Vito Pirrelli, Claudia Marzi, Marcello Ferro,
Franco Alberto Cardillo, Harald R. Baayen and Petar Milin

Psycho-computational modelling of the mental lexicon

A discriminative learning perspective

Abstract: Over the last decades, a growing body of evidence on the mechanisms governing lexical storage, access, acquisition and processing has questioned traditional models of language architecture and word usage based on the hypothesis of a direct correspondence between modular components of grammar competence (lexicon vs. rules), processing correlates (memory vs. computation) and neuro-anatomical localizations (prefrontal vs. temporo-parietal perisylvian areas of the left hemisphere). In the present chapter, we explore the empirical and theoretical consequences of a distributed, integrative model of the mental lexicon, whereby words are seen as emergent properties of the functional interaction between basic, language-independent processing principles and the language-specific nature and organization of the input. From this perspective, language learning appears to be inextricably related to the way language is processed and internalized by the speakers, and key to an interdisciplinary understanding of such a way, in line with Tomaso Poggio's suggestion that the development of a cognitive skill is causally and ontogenetically prior to its execution (and sits "on top of it"). In particular, we discuss conditions, potential and prospects of the epistemological continuity between psycholinguistic and computational modelling of word learning, and illustrate the yet largely untapped potential of their integration. We use David Marr's hierarchy to clarify the complementarity of the two viewpoints. Psycholinguistic models are informative about how speakers learn to use language (interfacing Marr's levels 1 and 2). When we move from the psycholinguistic analysis of the functional operations involved in language learning to an algorithmic description of how they are computed, computer simulations can help us explore the relation between speakers' behavior and general learning principles in more detail. In the end, psycho-computational

Vito Pirrelli, Claudia Marzi, Marcello Ferro, Franco Alberto Cardillo, Italian National Research Council (CNR), Institute for Computational Linguistics, Pisa, Italy
Harald R. Baayen, Eberhard Karls University, Seminar für Sprachwissenschaft/Quantitative Linguistics Sprachwissenschaft/Quantitative Linguistics, Tübingen, Germany
Petar Milin, University of Birmingham, Department of Modern Languages, Edgbaston, Birmingham, UK

https://doi.org/10.1515/9783110440577-002

models can be instrumental to bridge Marr's levels 2 and 3, bringing us closer to understanding the nature of word knowledge in the brain.

Keywords: mental lexicon, word storage and processing, psycholinguistics, computational linguistics, connectionist models, discriminative learning

1 Introduction

1.1 Motivation and historical background

Over the past 30 years, theoretical and applied linguistics, cognitive psychology and neuroscience have gradually shifted their research focus on language knowledge from discipline-specific issues to a broader range of shared interests, questions and goals. This has been particularly true in the domain of lexical knowledge since the mid-eighties, when the Parallel Distributed Processing (PDP) group simulated non-linear developmental trajectories in child acquisition of the English past tense, moving away from traditional box-and-arrow models to data-driven computer simulations of emergent phenomena (Rumelhart and McClelland 1986). The trend was concomitant with other important developments in this area. The dichotomy between data and programming, reflected in the contrast between static lexical items and dynamic rules of grammar (as in Pinker's "Words and Rules" approach, Pinker and Prince 1988, 1994) has progressively given way to more integrative views of the lexicon as a dynamic store of words in context, where basic levels of language representation (sound, syntax and meaning) are interfaced and co-organized into context-sensitive "chunks" (Jackendoff 2002; Goldberg 2006; Booij 2010). Accordingly, human brains must "contain" not only morphologically simple words, but also inflected and derived forms, compound words, light verb constructions, collocations, idioms, proverbs, social routine clichés and all sorts of ready-made, routinized sequences, maximizing processing opportunities (Libben 2005), augmenting the human working memory capacity (Baddeley 1986), and having distinct frequency/familiarity effects on processing (see Baayen et al. 2007; Kuperman et al. 2009; Tremblay and Baayen 2010, among others).

Probably, the best known assumption in morphological inquiry is the hypothesis that word processing is a form of algebraic calculus, based on the combination/composition of sublexical building blocks called "morphemes" (e.g. *will-*, *-ing*, *-ness*, *un-*), traditionally conceived of as minimal linguistic signs, or irreducible form-meaning pairs, according to an influential terminology whose roots can be

traced back to Bloomfield's work (1933).[1] Besides, the content of a morphologically complex word is assumed to be a function of the meaningful contribution of each of its internal morphemes. This assumption is part of a very influential view on language processing as the result of a staged sequence of processing steps and intermediate, hierarchically arranged representations: from sounds to syllables, morphemes, words and beyond. At each step, intermediate representations are output and fed into upper representation levels. In particular, morphemes are credited with playing an active role in word recognition and production.

These assumptions are usually bundled together. Effects of morpheme boundaries on word processing are often coupled with the hypothesis that morphemes are stored and accessed as independent, atomic linguistic signs, making the lexicon a redundancy-free store of simple, irreducible items. In addition, morphemes are assumed to be involved in processing prior to word identification/production. In fact, as we will see in the following sections, the involvement of morpheme-like structures in word processing is not necessarily staged prior to word access, and it does not imply, per se, further assumptions such as form-meaning pairing and strong compositionality. Besides, the linguistic status of the morpheme is confronted with a number of theoretical difficulties (Matthews 1991), suggesting that other relations than just the simple position of a sublexical constituent within an input word may influence human word knowledge. In particular, many studies in the framework of Word and Paradigm Morphology have challenged the idea that morphemes are the atomic units of morphological analysis, suggesting that full words represent basic building blocks in their own right (Anderson 1992; Aronoff 1994; Beard 1995; Booij 2010; Blevins 2016; Marzi et al. 2020, this volume). This has led to a radical reconceptualization of the role of morphemes in word processing that received indirect support by work in computational morphology (Pirrelli 2018). As we will see in more detail in the ensuing sections, computer modelling of morphological processes can shed light on dynamic aspects of language organization that would otherwise elude scientific inquiry. For example, the idea that linguistic structure can emerge through self-organization of unstructured input is nowadays key to our understanding of a number of issues in language acquisition (Bybee and Hopper 2001; Ellis and Larsen-Freeman 2006; MacWhinney 1999; MacWhinney and O'Grady 2015). Nonetheless, it had to await the challenging test of successful computer simulations before it could be given wide currency in the acquisitional literature. As will be argued more extensively in the following

1 Note, however, that only post-Bloomfieldian accounts translated Bloomfield's idea that complex lexical forms can be analyzed into simple constituents (morphemes) into the hypothesis that lexical forms can be reconstructed starting from their independently stored, simple parts (Blevins 2016; Blevins et al. 2016).

section, by giving center stage to processing issues, computational morphology and psycholinguistic approaches to word knowledge have in fact much more in common than ever acknowledged in the past.

