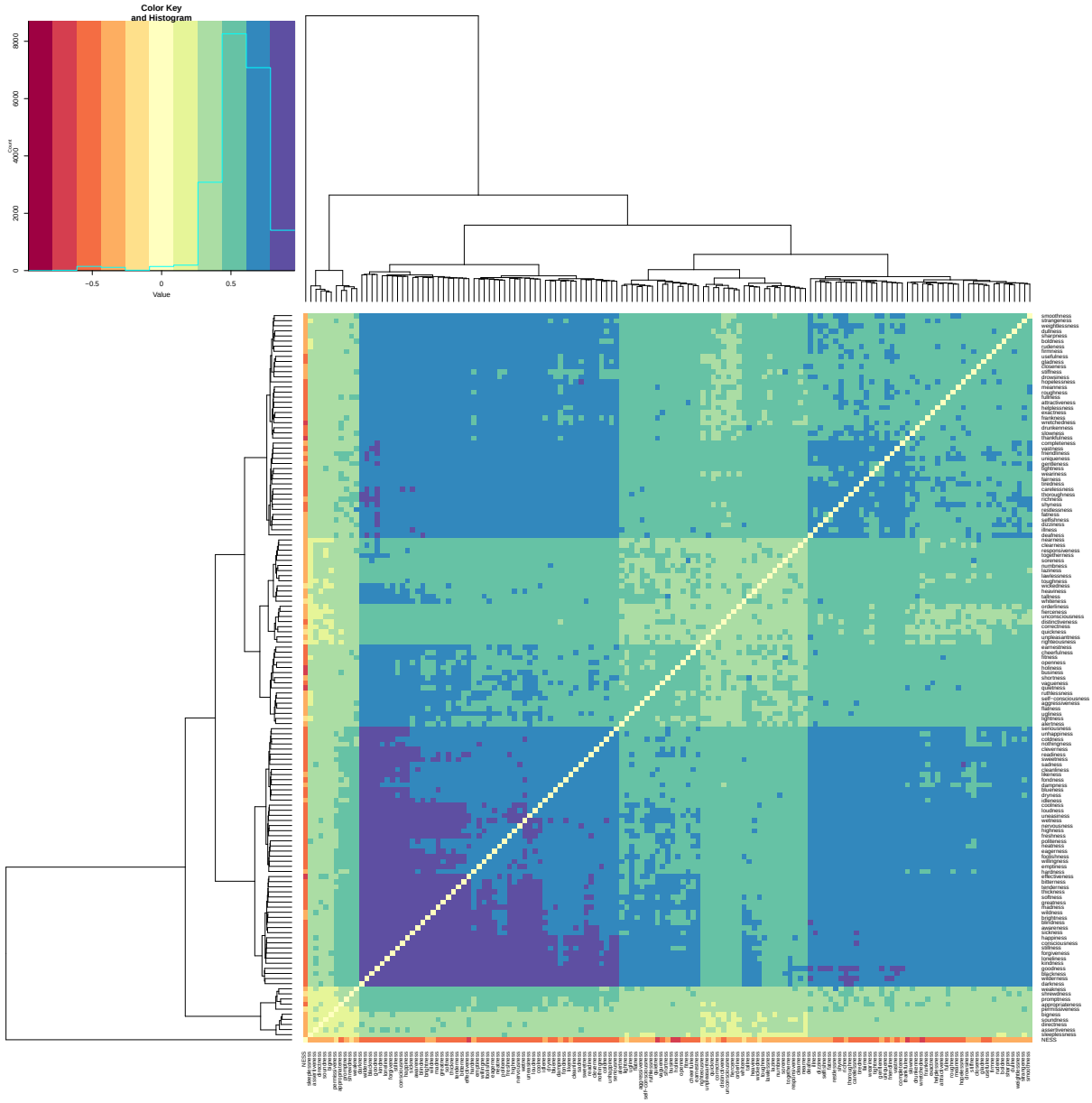

Figure 3: Heatmap for the correlation matrix of lexomes for content words with NESS, as well as the derivational lexome of NESS itself. This lexome is found at the very left edge of the dendrograms, and is negatively correlated with almost all content lexomes. (Individual words will become legible by zooming in on the figure at maximal magnification.)

The derivational lexomes show for a majority of pairs small positive correlations, and stronger correlations in the case of the verb-forming derivational lexomes MIS, OVER, UNDER, UNDO, which among themselves show the strongest positive correlations of all. The inflectional lexomes creating abstract nouns, ION, ENCE, ITY, NESS also form a subcluster. In other words, Figure 2 shows that there is some structure to the distribution of inflectional and derivational lexomes in the semantic

vector space.

Derived words (but not inflected words) have their own content lexomes, and hence their own row vectors in $\boldsymbol{S}$. Do the semantic vectors of these words cluster according the their formatives? To address this question, we extracted the row vectors from $\boldsymbol{S}$ for a total of 3500 derived words for 31 different formatives, resulting in a 3500 $\times$ 23,562 matrix D sliced out of $\boldsymbol{S}$. To this matrix, we added the semantic vectors for the 31 formatives. We then used linear discriminant analysis (LDA) to predict a lexome's formative from its semantic vector, using the `lda` function from the **MASS** package (Venables and Ripley, 2002). (For LDA to work, we had to remove columns from $\boldsymbol{D}$ with very small variance; as a consequence, the dimension of the matrix that was the input to LDA was 3531 $\times$ 814.) LDA accuracy for this classification task with 31 possible outcomes was 0.72. All 31 derivational lexomes were correctly classified.

When the formatives are randomly permuted, thus breaking the relation between the semantic vectors and their formatives, LDA accuracy was on average 0.528 (range across 10 permutations 0.519–0.537). From this, we conclude that derived words show more clustering in the semantic space of $\boldsymbol{S}$ than can be expected under randomness.

How are the semantic vectors of derivational lexomes positioned with respect to the clusters of their derived words? To address this question, we constructed heatmaps for the correlation matrices of the pertinent semantic vectors. An example of such a heatmap is presented for NESS in Figure 3. Apart from the existence of clusters within the cluster of NESS content lexomes, it is striking that the NESS lexome itself is found at the very left edge of the dendrogram, and at the very left column and bottom row of the heatmap. The color coding indicates that, surprisingly, the NESS derivational lexome is negatively correlated with almost all content lexomes that have *ness* as formative. Thus, the semantic vector of NESS is not a prototype at the center of its cloud of exemplars, but an *anti-prototype*. This vector is close to the cloud of semantic vectors, but it is outside its periphery. This pattern is not specific to NESS, but is found for the other derivational lexomes as well. It is intrinsic to our model.

The reason for this is straightforward. During learning, although for a derived word's lexome i and its derivational lexome NESS are co-present cues, the derivational lexome occurs in many other words j, and each time that another word j is encountered, weights are reduced from NESS to i. As this happens for all content lexomes, the derivational lexome is, during learning, slowly but steadily discriminated away from its content lexomes. We shall see that this is an important property for our model to capture morphological productivity for comprehension and speech production.

When the additive model of Mitchell and Lapata (2008) is used to construct a semantic vector for NESS, i.e., when the average vector is computed for the vectors obtained by subtracting the vector of the derived word from that of the base word, the result is a vector that is embedded inside the cluster of derived vectors, and hence inherits semantic idiosyncracies from all these derived words.

3.2.4 Semantic plausibility and transparency ratings for derived words

In order to obtain further evidence for the validity of inflectional and derivational lexomes, we re-analyzed the semantic plausibility judgements for word pairs consisting of a base and a novel derived form (e.g., *accent*, *accentable*) reported by Marelli and Baroni (2015) and available at `http://clic.cimec.unitn.it/composes/FRACSS`. The number of pairs available to us for analysis, 236, was restricted compared to the original 2,559 pairs due to the constraints that the base words had to occur in the TASA corpus. Furthermore, we also explored the role of words' emotional valence, arousal, and dominance, as available in Warriner et al. (2013), which restricted the number of items even further. The reason for doing so is that human judgements, such as those of age of acquisition,

may reflect dimensions of emotion (Baayen et al., 2016a).

A measure derived from the semantic vectors of the derivational lexomes, activation diversity, the L1-norm of the semantic vector (the sum of the absolute values of the vector elements, i.e., its city-block distance), turned out to be predictive for these plausibility ratings, as documented in Table 3 and the left panel of Figure 4. Activation diversity is a measure of lexicality (Milin et al., 2017b). In an auditory word identification task, for instance, speech that gives rise to a low activation diversity elicits fast rejections, whereas speech that generates high activation diversity elicits higher acceptance rates but at the cost of longer response times (Arnold et al., 2017).

Activation diversity interacted with word length. A greater word length had a strong positive effect on rated plausibility, but this effect progressively weakened as activation diversity increases. In turn, activation diversity had a negative effect on plausibility for shorter words, and a positive effect for longer words. Apparently, the evidence for lexicality that comes with higher L1-norms contributes positively when there is little evidence coming from word length. As word length increases and the relative contribution of the formative in the word decreases, the greater uncertainty that comes with higher lexicality (a greater L1-norm implies more strong links to many other lexomes) has a detrimental effect on the ratings. Higher arousal scores also contributed to higher perceived plausibility. Valence and dominance were not predictive, and were therefore left out of the specification of the model reported here. (Word frequency was not included as predictor because all derived words are neologisms with zero frequency; addition of base frequency did not improve model fit, $p > 0.91$.)

A. parametric coefficients	Estimate	Std. Error	t-value	p-value
Intercept	-1.8072	2.7560	-0.6557	0.5127
Word Length	0.5901	0.2113	2.7931	0.0057
Activation Diversity	1.1221	0.6747	1.6632	0.0976
Arousal	0.3847	0.1521	2.5295	0.0121
Word Length × Activation Diversity	-0.1318	0.0500	-2.6369	0.0089
B. smooth terms	edf	Ref.df	F-value	p-value
Random Intercepts Affix	3.3610	4.0000	55.6723	< 0.0001

Table 3: GAM fitted to plausibility ratings for derivational neologisms with the derivational lexomes ABLE, AGAIN, AGENT, IST, LESS, NOT; data from Marelli and Baroni (2015).

The degree to which a complex word is semantically transparent with respect to its base word is of both practical and theoretical interest. An evaluation of the semantic transparency of complex words using transformation matrices is developed in Marelli and Baroni (2015). Within the present framework, semantic transparency can be examined straightforwardly by comparing the correlations of (i) the semantic vectors of the base word to which the semantic vector of the affix is added, with (ii) the semantic vector of the derived word itself. The more distant the two vectors are, the more negative their correlation should be. This is exactly what we find. For NESS, for instance, the six most negative correlations are *business* ($r = -0.66$), *effectiveness* ($r = -0.51$), *awareness* ($r = -0.50$), *loneliness* ($r = -0.45$), *sickness* ($r = -0.44$), and *consciousness* ($r = -0.43$). Although NESS can have an anaphoric function in discourse (Kastovsky, 1986), words such as *business* and *consciousness* have a much deeper and richer semantics than just reference to a previously mentioned state of affairs. A simple comparison of the word's actual semantic vector in $\boldsymbol{S}$ (derived from the TASA corpus) and its semantic vector obtained by accumulating evidence over base and affix (i.e., summing the vectors of base and affix) brings this out straightforwardly.

However, evaluating correlations for transparency is prone to researcher bias. We therefore also investigated to what extent the semantic transparency ratings collected by Lazaridou et al. (2013)

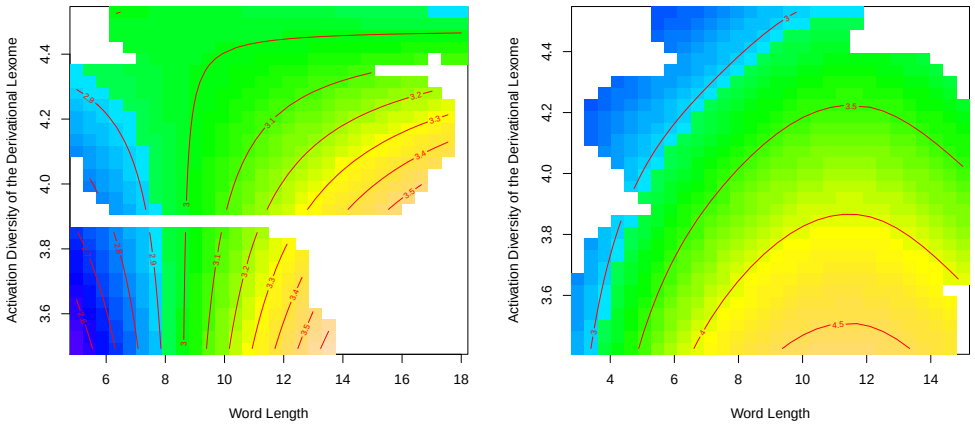

Figure 4: Interaction of word length by activation diversity in GAMs fitted to plausibility ratings for complex words (left) and to semantic transparency ratings (right).

can be predicted from the activation diversity of the semantic vector of the derivational lexome. The original dataset of Lazaridou and colleagues comprises 900 words, of which 330 meet the criterion of occurring more than 8 times in the TASA corpus and for which we have semantic vectors available. The summary of a Gaussian location-scale additive model is given in Table 4, and the right panel of Figure 4 visualizes the interaction of word length by activation diversity. Although modulated by word length, overall, transparency ratings decrease as activation diversity is increased. Apparently, the stronger the connections a derivational lexome has with other lexomes, the less clear it becomes what its actual semantic contribution to a novel derived word is. In other words, under semantic uncertainty, transparency ratings decrease.

A. parametric coefficients	Estimate	Std. Error	t-value	p-value
intercept [location]	5.5016	0.2564	21.4537	< 0.0001
intercept [scale]	-0.4135	0.0401	-10.3060	< 0.0001
B. smooth terms	edf	Ref.df	F-value	p-value
te(activation diversity, word length) [location]	5.3879	6.2897	23.7798	0.0008
s(derivational lexome) [location]	14.3180	15.0000	380.6254	< 0.0001
s(activation diversity) [scale]	2.1282	2.5993	28.4336	< 0.0001
s(word length) [scale]	1.7722	2.1403	132.7311	< 0.0001

Table 4: Gaussian location-scale additive model fitted to the semantic transparency ratings for derived words. te: tensor product smooth, s: thin plate regression spline.

4 Comprehension

Now that we have established that the present semantic vectors make sense, even though they are based on a small corpus, we next consider a comprehension model that has form vectors as input and semantic vectors as output.[5] We begin with introducing the central concepts underlying

[5] A package for R implementing the comprehension and production algorithms of linear discriminative learning is available at http://www.sfs.uni-tuebingen.de/~hbaayen/publications/WpmWithLdl_1.0.tar.gz. The package is

mappings from form to meaning. We then discuss visual comprehension, and then turn to auditory comprehension.

4.1 Setting up the mapping

Let $\boldsymbol{C}$ denote the cue matrix, a matrix that specifies for each word (rows) the form cues of that word (columns). For a toy lexicon with the words *one, two, three*, the $\boldsymbol{C}$ matrix is

$$\boldsymbol{C} = \begin{array}{c} \\ \text{one} \\ \text{two} \\ \text{three} \end{array} \begin{array}{c} \begin{array}{cccccccc} \text{\#wV} & \text{wVn} & \text{Vn\#} & \text{\#tu} & \text{tu\#} & \text{\#Tr} & \text{Tri} & \text{ri\#} \end{array} \\ \begin{pmatrix} 1 & 1 & 1 & 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 1 & 1 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 & 1 & 1 & 1 \end{pmatrix} \end{array}, \tag{6}$$

where we use the DISC keyboard phonetic alphabet[6] for triphones; the # symbol denotes the word boundary. Suppose that the semantic vectors for these words are the row vectors of the following matrix $\boldsymbol{S}$:

$$\boldsymbol{S} = \begin{array}{c} \\ \text{one} \\ \text{two} \\ \text{three} \end{array} \begin{array}{c} \begin{array}{ccc} \text{one} & \text{two} & \text{three} \end{array} \\ \begin{pmatrix} 1.0 & 0.3 & 0.4 \\ 0.2 & 1.0 & 0.1 \\ 0.1 & 0.1 & 1.0 \end{pmatrix} \end{array}. \tag{7}$$

We are interested in a transformation matrix F such that

$$\boldsymbol{CF} = \boldsymbol{S}. \tag{8}$$

The transformation matrix is straightforward to obtain. Let $\boldsymbol{C}'$ denote the Moore-Penrose generalized inverse of $\boldsymbol{C}$, available in R as the `ginv` function in the **MASS** package (Venables and Ripley, 2002). Then

$$\boldsymbol{F} = \boldsymbol{C}'\boldsymbol{S}. \tag{9}$$

For the present example,

$$\boldsymbol{F} = \begin{array}{c} \\ \text{\#wV} \\ \text{wVn} \\ \text{Vn\#} \\ \text{\#tu} \\ \text{tu\#} \\ \text{\#Tr} \\ \text{Tri} \\ \text{ri\#} \end{array} \begin{array}{c} \begin{array}{ccc} \text{one} & \text{two} & \text{three} \end{array} \\ \begin{pmatrix} 0.33 & 0.10 & 0.13 \\ 0.33 & 0.10 & 0.13 \\ 0.33 & 0.10 & 0.13 \\ 0.10 & 0.50 & 0.05 \\ 0.10 & 0.50 & 0.05 \\ 0.03 & 0.03 & 0.33 \\ 0.03 & 0.03 & 0.33 \\ 0.03 & 0.03 & 0.33 \end{pmatrix} \end{array}, \tag{10}$$

and for this simple example, $\boldsymbol{CF}$ is exactly equal to $\boldsymbol{S}$.

In the remainder of this section, we investigate how well this very simple end-to-end model performs for visual word recognition as well as auditory word recognition. For visual word recognition, we use the semantic matrix $\boldsymbol{S}$ developed in section 3, but we consider two different cue matrices $\boldsymbol{C}$, one using letter trigrams (following Milin et al., 2017b) and one using phone trigrams. A comparison of the performance of the two models sheds light on the role of words' "sound image" on word recognition in reading (cf. McBride-Chang, 1996; Van Orden et al., 2005; Jared and O'Donnell,

described in Baayen et al. (2018).

[6] The "DIstinct Single Character" representation with one character for each phoneme was introduced by CELEX (Burnage, 1988).

2017, for phonological effects in visual word recognition). For auditory word recognition, we make use of the acoustic features developed in Arnold et al. (2017).

Although the features that we selected as cues are based on domain knowledge, they do not have an ontological status in our framework, in contrast to units such as phonemes and morphemes in standard linguistic theories. In these theories, phonemes and morphemes are Russelian atomic units of formal phonological and syntactic calculi, and considerable research has been directed towards showing that these units are psychologically real. For instance, speech errors involving morphemes have been interpreted as clear evidence for the existence in the mind of morphemes (Dell, 1986). Research has been directed at finding behavioral and neural correlates of phonemes and morphemes (Kleinschmidt and Jaeger, 2015; Marantz, 2013; Bozic et al., 2007), ignoring fundamental criticisms of both the phoneme and the morpheme as theoretical constructs (Port and Leary, 2005; Matthews, 1991; Blevins, 2003).

The features that we use for representing aspects of form are heuristic features that have been selected or developed primarily because they work well as discriminators. We will gladly exchange the present features for other features, if these other features can be shown to afford higher discriminative accuracy. Nevertheless, the features that we use are grounded in domain knowledge. For instance, the letter trigrams used for modeling reading (see, e.g. Milin et al., 2017b) are motivated by the finding in stylometry that letter trigrams are outstanding features for discriminating between authorial hands (Forsyth and Holmes, 1996).[7] Letter n-grams are also posited by Cohen and Dehaene (2009) for the visual word form system, at the higher end of a hierarchy of neurons tuned to increasingly large visual features of words.