1.2 Computational Linguistics & Psycholinguistics: conditions for a methodological unification

Computational Linguistics (CL) and Psycholinguistics (PL) share a broad range of interests and goals. CL is chiefly concerned with computer-based simulations of how language is understood, produced and learned. Simulations are running models of language performance, implemented as sets of instructions performing specific tasks on a computer. They commonly require a precise algorithmic characterization of aspects of language processing that are often neglected by language theories, such as the encoding of input data, the structure of output representations, the basic operations of word segmentation, storage, access, retrieval and assembly of intermediate representations (e.g. Clark et al. 2010).

In a similar vein, PL focuses on the cognitive mechanisms and representations that are known to underlie language processing in the mind or brain of a speaker. Traditionally, PL uses experiments with human subjects to obtain measures of language behavior as response variables. In a typical lexical decision experiment, a speaker is asked to decide, as quickly and accurately as possible, whether a written form shown on a computer screen for a short time (or, alternatively, its acoustically rendered pronunciation) is a word in her language or not. The researcher controls and manipulates the factors that are hypothesized to be involved in the processing task, to measure the extent to which factor manipulation affects processing performance in terms of response time and accuracy. Of late, PL more and more often incorporates evidence from neural experimentation, measuring brain activity more directly as it unfolds during the task (e.g. Spivey et al. 2012; Marangolo and Papagno 2020, this volume).

In spite of their shared concerns, however, CL and PL have traditionally developed remarkably different approaches, principles and goals. The impact of information and communication technologies on language inquiry has spawned a myriad of successful commercial applications (from speech recognition and speech synthesis, to machine translation, information retrieval and knowledge extraction), laying more emphasis on optimizing the computational properties of parsing algorithms, such as their time and space complexity, and efficiency in task completion. This technological trend has, however, parceled out language usage into a fragmentary constellation of small sub-problems and ad hoc software solutions, proposed independently of one another.

Conversely, psycholinguistic models approach language as resulting from the interaction of both language specific functions (e.g. word co-activation and competition) and general-purpose cognitive functions (e.g. long-term storage, sensory-motor integration, rehearsal, executive control). Different global effects in the operation of low-level interactive processes are investigated as the by-products of specific levels of input representations (e.g. phonological, morpho-syntactic or semantic levels), giving rise to autonomous, self-organizing effects. Psycholinguistic models are also aimed to investigate under what conditions language processing can be found to perform sub-optimally, with inherent limitations, occasional errors and possible breakdowns of the human language processor being just as important to understand as processing efficiency and performance optimization (Berg 2020, this volume; Vulchanova, Saldaña and Baggio 2020, this volume).

The apparent divergence in the way CL and PL are concerned with issues of language performance, however, has not precluded growing awareness of their potential for synergy. We already mentioned the important role that seminal work by Rumelhart, McClelland and the Parallel Distributed Processing (PDP) group played in the mid-eighties in re-orienting the research focus on language processing away from algorithmic issues. We will consider the legacy of connectionism and its persisting influence on current models of lexical competence in the ensuing sections in more detail. Here, we would like to focus very briefly on the implications of the connectionist revolution for the methodological interaction between CL and PL.

Following the PDP success story, the question of how rules carry out computations in language, and what types of rules are needed for linguistic computations, stopped to be the exclusive concern of CL. In fact, emphasis on language learning slowly shifted the research spotlight on the more fundamental issue of how a speaker develops the computations and representations used by the brain from the experience of the natural world. This shift has two important methodological consequences. First, even if we assume (following traditional wisdom) that sentences are made of phrases, phrases of words and words of morphemes, and that language processing is an algebraic calculus combining smaller units into larger ones, the central question that must be addressed is how basic combinatorial units are acquired in the first place. Words, phrases and utterances are not given, but they should be investigated as dynamic processes, emerging from interrelated patterns of sensory experience, communicative and social interaction and psychological and neurobiological mechanisms (Elman 2009). Secondly, if both combinatorial rules and units are acquired, what are the principles underlying (i) rule learning and (ii) the intake/development of input representations during learning? In the scientific pursuit for ultimate explanatory mechanisms, learning principles informing our capacity to adaptively use regularities from

experience are better candidates than regularities themselves. In the end, we may ignore what rules consist of and what representations they manipulate, or even wonder whether rules and representations exist at all (questions that have animated much of the contemporary debate on language and cognition). Investigation of the basic neurocognitive functions (e.g., serial perception, storage, alignment, to mention but a few) that allow for the language input to be processed and acquired strikes us as an inescapable precondition to understanding what we know when we know a language. In this connection, learning represents a fundamental level of meta-cognition where PL and CL can successfully meet.

1.2.1 Marr's hierarchy

Tomaso Poggio, one the pioneers of computer vision, has recently suggested (2010, 2012) that *learning* should be added to Marr's classical hierarchy of levels of understanding of complex processing systems (Marr 1982). The original Marr's hierarchy defined three such levels:

(1) the *computational level*, answering the "semantic" question "what does it do?", by providing a precise characterization of what types of functions and operations are to be computed for a specific cognitive process to be carried out successfully;
(2) the *algorithmic level*, answering the "syntactic" question "how does it do it?", by specifying how computation takes place in terms of detailed equations and programming instructions;
(3) the *implementation level*, stating how representations and algorithms are actually realized at the physical level (e.g. as electronic circuits or patterns of neurobiological connectivity).[2]

Poggio argues that learning sits on top of Marr's computational level, as it allows us to replicate the ability of performing a particular task (e.g. object

2 Computer terminology plays, nowadays, a much more pervasive role than it did in the 70s and early 80s. Adjectives like "computational" and "implementational", which are common terminological currency in today's information sciences, were used by Marr in a different, more literal sense. In a contemporary adaptation of Marr's terminology, the "computational level" can arguably be translated into "functional" or "architectural level". Similarly, his "implementational level" could more readily be understood as referring to a "(bio-)physical level". This would avoid, among other things, the potential confusion arising when we ascribe "computer modelling" (and CL) to Marr's "algorithmic level" (rather than to his "computational level"). We decided to stick to Marr's original terminology nonetheless, and tried to avoid terminological clashes by using terms unambiguously in context.

identification) in machines "even without an understanding of which algorithms and specific constraints are exploited". This gives a special status to the study of machine learning and explains much of its influence in various areas of computer science and in today's computational neuroscience (Poggio 2010: 367). From our perspective, machine learning and statistical models of language have made an essential contribution in breaking a relatively new, interdisciplinary middle ground, for CL and PL to meet and profitably interact. But what is the ultimate goal of this interaction? Is it methodologically well founded?