Triphones, the units that we use for representing the 'acoustic image' or 'auditory verbal imagery' of canonical word forms, have the same discriminative potential as letter triplets, but have as additional advantage that they do better justice, compared to phonemes, to phonetic contextual interdependencies, such as plosives being differentiated primarily by formant transitions in adjacent vowels. The acoustic features that we use for modeling auditory comprehension, to be introduced in more detail below, are motivated in part by the sensitivity of specific areas on the cochlea to different frequencies in the speech signal.

Evaluation of model predictions proceeds by comparing the predicted semantic vector $\hat{\boldsymbol{s}}$ obtained by multiplying an observed cue vector $\boldsymbol{c}$ with the transformation matrix $\boldsymbol{F}$ ($\hat{\boldsymbol{s}} = \boldsymbol{cF}$) with the corresponding target row vector $\boldsymbol{s}$ of $\boldsymbol{S}$. A word i is counted as correctly recognized when $\hat{\boldsymbol{s}}_i$ is most strongly correlated with the target semantic vector $\boldsymbol{s}_i$ of all target vectors $\boldsymbol{s}_j$ across all words j.

Inflected words do not have their own semantic vectors in $\boldsymbol{S}$. We therefore created semantic vectors for inflected words by adding the semantic vectors of stem and affix, and added these as additional row vectors to $\boldsymbol{S}$ before calculating the transformation matrix $\boldsymbol{F}$.

We note here that there is only one transformation matrix, i.e., one discrimination network, that covers all affixes, inflectional and derivational, as well as derived and monolexomic (simple) words. This approach contrasts with that of Marelli and Baroni (2015), who pair every affix with its own transformation matrix.

[7] In what follows, we work with letter triplets and triphones, which are basically contextually enriched letter and phone units. For languages with strong phonotactic restrictions, such that the syllable inventory is quite small (e.g., Vietnamese), digraphs work appear to work better than trigraphs (Pham and Baayen, 2015). Baayen et al. (2018) show that working with four-grams may enhance performance, and current work in progress on many other languages shows that for highly inflecting languages with long words, 4-grams may outperform 3-grams. For computational simplicity, we have not experimented with mixtures of units of different length, nor with algorithms with which such units might be determined.

4.2 Visual comprehension

For visual comprehension, we first consider a model straightforwardly mapping form vectors onto semantic vectors. We then expand the model with an indirect route first mapping form vectors onto the vectors for the acoustic image (derived from words' triphone representations), and then mapping the acoustic image vectors onto the semantic vectors.

The dataset on which we evaluated our models comprised 3987 monolexomic English words, 6595 inflected variants of monolexomic words, and 898 derived words with monolexomic base words, to a total of 11480 words. These counts follow from the simultaneous constraints of (i) a word appearing with sufficient frequency in TASA, (ii) the word being available in the British Lexicon Project (BLP Keuleers et al., 2012), and (iii) the word being available in the CELEX database, and the word being of the abovementioned morphological type. In the present study, we did not include inflected variants of derived words, nor did we include compounds.

4.2.1 The direct route straight from orthography to semantics

To model single word reading as gauged by the visual lexical decision task, we used letter trigrams as cues. In our dataset, there were a total of 3465 unique trigrams, resulting in a 11480 × 3465 orthographic cue matrix $\boldsymbol{C}$. The semantic vectors of the monolexomic and derived words were taken from the semantic weight matrix described in section 3. For inflected words, semantic vectors were obtained by summation of the semantic vectors of base words and inflectional functions. For the resulting semantic matrix $\boldsymbol{S}$, we retained the 5030 column vectors with the highest variances, setting the cutoff value for the minimal variance to 0.34×10^{-7}. From $\boldsymbol{S}$ and $\boldsymbol{C}$, we derived the transformation matrix $\boldsymbol{F}$, which we used to obtain estimates $\hat{\boldsymbol{s}}$ of the semantic vectors $\boldsymbol{s}$ in $\boldsymbol{S}$.

For 59% of the words, $\hat{\boldsymbol{s}}$ had the highest correlation with the targeted semantic vector $\boldsymbol{s}$. (For the correctly predicted words, the mean correlation was 0.83, and the median was 0.86. With respect to the incorrectly predicted words, the mean and median correlation were both 0.48.) The accuracy obtained with naive discriminative learning, using orthogonal lexomes as outcomes instead of semantic vectors, was 27%. Thus, as expected, performance of linear discrimination learning (LDL) is substantially better than that of naive discriminative learning (NDL).

To assess whether this model is productive, in the sense that it can make sense of novel complex words, we considered a separate set of inflected and derived words which were not included in the original dataset. For an unseen complex word, both base word and the inflectional or derivational function appeared in the training set, but not the complex word itself. The network $\boldsymbol{F}$ therefore was presented with a novel form vector, which it mapped onto a novel vector in the semantic space. Recognition was succesful if this novel vector was more strongly correlated with the semantic vector obtained by summing the semantic vectors of base and inflectional or derivational function than with any other semantic vector.

Of 553 unseen inflected words, 43% were recognized successfully. The semantic vectors predicted from their trigrams by the network $\boldsymbol{F}$ were overall well correlated in the mean with the targeted semantic vectors of the novel inflected words (obtained by summing the semantic vectors of their base and inflectional function): $\bar{r} = 0.67, p < 0.0001$. The predicted semantic vectors also correlated well with the semantic vectors of the inflectional functions ($\bar{r} = 0.61, p < 0.0001$). For the base words, the mean correlation dropped to $\bar{r} = 0.28$ ($p < 0.0001$).

For unseen derived words (514 in total), we also calculated the predicted semantic vectors from their trigram vectors. The resulting semantic vectors $\hat{\boldsymbol{s}}$ had moderate positive correlations with the semantic vectors of their base words ($\bar{r} = 0.40, p < 0.0001$), but negative correlations with the semantic vectors of their derivational functions ($\bar{r} = -0.13, p < 0.0001$). They did not

correlate with the semantic vectors obtained by summation of the semantic vectors of base and affix ($\bar{r} = 0.01, p = 0.56$).

The reduced correlations for the derived words as compared to those for the inflected words likely reflects to some extent that the model was trained on many more inflected words (in all, 6595) than derived words (898), whereas the number of different inflectional functions (7) was much reduced compared to the number of derivational functions (24). However, the negative correlations of the derivational semantic vectors with the semantic vectors of their derivational functions fits well with the observation in section 3.2.3 that derivational functions are anti-prototypes that enter into negative correlations with the semantic vectors of the corresponding derived words. We suspect that the absence of a correlation of the predicted vectors with the semantic vectors obtained by integrating over the vectors of base and derivational function is due to the semantic idiosyncracies that are typical for derived words. For instance, *austerity* can denote harsh discipline, or simplicity, or a policy of deficit cutting, or harshness to the taste. And a *worker* is not just someone who happens to work, but someone earning wages, or a nonreproductive bee or wasp, or a thread in a computer program. Thus, even within the usages of one and the same derived word, there is a lot of semantic heterogeneity that stands in stark contrast to the straightforward and uniform interpretation of inflected words.

4.2.2 The indirect route from orthography via phonology to semantics

Although reading starts with orthographic input, it has been shown that phonology actually plays a role during the reading process as well. For instance, developmental studies indicate that children's reading development can be predicted by their phonological abilities (Wagner et al., 1994; McBride-Chang, 1996). Evidence of the influence of phonology on adults' reading has also been reported (Wong and Chen, 1999; Newman et al., 2012; Jared et al., 2016; Bitan et al., 2017; Jared and Bainbridge, 2017; Jared and O'Donnell, 2017; Amenta et al., 2017). As pointed out by Perrone-Bertolotti et al. (2012), silent reading often involves an imagery speech component: we hear our own "inner voice" while reading. Since written words produce a vivid auditory experience almost effortlessly, they argue that auditory verbal imagery should be part of any neuro-cognitive model of reading. Interestingly, in a study using intracranial EEG recordings with four epileptic neurosurgical patients, they observed that silent reading elicits auditory processing in the absence of any auditory stimulation, suggesting that auditory images are spontaneously evoked during reading (see also Yao et al., 2011).

To explore the role of auditory images in silent reading, we operationalized sound images by means of phone trigrams. (In our model, phone trigrams are independently motivated as intermediate targets of speech production, see section 5 for further discussion). We therefore ran the model again with phone trigrams instead of letter triplets. The semantic matrix $\boldsymbol{S}$ was exactly the same as above. However, a new transformation matrix $\boldsymbol{F}$ was obtained for mapping a 11480 $\times$ 5929 cue matrix $\boldsymbol{C}$ of phone trigrams onto $\boldsymbol{S}$. To our surprise, the triphone model outperformed the trigram model substantially. Overall accuracy rose to 78%, an improvement of almost 20%.

In order to integrate this finding into our model, we implemented an additional network $\boldsymbol{K}$ that maps orthographic vectors (coding the presence or absence of trigrams) onto phonological vectors (indicating the presence or absence of triphones). The result is a dual route model for (silent) reading, with a first route utilizing a network mapping straight from orthography to semantics, and a second, indirect, route utilizing two networks, one mapping from trigrams to triphones, and the other subsequently mapping from triphones to semantics.

The network $\boldsymbol{K}$, which maps trigram vectors onto triphone vectors, provides good support for the relevant triphones, but does not provide guidance as to their ordering. Using a graph-

based algorithm detailed in the Appendix (see also section 5.2), we found that for 92% of the words the correct sequence of triphones jointly spanning the word's sequence of letters received maximal support. The mean correlation for the correctly recognized words was 0.90, and the median correlation was 0.94. For words that were not recognized correctly, the mean and median correlation were 0.45 and 0.43 respectively.

We then constructed a matrix $\hat{\boldsymbol{T}}$ with the triphone vectors predicted from the trigrams by network $\boldsymbol{K}$, and defined a new network $\boldsymbol{H}$ mapping these predicted triphone vectors onto the semantic vectors $\boldsymbol{S}$. We now have, for each word, two estimated semantic vectors, a vector $\hat{\boldsymbol{s}}_1$ obtained with network $\boldsymbol{F}$ of the first route, and a vector $\hat{\boldsymbol{s}}_2$ obtained with network $\boldsymbol{H}$ from the auditory targets that themselves were predicted by the orthographic cues. The mean of the correlations of these pairs of vectors was $\bar{r} = 0.73$ ($p < 0.0001$).

To assess to what extent the networks that we defined are informative for actual visual word recognition, we made use of the reaction times in the visual lexical decision task available in the BLP. All 11,480 words in the current dataset are also included in BLP. We derived three measures from each of the two networks.

The first measure is the sum of the L1-norms of $\hat{\boldsymbol{s}}_1$ and $\hat{\boldsymbol{s}}_2$, to which we will refer as a word's *total activation diversity*. The total activation diversity is an estimate of the support for a word's lexicality as provided by the two routes. The second measure is the correlation of $\hat{\boldsymbol{s}}_1$ and $\hat{\boldsymbol{s}}_2$, to which we will refer as a word's *route congruency*. The third measure is the L1 norm of a lexome's column vector in $\boldsymbol{S}$, to which we refer as its *prior*. This last measure has previously been observed to be a strong predictor for lexical decision latencies (Baayen et al., 2016a). A word's prior is an estimate of a word's entrenchment and prior availability.

We fitted a generalized additive model to the inverse-transformed response latencies in the BLP with these three measures as key covariates of interest, including as control predictors word length and word type (derived, inflected, and monolexomic, with derived as reference level). As shown in Table 5, inflected words were responded to more slowly than derived words, and the same holds for monolexomic words, albeit to a lesser extent. Response latencies also increased with length. Total activation diversity revealed a U-shaped effect (Figure 5, left panel). For all but the lowest total activation diversities, we find that response times increase as total activation diversity increases. This result is consistent with previous findings for auditory word recognition (Arnold et al., 2017). As expected given the results of Baayen et al. (2016a), the prior was a strong predictor, with larger priors affording shorter response times. There was a small effect of route congruency, which emerged in a nonlinear interaction with the prior, such that for large priors, a greater route congruency afforded further reduction in response time (Figure 5, right panel). Apparently, when the two routes converge on the same semantic vector, uncertainty is reduced and a faster response can be initiated.

For comparison, the dual-route model was implemented with NDL as well. Three measures were derived from the networks and used to predict the same response latencies. These measures include activation, activation diversity, and prior, all of which have been reported to be reliable predictors for RT in visual lexical decision (Milin et al., 2017b; Baayen et al., 2011, 2016a). However, since here we are dealing with two routes, the three measures can be independently derived from both routes. We therefore summed up the measures of the two routes, obtaining three measures: total activation, total activation diversity, and total prior. With word type and word length as control factors, Table 6 shows that these measures participated in a three-way interaction, presented in Figure 6. Total activation showed a U-shaped effect on RT that is increasingly attenuated as total activation diversity is increased (left panel). Total activation also interacted with the total prior (right panel). For medium range values of total activation, RTs increased with total activation, and as expected decreased with the total prior. The center panel shows that RTs decrease with

total prior for most of the range of total activation diversity, and that the effect of total activation diversity changes sign going from low to high values of the total prior.

Model comparison revealed that the generalized additive model with LDL measures (Table 5) provides a substantially improved fit to the data compared to the GAM using NDL measures (Table 6), with a difference of no less than 651.01 AIC units. At the same time, the GAM based on LDL is less complex.

Given the substantially better predictability of LDL measures on human behavioral data, one remaining question is whether it is really the case that two routes are involved in silent reading. After all, the superior accuracy of the second route might be an artefact of a simple machine learning technique performing better for triphones than for trigrams. This question can be addressed by examining whether the fit of the GAM summarized in Table 5 improves or worsens depending on whether the activation diversity of the first or the second route is taken out of commission.

When the GAM is provided access to just the activation diversity of $\hat{\boldsymbol{s}}_2$, the AIC increased by 100.59 units. However, when the model is based only on the activation diversity of $\hat{\boldsymbol{s}}_1$, the model fit increased by no less than 147.90 AIC units. From this, we conclude that, at least for the visual lexical decision latencies in the BLP, the second route, first mapping trigrams to triphones, and then mapping triphones onto semantic vectors, plays the more important role.

The superiority of the second route may in part be due to the number of triphone features being larger than the number of trigram features (3465 trigrams versus 5929 triphones). More features, which mathematically amounts to more predictors, enable more precise mappings. Furthermore, the heavy use made in English of letter sequences such as *ough* (with 10 different pronunciations, https://www.dictionary.com/e/s/ough) reduces semantic discriminability compared to the corresponding spoken forms. It is noteworthy, however, that the benefits of the triphone-to-semantics mapping are possible only thanks to the high accuracy with which orthographic trigram vectors are mapped onto phonological triphone vectors (92%).

Table 5: Summary of a generalized additive model fitted to response latencies in visual lexical decision using measures based on LDL. s: thin plate regression spline smooth; te: tensor product smooth.

A. parametric coefficients	Estimate	Std. Error	t-value	p-value
intercept	-1.7774	0.0087	-205.414	<.0001
word type:Inflected	0.1110	0.0059	18.742	<.0001
word type:Monomorphemic	0.0233	0.0054	4.322	<.0001
word length	0.0117	0.0011	10.981	<.0001
B. smooth terms	edf	Ref.df	F	p-value
s(total activation diversity)	6.002	7.222	23.6	<.0001
te(route congruency, prior)	14.673	17.850	213.7	<.0001

It is unlikely that the 'phonological route' is always dominant in silent reading. Especially in fast 'diagonal' reading, the 'direct route' may be more dominant. There is remarkable, although for the present authors, unexpected, convergence with the dual route model for reading aloud of Coltheart et al. (1993); Coltheart (2005). However, while the text-to-phonology route of their model has as primary function to explain why nonwords can be pronounced, our results show that both routes can actually be active when silently reading real words. A fundamental difference is, however, that in our model, words' semantics play a central role.

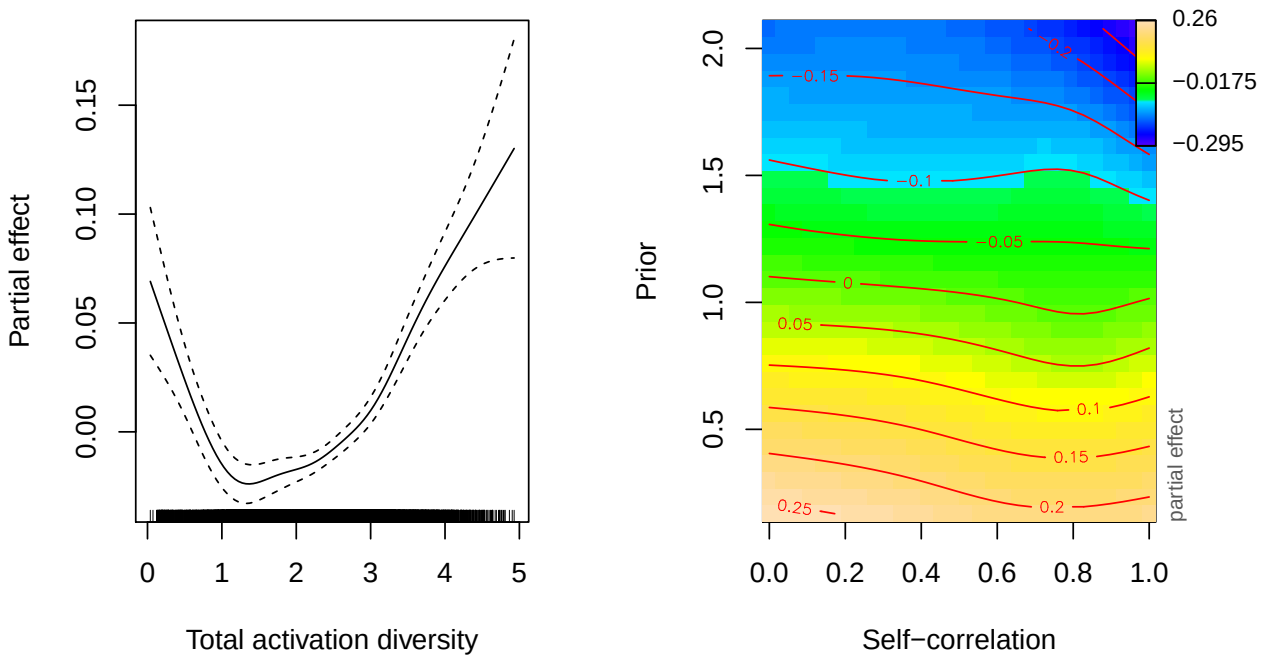

Figure 5: The partial effects of total activation diversity (left) and the interaction of route congruency and prior (right) on RT in the British Lexicon Project.

Table 6: Summary of a generalized additive model fitted to response latencies in visual lexical decision using measures from NDL. te: tensor product smooth.

A. parametric coefficients	Estimate	Std. Error	t-value	p-value
intercept	-1.6934	0.0086	-195.933	<.0001
word type:Inflected	0.0286	0.0052	5.486	<.0001
word type:Monomorphemic	0.0042	0.0055	0.770	= .4
word length	0.0066	0.0011	5.893	<.0001
B. smooth terms	edf	Ref.df	F	p-value
te(activation, activation diversity, prior)	41.34	49.72	108.4	<.0001

4.3 Auditory comprehension

For the modeling of reading, we made use of letter trigrams as cues. These cues abstract away from the actual visual patterns that fall on the retina, patterns that are already transformed at the retina before being sent to the visual cortex. Our hypothesis is that letter trigrams represent those high-level cells or cell assemblies in the visual system that are critical for reading, and we therefore leave the modeling, possibly with deep learning networks of how patterns on the retina are transformed into letter trigrams for further research.