Marr introduced his hierarchy to emphasize that explanations at different levels can be investigated largely independently of each other. A language engineer can automatically process large quantities of text data, disregarding how difficult they are for a human speaker to process. A neuroscientist can describe the biophysics of oscillations in the neural activity of cortical areas, and ignore how these oscillations can possibly map onto higher-level processing functions. However, full understanding of a complex system requires tight inter-level interaction. In the spirit of computational neuroscience, one must eventually understand what kind of computations are performed by oscillations, and what algorithm controls them.

We agree with Poggio (2012) that it is time to clarify the potential for between-level interaction in Marr's hierarchy, and investigate the methodological conditions for their appropriate integration. It has been observed (Alvargonzáles 2011) that interdisciplinary convergence requires operational, material continuity between the objects of investigation of neighboring scientific fields. Trivially, using the same battery of formal/mathematical methods and functions to model as diverse empirical domains as mechanics, economy or epidemiology, does not make the boundaries between these domains any closer. Only if we can clarify the role of formal psycholinguistic models of language processing and computer simulations along Marr's hierarchy, we can establish a material common ground between PL and CL, and, ultimately, assess the potential for their unification.

1.2.2 Complementarity and integration

In a classical psycholinguistic experiment, scholars aim to understand more of the architecture and functioning principles of the human language processor by investigating human language behavior in highly controlled conditions. From this standpoint, the human processor represents a "black box" (the research *explanandum*), whose internal organization and principles are inferred through observation of overt behavioral variables (the *explanans*). The approach of psycholinguistic

inquiry can thus be described in terms of *abductive inference*, whereby underlying causes are studied and understood by analyzing their overt effects.[3]

Conversely, experiments conducted by implementing and running computer simulations of a specific language task can be used to understand more of the human processing behavior by testing the mechanisms that are assumed to be the *cause* of this behavior. Suppose that we want to model how speakers learn to process words as a dynamic process of optimal resolution of multiple, parallel (and possibly conflicting) constraints on complex lexical structures (Seidenberg and MacDonald 1999). In this case, a parallel processing architecture represents our *explanans*, designed and implemented to combine top-down expectations (based on past input evidence) with the on-line bottom-up requirements of current input stimuli. If successful, the simulator should be able to replicate aspects of human language behavior.

Such a methodological complementarity between CL and PL enables us to establish an effective continuity between observations and hypotheses. Abductively inferred functions in the human processor can be simulated through a piece of programming code replicating human results on a comparable set of test data. But replicating results is of little explanatory power unless we understand why and how simulations are successful (Marzi and Pirrelli 2015). The real insights often come from examining the way problems are solved algorithmically, how they are affected by changes in data distribution or parameter setting, and by observing the interaction between these changes and principles that were not specified by the original psycholinguistic model, but had to be implemented for the computational model to carry out a specific task. We can then check these new insights back on human subjects, and make abductive reasoning and computer modelling interact for our level of knowledge to scale up along Marr's hierarchy. Ultimately, simulations should be able to incorporate requirements coming from Marr's implementational level, and make processing mechanisms match what is known about the neurophysiological principles supporting language processing. From this perspective, computational modelling cannot only provide a framework for psycholinguistic theories to be tested, but can also bridge the gap between high-level psycholinguistic and cognitive functions, and low-level interactive brain processes.

3 Abductive inference, also known as "inference to the best explanation", must be distinguished from both deductive and inductive inference. Deductive reasoning allows deriving *b* from *a* only when *b* is a formal logical consequence of *a*. Inductive reasoning allows inferring *b* from *a*, by way of a logically unnecessary generalization: if one has experience of white swans only, one can (wrongly) believe that all swans are white. Abductive reasoning allows inferring *a* as a possible explanation of *b*. If you glance an apple falling from a tree, you can abduce (rather uneconomically) that someone hidden in the tree leaves is dropping apples to the ground.

To sum up, by describing and interpreting the behavior of a speaker performing a certain task, psycholinguistic models help us bridge the gap between Marr's computational (i.e. "what the speaker does") and algorithmic level (i.e. "how she does it"). On the other hand, by simulating how the same problems are solved by a computer, machine learning models can help us test psycholinguistic models algorithmically. If algorithmic results prove to match human results, and if the implemented mechanisms can be mapped onto high-level aspects of human behavior to make independent predictions about it, progress is made. Finally, if algorithmic models are implemented to incorporate neurobiologically grounded processing principles, we make progress in filling the gap between Marr's algorithmic and implementation levels.

In this section, we discussed the methodological conditions for a fruitful interaction between PL and CL approaches to language processing, in line with Marr's original idea that a full scientific theory of a complex processing system requires understanding its computational, algorithmic and biophysical levels and making predictions at all such levels. In the following section, we will selectively overview a few psycholinguistic and algorithmic models of the mental lexicon, with a view to exploring concrete prospects for methodological unification in the context of language learning. As a final methodological remark, it is important to be clear on where we agree and where we disagree with Poggio's claims. We think that Poggio is right in emphasizing that, from an ontogenetic perspective, learning how to execute a cognitive task is temporally and causally prior to task execution. Besides, understanding how the task is learned is inextricably related to the way the task is executed, and is key to understanding such a way. However, this hierarchy of (meta-)cognitive levels is concerned with their ontogenetic and possibly phylogenetic relationships (e.g. in connection with evolutionary changes of biological processing systems), and has little to do with Marr's hierarchy. In our view (unlike Poggio's), learning does not sit on top of Marr's levels, but can better be analyzed and understood *through* each of them.

2 Psycho-computational models of the mental lexicon: A selective overview

2.1 Morpheme-based and a-morphous models

For decades, issues of lexical processing, access and organization have been investigated by focusing on aspects of the internal structure of complex words (Bloomfield 1933; Bloch 1947; Chomsky and Halle 1968; Lieber 1980; Selkirk 1984).

According to the classical generative view, words are made up out of morphemes. A repository of sublexical constituents accounts for the ways morphologically complex words are mutually related in the speaker's' mind. For example, the theory of speech production developed by Levelt et al. (1999) assumes that only irreducible forms are stored in the lexicon as separate entries, thus providing a psycholinguistic model of this view.

The generative approach goes back to an "Item and Arrangement" view of morphological competence (Hockett 1954), and was influenced by the dominant computer metaphor of the 50s, equating the human language processor to a processing device coupled with highly efficient retrieval procedures (Baayen 2007). Since morphemes were understood as sign-based units, which capture the minimal patterns of recurrence of form and meaning in our vocabulary, they were conceived of as potential access units of the mental lexicon. These assumptions boil down to what Blevins (2006) termed a *constructive approach* to morphological theory, where roots/stems (and possibly affixes) are the basic building blocks of morphological competence, in a largely redundancy-free lexicon. This is contrasted with an *abstractive approach*, according to which full word forms are the building blocks of morphological competence, and recurrent sublexical parts define *abstractions* over full forms.