As a consequence of the high level of abstraction of the trigrams, a word form is represented by a unique vector specifying which of a fixed set of letter trigrams is present in the word. Although one might consider modeling auditory comprehension with phone triplets (triphones), replacing the letter trigrams of visual comprehension, such an approach would not do justice to the enormous variability of actual speech. Whereas the modeling of reading printed words can depart from the assumption that the pixels of a word's letters on a computer screen are in a fixed configuration, independently of where the word is shown on the screen, the speech signal of the same word type

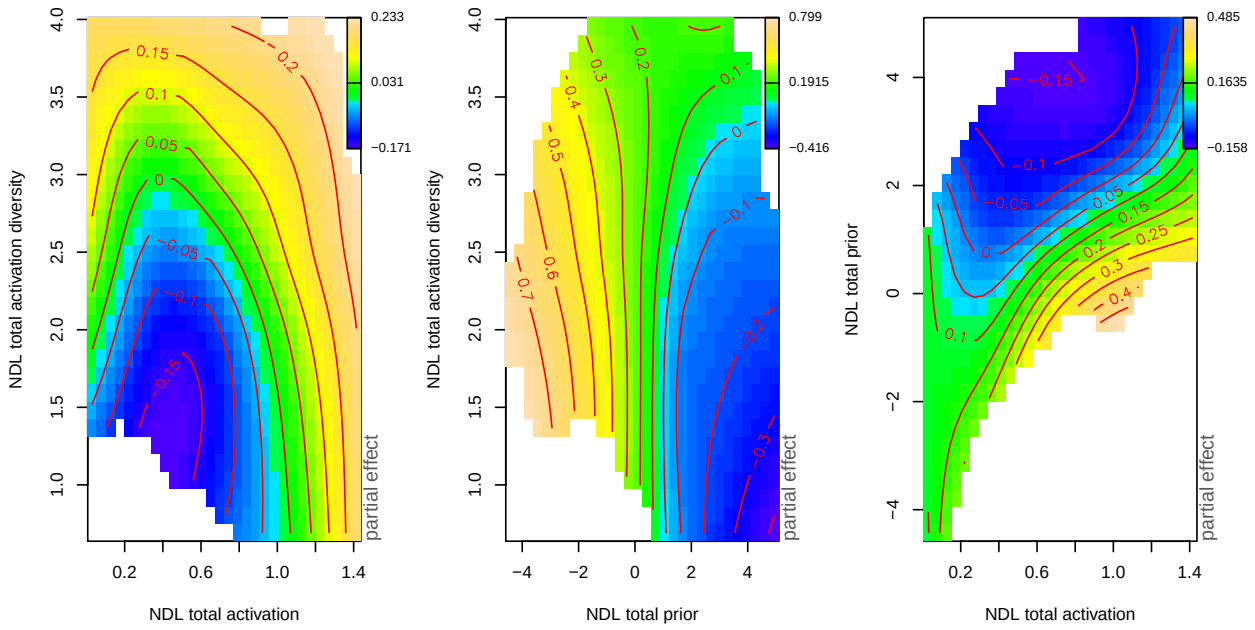

Figure 6: The interaction of total activation and total activation diversity (left), of total prior by total activation diversity (center), and of total activation by total prior (right) on RT in the British Lexicon Project, based on NDL.

varies from token to token, as illustrated in Figure 7 for the English word *crisis*. A survey of the Buckeye corpus (Pitt et al., 2005) of spontaneous conversations recorded at Columbus, Ohio (Johnson, 2004) indicates that around 5% of the words are spoken with one syllable missing, and that a little over 20% of words have at least one phone missing.

It is widely believed that the phoneme, as an abstract unit of sound, is essential for coming to grips with the huge variability that characterizes the speech signal (Phillips, 2001; Diehl et al., 2004; Norris and McQueen, 2008). However, the phoneme as theoretical linguistic construct is deeply problematic (Port and Leary, 2005), and for many spoken forms, canonical phonemes do not do justice to the phonetics of the actual sounds (Hawkins, 2003). Furthermore, if words are defined as sequences of phones, the problem arises what representations to posit for words with two or more reduced variants. Adding entries for reduced forms to the lexicon turns out not to afford better overal recognition (Cucchiarini and Strik, 2003). Although exemplar models have been put forward to overcome this problem (Johnson, 1997), we take a different approach here, and, following Arnold et al. (2017), lay out a discriminative approach to auditory comprehension.

The cues that we make use of to represent the acoustic signal are the Frequency Band Summary Features (FBSFs) introduced by Arnold et al. (2017) as input cues. FBSFs summarize the information present in the spectrogram of a speech signal. The algorithm that derives FBSFs first chunks the input at the minima of the Hilbert amplitude envelope of the signal's oscillogram (see the upper panel of Figure 8). For each chunk, the algorithm distinguishes 21 frequency bands in the MEL scaled spectrum, and intensities are discretized into 5 levels (lower panel in Figure 8) for small intervals of time. For each chunk, and for each frequency band in these chunks, a discrete feature is derived that specifies chunk number, frequency band number, and a description of the temporal variation in the band bringing together minimum, maximum, median, initial, and final intensity values. The 21 frequency bands are inspired by the 21 receptive areas on the cochlear membrane that are sensitive to different ranges of frequencies (Fletcher, 1940). Thus, a given FBSF is a proxy

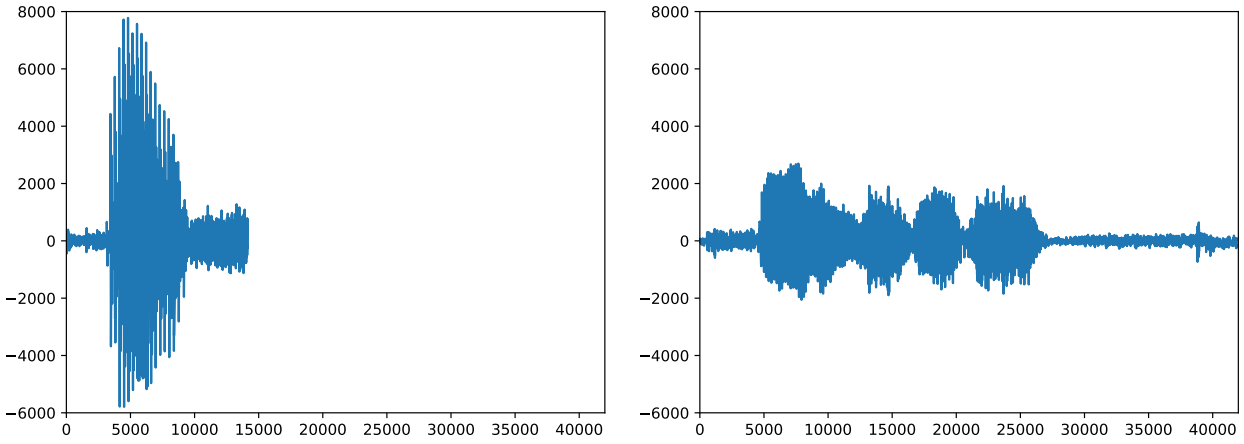

Figure 7: Oscillogram for two different realizations of the word *crisis* with different degrees of reduction in the NewsScape archive: [kraɪz](left) and [k^{h}raɪsɪs](right).

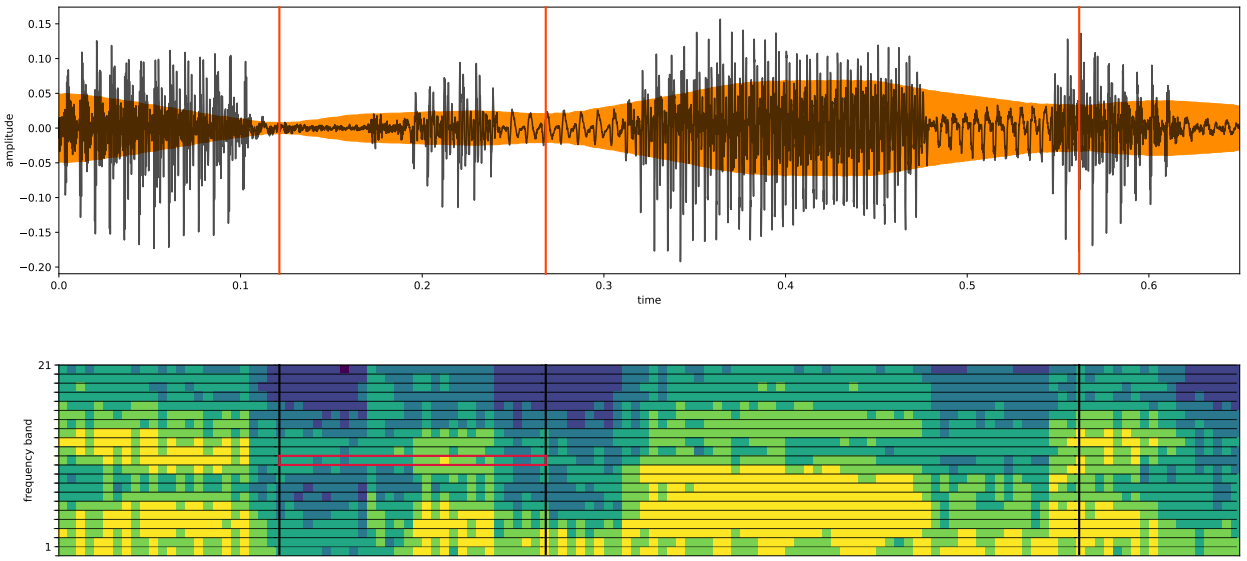

Figure 8: Oscillogram of the speech signal for a realization of the word *economic* with Hilbert envelope (in orange) outlining the signal is shown on the top panel. The lower panel depicts the discretized MEL scaled spectrum of the signal. The vertical bars are the boundary points that partition the signal into 4 chunks. For one chunk, horizontal lines (in red) highlight one of the frequency bands for which the FBSFs provide summaries of the variation over time in that frequency band. The FBSF for the highlighted band is "band11-start3-median3-min2-max5-end3-part2".

for cell assemblies in the auditory cortex that respond to a particular pattern of changes over time in spectral intensity. The `AcousticNDLCodeR` package (Arnold, 2017) for R (R Core Team, 2016) was employed to extract the FBSFs from the audio files.

We tested LDL on 20 hours of speech sampled from the audio files of the UCLA Library Broadcast NewsScape data, a vast repository of multi-modal TV news broadcasts, provided to us by the Distributed Little Red Hen Lab. The audio files of this resource were automatically classified as *clean* for relatively clean parts where there is speech without background noise or music,

and *noisy* for speech snippets where background noise or music is present. Here, we report results for 20 hours of clean speech, to a total of 131,673 word tokens (representing 4779 word types) with in all 40,639 distinct FBSFs.

The FBSFs for the word tokens are brought together in a matrix $\boldsymbol{C}_a$, with dimensions 131,673 audio tokens × 40,639 FBSFs. The targeted semantic vectors are taken from the $\boldsymbol{S}$ matrix, which is expanded to a matrix with 131,673 rows, one for each audio token, and 4,609 columns, the dimension of the semantic vectors. Although the transformation matrix $\boldsymbol{F}$ could be obtained by calculating $\boldsymbol{C}'\boldsymbol{S}$, the calculation of $\boldsymbol{C}'$ is numerically expensive. To reduce computational costs, we calculated $\boldsymbol{F}$ as follows:[8]

$$\begin{aligned} \boldsymbol{C}\boldsymbol{F} &= \boldsymbol{S} \\ \boldsymbol{C}^T\boldsymbol{C}\boldsymbol{F} &= \boldsymbol{C}^T\boldsymbol{S} \\ (\boldsymbol{C}^T\boldsymbol{C})^{-1}(\boldsymbol{C}^T\boldsymbol{C})\boldsymbol{F} &= (\boldsymbol{C}^T\boldsymbol{C})^{-1}\boldsymbol{C}^T\boldsymbol{S} \\ \boldsymbol{F} &= (\boldsymbol{C}^T\boldsymbol{C})^{-1}\boldsymbol{C}^T\boldsymbol{S}. \end{aligned} \tag{11}$$

In this way, matrix inversion is required for a much smaller matrix $\boldsymbol{C}^T\boldsymbol{C}$, which is a square matrix of size 40,639 × 40,639.

To evaluate model performance, we compared the estimated semantic vectors with the targeted semantic vectors using the Pearson correlation. We therefore calculated the 131,673 × 131,673 correlation matrix for all possible pairs of estimated and targeted semantic vectors. Precision was calculated by comparing predicted vectors with the gold standard provided by the targeted vectors. Recognition was defined to be successful if the correlation of the predicted vector with the targeted gold vector was the highest of all the pairwise correlations of this predicted vector with any of the other gold semantic vectors. Precision, defined as the proportion of correct recognitions divided by the total number of audio tokens, was at 33.61%. For the correctly identified words, the mean correlation was 0.72, for the incorrectly identified words, it was 0.55. To place this in perspective, a naive discrimination model with discrete lexomes as output performed at 12%, and a deep convolution network, Mozilla DeepSpeech (`https://github.com/mozilla/DeepSpeech`, based on Hannun et al. (2014)) performed at 6%. The low performance of Mozilla DeepSpeech is due primarily due to its dependence on a language model. When presented with utterances instead of single words, it performs remarkably well.

4.4 Discussion

A problem with naive discriminative learning that has been puzzling for a long time is that measures based on NDL performed well as predictors of processing times (Baayen et al., 2011; Milin et al., 2017b), whereas accuracy for lexical decisions was low. An accuracy of 27% for a model performing an 11,480-classification task is perhaps reasonable, but the lack of precision is unsatisfactory when the goal is to model human visual lexicality decisions. By moving from NDL to LDL, model accuracy is substantially improved (to 59%). At the same time, predictions for reaction times improved

[8] The transpose of a square matrix $\boldsymbol{X}$, denoted by $\boldsymbol{X}^T$ is obtained by replacing the upper triangle of the matrix by the lower triangle, and vice versa. Thus,

$$\begin{pmatrix} 3 & 8 \\ 7 & 2 \end{pmatrix}^T = \begin{pmatrix} 3 & 7 \\ 8 & 2 \end{pmatrix}.$$

For a non-square matrix, the transpose is obtained by switching rows and columns. Thus, a 4 × 2 matrix becomes a 2 × 4 matrix when transposed.

considerably as well. As lexicality decisions do not require word identification, further improvement in predicting decision behavior is expected to be possible by considering not only whether the predicted semantic is closest to the targeted vector, but also measures such as how densely the space around the predicted semantic vector is populated.

Accuracy for auditory comprehension is lower, for the data we considered above at around 33%. Interestingly, a series of studies indicates that recognizing isolated words taken out of running speech is a non-trivial task also for human listeners (Pickett and Pollack, 1963; Shockey, 1998; Ernestus et al., 2002). Correct identification by native speakers of 1000 randomly sampled word tokens from a German corpus of spontaneous conversational speech ranged between 21% and 44% (Arnold et al., 2017). For both human listeners and automatic speech recognition systems, recognition improves considerably when words are presented in their natural context. Given that LDL with FBSFs performs very well on isolated word recognition, it seems worth investigating further whether the present approach can be developed into a fully-fledged model of auditory comprehension that can take full utterances as input. For a blueprint of how we plan to implement such a model, see Baayen et al. (2016b).

5 Speech production

This section examines whether we can predict words' forms from the semantic vectors of $\boldsymbol{S}$. If this is possible for the present dataset with reasonable accuracy, we have a proof of concept that discriminative morphology is feasible not only for comprehension, but also for speech production. The first subsection introduces the computational implementation. The next subsection reports on the model's performance, which is evaluated first for monomorphemic words, then for inflected words, and finally for derived words. Subsection 5.3 provides further evidence for the production network by showing that as the support from the semantics for the triphones becomes weaker, the amount of time required for articulating the corresponding segments increases.

5.1 Computational implementation

For a production model, some representational format is required for the output that in turn drives articulation. In what follows, we make use of triphones as output features. Triphones capture part of the contextual dependencies that characterize speech and that render problematic the phoneme as elementary unit of a phonological calculus (Port and Leary, 2005). Triphones are in many ways not ideal, in that they inherit the limitations that come with discrete units. Other output features, structured along the lines of gestural phonology (Browman and Goldstein, 1992), or time series of movements of key articulators registered with electromagnetic articulography or ultrasound are on our list for further exploration. For now, we use triphones as a convenience construct, and we will show that given the support from the semantics for the triphones, the sequence of phones can be reconstructed with high accuracy. As some models of speech production assemble articulatory targets from phone segments (e.g., Tourville and Guenther, 2011), the present implementation can be seen as a front-end for this type of model.

Before putting the model to the test, we first clarify the way the model works by means of our toy lexicon with the words *one, two, three*. Above, we introduced the semantic matrix $\boldsymbol{S}$ (equation 7), which we repeat here for convinience,

$$\boldsymbol{S} = \begin{array}{c} \\ \text{one} \\ \text{two} \\ \text{three} \end{array} \begin{array}{c} \begin{array}{ccc} \text{one} & \text{two} & \text{three} \end{array} \\ \begin{pmatrix} 1.0 & 0.3 & 0.4 \\ 0.2 & 1.0 & 0.1 \\ 0.1 & 0.1 & 1.0 \end{pmatrix} \end{array}. \tag{12}$$

as well as an $\boldsymbol{C}$ indicator matrix specifying which triphones occur in which words (equation 6). As in what follows this matrix specifies the triphones targeted for production, we hence forth refer to this matrix as the $\boldsymbol{T}$ matrix.

$$\boldsymbol{T} = \begin{array}{c} \\ \text{one} \\ \text{two} \\ \text{three} \end{array} \begin{array}{c} \begin{array}{cccccccc} \text{\#wV} & \text{wVn} & \text{Vn\#} & \text{\#tu} & \text{tu\#} & \text{\#Tr} & \text{Tri} & \text{ri\#} \end{array} \\ \begin{pmatrix} 1 & 1 & 1 & 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 1 & 1 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 & 1 & 1 & 1 \end{pmatrix} \end{array}. \tag{13}$$

For production, our interest is in the matrix $\boldsymbol{G}$ that transforms the row vectors of $\boldsymbol{S}$ into the row vectors of $\boldsymbol{T}$, i.e., we need to solve

$$\boldsymbol{SG} = \boldsymbol{T}. \tag{14}$$

Given $\boldsymbol{G}$, we can predict for any semantic vector $\boldsymbol{s}$ the vector of triphones $\hat{\boldsymbol{t}}$ that quantifies the support for the triphones provided by $\boldsymbol{s}$, simply by multiplying $\boldsymbol{s}$ with $\boldsymbol{G}$.

$$\hat{\boldsymbol{t}} = \boldsymbol{sG}. \tag{15}$$

As before, the transformation matrix $\boldsymbol{G}$ is straightforward to estimate. Let $\boldsymbol{S}'$ denote the Moore-Penrose generalized inverse of $\boldsymbol{S}$. Since

$$\boldsymbol{S}'\boldsymbol{SG} = \boldsymbol{S}'\boldsymbol{T}$$

we have that

$$\boldsymbol{G} = \boldsymbol{S}'\boldsymbol{T}.$$

Given $\boldsymbol{G}$, we can predict the triphone matrix $\hat{\boldsymbol{C}}$ from the semantic matrix $\boldsymbol{S}$:

$$\boldsymbol{SG} = \hat{\boldsymbol{T}}.$$

For the present example, the inverse of $\boldsymbol{S}$, $\boldsymbol{S}'$, is

$$\boldsymbol{S}' = \begin{array}{c} \\ \text{one} \\ \text{two} \\ \text{three} \end{array} \begin{array}{c} \begin{array}{ccc} \text{one} & \text{two} & \text{three} \end{array} \\ \begin{pmatrix} 1.10 & -0.29 & -0.41 \\ -0.21 & 1.07 & -0.02 \\ -0.09 & -0.08 & 1.04 \end{pmatrix} \end{array} \tag{16}$$

and the transformation matrix $\boldsymbol{G}$ is

$$\boldsymbol{G} = \begin{array}{c} \\ \text{one} \\ \text{two} \\ \text{three} \end{array} \begin{array}{c} \begin{array}{cccccccc} \text{\#wV} & \text{wVn} & \text{Vn\#} & \text{\#tu} & \text{tu\#} & \text{\#Tr} & \text{Tri} & \text{ri\#} \end{array} \\ \begin{pmatrix} 1.10 & 1.10 & 1.10 & -0.29 & -0.29 & -0.41 & -0.41 & -0.41 \\ -0.21 & -0.21 & -0.21 & 1.07 & 1.07 & -0.02 & -0.02 & -0.02 \\ -0.09 & -0.09 & -0.09 & -0.08 & -0.08 & 1.04 & 1.04 & 1.04 \end{pmatrix} \end{array}. \tag{17}$$

For this simple example, $\hat{\boldsymbol{T}}$ is virtually identical to $\boldsymbol{T}$. For realistic data, $\hat{\boldsymbol{T}}$ will not be identical to $\boldsymbol{T}$, but will be an approximation of it that is optimal in the least squares sense. The triphones with the strongest support are expected to be the most likely to be the triphones defining a word's form.

We made use of the $\boldsymbol{S}$ matrix, which we derived from the TASA corpus as described in section 3. The majority of columns of the 23,561 × 23,561 matrix $\boldsymbol{S}$ show very small deviations from zero, and hence are uninformative. As before, we reduced the number of columns of $\boldsymbol{S}$ by removing columns with very low variance. Here, one option is to remove all columns with a variance below a preset threshold θ. However, to avoid adding a threshold as a free parameter, we set the number of columns retained to the number of different triphones in a given dataset, a number which is around $n = 4500$. In summary, $\boldsymbol{S}$ denotes a $w \times n$ matrix that specifies, for each of w words, an n-dimensional semantic vector.

5.2 Model performance

5.2.1 Performance on monolexomic words

We first examined model performance on a dataset comprising monolexomic words that did not carry any inflectional exponents. This dataset of 3987 words comprised three irregular comparatives (*elder, less, more*), two irregular superlatives (*least, most*), as well as 28 irregular past tense forms, and one irregular past participle (*smelt*). For this set of words, we constructed a 3987 $\times$ 4446 matrix of semantic vectors $\boldsymbol{S}_m$ (a submatrix of the $\boldsymbol{S}$ matrix introduced above in section 3) and a 3987 $\times$ 4446 triphone matrix $\boldsymbol{T}_m$ (a submatrix of $\boldsymbol{T}$). The number of colums of $\boldsymbol{S}_m$ was set to the number of different triphones (the column dimension of $\boldsymbol{T}_m$). Those column vectors of $\boldsymbol{S}_m$ were retained that had the 4446 highest variances. We then estimated the transformation matrix $\boldsymbol{G}$ and used this matrix to predict the triphones that define words' forms.

We evaluated model performance in two ways. First, we inspected whether the triphones with maximal support were indeed the targeted triphones. This turned out to be the case for all words. Targeted triphones had an activation value close to one, and non-targeted triphones an activation close to zero. As the triphones are not ordered, we also investigated whether the sequence of phones could be constructed correctly for these words. To this end, we set a threshold of 0.99, extracted all triphones with an activation exceeding this threshold, and used the `all_simple_paths`[9] function from the **igraph** package (Csardi and Nepusz, 2006) to calculate all paths starting with any left-edge triphone in the set of extracted triphones. From the resulting set of paths, we selected the longest path, which invariably was perfectly aligned with the sequence of triphones that defines words' forms.

We also evaluated model performance with a second, more general, heuristic algorithm that also makes use of the same algorithm from graph theory. Our algorithm sets up a graph with vertices collected from the triphones that are best supported by the relevant semantic vectors, and considers all paths it can find that lead from an initial triphone to a final triphone. This algorithm, which is presented in more detail in the appendix, and which is essential for novel complex words, produced the correct form for 3982 out of 3987 words. It selected a shorter form for five words, *int* for *intent*, *lin* for *linnen*, *mis* for *mistress*, *oint* for *ointment*, and *pin* for *pippin*. The correct forms were also found, but ranked second due to a simple length penalty that is implemented in the algorithm.

From these analyses, it is clear that mapping nearly 4000 semantic vectors on their corresponding triphone paths can be accomplished with very high accuracy for English monolexomic words. The question to be addressed next is how well this approach works for complex words. We first address inflected forms, and limit ourselves here to the inflected variants of the present set of 3987 monolexomic words.

5.2.2 Performance on inflected words

Following the classic distinction between inflection and word formation, inflected words did not receive semantic vectors of their own. Nevertheless, we can create semantic vectors for inflected words by adding the semantic vector of an inflectional function to the semantic vector of its base. However, there are several ways in which the mapping from meaning to form for inflected words can be set up. To explain this, we need some further notation.

Let $\boldsymbol{S}_m$ and $\boldsymbol{T}_m$ denote the submatrices of $\boldsymbol{S}$ and $\boldsymbol{T}$ that contain the semantic and triphone vectors of monolexomic words. Assume that a subset of k of these monolexomic words is attested in the training corpus with inflectional function a. Let $\boldsymbol{T}_a$ denote the matrix with the triphone vectors

[9] A path is simple if the vertices it visits are not visited more than once.

of these inflected words, and let $\boldsymbol{S}_a$ denote the corresponding semantic vectors. To obtain $\boldsymbol{S}_a$, we take the pertinent submatrix $\boldsymbol{S}_{m_a}$ from $\boldsymbol{S}_m$ and add the semantic vector $\boldsymbol{s}_a$ of the affix:

$$\boldsymbol{S}_a = \boldsymbol{S}_{m_a} + \boldsymbol{i} \bigotimes \boldsymbol{s}_a. \tag{18}$$

Here, $\boldsymbol{i}$ is a unit vector of length k and $\bigotimes$ is the generalized Kronecker product, which in (18) stacks k copies of $\boldsymbol{s}_a$ row-wise. As a first step, we could define a separate mapping $\boldsymbol{G}_a$ for each inflectional function a,

$$\boldsymbol{S}_a \boldsymbol{G}_a = \boldsymbol{T}_a, \tag{19}$$

but in this set-up, learning of inflected words does not benefit from the knowledge of the base words. This can be remedied by a mapping for augmented matrices that contain the row vectors for both base words and inflected words:

$$\begin{bmatrix} \boldsymbol{S}_m \\ \boldsymbol{S}_a \end{bmatrix} \boldsymbol{G}_a = \begin{bmatrix} \boldsymbol{T}_m \\ \boldsymbol{T}_a \end{bmatrix}. \tag{20}$$

The dimension of $\boldsymbol{G}_a$ (length of semantic vector by length of triphone vector) remains the same, so this option is not more costly than the preceding one. Nevertheless, for each inflectional function, a separate large matrix is required. A much more parsimonious solution is to build augmented matrices for base words and all inflected words jointly:

$$\begin{bmatrix} \boldsymbol{S}_m \\ \boldsymbol{S}_{a_1} \\ \boldsymbol{S}_{a_2} \\ \vdots \\ \boldsymbol{S}_{a_n} \end{bmatrix} \boldsymbol{G} = \begin{bmatrix} \boldsymbol{T}_m \\ \boldsymbol{T}_{a_1} \\ \boldsymbol{T}_{a_2} \\ \vdots \\ \boldsymbol{T}_{a_n} \end{bmatrix} \tag{21}$$

The dimension of $\boldsymbol{G}$ is identical to that of $\boldsymbol{G}_a$, but now all inflectional functions are dealt with by a single mapping. In what follows, we report the results obtained with this mapping.

We selected 6595 inflected variants which met the criterion that the frequency of the corresponding inflectional function was at least 50. This resulted in a dataset with 91 comparatives, 97 superlatives, 2401 plurals, 1333 continuous forms (e.g., *walking*), 859 past tense forms and 1086 forms classed as perfective (past participles), as well as 728 third person verb forms (e.g., *walks*). Many forms can be analyzed as either past tenses or as past participles. We followed the analyses of the treetagger, which resulted in a dataset in which both inflectional functions are well-attested.

Following equation (21), we obtained an (augmented) 10582 $\times$ 5483 semantic matrix $\boldsymbol{S}$, where as before we retained the 5483 columns with the highest column variance. The (augmented) triphone matrix $\boldsymbol{T}$ for this dataset had the same dimensions.

Inspection of the activations of the triphones revealed that targeted triphones had top activations for 85% of the monolexomic words and 86% of the inflected words. The proportion of words with at most one intruding triphone was 97% for both monolexomic and inflected words. The graph-based algorithm performed with an overall accuracy of 94%, accuracies broken down by morphology revealed an accuracy of 99% for the monolexomic words and an accuracy of 92% for inflected words. One source of errors for the production algorithm is inconsistent coding in the CELEX database. For instance, the stem of *prosper* is coded as having a final schwa followed by r, but the inflected forms are coded without the r, creating a mismatch between a partially rhotic stem and completely non-rhotic inflected variants.

We next put model performance to a more stringent test by using 10-fold cross-validation for the inflected variants. For each fold, we trained on all stems and 90% of all inflected forms and then evaluated performance on the 10% of inflected forms that were not seen in training. In this way,

we can ascertain the extent to which our production system (network plus graph-based algorithm for ordering triphones) is productive. We excluded from the cross-validation procedure irregular inflected forms, forms with CELEX phonological forms with inconsistent within-paradigm rhoticism, as well as forms the stem of which was not available in the training set. Thus, cross-validation was carried out for a total of 6236 inflected forms.

As before, the semantic vectors for inflected words were obtained by addition of the corresponding content and inflectional semantic vectors. For each training set, we calculated the transformation matrix $\boldsymbol{G}$ from the $\boldsymbol{S}$ and $\boldsymbol{T}$ matrices of that training set. For an out-of-bag inflected form in the test set, we calculated its semantic vector and multiplied this vector with the transformation matrix (using equation 15) to obtain the predicted triphone vector $\hat{\boldsymbol{t}}$.

The proportion of forms that were predicted correctly was 0.62. The proportion of forms ranked second was 0.25. Forms that were incorrectly ranked first typically were other inflected forms (including bare stems) that happened to receive stronger support than the targeted form. Such errors are not uncommon in spoken English. For instance, in the Buckeye corpus (Pitt et al., 2005), *closest* is once reduced to *clos*. Furthermore, Dell (1986) classified forms such as *concludement* for *conclusion*, and *he relax* for *he relaxes*, as (noncontextual) errors.

The algorithm failed to produce the targeted form for 3% of the cases. Examples of the forms produced instead are the past tense for *blazed* being realized with [zId] instead of [zd], the plural *mouths* being predicted as having [Ts] as coda rather than [Ds], and *finest* being reduced to *finst*. The voiceless production for *mouth* does not follow the dictionary norm, but is used as attested by on-line pronunciation dictionaries. Furthermore, the voicing alternation is partly unpredictable (see, e.g., Ernestus and Baayen, 2003, for final devoicing in Dutch), and hence model performance here is not unreasonable. We next consider model accuracy for derived words.

5.2.3 Performance on derived words

When building the vector space model, we distinguished between inflection and word formation. Inflected words did not receive their own semantic vectors. By contrast, each derived word was assigned its own lexome, together with a lexome for its derivational function. Thus, *happiness* was paired with two lexomes, HAPPINESS and NESS. Since the semantic matrix for our dataset already contains semantic vectors for derived words, we first investigated how well forms are predicted when derived words are assessed along with monolexomic words, without any further differentiation between the two. To this end, we constructed a semantic matrix $\boldsymbol{S}$ for 4885 words (rows) by 4993 (columns), and constructed the transformation matrix $\boldsymbol{G}$ from this matrix and the corresponding triphone matrix $\boldsymbol{T}$. The predicted form vectors of $\hat{\boldsymbol{T}} = \boldsymbol{SG}$ supported the targeted triphones above all other triphones without exception. Furthermore, the graph-based algorithm correctly reconstructed 99% of all forms, with only 5 cases where it assigned the correct form second rank.

Next, we inspected how the algorithm performs when the semantic vectors of derived forms are obtained from the semantic vectors of their base words and those of their derivational lexomes, instead of using the semantic vectors of the derived words themselves. To allow subsequent evaluation by cross-validation, we selected those derived words that contained an affix that occured at least 30 times in our data set (AGAIN (38), AGENT (177), FUL (45), INSTRUMENT (82), LESS (54), LY (127) AND NESS (57)), to a total of 770 complex words.

We first combined these derived words with the 3987 monolexomic words. For the resulting 4885 words, the $\boldsymbol{T}$ and $\boldsymbol{S}$ matrices were constructed, from which we derived the transformation matrix $\boldsymbol{G}$ and subsequently the matrix of predicted triphone strengths $\hat{\boldsymbol{T}}$. The proportion of words for which the targeted triphones were the best supported triphones was 0.96, and the graph algorithm performed with an accuracy of 98.9%.

In order to assess the productivity of the system, we evaluated performance on derived words by means of 10-fold cross-validation. For each fold, we made sure that each affix was present proportionally to its frequency in the overall dataset, and that a derived word's base word was included in the training set.

We first examined performance when the transformation matrix is estimated from a semantic matrix that contains the semantic vectors of the derived words themselves, whereas semantic vectors for unseen derived words are obtained by summation of the semantic vectors of base and affix. It turns out that this set-up results in a total failure. In the present framework, the semantic vectors of derived words are too idiosyncratic, and too finely tuned to their own collocational preferences. They are too scattered in semantic space to support a transformation matrix that supports the triphones of both stem and affix.

We then used exactly the same cross-validation procedure as outlined above for inflected words, constructing semantic vectors for derived words from the semantic vectors of base and affix, and calculating the transformation matrix $\boldsymbol{G}$ from these summed vectors. For unseen derived words, $\boldsymbol{G}$ was used to transform the semantic vectors for novel derived words (obtained by summing the vectors of base and affix) into triphone space.

For 75% of the derived words, the graph algorithm reconstructed the targeted triphones. For 14% of the derived nouns, the targeted form was not retrieved. These include cases such as *resound*, which the model produced with [s] instead of the (unexpected) [z], *sewer*, which the model produced with `@R` instead of only `R`, and *tumbler*, where the model used syllabic [l] instead of nonsyllabic [l] given in the CELEX target form.

5.3 Weak links in the triphone graph and delays in speech production

An important property of the model is that the support for trigrams changes where a word's graph branches out for different morphological variants. This is illustrated in Figure 9 for the inflected form *blending*. Support for the stem-final triphone `End` is still at 1, but then *blend* can end, or continue as *blends*, *blended*, or *blending*. The resulting uncertainty is reflected in the weights on the edges leaving `End`. In this example, the targeted *ing* form is driven by the inflectional semantic vector for CONTINUOUS, and hence the edge to `ndI` is best supported. For other forms, the edge weights will be different, and hence other paths will be better supported.

The uncertainty that arises at branching points in the triphone graph are of interest in the light of several experimental results. For instance, inter keystroke intervals in typing become longer at syllable and morph boundaries (Weingarten et al., 2004; Bertram et al., 2015). Evidence from articulography suggests variability is greater at morph boundaries (Cho, 2001). Longer keystroke execution times and greater articulatory variability are exactly what is expected under reduced edge support. We therefore examined whether the edge weights at the first branching point are predictive for lexical processing. To this end, we investigated the acoustic duration of the segment at the center of the first triphone with a reduced LDL edge weight (in the present example, `d` in `ndI`, henceforth 'branching segment') for those words in our study that are attested in the Buckeye corpus (Pitt et al., 2005).

The dataset we extracted from the Buckeye corpus comprised 15105 tokens of a total of 1327 word types, collected from 40 speakers. For each of these words, we calculated the relative duration of the branching segment, calculated by dividing segment duration by word duration. For the purposes of statistical evaluation, the distribution of this relative duration was brought closer to normality by means of a logarithmic transformation. We fitted a generalized additive mixed model (Wood, 2017) to log relative duration with random intercepts for word and speaker, log NDL edge weight as predictor of interest, and local speech rate (Phrase Rate), log neighborhood density (Ncount)

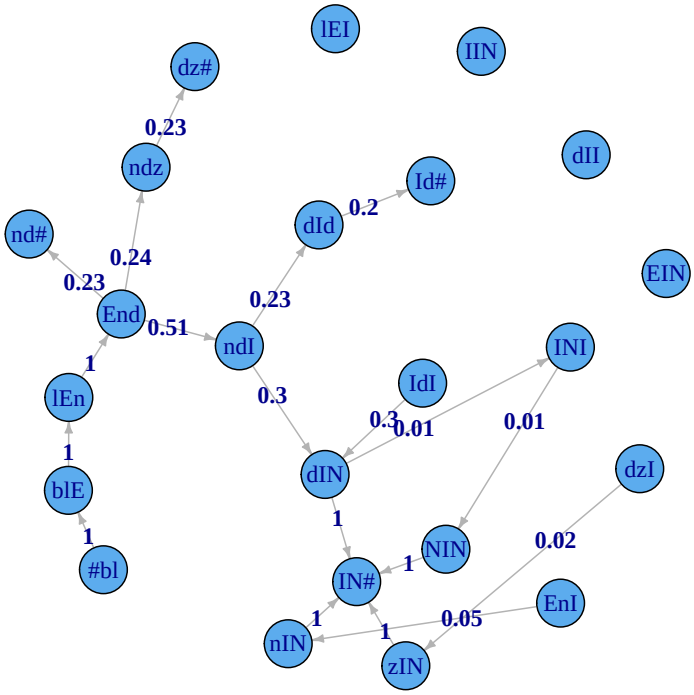

Figure 9: The directed graph for *blending*. Vertices represent triphones, including triphones such as `IIN` that were posited by the graph algorithm to bridge potentially relevant transitions that are not instantiated in the training data. Edges are labelled with their edge weights. The graph, the edge weights of which are specific to the lexomes BLEND and CONTINUOUS, incorporates not only the path for *blending*, but also the paths for *blend*, *blends*, and *blended*.

and log word frequency as control variables. A summary of this model is presented in Table 7. As expected, an increase in LDL Edge Weight goes hand in hand with a reduction in the duration of the branching segment. In other words, when the edge weight is reduced, production is slowed.

The adverse effects of weaknesses in a word's path in the graph where the path branches is of interest against the discussion about the function of weak links in diphone transitions in the literature on comprehension (Seidenberg, 1987; McQueen, 1998; Hay, 2003, 2002; Hay and Baayen, 2003; Hay et al., 2004). For comprehension, it has been argued that bigram or diphone 'troughs', i.e., low transitional probabilities in a chain of high transitional probabilities, provide points where sequences are segmented and parsed into their constituents. From the perspective of discrimination learning, however, low-probability transitions function in exactly the opposite way for comprehension (Baayen et al., 2011, 2016b). Naive discriminative learning also predicts that troughs should give rise to shorter processing times in comprehension, but not because morphological decomposition would proceed more effectively. Since high-frequency boundary bigrams and diphones are typically

A. parametric coefficients	Estimate	Std. Error	t-value	p-value
Intercept	-1.9165	0.0592	-32.3908	< 0.0001
Log NDL Edge Weight	-0.1358	0.0400	-3.3960	0.0007
Phrase Rate	0.0047	0.0021	2.2449	0.0248
Log Ncount	0.1114	0.0189	5.8837	< 0.0001
B. smooth terms	edf	Ref.df	F-value	p-value
TPRS log Frequency	1.9007	1.9050	7.5495	0.0004
random intercepts word	1149.8839	1323.0000	31.4919	< 0.0001
random intercepts speaker	30.2995	39.0000	34.5579	< 0.0001

Table 7: Statistics for the partial effects of a generalized additive mixed model fitted to the relative duration of edge segments in the Buckeye corpus. The trend for log frequency is positive accelerating. TPRS: thin plate regression spline.

used word-internally across many words, they have a low cue validity for these words. Conversely, low-frequency boundary bigrams are much more typical for very specific base+affix combinations, and hence are better discriminative cues that afford enhanced activation of words' lexomes. This, in turn, gives rise to faster processing (see also Ramscar et al. (2014) for the discriminative function of low-frequency bigrams in low-frequency monolexomic words).