Since early work in the lexicalist framework (Halle 1973; Jackendoff 1975; Aronoff 1976; Scalise 1984), it was clear that morphological rules might not be heavily involved in *on-line* word processing (see Fábregas and Penke 2020, this volume). Besides, despite its attractiveness and simplicity, the constructive idea that morphemes play a fundamental role as representational units in the mental lexicon has met a number of theoretical, computational and psycholinguistic difficulties (Blevins 2016). In the psycholinguistic literature, this awareness led to a sweeping reappraisal of the role of morphemes in language usage, and prompted a flourishing number of diverse theoretical perspectives on the mental lexicon.

Psycholinguistic models in the '70s (Becker 1980; Rubenstein et al. 1970, 1971; Snodgrass and Jarvella 1972) investigated the idea that lexical units compete for recognition. Token frequency of single input forms, type frequency of related forms (size of morpho-lexical families) and their relative probabilistic distribution, were shown to affect the way lexical units are matched against an input stimulus, with high-frequency units being checked earlier for matching than low-frequency units are. In line with this evidence, it was suggested that morpheme-based representations do not provide an alternative to full word listing in lexical organization, but are rather complementary access units to whole words. We can mention at least four different views of the role of sublexical units in the morphological lexicon:

(i) as permanent access units to full words, speeding up lexical access/retrieval (Taft and Forster 1975; Taft 1994, 2004);
(ii) as fallback processing routes, in case of failure to access fully-inflected lexical entries (Caramazza et al. 1988);
(iii) as pre-lexical processing routes, running in parallel with full-word access routes, and competing with the latter in a race for lexical access (Schreuder and Baayen 1995);
(iv) as post-lexical meaningful formal cores reflecting inter-word relationships in so-called morphological families (Giraudo and Grainger 2000; Grainger et al. 1991).

As a radical departure from a morpheme-centered view of the mental lexicon, other lexical models were put forward that appeared to dispense altogether with the idea that lexical access is mediated by sublexical constituents. Morton's (1969, 1970, 1979) original logogen model and its updates were apparently influenced by feature detection models of visual object recognition, based on the parallel activation of competing "demons" (neurons), dedicated to perform processing of specific input features, and "yelling" for primacy (Selfridge 1959). Morton's demons, named "logogens", were conceived of as specialized word receptors, accumulating sensory properties of linguistic stimuli and outputting their own response (e.g. a single word form) when accumulated properties (e.g. semantic, visual or acoustic features) rose above a threshold value.

The Parallel Distributed Processing (or PDP) way to connectionism in the eighties (Rumelhart et al. 1986) followed in Morton's footsteps, to popularize the idea that the lexical processor consists of a network of parallel processing nodes (functionally equivalent to neuron clusters) selectively firing in response to sensory stimuli (McClelland and Elman 1986; Norris 1994; Rumelhart and McClelland 1986). Accordingly, word production was modelled as a mapping function between two levels of representation, consisting of the input and output layers of processing nodes in a multi-layered neural network: namely, the level of morpho-lexical content (consisting of lexical meanings and morpho-syntactic features), and the level of surface form (strings of letters or sounds). For example, given an appropriate encoding of the base form *go* and the feature PAST on the input layer, this representation is mapped onto the string *went* on the output layer.

The PDP model explicitly implemented an assumption that was common to most psycholinguistic models of the lexicon; namely the idea that, when a word is input, multiple access units are activated in parallel. Levels of co-activation depend on the degree of fit between the incoming input and each lexical unit represented in the lexicon, and is modulated by the prior probability of input

representations, estimated with their relative frequency of occurrence. Word recognition and production are guided by competition among similar representations (or lexical neighbors), whose influence on the process is a function of their number (or neighborhood density), their independent token frequencies, and their uniqueness recognition points in input/output words (Marslen-Wilson 1984).[4]

Each of these principles is quite general, and allows for considerable cross-model variation (Dahan and Magnuson 2006). For example, frequency can directly affect the activation of processing units by modulating either the units' threshold for response (as in Morton's logogen model), or the units' resting activation level (as in Marslen-Wilson's cohort model), or the strength of connections between sublexical and lexical representations (MacKay 1982). Alternatively, frequency can act as a post-activation bias, thus influencing lexical selection, as in the NAM model (Luce 1986; Luce and Pisoni 1998). Besides, theories may differ in their similarity metrics and/or bottom-up activation mechanisms (which determine degree of fit), information flow (e.g. only bottom-up or top-down as well), and the nature of the competition mechanisms they assume (e.g. decision rule, lateral inhibition, or interference).

Differences and similarities notwithstanding, the PDP connectionism brought to the fore a factor missing in all previous models: the temporal dynamic of *learning*. In fact, non-connectionist models simply assumed the existence of a representational level made up out of access units, and an independent access procedure, mapping the input signal onto lexical representations. However, very little was said about how representations develop in the first place: how do children come to the decision of storing an irregular form as an unsegmented access unit, and a regular form as consisting of distinct access units? Even for those approaches where the decision does not have to be yes-or-no (since both hypotheses can be entertained at the same time, as in race models of lexical access), questions about how this is implemented (e.g., how does a child come up with the appropriate segmentation of a word form into sub-lexical units?) are left open.

In classical multi-layered perceptrons, internalized representations develop as the result of learning. The mapping of an input full form onto its morphological constituents is a continuous function of the statistical regularities in the

4 A uniqueness point (or UP) refers to the word-internal point (e.g. a sound, or a letter) at which an input form is uniquely identified among all its morphologically unrelated competitors (e.g. *k* in *walk* compared with *wall*). More recently, Balling and Baayen (2008) define a second uniqueness point, or Complex Uniqueness Point (CUP), where morphologically related competitors become incompatible with the input word (e.g. *i* in *walking* compared with *walk, walks*, *walked* etc.).

form and meaning of different words. Since form-meaning mapping is predicted to be a graded phenomenon, perception of morphological boundaries by a connectionist network may vary as a result of the probabilistic support the boundaries receive from frequency distributions of acquired exemplars (e.g., Hay and Baayen 2005; Plaut and Gonnerman 2000; Rueckl and Raveh 1999). This mechanism is key to what is arguably the most important legacy of connectionism for models of the mental lexicon: both regular and irregular words are processed by the same underlying mechanism and supported by the same memory resources. Pace Pinker and Ullman (2002), perception of morphological structure is not the by-product of the design of the human word processor, purportedly segregating exceptions from rules. Rather, it is an emergent property of the dynamic self-organization of lexical representations, contingent on the processing history of past input word forms.