There is remarkable convergence between the directed graphs for speech production such as illustrated in Figure 9 and computational models using temporal self-organizing maps (TSOMs, Ferro et al., 2011; Chersi et al., 2014; Pirrelli et al., 2015). TSOMs also use error-driven learning, and in recent work attention is also drawn to weak edges in words' paths in the TSOM and the consequences thereof for lexical processing (Marzi et al., 2018). We are not using TSOMs, however, one reason being that in our experience they do not scale up well to realistically sized lexicons. A second reason is that we are pursuing the hypothesis that form serves meaning, and that self-organization of form by itself is not necessary. This hypothesis is likely too strong, especially as we do not provide a rationale for the trigram and triphone units that we use to represent aspects of form. It is worth noting that spatial organization of form features is not restricted to TSOMs. Although in our approach, the spatial organization of triphone features is left unspecified, such an organization can be easily enforced (see Baayen and Blevins, 2018, for further details) by using an algorithm such as graphopt, which self-organizes the vertices of a graph into a two-dimensional plane. Figure 9 was obtained with this algorithm.[10]

5.4 Discussion

Our production network reconstructs known words' forms with a high accuracy, 99.9% for monolexomic words, 92% for inflected words, and 99% for derived words. For novel complex words, accuracy under 10-fold cross-validation was 62% for inflected words and 75% for derived words.

The drop in performance for novel forms is perhaps unsurprising, given that speakers understand many more words than they themselves produce, even though they hear novel forms on a fairly regular basis as they proceed through life (Keuleers et al., 2015; Ramscar et al., 2017)

However, we also encountered several technical problems that are not due to the algorithm

[10] Graphopt was developed for the layout of large graphs `http://www.schmuhl.org/graphopt/` and is implemented in the **igraph** package. Graphopt uses basic principles of physics to iteratively determine an optimal layout. Each node in the graph is given both mass and an electric charge, and edges between nodes are modeled as springs. This sets up a system in which there are attracting and repelling forces between the vertices of the graph, and this physical system is simulated until it reaches an equilibrium.

but to the representations that the algorithm has had to work with. First, it is surprising that accuracy is as high as it is given that the semantic vectors are constructed from a small corpus with a very simple discriminative algorithm. Second, we encountered inconsistencies in the phonological forms retrieved from the CELEX database, inconsistencies that in part are due to the use of discrete triphones. Third, many cases where the model predictions do not match the targeted triphone sequence, the targeted forms have a minor irregularity (e.g., *resound* with [z] instead of [s]). Fourth, several of the typical errors that the model makes are known kinds of speech errors or reduced forms that one might encounter in engaged conversational speech.

It is noteworthy that the model is almost completely data-driven. The triphones are derived from CELEX, and other than some manual corrections for inconsistencies, are derived automatically from words' phone sequences. The semantic vectors are based on the TASA corpus, and were not in any way optimized for the production model. Given the triphone and semantic vectors, the transformation matrix is completely determined. No by-hand engineering of rules and exceptions is required, nor is it necessary to specify with hand-coded links what the first segment of a word is, what its second segment is, etc., as in the WEAVER model (Levelt et al., 1999). It is only in the heuristic graph algorithm that three thresholds are required, in order to avoid that graphs become too large to remain computationally tractable.

The speech production system that emerges from this approach comprises first of all an excellent memory for forms that have been encountered before. Importantly, this memory is not a static memory, but a dynamic one. It is not a repository of stored forms, but it reconstructs the forms from the semantics it is requested to encode. For regular unseen forms, however, a second network is required that projects a regularized semantic space (obtained by accumulation of the semantic vectors of content and inflectional or derivational functions) onto the triphone output space. Importantly, it appears that no separate transformations or rules are required for individual inflectional or derivational functions.

Jointly, the two networks, both of which can also be trained incrementally using the learning rule of Widrow-Hoff (Widrow and Hoff, 1960), define a dual route model; attempts to build a single integrated network were not successful. The semantic vectors of derived words are too idiosyncratic to allow generalization for novel forms. It is the property of the semantic vectors of derived lexomes described in section 3.2.3, namely, that they are close to but not inside the cloud of their content lexomes, that makes them productive. Novel forms do not partake in the idiosyncracies of lexicalized existing words, but instead remain bound to base and derivational lexomes. It is exactly this property that makes them pronounceable. Once produced, a novel form will then gravitate as experience accumulates towards the cloud of semantic vectors of its morphological category, meanwhile also developing its own idiosyncracies in pronunciation, including sometimes highly reduced pronunciation variants (e.g., /tyk/ for Dutch *natuurlijk* ([natyrl@k]) (Johnson, 2004; Ernestus, 2000; Kemps et al., 2004). It is perhaps possible to merge the two routes into one, but for this, much more fine-grained semantic vectors are required that incorporate information about discourse and the speaker's stance with respect to the adressee (see, e.g. Hawkins, 2003, for detailed discussion of such factors).

6 Bringing in time

Thus far, we have not considered the role of time. The audio signal comes in over time, longer words are typically read with more than one fixation, and likewise, articulation is a temporal process. In this section, we briefly outline how time can be brought into the model. We do so by discussing the reading of proper names.

John Clark wrote great books about ants.
John Clark published a great book about dangerous ants.
John Welsch makes great photographs of airplanes.
John Welsch has a collection of great photographs of airplanes landing.
John Wiggam will build great harpsichords.
John Wiggam will be a great expert in tuning harpsichords.
Anne Hastie has taught statistics to great second year students.
Anne Hastie has taught probability to great first year students.
Janet Clark teaches graph theory to second year students.
JohnClark, write, great, book, about, ant
JohnClark, publish, a, great, book, about, dangerous, ant
JohnWelsch, make, great, photograph, of, airplane
JohnWelsch, have, a, collection, of, great, photograph, airplane, landing
JohnWiggam, future, build, great, harpsichord
JohnWiggam, future, be, a, great, expert, in, tuning, harpsichord
AnneHastie, have, teach, statistics, to, great, second, year, student
AnneHastie, have, teach, probability, to, great, first, year, student
JanetClark, teach, graphtheory, to, second, year, student

Table 8: Example sentences for the reading of proper names. To keep the example simple, inflectional lexomes are not taken into account.

Proper names (morphologically compounds) pose a challenge to compositional theories, as the people referred to by names such as Richard Dawkins, Richard Nixon, Richard Thompson, and Sandra Thompson are not obviously semantic composites of each other. We therefore assign lexomes to names, irrespective of whether the individuals referred to are real or fictive, alive or dead. Furthermore, we assume that when the personal name and family name receive their own fixations, both names are understood as pertaining to the same named entity, which is therefore coupled with its own unique lexome. Thus, for the example sentences in Table 8, the letter trigrams of *John* are paired with the lexome JohnClark, and likewise the letter trigrams of *Clark* are paired with this lexome.

By way of illustration, we obtained the matrix of semantic vectors training on 100 randomly ordered tokens of the sentences of Table 8. We then constructed a trigram cue matrix $\boldsymbol{C}$ specifying, for each word (*John, Clark, wrote, ...*), which trigrams it contains. In parallel, a matrix $\boldsymbol{L}$ specifying for each word its corresponding lexomes (JohnClark, JohnClark, write, ...) was set up. We then calculated the matrix $\boldsymbol{F}$ by solving $\boldsymbol{CF} = \boldsymbol{L}$, and used $\boldsymbol{F}$ to calculate estimated (predicted) semantic vectors $\hat{\boldsymbol{L}}$. Figure 10 presents the correlations of the estimated semantic vectors for the word forms *John*, *John Welsch*, *Janet*, and *Clark* with the targeted semantic vectors for the named entities JohnClark, JohnWelsch, JohnWiggam, AnneHastie, and JaneClark. For the sequentially read words *John* and *Welsch*, the semantic vector generated by *John* and that generated by *Welsch* were summed to obtain the integrated semantic vector for the composite name.

Figure 10, upper left panel, illustrates that upon reading the personal name *John*, there is considerable uncertainty about which named entity is at issue. When subsequently the family name is read (upper left panel), uncertainty is reduced and *John Welsch* now receives full support, whereas other named entities have correlations close to zero, or even negative correlations. As in the small world of the present toy example *Janet* is a unique personal name, there is no uncertainty about what named entity is at issue when *Janet* is read. For the family name *Clark* on its own, by contrast, *John Clark* or *Janet Clark* are both viable options.

For the present example, we assumed that every orthographic word received one fixation. However, depending on whether the eye lands far enough into the word, names such as *John Clark* and

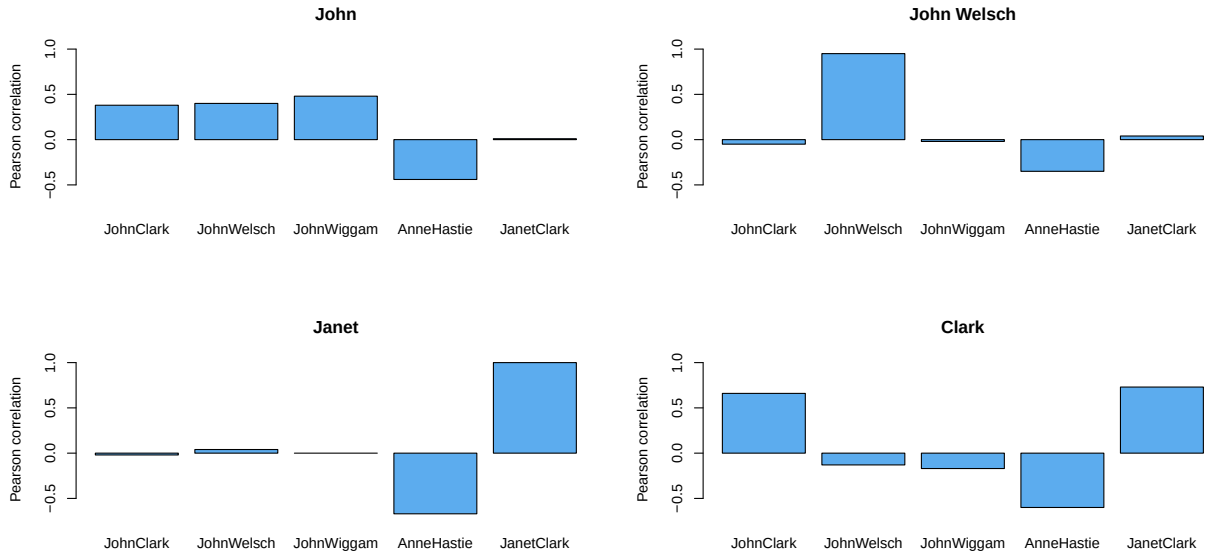

Figure 10: Correlations of estimated semantic vectors for *John*, *John Welsch*, *Janet*, and *Clark* with the targeted semantic vectors of JohnClark, JohnWelsch, JohnWiggam, AnneHastie, and JaneClark.

compounds such as *graph theory* can also be read with a single fixation. For single-fixation reading, learning events should comprise the joint trigrams of the two orthographic words, which are then simultaneously calibrated against the targeted lexome (JohnClark, graphtheory).

Obviously, the true complexities of reading, such as parafoveal preview, are not captured by this example. Nevertheless, as a rough first approximation, this example shows a way forward for integrating time into the discriminative lexicon.

7 General discussion

We have presented a unified discrimination-driven model for visual and auditory word comprehension and word production: the discriminative lexicon. The layout of this model is visualized in Figure 11. Input (cochlea and retina) and output (typing and speaking) systems are presented in light gray, the representations of the model are shown in dark gray. For auditory comprehension, we make use of frequency band summary (FBS) features, low-level cues that we derive automatically from the speech signal. The FBS features serve as input to the auditory network $\boldsymbol{F}_a$ that maps vectors of FBS features onto the semantic vectors of $\boldsymbol{S}$. For reading, letter trigrams represent the visual input at a level of granularity that is optimal for a functional model of the lexicon (see Cohen and Dehaene, 2009, for a discussion of lower-level visual features). Trigrams are relatively high-level features compared to the FBS features. For features implementing visual input at a much lower level of visual granularity, using histograms of oriented gradients (Dalal and Triggs, 2005), see Linke et al. (2017). Letter trigrams are mapped by the visual network $\boldsymbol{F}_o$ onto $\boldsymbol{S}$.

For reading, we also implemented a network, here denoted by $\boldsymbol{K}_a$, that maps trigrams onto auditory targets (auditory verbal images), represented by triphones. These auditory triphones in turn are mapped by network $\boldsymbol{H}_a$ onto the semantic vectors $\boldsymbol{S}$. This network is motivated not only by our observation that for predicting visual lexical decision latencies, triphones outperform trigrams

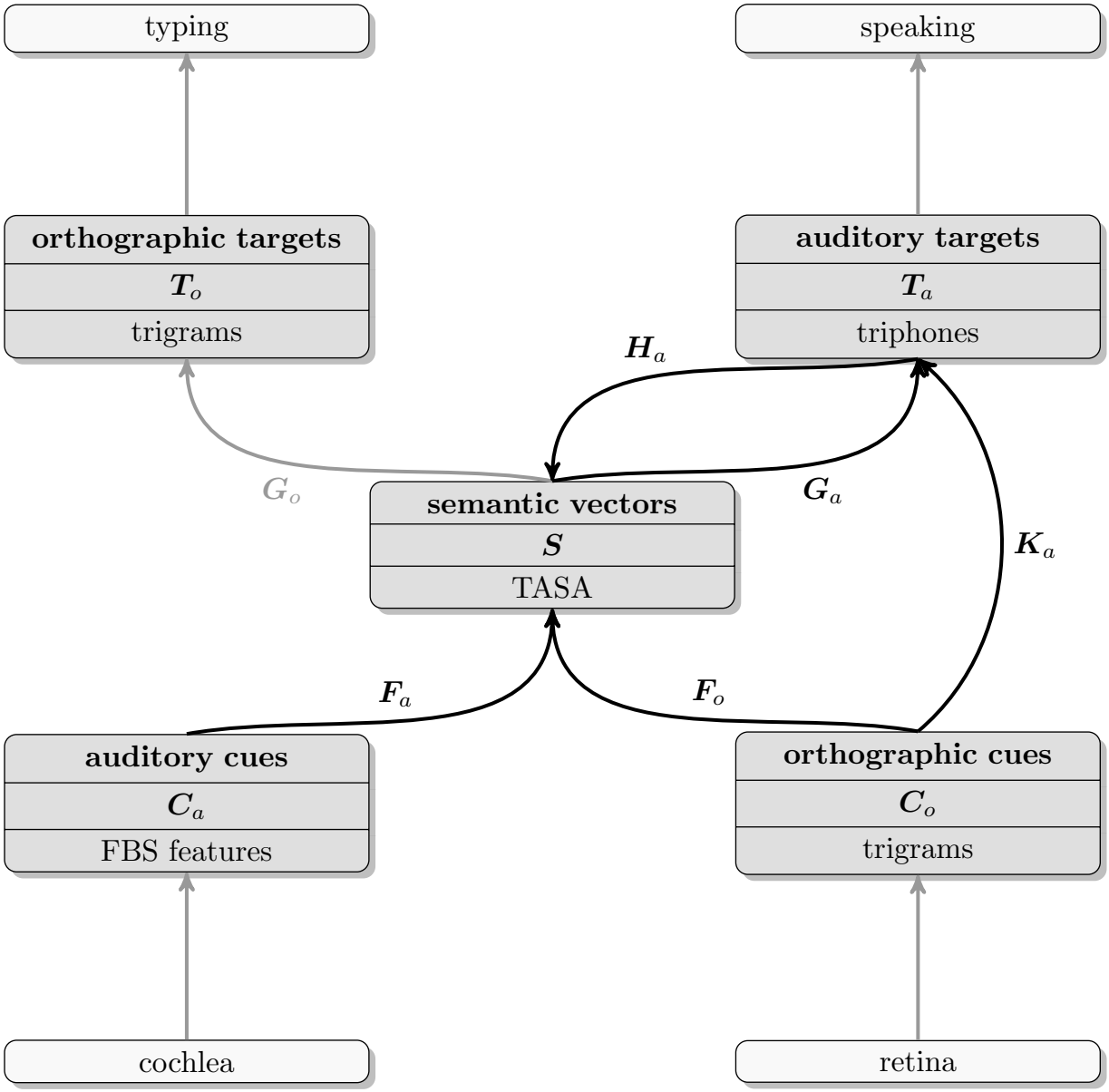

Figure 11: Overview of the discriminative lexicon. Input and output systems are presented in light gray, the representations of the model are shown in dark gray. Each of these representations is a matrix with its own set of features. Arcs are labeled with the discriminative networks (transformation matrices) that map matrices onto each other. The arc from semantic vectors to orthographic vectors is in gray, as this mapping is not explored in the present study.

by a wide margin. Network $\boldsymbol{H}_a$ is also part of the control loop for speech production, see Hickok (2014) for discussion of this control loop and the necessity of auditory targets in speech production. Furthermore, the acoustic durations with which English speakers realize the stem vowels of English verbs (Tucker et al., 2018), as well as the duration of word final *s* in English (Plag et al., 2017) can be predicted from NDL networks using auditory targets to discriminate between lexomes (Tomaschek et al., 2018).

Speech production is modeled by network $\boldsymbol{G}_a$, which maps semantic vectors onto auditory

targets, which in turn serve as input for articulation. Although current models of speech production assemble articulatory targets from phonemes (see, e.g., Tourville and Guenther, 2011; Hickok, 2014), a possibility that is certainly compatible with our model when phones are recontextualized as triphones, we think that it is worthwhile investigating whether a network mapping semantic vectors onto vectors of articulatory parameters that in turn drive the articulators can be made to work, especially since many reduced forms are not straightforwardly derivable from the phone sequences of their 'canonical' forms (Ernestus, 2000; Johnson, 2004).

We have shown that the networks of our model have reasonable to high production and recognition accuracies, and also generate a wealth of well-supported predictions for lexical processing. Central to all networks is the hypothesis that the relation between form and meaning can be modeled discriminatively with large but otherwise surprisingly simple linear networks, the underlying mathematics of which (linear algebra) is well-understood. Given the present results, it is clear that the potential of linear algebra for understanding the lexicon and lexical processing has thus far been severely underestimated.[11] In what follows, we touch upon a series of issues that are relevant for evaluating our model.

Incremental learning. In the present study, we estimated the weights on the connections of these networks with matrix operations, but, importantly, these weights can also be estimated incrementally, using the learning rule of Widrow and Hoff (1960); further improvements in accuracy are expected when using the Kalman filter (Kalman, 1960). As all matrices can be updated incrementally (and this holds as well for the matrix with semantic vectors, which are also time-variant), and in theory should be updated incrementally whenever information about the order of learning events is available, the present theory has potential for studying lexical acquisition and the continuing development of the lexicon over the lifetime (Ramscar et al., 2017).

Morphology without compositional operations. We have shown that our networks are predictive for a range of experimental findings. Important from the perspective of discriminative linguistics is that there are no compositional operations in the sense of Frege and Russell (Frege, 1879; Russell, 1942, 1905). The work on compositionality in logic has deeply influenced formal linguistics (e.g. Montague, 1973; Hornstein, 1995), and has led to the belief that the "architecture of the language faculty" is grounded in a homomorphism between a calculus (or algebra) of syntactic representations and a calculus (or algebra) based on semantic primitives. Within this tradition, compositionality arises when rules combining representations of form are matched with rules combining representations of meaning. The approach of Marelli and Baroni (2015), who derive the semantic vector of a complex word from the semantic vector of its base and a dedicated affix-specific mapping from the pertinent base words to the corresponding derived words, is in the spirit of compositionality, in that a change in form (through affixation) goes hand in hand with a change in meaning (modeled with a linear mapping in semantic space).

The perspective on morphological processing developed in the present study breaks with this tradition. Although semantic vectors for inflected words are obtained by summation of the semantic vectors of the base word and those of its inflectional functions (see Baayen et al., 2018, for modeling of the rich verbal paradigms of Latin), the form vectors corresponding to the resulting semantic

[11] The model as outlined in Figure 11 is a modular one, for reasons of computational convenience and interpretability. The different networks probably interact (see for some evidence Harm and Seidenberg, 2004). One such interaction is explored in Baayen et al. (2018). In their production model for Latin inflection, feedback from the projected triphone paths back to the semantics (synthesis by analysis) is allowed, so that of otherwise equally well supported paths, the path that best expresses the targeted semantics is selected. For information flows between systems in speech production, see Hickok (2014).

vectors are not obtained by the summation of vectors for chunks of form, nor by any other operations on forms. Attempts to align chunks of words' form with their semantic structures — typically represented by graphs constructed over semantic primitives (see, e.g., Jackendoff, 1990) — are destined to end in quagmires of arbitrary decisions about the number of inflectional classes required (Estonian, 30 or 300?), what the chunks should look like (does German have a morpheme *d* in its articles and demonstratives *der, die, das, diese, ...*?), whether Semitic morphology requires operations on form on one or two tiers (McCarthy, 1981; Ussishkin, 2005, 2006), and how many affixal position slots are required for Athebascan languages such as Navajo (Young and Morgan, 1980). A language that is particularly painful for compositional theories is Yélî Dnye, an Indo-Pacific language spoken on Rossel island (in the south-east of Papua New Guinea). The language has a substantial inventory of over a 1000 function words used for verb agreement. Typically, these words are monosyllables that simultaneously specify values for negation, tense, aspect, person and number of subject, deixis, evidentiality, associated motion, and counterfactuality. Furthermore, verbs typically have no less than eight different stems, the use of which depends on particular constellations of inflectional functions (Levinson and Majid, 2013; Levinson, 2006). Yélî Dnye thus provides a striking counterexample to the hypothesis that a calculus of meaning would be paralleled by a calculus of form.

Whereas the theory of linear discriminative learning radically rejects the idea that morphological processing is grounded in compositionality as defined in mainstream formal linguistic theory, it does allow for the conjoint expression in form of lexical meanings and inflectional functions, and for forms to map onto such conjoint semantic vectors. But just as multiple inflectional functions can be expressed jointly in a single word, multiple words can map onto non-conjoint semantic vectors, as was illustrated above for named entity recognition: *John Wiggam* can be understood to refer to a very specific person without this person being a compositional function of the proper name *John* and the surname *Wiggam*. In both these cases, there is no isomorphy between inflectional functions and lexical meanings on the one hand, and chunks of form on the other hand.

Of course, conjoint vectors for lexical meanings and inflectional functions can be implemented and made to work only because a Latin verb form such as *sapīvissēmus* is understood, when building the semantic space $\boldsymbol{S}$, to express lexomes for person (first), number (plural), voice (active), tense (pluperfect), mood (subjunctive), and lexical meaning (to know) (see Baayen et al., 2018, for computational modeling details). Although this could loosely be described as a decompositional approach to inflected words, we prefer not to use the term 'decompositional' as in formal linguistics this term, as outlined above, has a very different meaning. A more adequate terminology would be that the system is conjoint realizational for production and conjoint inferential for comprehension. (We note here that we have focused on temporally conjoint realization and inference, leaving temporally disjoint realization and inference for sequences of words, and possibly compounds and fixed expression, for further research.)

Naive versus discriminative learning. There is one important technical difference between the learning engine of the current study, LDL, and naive discriminative learning (NDL). Because NDL makes the simplifying assumption that outcomes are orthogonal, the weights from the cues to a given outcome are independent of the weights from the cues to the other outcomes. This independence assumption, which motivates the word *naive* in naive discriminative learning, makes it possible to very efficiently update network weights during incremental learning. In LDL, by contrast, this independence no longer holds: learning is no longer naive. We therefore refer to our networks not as NDL networks but simply as *linear* discriminative learning networks. Actual incremental updating of LDL networks is computationally more costly, but further numerical optimization is possible and implementations are under way.

Scaling. An important property of the present approach is that good results are obtained already for datasets of modest size. Our semantic matrix is derived from the TASA corpus, which with only 10 million words is dwarfed by the 2.8 billion words of the ukWaC corpus used by Marelli and Baroni (2015), more words than any individual speaker can ever encounter in their lifetime. Similarly, Arnold et al. (2017) report good results for word identification when models are trained on 20 hours of speech, which contrasts with the huge volumes of speech required for training deep learning networks for speech recognition. Although performance increases somewhat with more training data (Shafaei-Bajestan and Baayen, 2018), it is clear that considerable headway can be made with relatively moderate volumes of speech. It thus becomes feasible to train models on, for instance, Australian English and New Zealand English, using already existing speech corpora, and to implement networks that make precise quantitative predictions for how understanding is mediated by listeners' expectations about their interlocutors (see, e.g., Hay and Drager, 2010).

In this study, we have obtained good results with a semantic matrix with a dimensionality of around 4000 lexomes. By itself, we find it surprising that a semantic vector space can be scaffolded with a mere 4000 words. But this finding may also shed light on the phenomenon of childhood amnesia (Henri and Henri, 1895; Freud, 1953). Bauer and Larkina (2014) pointed out that young children have autobiographical memories that they may not remember a few years later. They argue that the rate of forgetting diminishes as we grow older, and stabilizes in adulthood. Part of this process of forgetting may relate to the changes in the semantic matrix. Crucially, we expect the correlational structure of lexomes to change considerably over time as experience accumulates. If this is indeed the case, the semantic vectors for the lexomes for young children are different from the corresponding lexomes for older children, which will again differ from those of adults. As a consequence, autobiographical memories that were anchored to the lexomes at a young age can no longer be accessed as the semantic vectors to which these memories were originally anchored no longer exist.

No examplars. Given how we estimate the transformation matrices with which we navigate between form and meaning, it might seem that our approach is in essence an implementation of an exemplar-based theory of the mental lexicon. After all, for instance the $\boldsymbol{C}_a$ matrix specifies, for each auditory word form (exemplar), its pertinent frequency band summary features. However, the computational implementation of the present study should not be confused with the underlying algorithmic conceptualization. The underlying conceptualization is that learning proceeds incrementally, trial by trial, with the weights of transformation networks being updated with the learning rule of Widrow and Hoff (1960). In the mean, trial-by-trial updating with the Widrow-Hoff learning rule results in the same expected connection weights as obtained by the present estimates using the generalized inverse. The important conceptual advantage of incremental learning is, however, that there is no need to store exemplars. By way of example, for auditory comprehension, acoustic tokens are given with the input to the learning system. These tokens, the ephemeral result of interacting with the environment, leave their traces in the transformation matrix $\boldsymbol{F}_a$, but are not themselves stored. The same holds for speech production. We assume that speakers dynamically construct the semantic vectors from their past experiences, rather than retrieving these semantic vectors from some dictionary-like fixed representation familiar from current file systems. The dynamic nature of these semantic vectors emerged loud and clear from the modeling of novel complex words for speech production under cross-validation. In other words, our hypothesis is that all input and output vectors are ephemeral, in the sense that they represent temporary states of the cognitive system that are not themselves represented in that system but that leave their traces in the transformation matrices that jointly characterize and define the cognitive system that we approximate

with the discriminative lexicon.

This dynamic perspective on the mental lexicon as consisting of several coupled networks that jointly bring the cognitive system into states that in turn feed into further networks (not modeled here) for sentence integration (comprehension) and articulation (production) raises the question about the status of units in this theory. Units play an important role when constructing numeric vectors for form and meaning. Units for letter trigrams and phone trigrams are, of course, context-enriched letter and phone units. Content lexomes as well as lexomes for derivational and inflectional functions are crutches for central semantic functions ranging from functions realizing relatively concrete onomasiological functions ('this is a dog, not a cat') to abstract situational functions allowing entities, events, and states to be specified on the dimensions of time, space, quantity, aspect, etc. However, crucial to the central thrust of the present study, there are no morphological symbols (units combining form and meaning) nor morphological form units such as stems and exponents of any kind in the model. Above, we have outlined the many linguistic considerations that have led us not to want to incorporate such units as part of our theory. Importantly, as there are no hidden layers in our model, it is not possible to seek for 'subsymbolic' reflexes of morphological units in patterns of activation over hidden layers. Here, we depart from earlier connectionist models, which rejected symbolic representations but retained the belief that there should be morphological represention, albeit subsymbolic ones. Seidenberg and Gonnerman (2000), for instance, discuss morphological representations as being 'interlevel representations' that are emergent reflections of correlations among orthography, phonology, and semantics. This is not to say that the more agglutinative a language is, the more the present model will seem, to the outside observer, to be operating with morphemes. But the more a language tends towards fusional or polysynthetic morphology, the less strong this impression will be.

Model complexity Next consider the question of how difficult lexical processing actually is. Sidestepping the complexities of the neural embedding of lexical processing (Tourville and Guenther, 2011; Hickok, 2014), here we narrow down this question to algorithmic complexity. State-of-the-art models in psychology (Levelt et al., 1999; Norris and McQueen, 2008; Coltheart et al., 2001) implement many layers of hierarchically organized units, and many hold it for an established fact that such units are in some sense psychologically real (see, e.g., Zwitserlood, 2018; Butz and Kutter, 2016; Marantz, 2013). However, empirical results can be equally well understood without requiring the theoretical construct of the morpheme (see, e.g., Harm and Seidenberg, 1999; Seidenberg and Gonnerman, 2000; Harm and Seidenberg, 2004; Gonnerman et al., 2007; Baayen et al., 2011; Milin et al., 2017b). The present study illustrates this point for 'boundary' effects in written and oral production. When paths in the triphone graph branch, uncertainty increases and processing slows down. Although observed 'boundary' effects are compatible with a post-Bloomfieldian lexicon, the very same effects arise in our model, even though morphemes do not figure in any way in our computations.

Importantly, the simplicity of the linear mappings in our model of the discriminative lexicon allows us to sidestep the problem that typically arise in computationally implemented full-scale hierarchical systems, namely, that errors arising at lower levels propagate to higher levels. The major steps forward made in recent end-to-end models in language engineering suggest that end-to-end cognitive models are also likely to perform much better. One might argue that the hidden layers of deep learning networks represent the traditional 'hidden layers' of phonemes, morphemes, and word forms mediating between form and meaning in linguistic models and offshoots thereof in psychology. However, the case of baboon lexical learning (Grainger et al., 2012; Hannagan et al., 2014) illustrates that this research strategy is not without risk, and that simpler discriminative networks can substantially outperform deep learning networks (Linke et al., 2017). We grant, however, that

it is conceivable that model predictions will improve when the present linear networks are replaced by deep learning networks when presented with the same input and output representations used here (see Zhao et al. (2017), for a discussion of loss functions for deep learning targeting numeric instead of categorical output vectors). What is unclear is whether such networks will improve our understanding — the current linear networks offer the analyst connection weights between input and output representations that are straightforwardly linguistically interpretable. Where we think deep learning networks really will come into their own is bridging the gap between for instance the cochlea and retina on the one hand, and the heuristic representations (letter trigrams) that we have used as a starting point for our functional theory of the mappings between form and meaning.

Network flexibility An important challenge for deep learning networks designed to model human lexical processing is to keep the networks flexible and open to rapid updating. This openness to continuous learning is required not only by findings on categorization (Love et al., 2004; Marsolek, 2008; Ramscar and Port, 2015; Marsolek, 2008), including phonetic categorization (Clarke and Luce, 2005; Norris et al., 2003), but is also visible in time series of reaction times. Baayen and Hendrix (2017) reported improved prediction of visual lexical decision reaction times in the British Lexicon Project when a naive discriminative learning network is updated from trial to trial, simulating the within-experiment learning that goes on as subjects proceed through the experiment. In natural speech, the phenomenon of phonetic convergence between interlocutors (Schweitzer and Lewandowski, 2014) likewise bears witness to a lexicon that is constantly recalibrated in interaction with its environment (for the many ways in which interlocutors' speech aligns, see Pickering and Garrod, 2004). Likewise, the rapidity with which listeners adapt to foreign accents (Clarke and Garrett, 2004), within a few minutes, shows that lexical networks exhibit fast local optimization. As deep learning networks typically require huge numbers of cycles through the training data, once trained, they are at risk of not having the required flexibility for fast local adaptation. This risk is reduced for (incremental) wide learning, as illustrated by the phonetic convergence evidenced by the model of Arnold et al. (2017) for auditory comprehension.[12]

Open questions. This initial study necessarily leaves many questions unanswered. Compounds and inflected derived words have not been addressed, and morphological operations such as reduplication, infixation, and non-concatenative morphology provide further testing grounds for the present approach. We also need to address the processing of compounds with multiple constituents, ubiquitous in languages as different as German, Mandarin, and Estonian. Furthermore, in actual utterances, words undergo assimilation at their boundaries, and hence a crucial test case for the discriminative lexicon is to model the production and comprehension of words in multi-word utterances. For our production model, we have also completely ignored stress, syllabification, and pitch, which, however, likely has rendered the production task more difficult than necessary.

A challenge for discriminative morphology is the development of proper semantic matrices. For languages with rich morphology, such as Estonian, current morphological parsers will identify case-inflected words as nominatives, genitives, or partitives (among others), but these labels for inflectional forms do not correspond to the semantic functions encoded in these forms (see, e.g., Kostić, 1995; Blevins, 2016). What is required are computational tools that detect the appropriate inflectional lexomes for these words in the sentences or utterances in which they are embedded.

A related problem specifically for the speech production model is how to predict the strongly reduced word forms that are rampant in conversational speech (Johnson, 2004). Our auditory comprehension network $\boldsymbol{F}_a$ is confronted with reduced forms in training, and the study of Arnold

[12] A very similar point was made by Levelt (1991) in his critique of connectionist models in the nineties.

et al. (2017) indicates that present identification accuracy may approximate human performance for single-word recognition. Whereas for auditory comprehension we are making progress, what is required for speech production is a much more detailed understanding of the circumstances under which speakers actually use reduced word forms. The factors driving reduction may be in part captured by the semantic vectors, but we anticipate that aspects of the unfolding discourse play an important role as well, which brings us back to the unresolved question of how to best model not isolated words but words in context.

An important challenge for our theory, which formalizes incremental implicit learning, is how to address and model the wealth of explicit knowledge that speakers have about their language. We play with the way words sound in poetry, we are perplexed by words' meanings and reinterpret them so that they make more sense (e.g., the folk etymology (Förstemann, 1852) reflected in modern English *crawfish*, a crustacean and not a fish, originally Middle English *crevis*, compare *écrevisse* in French), we teach morphology in schools, and we even find that making morphological relations explicit may be benificial for aphasic patients (Nault, 2010). The rich culture of word use cannot but interact with the implicit system, but how exactly this interaction unfolds and what its consequences are at both sides of the conscious/unconscious divide is presently unclear.

Although many questions remain, our model is remarkably successful in capturing many aspects of lexical and morphological processing in both comprehension and production. Since the model builds on a combination of linguistically motivated 'smart' low-level representations for form and meaning and large but otherwise straightforward and mathematically well-understood two-layer linear networks (to be clear, networks without any hidden layers), the conclusion seems justified that, algorithmically, the "mental lexicon" may be much simpler than previously thought.

Appendix

A Vectors, matrices, and matrix multiplication

Figure 12 presents two data points, a and b, with coordinates (3, 4) and (-2, -3). Arrows drawn from the origin to these points are shown in blue with a solid and dashed line respectively. We refer to these points as vectors. Vectors are denoted with lower case letters in bold, and we place the x coordinate of a point next to the y coordinate:

$$\begin{aligned} \boldsymbol{a} &= \begin{pmatrix} 3 & 4 \end{pmatrix}, \\ \boldsymbol{b} &= \begin{pmatrix} -2 & -3 \end{pmatrix}. \end{aligned}$$

We can represent $\boldsymbol{a}$ and $\boldsymbol{b}$ jointly by placing them next to each other in a matrix. For matrices, we use capital letters in bold font.

$$\boldsymbol{A} = \begin{pmatrix} 3 & 4 \\ -2 & -3 \end{pmatrix}.$$

In addition to the points a and b, Figure 12 also shows two other datapoints, x and y. The matrix for these two points is

$$\boldsymbol{B} = \begin{pmatrix} -5 & 2 \\ 4 & -1 \end{pmatrix}.$$

Let's assume that points a and b represent the forms of two words ω_1 and ω_2, and that the points x and y represent the meanings of these two words. (Below, we discuss in detail how exactly words'

forms and meanings can be represented as vectors of numbers.) As morphology is the study of the relation between words' forms and their meanings, we are interested in how to transform points a and b into points x and y, and likewise, in how to transform points x and y back into points a and b. The first transformation is required for comprehension, and the second transformation for production.

We begin with the transformation for comprehension. Formally, we have matrix $\boldsymbol{A}$, which we want to transform into matrix $\boldsymbol{B}$. A way of mapping $\boldsymbol{A}$ onto $\boldsymbol{B}$ is by means of a linear transformation using matrix multiplication. For 2×2 matrices, matrix multiplication is defined as follows:

$$\begin{pmatrix} a_1 & a_2 \\ b_1 & b_2 \end{pmatrix} \begin{pmatrix} x_1 & x_2 \\ y_1 & y_2 \end{pmatrix} = \begin{pmatrix} a_1 x_1 + a_2 y_1 & a_1 x_2 + a_2 y_2 \\ b_1 x_1 + b_2 y_1 & b_1 x_2 + b_2 y_2 \end{pmatrix}.$$

Such a mapping of $\boldsymbol{A}$ onto $\boldsymbol{B}$ is given by a matrix $\boldsymbol{F}$,

$$\boldsymbol{F} = \begin{pmatrix} 1 & 2 \\ -2 & -1 \end{pmatrix},$$

and it is straightforward to verify that indeed

$$\begin{pmatrix} 3 & 4 \\ -2 & -3 \end{pmatrix} \begin{pmatrix} 1 & 2 \\ -2 & -1 \end{pmatrix} = \begin{pmatrix} -5 & 2 \\ 4 & -1 \end{pmatrix}. \tag{22}$$

Using matrix notation, we can write

$$\boldsymbol{AF} = \boldsymbol{B}.$$

Given $\boldsymbol{A}$ and $\boldsymbol{B}$, how do we obtain $\boldsymbol{F}$? The answer is straightforward, but we need two further concepts: the identity matrix and the matrix inverse. For multiplication of numbers, the identity multiplication is multiplication with 1. For matrices, multiplication with the identity matrix $\boldsymbol{I}$

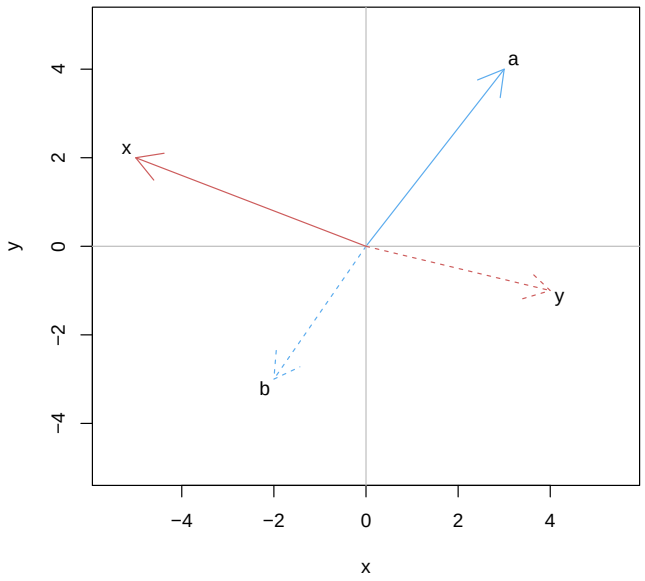

Figure 12: Points a and b can be transformed into points p and q by a linear transformation.

leaves the locations of the points unchanged. The identity matrix has ones on the main diagonal, and zeros elsewhere. It is easily verified that indeed

$$\begin{pmatrix} 1 & 0 \\ 0 & 1 \end{pmatrix} \begin{pmatrix} x_1 & x_2 \\ y_1 & y_2 \end{pmatrix} = \begin{pmatrix} x_1 & x_2 \\ y_1 & y_2 \end{pmatrix} \begin{pmatrix} 1 & 0 \\ 0 & 1 \end{pmatrix} = \begin{pmatrix} x_1 & x_2 \\ y_1 & y_2 \end{pmatrix}.$$

For numbers, the inverse of multiplication with s is dividing by s:

$$(1/s)(sx) = x.$$

For matrices, the inverse of a square matrix $\boldsymbol{X}$ is that matrix $\boldsymbol{X}^{-1}$ such that their product is the identity matrix:

$$\boldsymbol{X}^{-1}\boldsymbol{X} = \boldsymbol{X}\boldsymbol{X}^{-1} = \boldsymbol{I}.$$

For non-square matrices, the inverse $\boldsymbol{Y}^{-1}$ is defined such that

$$\boldsymbol{Y}^{-1}(\boldsymbol{Y}\boldsymbol{X}) = (\boldsymbol{X}\boldsymbol{Y})Y^{-1} = \boldsymbol{X}.$$

We find the square matrix $\boldsymbol{F}$ that maps the square matrices $\boldsymbol{A}$ onto $\boldsymbol{B}$ as follows.

$$\begin{aligned} \boldsymbol{A}\boldsymbol{F} &= \boldsymbol{B} \\ \boldsymbol{A}^{-1}\boldsymbol{A}\boldsymbol{F} &= \boldsymbol{A}^{-1}\boldsymbol{B} \\ \boldsymbol{I}\boldsymbol{F} &= \boldsymbol{A}^{-1}\boldsymbol{B} \\ \boldsymbol{F} &= \boldsymbol{A}^{-1}\boldsymbol{B}. \end{aligned}$$

Since for the present example, $\boldsymbol{A}$ and its inverse happen to be identical, ($\boldsymbol{A}^{-1} = \boldsymbol{A}$), we obtain equation (22).