However, as correctly observed by Baayen (2007), classical connectionist simulations model word acquisition as the mapping of a base input form onto its inflected output form (e.g. *go* → *went*). This protocol is in fact compatible with the view of a redundancy-free lexicon, and seems to adhere to a *derivational* approach to morphological competence, reminiscent of classical generative theories. Nonetheless, since network-internal representations (encoded in hidden layers of processing nodes) are dependent on the temporal dynamics of input-output mapping steps, connectionist principles are conducive to the idea that sublexical constituents dynamically *emerge* from the lexical store. Emergence of morphological structure is the result of morphologically complex words being redundantly memorized and mutually related as full forms.

2.2 Morphological emergence and paradigm-based models

The general idea that word structure emerges from lexical self-organization allows for considerable variation in matters of detail. According to Bybee (1995), stored words presenting overlapping parts with shared meaning are mutually related through lexical connections. Connection strength correlates positively with the number of related words (their family size) and negatively with their token frequency (see Bybee and McClelland 2005 for a more connectionist rendering of these ideas). Burzio (1998) interprets lexical connections as global lexical entailments, which may redundantly specify multiple surface bases. In line with this view, Word and Paradigm Morphology (Matthews 1991; Blevins 2006, 2016) conceives of mastering the morphology of a language as the acquisition of an increasing number of paradigmatic constraints on how paradigm cells are filled in (or *cell-filling problem*: Ackerman et al. 2009; Cardillo et al. 2018; Finkel and

Stump 2007; Pirrelli and Battista 2000; Pirrelli and Yvon 1999). What all these approaches have in common is the assumption that full word forms are the building blocks of morphological competence, and recurrent sublexical parts define *abstractions* over full forms (Blevins 2006).

The extent to which abstracted sublexical parts play a role in word processing remains a highly debated point in the psycholinguistic literature (see Schmidtke et al. 2017, for a recent, concise overview). Nonetheless, there seems to be a general consensus on the idea that the organization of items into morphologically natural classes (be they inflectional paradigms, inflectional classes, derivational families or compound families) has a direct influence on morphological processing, and that surface word relations constitute a fundamental domain of morphological competence. Of late, the emphasis on lexical families prompted a growing interest in information-theoretic measures of their degree of complexity. Once more, the connection between self-organization of word forms into morphological families and Shannon's information theory (Shannon 1948) is mainly provided by the relation between lexical knowledge and learning. Due to the Zipfian distribution of word forms in the speaker's input, inflectional paradigms happen to be attested only partially also for high-frequency lexemes (Blevins et al. 2017). Speakers must then be able to generalize available knowledge, and infer the inflectional class to which a partially attested paradigm belongs, for non-attested cells to be filled in accordingly.

Inferring non-attested forms of a paradigm on the basis of a few attested forms only thus requires that some word forms be diagnostic for inflectional class. Some forms can be more diagnostic than others, but it is often the case that no single form exists in a paradigm from which all other forms of the same paradigm can be inferred. This is not only true of irregular verb paradigms, but also of regular ones, where some inflected forms may neutralize class-membership diacritics (e.g. theme vowels for verb inflectional classes, see Albright 2002). Different forms can be instrumental for filling in specific subsets of paradigm cells (irrespective of their degree of morphological or phonological predictability), and more forms can be interchangeably used to predict the same subclass. On the one hand, this strategy calls for more evidence to be stored (so-called exemplary diagnostic forms, also referred to as "principal parts" in classical grammars). On the other hand, a speaker does not have to wait for one specific form (a "base" form) to be input, or abstract away an appropriate representation from available evidence. More forms can be used interchangeably for class assignment.

2.3 The "disappearing" lexicon

Paradigm-based approaches prompt a significant shift of emphasis away from traditional computational work on morphology, chiefly based on finite state technology and concerned with cognitively neutral, rule-like representations and analyses (Corbett and Fraser 1993; Karttunen 2003; Pirrelli 2018). A way to understand the difference between classical morpheme-based approaches and paradigm-based approaches to morphology is by looking at analogical proportions between paradigmatically-related word forms like the following:

(*drink*, PRES) : (*drank*, PAST) :: (*sink*, PRES) : (*sank*, PAST)

Given some computational constraints, one can infer any of the forms in the proportion above on the basis of the remaining three forms (Pirrelli and Yvon 1999). To illustrate, from the relation between *drink* and *drank*, one can infer that, by changing *i* into *a*, PRES is turned into PAST. Given (*drink*, PRES), (*drank*, PAST) and (*sink*, PRES), we can thus infer (*sank*, PAST). Note that, for a proportion to apply consistently, proportional relations must obtain concurrently and independently *within* each representation level (in our example, lexical form and grammatical content). Nothing explicit is stated about inter-level relations, i.e. about what substring in *drink* is associated with PRES. We could have stated the formal relationship between *drink* and *drank* as a (redundant) change of *ink* into *ank*, and the same inference would obtain. In fact, by the *principle of contrast* (Clark 1987, 1990), any formal difference can be used to mark a grammatical opposition as long as it obtains within one minimal pair of paradigmatically-related forms. This principle solves many of the paradoxes in the traditional notion of morpheme as a minimal linguistic sign: e.g. morphemes with no meanings (or empty morphemes), meanings with no morphemes (or zero morphemes), bracketing paradoxes etc.

It is noteworthy that the time-honored principle of contrast in linguistics is fully in line with principles of *discriminative learning*, whose roots can be traced back to philosophical pragmatism (particularly James 1907, 1909; and later Wittgenstein 1953; Quine 1960), functional psychology (James 1890) and behaviorism (Tolman 1932, 1951; Osgood 1946, 1949, 1966; Skinner 1953, 1957). Discriminative principles received their formal and mathematical modeling in the work of Rescorla and Wagner on classical conditioning, also known as error-driven learning (Rescorla 1988; Rescorla and Wagner 1972). More recently, work of Ramscar and collaborators (Ramscar and Yarlett 2007; Ramscar et al. 2010) and Ellis (2006a, also see Ellis and Larsen-Freeman 2006) laid the foundations of error-driven learning in the context of language learning. Baayen et al. (2011)

and Milin, Feldman et al. (2017) provide the discriminative approach with its computational platform, dubbed Naive Discrimination Learning (NDL).