$\boldsymbol{F}$ maps words' form vectors onto words' semantic vectors. Let's now consider the reverse, and see how we can obtain words' form vectors from words' semantic vectors. That is, given $\boldsymbol{B}$, we want to find that matrix $\boldsymbol{G}$ which maps $\boldsymbol{B}$ onto $\boldsymbol{A}$, i.e.,

$$\boldsymbol{B}\boldsymbol{G} = \boldsymbol{A}.$$

Solving this equation proceeds exactly as above:

$$\begin{aligned} \boldsymbol{B}\boldsymbol{G} &= \boldsymbol{A} \\ \boldsymbol{B}^{-1}\boldsymbol{B}\boldsymbol{G} &= \boldsymbol{B}^{-1}\boldsymbol{A} \\ \boldsymbol{I}\boldsymbol{G} &= \boldsymbol{B}^{-1}\boldsymbol{A} \\ \boldsymbol{G} &= \boldsymbol{B}^{-1}\boldsymbol{A}. \end{aligned}$$

The inverse of $\boldsymbol{B}$ is

$$\boldsymbol{B}^{-1} = \begin{pmatrix} 1/3 & 2/3 \\ 4/3 & 5/3 \end{pmatrix},$$

and hence we have

$$\begin{pmatrix} 1/3 & 2/3 \\ 4/3 & 5/3 \end{pmatrix} \begin{pmatrix} 3 & 4 \\ -2 & -3 \end{pmatrix} = \begin{pmatrix} -1/3 & -2/3 \\ 2/3 & 1/3 \end{pmatrix} = \boldsymbol{G}.$$

The inverse of a matrix needs not exist. A square matrix that does not have an inverse is referred to as a singular matrix. Almost all matrices with which we work in the remainder of this study are singular. For singular matrices, an approximation of the inverse can be used, such as the

Moore-Penrose generalized inverse.[13] In this study, we denote the generalized inverse of a matrix $\boldsymbol{X}$ by $\boldsymbol{X}'$.

The examples presented here are restricted to 2×2 matrices, but the mathematics generalize to matrices of larger dimensions. When multiplying two matrices $\boldsymbol{X}$ and $\boldsymbol{Y}$, the only constraint is that the number of columns of $\boldsymbol{X}$ is the same as the number of rows of $\boldsymbol{Y}$. The resulting matrix has the same number of rows as $\boldsymbol{X}$ and the same number of columns as $\boldsymbol{Y}$.

In the present study, the rows of matrices represent words and columns the features of these words. For instance, the row vectors of $\boldsymbol{A}$ can represent words' form features and the row vectors of and $\boldsymbol{B}$ the corresponding semantic features. The first row vector of $\boldsymbol{A}$ is mapped onto the first row vector of $\boldsymbol{B}$ as follows:

$$\begin{pmatrix} 3 & 4 \end{pmatrix} \begin{pmatrix} 1 & 2 \\ -2 & -1 \end{pmatrix} = \begin{pmatrix} -5 & 2 \end{pmatrix}$$

A transformation matrix such as $\boldsymbol{F}$ can be represented as a two-layer network. The network corresponding to $\boldsymbol{F}$ is shown in Figure 13. When the input vector $(3, 4)$ is presented to the network, the nodes i_1 and i_2 are set to 3 and 4 respectively. To obtain the activations $o_1 = -5$ and $o_2 = 2$ for the output vector $(-5, 2)$, we cumulate the evidence at the input, weighted by the connection strengths specified on the edges of the graph. Thus, for o_1, we obtain the value $3 \times 1 + 4 \times -2 = -5$ and for o_2, we have $3 \times 2 + 4 \times -1 = 2$. In other words, the network maps the input vector $(3, 4)$ onto the output vector $(-5, 2)$.

An important property of linear maps is that they are productive. We can present a novel form vector, say,

$$\boldsymbol{s} = \begin{pmatrix} 2 & 2 \end{pmatrix},$$

to the network, and it will map this vector onto a novel semantic vector. Using matrices, this new semantic vector is straightforwardly calculated:

$$\begin{pmatrix} 2 & 2 \end{pmatrix} \begin{pmatrix} 1 & 2 \\ -2 & -1 \end{pmatrix} = \begin{pmatrix} -2 & 2 \end{pmatrix}.$$

B Graph-based triphone sequencing

The hypothesis underlying graph-based triphone sequencing is that in the directed graph that has triphones as vertices and an edge between any two vertices that overlap (i.e., segments 2 and 3 of triphone 1 are identical to segments 1 and 2 of triphone 2), the path from a word's initial triphone (the triphone starting with `#`) to a word's final triphone (the triphone ending with a `#`) is the path receiving the best support of all possible paths from the initial to the final triphone. Unfortunately, the directed graph containing all vertices and edges is too large to make computations tractable in a reasonable amount of computation time. The following algorithm is a heuristic algorithm that first collects potentially relevant vertices and edges, and then calculates all paths starting from any initial vertex, using the `all_simple_paths` function from the **igraph** package (Csardi and Nepusz, 2006). As a first step, all vertices that are supported by the stem and whole word with network support exceeding 0.1 are selected. To the resulting set, those vertices are added that are supported by the affix, conditional on these vertices having a network support exceeding 0.95. A directed graph can now be constructed, and for each initial vertex, all paths starting at these vertices are calculated. The subset of those paths reaching a final vertex is selected, and the support for each

[13] In R, the generalized inverse is available in the **MASS** package, function `ginv`. The numeric library that is used by `ginv` is LAPACK, available at `http://www.netlib.org/lapack`.

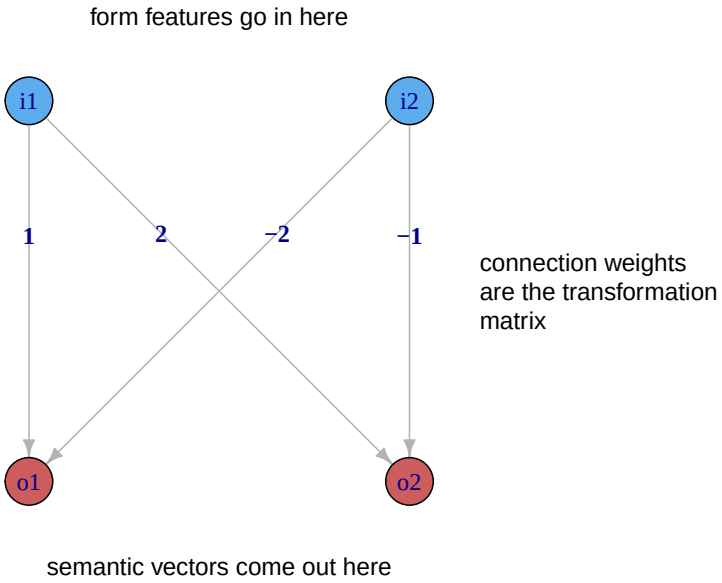

Figure 13: The linear network corresponding to the transformation matrix $\boldsymbol{F}$. There are two input nodes, i_1 and i_2 (in blue), and two output nodes, o_1 and o_2, in red. When the input vector $(3, 4)$ is presented to the network, the value at o_1 is $1 \times 3 + -2 \times 4 = -5$, and that at o_2 is $3 \times 2 + -1 \times 4 = 2$. The network maps the point $(3, 4)$ onto the point $(-5, 2)$.

path is calculated. As longer paths trivially will have greater support, the support for a path is weighted for its length, simply by dividing the raw support by the path length. Paths are ranked by weighted path support, and the path with maximal support is selected as the acoustic image driving articulation.

It is possible that no path from an initial vertex to a final vertex is found, due to critical boundary triphones not being instantiated in the data on which the model was trained. This happens when the model is assessed on novel inflected and derived words under cross-validation. Therefore, vertices and edges are added to the graph for all non-final vertices that are at the end of any of the paths starting at any initial triphone. Such novel vertices and edges are assigned zero network support. Typically, paths from the initial vertex to the final vertex can now be constructed, and path support is evaluated as above.

For an algorithm that allows triphone paths to include cycles, which may be the case in languages with much richer morphology than English, see Baayen et al. (2018) and for code the R package **WpmWithLdl** described therein.

C A heatmap for function words (pronouns and prepositions)

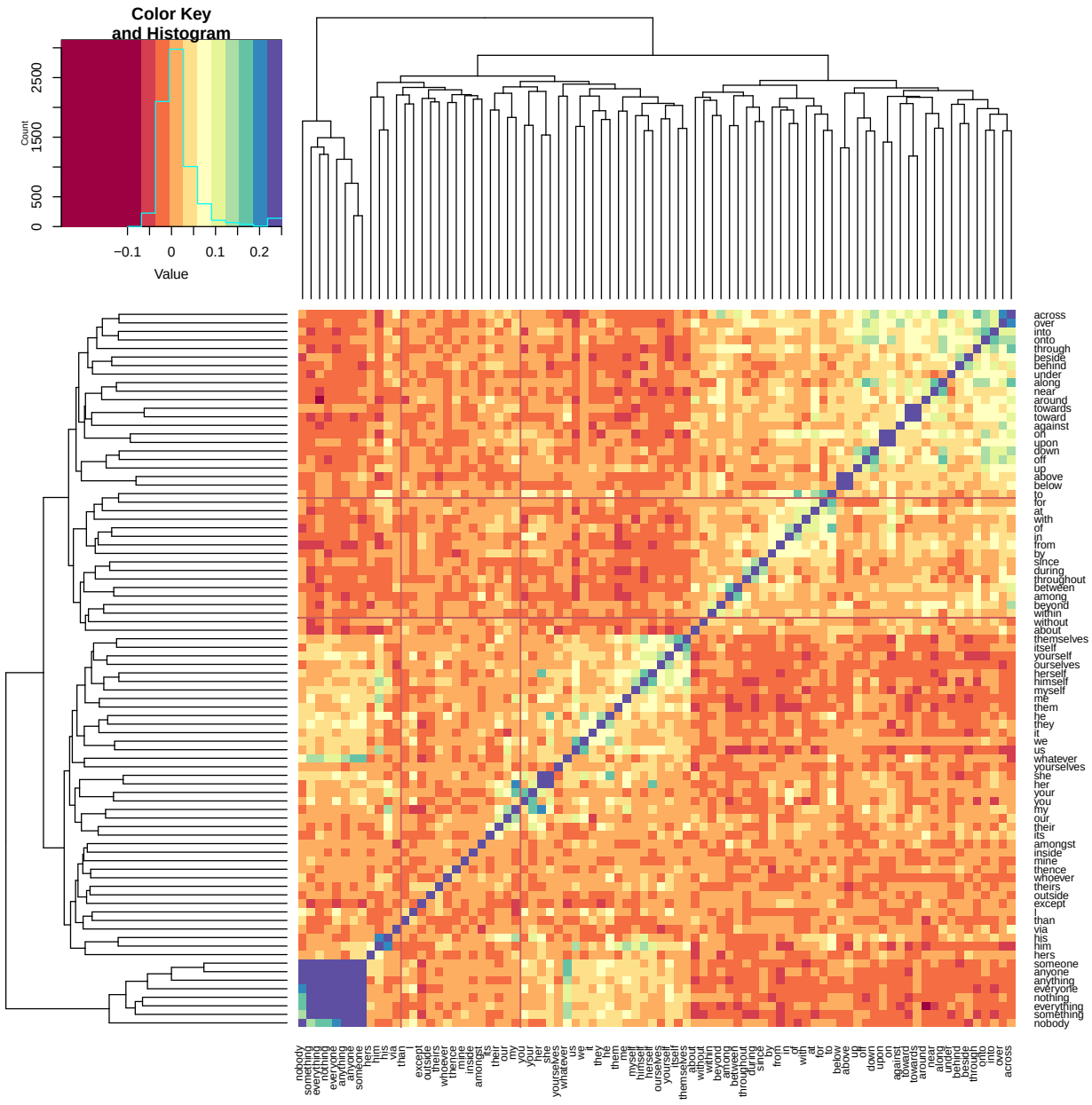

Figure 14: Heatmap for the correlations of the semantic vectors of pronouns and prepositions. Note that quantifier pronouns form a group of their own, that prepositions and the other pronouns form distinguisable groups, and that within groups, further subclusters are visible, e.g., for *we* and *us*, *she* and *her*, and *across* and *over*. All diagonal elements represent correlations equal to 1. (This is not brought out by the heatmap, which color-codes diagonal elements with the color for the highest correlation of the range we specified, 0.25).

Acknowledgements

The authors are indebted to Geoff Hollis, Jessie Nixon, Chris Westbury, and Luis Mienhardt for their constructive feedback on earlier versions of this manuscript, as well as to the three anonymous reviewers who helped us improve the paper. This research was supported by an ERC advanced grant (742545) to the first author.

References

Amenta, S., Marelli, M., and Sulpizio, S. (2017). From sound to meaning: Phonology-to-semantics mapping in visual word recognition. *Psychonomic bulletin & review*, 24(3):887–893.

Arnold, D. (2017). Acousticndlcoder: Coding sound files for use with ndl. R package version 1.0.1.

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. and Blevins, J. P. (2018). Inflectional morphology with linear mappings. *The Mental Lexicon*, page submitted.

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

Baayen, R. H. and Hendrix, P. (2017). Two-layer networks, non-linear separation, and human learning. In Wieling, M., Kroon, M., Van Noord, G., and Bouma, G., editors, *From Semantics to Dialectometry. Festschrift in honor of John Nerbonne. Tributes 32.*, pages 13–22. College Publications.

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. (2016a). Frequency in lexical processing. *Aphasiology*, 30(11):1174–1220.

Baayen, R. H., Piepenbrock, R., and Gulikers, L. (1995). *The CELEX lexical database (CD-ROM)*. Linguistic Data Consortium, University of Pennsylvania, Philadelphia, PA.

Baayen, R. H., Schreuder, R., and Sproat, R. (2000). Morphology in the mental lexicon: a computational model for visual word recognition. In van Eynde, F. and Gibbon, D., editors, *Lexicon Development for Speech and Language Processing*, pages 267–291. Kluwer Academic Publishers.

Baayen, R. H., Shaoul, C., Willits, J., and Ramscar, M. (2016b). Comprehension without segmentation: A proof of concept with naive discriminative learning. *Language, Cognition, and Neuroscience*, 31(1):106–128.

Bauer, L. (1983). *English Word Formation*. CUP, Cambridge.

Bauer, P. J. and Larkina, M. (2014). The onset of childhood amnesia in childhood: a prospective investigation of the course and determinants of forgetting of early-life events. *Memory*, 22(8):907–924.

Beard, R. (1977). On the extent and nature of irregularity in the lexicon. *Lingua*, 42:305–341.

Beard, R. (1995). *Lexeme-morpheme base morphology: A general theory of inflection and word formation*. State University of New York Press, Albany, NY.

Bergen, B. K. (2004). The psychological reality of phonaesthemes. *Language*, 80:290–311.

Bertram, R., Tønnessen, F. E., Strömqvist, S., Hyönä, J., and Niemi, P. (2015). Cascaded processing in written compound word production. *Frontiers in human neuroscience*, 9:207.

Bitan, T., Kaftory, A., Meiri-Leib, A., Eviatar, Z., and Peleg, O. (2017). Phonological ambiguity modulates resolution of semantic ambiguity during reading: An fMRI study of Hebrew. *Neuropsychology*, 31(7):759.

Blevins, J. P. (2003). Stems and paradigms. *Language*, 79:737–767.

Blevins, J. P. (2006). Word-based morphology. *Journal of Linguistics*, 42(03):531–573.

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

Booij, G. (2010). Construction morphology. *Language and linguistics compass*, 4(7):543–555.

Bozic, M., Marslen-Wilson, W. D., Stamatakis, E. A., Davis, M. H., and Tyler, L. K. (2007). Differentiating morphology, form, and meaning: Neural correlates of morphological complexity. *Journal of cognitive neuroscience*, 19(9):1464–1475.

Breiman, L. (2001). Random forests. *Machine Learning*, 45:5–32.

Browman, C. and Goldstein, L. (1992). Articulatory Phonology: An Overview. *Phonetica*, 49:155–180.

Bruni, E., Tran, N., and Baroni, M. (2014). Multimodal distributional semantics. *Journal of Artificial Intelligence Research*, 49:1–47.

Burnage (1988). *CELEX; A guide for users*. Centre for Lexical Information, Nijmegen.

Butz, M. V. and Kutter, E. F. (2016). *How the mind comes into being: Introducing cognitive science from a functional and computational perspective*. Oxford University Press.

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.

Cho, T. (2001). Effects of morpheme boundaries on intergestural timing: Evidence from Korean. *Phonetica*, 58(3):129–162.

Chomsky, N. (1957). *Syntactic Structures*. Mouton, The Hague.

Chomsky, N. and Halle, M. (1968). *The sound pattern of English*. Harper and Row, New York.

Clarke, C. and Luce, P. (2005). Perceptual adaptation to speaker characteristics: Vot boundaries in stop voicing categorization. In *ISCA workshop on plasticity in speech perception*.

Clarke, C. M. and Garrett, M. F. (2004). Rapid adaptation to foreign-accented english. *The Journal of the Acoustical Society of America*, 116(6):3647–3658.

Cohen, L. and Dehaene, S. (2009). Ventral and dorsal contributions to word reading. In Gazzaniga, M. S., editor, *The cognitive neurosciences*, pages 789–804. Massachusetts Institute of Technology.

Coltheart, M. (2005). Modeling reading: The dual-route approach. *The science of reading: A handbook*, pages 6–23.

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.

Coltheart, M., Rastle, K., Perry, C., Langdon, R., and Ziegler, J. (2001). The DRC model: A model of visual word recognition and reading aloud. *Psychological Review*, 108:204–258.

Cotterell, R., Schütze, H., and Eisner, J. (2016). Morphological smoothing and extrapolation of word embeddings. In *Proceedings of the 54th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, volume 1, pages 1651–1660.

Csardi, G. and Nepusz, T. (2006). The igraph software package for complex network research. *InterJournal*, Complex Systems:1695.