Unlike strongly compositional and associative approaches to learning, error-driven or discriminative learning assumes that learning "results from exposure to relations among events in the environment", and, as such, it is "a primary means by which the organism represents the structure of its world" (Rescorla 1988: 152). Learning proceeds not by *associating* co-occurring cues and outcomes, but by *discriminating* between multiple cues that are constantly in competition for their predictive value for a given outcome. Furthermore, cues are not fixed in advance, but they emerge dynamically within an environment, shaped up by adaptive pressures. According to this view, human lexical information is never stable, time- or context-independent. Its content is continuously updated and reshaped as a function of when, why and how often it is accessed and processed, with activation spreading to neighboring patterns of connectivity. Such flowing activation states are more reminiscent of the wave/particle duality in quantum physics (Libben 2016) or the inherently adaptive, self-organizing behavior of biological dynamic systems (Beckner et al. 2009; Larsen-Freeman and Cameron 2008) than ever thought in the past. From this perspective, the very notion of the mental lexicon is challenged; it may represent, at best, a metaphorical device or a convenient terminological shortcut (Elman 2009).

We saw that, from a theoretical linguistic perspective, the discriminative view fits in very well with Word and Paradigm Morphology (Blevins 2016), according to which morphemes and words are set-theoretic constructs. In a more computational perspective, it appears to support the view that storage and processing are not functionally and physically independent components of an information processing architecture (as with familiar desktop computers). Rather, they are better conceived of as two interdependent long-term and short-term dynamics of the same underlying process: learning.

Ultimately, we believe that understanding more of the far-reaching implications of (human) learning and adaptive behavior pushes us into a profound reassessment of traditional linguistic notions and processing requirements. This calls for more advanced computer models of human language behavior. In this section, we reviewed converging evidence of the role of morphological families and paradigmatic relations in the developmental course of lexical acquisition. The evidence bears witness to a fundamental interdependency between mechanisms of lexical activation/competition and effects of lexical token frequency, paradigm frequency, and paradigm regularity in word processing and learning. However, there have been comparatively few attempts to simulate this interdependency algorithmically. Most existing computational models of word recognition and production (Chen and Mirman 2012; Gaskell and Marslen-Wilson 2002;

Levelt et al. 1999; McClelland and Elman 1986; Norris, McQueen and Cutler 1995; among others) either focus on processing issues, by analyzing how input patterns can be mapped onto stored exemplars during processing; or focus on storage, by entertaining different hypotheses concerning the nature of stored representations (e.g. Henson 1998; Davis 2010, among others). Much less work is devoted to more "integrative" (neuro)computational accounts, where (i) "memory units that are repeatedly activated in processing an input word are the same units responsible for its stored representation" (Marzi and Pirrelli 2015: 495), and (ii) "memory units are made develop dynamically as the result of learning" (Marzi et al. 2016: 80). Truly integrative models would lead to an effective implementation and a better understanding of the dynamic interaction between processing and storage, and make room for a careful analysis of the empirical consequences of such a mutual implication on realistically distributed lexical data.

In the ensuing sections, we investigate what can be learned about the impact of principles of discriminative learning on lexical acquisition, access and production, by running computer simulations of models of dynamic lexical storage. We start with a general introduction of the Naive Discriminative Learning framework, its mathematical underpinnings and general philosophy, moving from the basics to advanced applications. Then, we investigate the time-bound dynamics of co-activation and competition in the acquisition of families of inflected forms, with a view to providing a unitary account of paradigm-based lexical acquisition and effects of neighbor families on lexical processing. This will be done using a family of recurrent neural networks known as Temporal Self-Organizing Maps. We will show that self-organizing memories provide a biologically inspired explanatory framework accounting for the interconnection between Word and Paradigm Morphology and principles of Discriminative Learning.

3 Computer models of discriminative learning

3.1 Naive Discriminative Learning

Naive Discriminative Learning (NDL) represents a computational modelling approach to language processing, providing theoretical and methodological grounding of research on diverse language phenomena. The NDL computational model itself implements the simplest possible error-driven learning rule, originally proposed by Rescorla and Wagner (1972), which since then has been shown to make powerful predictions for a range of phenomena in language learning and language comprehension (Ellis 2006a, 2006b; Ramscar, Dye and McCauley 2013;

Ramscar and Yarlett 2007; Ramscar et al. 2010). The first study to apply discrimination learning to predict reaction times by training a network on large corpora is Baayen, Milin et al. (2011), and the term NDL was coined and first used in this study.

3.1.1 NDL – The Basics

The Rescorla-Wagner learning rule updates the weights on connections from input features (henceforth cues) to output classes (henceforth outcomes) in a simple two-layer network. Outcomes are word-like units that are labelled "lexomes" in the NDL terminology (e.g. the unit *something*), cues are typically letter bigrams, trigrams or even word forms (like *#so*, *som*, *ome*, *met*, *eth*, *thi*, *hin*, *ing*, *ng#* for the word *something*; with the '#' symbol replacing start-of-word and end-of-word spaces). The relationship between cues and outcomes is incremental, and develops in discrete time steps. Presence of a cue C_i in a given learning event E_t taking place at time t is indicated by $PRESENT(C_i, t)$, and presence of an outcome O_j in E_t by $PRESENT(O_J, t)$. The weight w_{ij}^t is defined on the connection between a given cue C_i and specific outcome O_j at time t, and at the subsequent timestep w_{ij}^{t+1} this weight is defined as:

$$w_{ij}^{t+1} = w_{ij}^t + \Delta w_{ij}^t \tag{1}$$

where the change in weight Δw_{ij}^t is specified as:

$$\Delta w_{ij}^t = \begin{cases} = 0 \quad ; \quad \text{if } PRESENT(C_{i,t}) \text{ is } \textbf{false} \\ = \eta_i \left(\lambda_j - \sum_{present(C_k, t)} w_{kj} \right); \text{if } PRESENT(C_{i,t}) \text{ is } \textbf{true} \,\&\, PRESENT(O_{j,t}) \text{ is } \textbf{true} \\ = \eta_i \left(0 - \sum_{present(C_k, t)} w_{kj} \right); \text{if } PRESENT(C_{i,t}) \text{ is } \textbf{true} \,\&\, PRESENT(O_{j,t}) \text{ is } \textbf{false} \end{cases} \tag{2}$$

Weights on connections from cues that are absent in the input are left unchanged. For cues that are present in the input, the weights to a given outcome are updated, depending on whether the outcome was correctly predicted. The prediction strength or activation a for an outcome is defined as the sum of the weights on the connections from the cues in the input to the outcome. If the outcome is present in a learning event, together with the cues, then the weights are increased by a proportion η of the difference between the maximum prediction

strength (λ, set at 1 in *NDL* simulations) and a. The proportionality constant η defines the learning rate of the model. Thus, the adjustment to the weights when the outcome is indeed present is $\eta(\lambda - a)$. When the outcome is not present, the weights are decreased by $\eta(0 - a)$. For networks trained on large corpora, setting η to 0.001 appears optimal. In general, learning rate η should be set to a small value (commonly between 0.1 and 0.001) to allow for learning to be incremental (Rescorla and Wagner 1972; Blough 1975; Baayen et al. 2011; Ghirlanda 2005; Ghirlanda et al. 2017). The learning rate η is the only free parameter of the NDL implementation of the Rescorla-Wagner learning rule.[5]

3.1.2 Current results

Naive Discriminative Learning has been used successfully to model the results of a range of experiments. Baayen, Milin et al. (2011), Pham and Baayen (2015), and Milin, Feldman et al. (2017) investigated primed and unprimed lexical decision. Arnold et al. (2017) developed a model of spoken word recognition using input cues derived from the speech signal. Linke et al. (2017) modeled (supposed) lexical learning in baboons. Geeraert et al. (2017) used NDL to clarify idiom variation; and Ramscar et al. (2014, 2017) used NDL to study the consequences of the accumulation of knowledge over a lifetime.

The 2011 study applying Naive Discriminative Learning to lexical decision latencies used cues consisting of individual letters and letter pairs. It has since been shown that letter triplets provide better cues for modelling reading. In the same paper, outcomes were conceptualized as "semantic units". In subsequent

5 Implementations of NDL are available for R (package ndl, Arppe et al. 2015) and as a Python library (pyndl: Weitz et al. 2017). The first study to explore the potential of discrimination learning for understanding reaction times (Baayen, Milin et al. 2011) did not make use of the Rescorla-Wagner equations themselves, but instead used the equations developed by Danks (2003). Danks developed equations for estimating the weights under the assumption that the system has reached a state of equilibrium in which no further learning takes place. Although the option of using Danks' equilibrium equations is implemented in the available software packages, subsequent research strongly suggests it is preferable to use the original equations and apply them step by step to the sequence of learning events. NDL networks appear quite sensitive to the order in which sets of cues and outcomes are presented for learning. Hence, if order information is available (as when models are trained on corpora), it is advisable to let this order co-determine learning. The available software implements optimized algorithms that can utilize multiple cores in parallel to speed up the incremental updating of the weights. For large data sets, estimating the weights is actually accomplished substantially more quickly for 'incremental' learning as compared to the estimation method based on the Danks equations.

work, the nature of these units was clarified: they are now conceptualized as pointers to locations in a multidimensional semantic space. To avoid confusing these pointers with contentful lexical representations, we labelled these pointers 'lexomes' (c.f., Baayen, Shaoul et al. 2016; Milin, Divjak, and Baayen 2017; Milin, Feldman et al. 2017). Lexomes thus link lexical contrasts in form to lexical contrasts in semantic space. Figure 1 clarifies the role of this theoretical construct in the model. This figure simultaneously represents three discrimination networks, each of which is trained independently. The three networks have all been used for successfully predicting data from experimental studies.

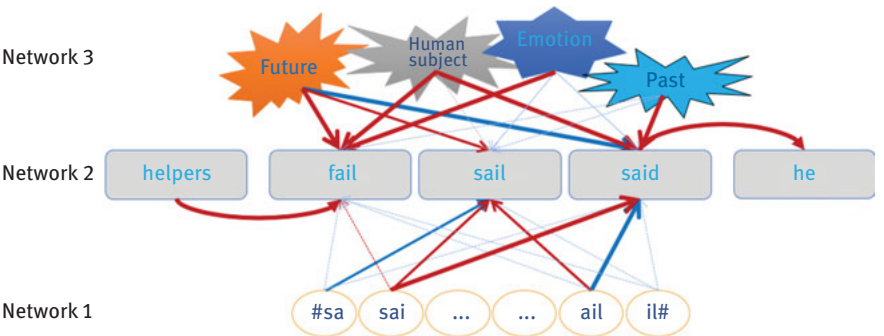

Figure 1: NDL network layout obtained with the iterative application of Rescorla-Wagner rule, for the lexomes *fail*, *sail*, and *said*. Red arrows represent positive associations, while blue arrows represent negative associations. Arrow width reflects the absolute magnitude of the weights on the connections. Networks are trained independently of each other.

Of the three networks in Figure 1, the first one represents bottom-up associations from perceptual input cues (here letter trigrams) to lexomes. This network is referred to as a 'Grapheme-to-Lexome network' (or G2L-network). Milin, Feldman et al. (2017) trained such a network on utterances from a 1.1 billion word corpus of English subtitles (Tiedemann 2012), using letter trigrams such as *#sa*, *sai*, *ail* and *il#*, or *#he* and *he#*, to lexomes such as *sail* and *he*. Three measures that can be derived from such G2L networks have been found to be predictive for experimental measures gauging lexical processing costs. First, the *Activation* of a lexome is defined as the sum of the weights on the connections from the cues in the input to that lexome. Second, the *Prior* availability of a lexome is estimated by the L1-norm of the weights on the connections from all cues to that lexome.[6] Whereas the

6 The L1-norm of a numeric vector is the sum of its absolute values. Like the Euclidean distance (the L2-norm), the L1-norm is a distance measure. It is the distance between two points

activation measure, given the network, is determined by the input, the prior availability is a systemic measure that is independent from the input and is determined by the network only. Prior availability can be understood as a measure of network entrenchment, and hence is reminiscent of the priors in Bayesian models of word recognition (Norris 1994, 2006; Norris and McQueen 2008). The activation *Diversity*, finally, is the L1-norm of the activations of all lexomes generated by the input. It gauges the extent to which other lexomes are co-activated by the input. All three measures have been found to be good predictors for a number of experimental tasks across languages (cf. Baayen et al. 2011; Baayen, Milin, and Ramscar 2016 for visual lexical decision; Milin, Divjak et al. 2017 for self-paced reading in Russian, Hendrix, Bolger, and Baayen 2017 for ERPs, and Arnold et al. 2017 for spoken word identification).

The second learning network, partially represented in the middle row in Figure 1, has lexomes both as input cues and as output outcomes. In Figure 1, only two connections are indicated: the connection from the lexome *helpers* (cue) to the lexome *fail* (outcome), and from the cue *said* to the outcome *he*. Weights estimated from the corpus of English subtitles suggest that these two connections have strong and positive association strengths. From this 'Lexome-to-Lexome network' (L2L-network), several further measures can be derived. In parallel to the Diversity and outcome Prior availability based on a G2L network, an L2L Diversity of activations as well as an L2L Prior availability can be derived, again using L1-norms. Both measures are strong predictors of lexical processing costs, alongside the G2L measures.