Cucchiarini, C. and Strik, H. (2003). Automatic Phonetic Transcription: An overview. In *Proceedings of the 15th ICPhS*, pages 347–350, Barcelona, Spain.

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.

Dell, G. (1986). A spreading-activation theory of retrieval in sentence production. *Psychological review*, 93(3):283–321.

desRosiers, G. and Ivison, D. (1988). Paired associate learning: Form 1 and form 2 of the wechsler memory scale. *Archives of Clinical Neuropsychology*, 3(1):47–67.

Diehl, R. L., Lotto, A. J., and Holt, L. L. (2004). Speech perception. *Annu. Rev. Psychol.*, 55:149–179.

Erelt, M., editor (2003). *Estonian language*. Estonian academy publishers, Tallinn.

Ernestus, M. (2000). *Voice assimilation and segment reduction in casual Dutch. A corpus-based study of the phonology-phonetics interface*. LOT, Utrecht.

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.

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.

Firth, J. R. (1968). *Selected papers of J R Firth, 1952-59*. Indiana University Press.

Fletcher, H. (1940). Auditory patterns. *Rev. Mod. Phys.*, 12:47–65.

Förstemann, E. (1852). Über Deutsche volksetymologie. *Zeitschrift für vergleichende Sprachforschung auf dem Gebiete des Deutschen, Griechischen und Lateinischen*, 1(1. H):1–25.

Forsyth, R. and Holmes, D. (1996). Feature-finding for test classification. *Literary and Linguistic Computing*, 11(4):163–174.

Frege, G. (1879). Begriffsschrift, a formula language, modeled upon that of arithmetic, for pure thought. *From Frege to Gödel: A source book in mathematical logic*, 1931:1–82.

Freud, S. (1905/1953). Childhood and concealing memories. In Brill, A. A., editor, *The basic writings of Sigmund Freud*. The Modern Library, New York.

Geeraert, K., Newman, J., and Baayen, R. H. (2017). Idiom variation: Experimental data and a blueprint of a computational model. *Topics in Cognitive Science*, 9(3):653–669.

Gonnerman, L. M., Seidenberg, M. S., and Andersen, E. S. (2007). Graded semantic and phonological similarity effects in priming: Evidence for a distributed connectionist approach to morphology. *Journal of experimental psychology: General*, 136(2):323.

Grainger, J., Dufau, S., Montant, M., Ziegler, J. C., and Fagot, J. (2012). Orthographic processing in baboons (papio papio). *Science*, 336(6078):245–248.

Hannagan, T., Ziegler, J. C., Dufau, S., Fagot, J., and Grainger, J. (2014). Deep learning of orthographic representations in baboons. *PLOS-one*, 9:e84843.

Hannun, A., Case, C., Casper, J., Catanzaro, B., Diamos, G., Elsen, E., Prenger, R., Satheesh, S., Sengupta, S., Coates, A., et al. (2014). Deep speech: Scaling up end-to-end speech recognition. *arXiv preprint arXiv:1412.5567*.

Harm, M. W. and Seidenberg, M. S. (1999). Phonology, reading acquisition, and dyslexia: Insights from connectionist models. *Psychological Review*, 106:491–528.

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.

Hawkins, S. (2003). Roles and representations of systematic fine phonetic detail in speech understanding. *Journal of Phonetics*, 31:373–405.

Hay, J. and Drager, K. (2010). Stuffed toys and speech perception. *Linguistics*, 48(4):865–892.

Hay, J., Pierrehumbert, J., and Beckman, M. (2004). Speech perception, well-formedness, and the statistics of the lexicon. *Papers in laboratory phonology VI*, pages 58–74.

Hay, J. B. (2002). From speech perception to morphology: Affix-ordering revisited. *Language*, 78:527–555.

Hay, J. B. (2003). *Causes and Consequences of Word Structure*. Routledge, New York and London.

Hay, J. B. and Baayen, R. H. (2003). Phonotactics, parsing and productivity. *Italian Journal of Linguistics*, 1:99–130.

Henri, V. and Henri, C. (1895). On our earliest recollections of childhood. *Psychological Review*, 2:215–216.

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

Hockett, C. (1960). The origin of speech. *Scientific American*, 203:89–97.

Hornstein, N. (1995). *Logical form: From GB to minimalism*. Blackwell.

Hubel, D. H. and Wiesel, T. N. (1962). Receptive fields, binocular interaction and functional architecture in the cat's visual cortex. *The Journal of physiology*, 160(1):106–154.

Ivens, S. H. and Koslin, B. L. (1991). *Demands for Reading Literacy Require New Accountability Methods*. Touchstone Applied Science Associates.

Jackendoff, R. (1990). *Semantic Structures*. MIT Press, Cambridge.

Jared, D., Ashby, J., Agauas, S. J., and Levy, B. A. (2016). Phonological activation of word meanings in grade 5 readers. *Journal of Experimental Psychology: Learning, Memory, and Cognition*, 42(4):524.

Jared, D. and Bainbridge, S. (2017). Reading homophone puns: Evidence from eye tracking. *Canadian Journal of Experimental Psychology/Revue canadienne de psychologie expérimentale*, 71(1):2.

Jared, D. and O'Donnell, K. (2017). Skilled adult readers activate the meanings of high-frequency words using phonology: Evidence from eye tracking. *Memory & cognition*, 45(2):334–346.

Johnson, K. (1997). The auditory/perceptual basis for speech segmentation. *Ohio State University Working Papers in Linguistics*, 50:101–113.

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.

Kalman, R. E. (1960). A new approach to linear filtering and prediction problems. *Journal of basic Engineering*, 82(1):35–45.

Kamin, L. J. (1969). Predictability, surprise, attention, and conditioning. In Campbell, B. A. and Church, R. M., editors, *Punishment and Aversive Behavior*, pages 276–296. Appleton-Century-Crofts, New York.

Kastovsky, D. (1986). Productivity in word formation. *Linguistics*, 24:585–600.

Kaye, R. and Wilson, R. (1998). *Linear Algebra*. Oxford University Press.

Kemps, R., Ernestus, M., Schreuder, R., and Baayen, R. H. (2004). Processing reduced word forms: The suffix restoration effect. *Brain and Language*, 19:117–127.

Keuleers, E., Lacey, P., Rastle, K., and Brysbaert, M. (2012). The British Lexicon Project: Lexical decision data for 28,730 monosyllabic and disyllabic English words. *Behavior Research Methods*, 44:287–304.

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.

Kleinschmidt, D. F. and Jaeger, T. F. (2015). Robust speech perception: Recognize the familiar, generalize to the similar, and adapt to the novel. *Psychological Review*, 122(2):148.

Kostić, A. (1995). Informational load constraints on processing inflected morphology. In Feldman, L. B., editor, *Morphological Aspects of Language Processing*. Lawrence Erlbaum Inc. Publishers, New Jersey.

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.

Landauer, T. K., Foltz, P. W., and Laham, D. (1998). Introduction to latent semantic analysis. *Discourse Processes*, 25:259–284.

Lazaridou, A., Marelli, M., Zamparelli, R., and Baroni, M. (2013). Compositionally derived representations of morphologically complex words in distributional semantics. In *ACL (1)*, pages 1517–1526.

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

Levelt, W. J. (1991). Die konnektionistische mode. *Sprache und Kognition*, 10(2):61–72.

Levelt, W. J. M. (2008). *Formal grammars in linguistics and psycholinguistics: Volume 1: An introduction to the theory of formal languages and automata, Volume 2: Applications in linguistic theory; Volume 3: Psycholinguistic applications*. John Benjamins Publishing.

Levinson, S. C. (2006). The language of space in Yélî Dnye. In *Grammars of space: Explorations in cognitive diversity*, pages 157–203. Cambridge University Press.

Levinson, S. C. and Majid, A. (2013). The island of time: Yélî Dnye, the language of Rossel island. *Frontiers in psychology*, 4.

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

Love, B. C., Medin, D. L., and Gureckis, T. M. (2004). Sustain: a network model of category learning. *Psychological Review*, 111(2):309.

Lund, K. and Burgess, C. (1996). Producing high-dimensional semantic spaces from lexical co-occurrence. *Behaviour Research Methods, Instruments, and Computers*, 28(2):203–208.

Marantz, A. (2013). No escape from morphemes in morphological processing. *Language and Cognitive Processes*, 28(7):905–916.

Marchand, H. (1969). *The Categories and Types of Present-Day English Word Formation. A Synchronic-Diachronic Approach*. Beck'sche Verlagsbuchhandlung, München.

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

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

Marzi, C., Ferro, M., and Pirrelli, V. (2018). Is inflectional irregularity dysfunctional to human processing? In Kuperman, V., editor, *Abstract booklet, The Mental Lexicon 2018*, page 60. University of Alberta, Edmonton.

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

Matthews, P. H. (1991). *Morphology. An Introduction to the Theory of Word Structure*. Cambridge University Press, Cambridge.

Matthews, P. H. (1993). *Grammatical theory in the United States from Bloomfield to Chomsky*. Cambridge University Press, Cambridge.

McBride-Chang, C. (1996). Models of speech perception and phonological processing in reading. *Child development*, 67(4):1836–1856.

McCarthy, J. J. (1981). A prosodic theory of non-concatenative morphology. *Linguistic Inquiry*, 12:373–418.

McQueen, J. M. (1998). Segmentation of continuous speech using phonotactics. *Journal of Memory and Language*, 39(1):21–46.

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

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

Milin, P., Feldman, L. B., Ramscar, M., Hendrix, P., and Baayen, R. H. (2017b). Discrimination in lexical decision. *PLOS-one*, 12(2):e0171935.

Milin, P., Filipović Durdević, D., and Moscoso del Prado Martín, F. (2009). The simultaneous effects of inflectional paradigms and classes on lexical recognition: Evidence from Serbian. *Journal of Memory and Language*, 60(1):50–64.

Miller, R. R., Barnet, R. C., and Grahame, N. J. (1995). Assessment of the Rescorla-Wagner model. *Psychological Bulletin*, 117(3):363–386.

Mitchell, J. and Lapata, M. (2008). Vector-based models of semantic composition. In *ACL*, pages 236–244.

Montague, R. (1973). The proper treatment of quantification in ordinary english. In *Approaches to natural language*, pages 221–242. Springer.

Moscoso del Prado Martín, F., Bertram, R., Häikiö, T., Schreuder, R., and Baayen, R. H. (2004). Morphological family size in a morphologically rich language: The case of Finnish compared to Dutch and Hebrew. *Journal of Experimental Psychology: Learning, Memory and Cognition*, 30:1271–1278.

Nault, K. (2010). *Morphological Therapy Protocal*. Ph. D. thesis, University of Alberta, Edmonton.

Newman, R. L., Jared, D., and Haigh, C. A. (2012). Does phonology play a role when skilled readers read high-frequency words? evidence from erps. *Language and Cognitive Processes*, 27(9):1361–1384.

Norris, D. and McQueen, J. (2008). Shortlist B: A Bayesian model of continuous speech recognition. *Psychological Review*, 115(2):357–395.

Norris, D., McQueen, J. M., and Cutler, A. (2003). Perceptual learning in speech. *Cognitive psychology*, 47(2):204–238.

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

Pham, H. and Baayen, R. H. (2015). Vietnamese compounds show an anti-frequency effect in visual lexical decision. *Language, Cognition, and Neuroscience*, 30(9):1077–1095.

Phillips, C. (2001). Levels of representation in the electrophysiology of speech perception. *Cognitive Science*, 25(5):711–731.

Pickering, M. J. and Garrod, S. (2004). Toward a mechanistic psychology of dialogue. *Behavioral and brain sciences*, 27(2):169–190.

Pickett, J. and Pollack, I. (1963). Intelligibility of excerpts from fluent speech: Effects of rate of utterance and duration of excerpt. *Language and Speech*, 6:151–164.

Pirrelli, V., Ferro, M., and Marzi, C. (2015). Computational complexity of abstractive morphology. In Bearman, M., Brown, D., and Corbett, G. G., editors, *Understanding and measuring morphological complexity*, pages 141–166. Oxford University Press Oxford.

Pitt, M., Johnson, K., Hume, E., Kiesling, S., and Raymond, W. (2005). The Buckeye corpus of conversational speech: labeling conventions and a test of transcriber reliability. *Speech Communication*, 45(1):89–95.

Plag, I. (2003). *Word-formation in English*. Cambridge University Press, Cambridge.

Plag, I., Homann, J., and Kunter, G. (2017). Homophony and morphology: The acoustics of word-final S in English. *Journal of Linguistics*, 53(1):181–216.

Port, R. F. and Leary, A. P. (2005). Against formal phonology. *Language*, 81:927–964.

R Core Team (2016). *R: A Language and Environment for Statistical Computing*. R Foundation for Statistical Computing, Vienna, Austria.

Ramscar, M., Hendrix, P., Shaoul, C., Milin, P., and Baayen, R. H. (2014). Nonlinear dynamics of lifelong learning: the myth of cognitive decline. *Topics in Cognitive Science*, 6:5–42.

Ramscar, M. and Port, R. (2015). Categorization (without categories). In Dabrowska, E. and Divjak, D., editors, *Handbook of Cognitive Linguistics*, pages 75–99. De Gruyter, Berlin.

Ramscar, M., Sun, C. C., Hendrix, P., and Baayen, R. H. (2017). The mismeasurement of mind: Life-span changes in paired-associate-learning scores reflect the "cost" of learning, not cognitive decline. *Psychological Science*. https://doi.org/10.1177/0956797617706393.

Ramscar, M. and Yarlett, D. (2007). Linguistic self-correction in the absence of feedback: A new approach to the logical problem of language acquisition. *Cognitive Science*, 31(6):927–960.

Ramscar, M., Yarlett, D., Dye, M., Denny, K., and Thorpe, K. (2010). The effects of feature-label-order and their implications for symbolic learning. *Cognitive Science*, 34(6):909–957.

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

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

Roelofs, A. (1997). The WEAVER model of word-form encoding in speech production. *Cognition*, 64:249–284.

Russell, B. (1905). On denoting. *Mind*, 14(56):479–493.

Russell, B. (1942). *An inquiry into meaning and truth*. Allen and Unwin, London.

Scarf, D., Boy, K., Reinert, A. U., Devine, J., Güntürkün, O., and Colombo, M. (2016). Orthographic processing in pigeons (columba livia). *Proceedings of the National Academy of Sciences*, 113(40):11272–11276.

Schmid, H. (1995). Improvements in part-of-speech tagging with an application to german. In *Proceedings of the ACL SIGDAT-Workshop*, Dublin, Ireland.

Schmidhuber, J. (2015). Deep learning in neural networks: An overview. *Neural Networks*, 61:85–117.

Schreuder, R. and Baayen, R. H. (1995). Modeling morphological processing. In Feldman, L. B., editor, *Morphological Aspects of Language Processing*, pages 131–154. Lawrence Erlbaum, Hillsdale, New Jersey.

Schweitzer, A. and Lewandowski, N. (2014). Social factors in convergence of f1 and f2 in spontaneous speech. In *Proceedings of the th International Seminar on Speech Production, Cologne. pp.*

Seidenberg, M. (1987). Sublexical structures in visual word recognition: Access units or orthographic redundancy. In Coltheart, M., editor, *Attention and Performance XII*, pages 245–264. Lawrence Erlbaum Associates, Hove.

Seidenberg, M. S. and Gonnerman, L. M. (2000). Explaining derivational morphology as the convergence of codes. *Trends in Cognitive Sciences*, 4(9):353–361.

Sering, T., Milin, P., and Baayen, R. H. (2018). Language comprehension as a multiple label classification problem. *Statistica Neerlandica*, pages 1–15.

Shafaei-Bajestan, E. and Baayen, R. H. (2018). Wide learning for auditory comprehension. In *Proceedings of Interspeech 2018*.

Shaoul, C., Bitschau, S., Schilling, N., Arppe, A., Hendrix, P., Milin, P., and Baayen, R. H. (2015). ndl2: Naive discriminative learning: an implementation in R. R package.

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

Shockey, L. (1998). Perception of reduced forms by non-native speakers of English. In Duez, D., editor, *Sound Patterns of Spontaneous Speech*, pages 97–100. ESCA, Aix.

Strobl, C., Boulesteix, A.-L., Zeileis, A., and Hothorn, T. (2007). Bias in random forest variable importance measures: Illustrations, sources and a solution. *BMC Bioinformatics*, 8.

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

Taft, M. (1988). A morphological decomposition model of lexical representation. *Linguistics*, 26:657–667.

Taft, M. (1994). Interactive-activation as a framework for understanding morphological processing. *LCP*, 9(3):271–294.

Tomaschek, F., Plag, I., Ernestus, M., and Baayen, R. H. (2018). Modeling the duration of word-final s in English with naive discriminative learning. *Manuscript, University of Siegen/Tübingen/Nijmegen*.

Tourville, J. A. and Guenther, F. H. (2011). The diva model: A neural theory of speech acquisition and production. *Language and cognitive processes*, 26(7):952–981.

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.

Tucker, B. V., Sims, M., and Baayen, R. H. (2018). Opposing forces on acoustic duration. *Manuscript, University of Alberta and University of Tübingen*.

Tüske, Z., Golik, P., Schlüter, R., and Ney, H. (2014). Acoustic modeling with deep neural networks using raw time signal for lvcsr. In *INTERSPEECH*, pages 890–894.

Ussishkin, A. (2005). A Fixed Prosodic Theory of Nonconcatenative Templaticmorphology. *Natural Language & Linguistic Theory*, 23(1):169–218.

Ussishkin, A. (2006). Affix-favored Contrast Inequity and Psycholinguistic Grounding for Non-concatenative Morphology. *Morphology*, 16(1):107–125.

Van Orden, G. C., Kloos, H., et al. (2005). The question of phonology and reading. *The science of reading: A handbook*, pages 61–78.

Venables, W. N. and Ripley, B. D. (2002). *Modern Applied Statistics with S-Plus*. Springer, New York.

Wagner, R. K., Torgesen, J. K., and Rashotte, C. A. (1994). Development of reading-related phonological processing abilities: New evidence of bidirectional causality from a latent variable longitudinal study. *Developmental psychology*, 30(1):73.

Warriner, A. B., Kuperman, V., and Brysbaert, M. (2013). Norms of valence, arousal, and dominance for 13,915 English lemmas. *Behavior Research Methods*, 45(4):1191–1207.

Weaver, W. (1955). Translation. In Locke, W. N. and Booth, A. D., editors, *Machine Translation of Languages: Fourteen Essays*, pages 15–23. MIT Press, Cambridge.

Weingarten, R., Nottbusch, G., and Will, U. (2004). Morphemes, syllables and graphemes in written word production. In Pechmann, T. and Habel, C., editors, *Multidisciplinary approaches to speech production*, pages 529–572. Mouton de Gruyter, Berlin.

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

Wong, K. F. E. and Chen, H.-C. (1999). Orthographic and phonological processing in reading chinese text: Evidence from eye fixations. *Language and Cognitive Processes*, 14(5-6):461–480.

Wood, S. N. (2017). *Generalized Additive Models*. Chapman & Hall/CRC, New York.

Yao, B., Belin, P., and Scheepers, C. (2011). Silent reading of direct versus indirect speech activates voice-selective areas in the auditory cortex. *Journal of Cognitive Neuroscience*, 23(10):3146–3152.

Young, R. and Morgan, W. (1980). *The Navajo language: A grammar and colloquial dictionary*. University of New Mexico Press.

Zeller, B., Padó, S., and Šnajder, J. (2014). Towards semantic validation of a derivational lexicon. In *Proceedings of COLING 2014, the 25th International Conference on Computational Linguistics: Technical Papers*, pages 1728–1739.

Zhao, H., Gallo, O., Frosio, I., and Kautz, J. (2017). Loss functions for image restoration with neural networks. *IEEE Transactions on Computational Imaging*, 3(1):47–57.

Zwitserlood, P. (2018). Processing and representation of morphological complexity in native language comprehension and production. In Booij, G. E., editor, *The construction of words*, pages 583–602. Springer.