L2L networks define semantic vector spaces (cf. Baayen, Milin et al. 2016; Milin, Feldman et al. 2017; see Marelli and Baroni 2015; Acquaviva et al. 2020, this volume for an overview of distributed semantic models). The rows of the L2L weight matrix that defines the L2L network constitute the semantic vectors of the model. Importantly, it is these semantic vectors that the lexome units in the G2L and L2L networks identify (or "point" to). From the cosine similarity matrix of the L2L row vectors, two further measures have been derived and tested against empirical data: a lexome's Semantic Density and a lexome's Semantic Typicality. A lexome's Semantic Density is defined as the number of all lexomes that have a very high cosine similarity with the target lexome. Similarly, a lexome's Semantic Typicality is defined as the cosine similarity of that lexome's semantic vector and the average semantic vector (see also Marelli and Baroni 2015; Shaoul

on a grid when one can move only in the direction of the axes. Thus, whereas the L2-norm of the point (3,–4) is 5, the L1-norm is 7.

and Westbury 2010). Milin, Feldman et al. (2017) observe inhibition from semantic density and facilitation from semantic typicality for lexical decision latencies.

Milin, Divjak et al. (2017) introduced a third NDL network with content lexomes as outcomes, and as cues what we call 'experiential' lexomes. This third network was labeled the BP2L network. Relying on the Behavioural Profiles developed by Divjak & Gries (2006) and later publications, it indexes dimensions of experience, including those that are marked grammatically, such as aspect, tense, mood and number. The authors show that the activations that lexomes of 'try'- verbs receive from such grammatical lexomes are predictive for reading latencies obtained in self-paced sentence reading in Russian. Statistical analyses also revealed that participants optimized their responses in the course of the experiment: the activations had an inhibitory effect on reading latencies at the beginning of the experiment, that later reversed into facilitation. The results from the Milin, Divjak et al. (2017) study are especially interesting as they show that the linguistic profiling of words or constructions (Divjak and Gries 2006; see also Bresnan et al. 2005) can be integrated within a computationally exact approach to learning to yield novel insights into language processing.

Baayen, Milin, and Ramscar (2016), for example, demonstrated and discussed how empirically well-established yet theoretically neglected frequency effects emerge naturally from discriminative learning. The Activation and Prior availability measure are strongly correlated with frequency of occurrence in the corpus on which the network is trained. They can be viewed as measures of frequency that have been molded by discriminative learning. At the same time, interactive activation models account for frequency effects by coding frequency of use into resting activation levels, and Bayesian models build them in by means of priors. Both approaches in effect assume some kind of counter in the head.

3.1.3 Recent developments

In principle, any activation-based computer model of utterance comprehension should be able to discriminate, based on levels of activation, between the intended words actually found in an input utterance and the tens of thousands of other irrelevant words that are potentially available. For example, upon being exposed to *Bill ate the apple pie*, the model should perceive, as the most highly activated units, the individual forms corresponding to the following lexical and grammatical categories: BILL, EAT_PAST and DEF_APPLEPIE. In practice, the two individual forms *apple* and *pie* may be the most highly activated units, and may (wrongly) be perceived as associated with APPLE and PIE respectively, rather than with APPLEPIE as one 'meaning' contrast. In the context of *Bill ate*

the apple pie this would be a case of misclassification of input data. The correct interpretation of an utterance thus requires that all and only its intended word units are classified correctly, by discarding all other irrelevant units that may possibly get activated.

NDL models trained on large corpora may not always achieve this. This is perhaps unsurprising, as a 10 million word corpus such as the TASA (Landauer, Foltz, and Laham 1998) can easily contain 50,000 words that occur at least twice. Hence, classification of these 50,000 words, given their large number and rare occurrence, is a formidable task. In that sense, if these words would be among the first 300 most highly activated candidates, such result would be respectable. Nevertheless, human performance is typically more precise. Baayen et al. (2017) show that classification accuracy can be improved considerably, to human-like levels, by working with coupled error-driven networks. The weights of the two networks are estimated independently, i.e. the same error is 'injected' twice. The first network takes sublexical orthographic or auditory features as input cues, and has lexomes as outcomes. The second network takes as input the output of the first network, i.e. a vector of activations over all lexomes. The outcomes of the second network are again lexomes. The second network thus implements a second try, taking the results from the first network and attempting to predict once again the lexomes that are actually present in the learning event.

We illustrate the coupled networks by means of a simple example, which we also use to lay out the novel way in which the discriminative perspective addresses lexical access. Table 1 lists 10 sentences together with their (randomly generated) frequency of occurrence and a list of the lexomes occurring in each sentence. This list is not intended to be comprehensive, but to illustrate some modelling strategies while keeping the complexity of the example low.

Table 1: Sentences, selected lexomes in the message, and frequency of occurrence, totaling 771.

no.	Sentence	Lexomes (lexical meanings)	Frequency
1	Mary passed away	MARY DIE PAST	40
2	Bill kicked the ball	BILL KICK PAST DEF BALL	100
3	John kicked the ball away	JOHN KICK PAST DEF BALL AWAY	120
4	Mary died	MARY DIE PAST	300
5	Mary bought some flowers	MARY BUY PAST SOME FLOWERS	20
6	Ann bought a ball	ANN BUY PAST INDEF BALL	45
7	John filled the bucket	JOHN FILL PAST DEF BUCKET	100
8	John kicked the bucket	JOHN DIE PAST	10
9	Bill ate the apple pie	BILL EAT DEF APPLEPIE	3
10	Ann tasted an apple	ANN TASTE PAST INDEF APPLE	33

Several aspects of the choice of lexomes are important. First, the sentences with "kicked the bucket", "passed away", and "died", are all associated with the same lexome DIE. This is a many-forms-to-one-lexome mapping (for a discussion of idiom comprehension in this framework, see Geeraert et al. 2017). Second, past-tense word forms such as regular "passed" and irregular "ate" are mapped onto two lexomes, PASS and PAST, and EAT and PAST respectively. One might want to add further grammatical lexomes here, such as a lexomes for person and number. Here, we have a one-form-to-multiple-lexomes mapping. Third, the compound "apple pie" is represented as a single onomasiological entity with the lexome APPLEPIE.

The task of the network is to identify all lexomes that are encoded in the input. This multi-label classification task is one that has to be accomplished solely on the basis of the letter trigrams in the input. For the sentence *John kicked the bucket*, the unique trigraphs that constitute the input cues are #Jo, Joh, ohn, hn#, n#k, #ki, kic, ick, cke, ked, ed#, d#t, #th, the, he#, e#b, #bu, buc, uck, ket, et# (duplicate triplets like *cke* are included only once; again, the # symbol represents the space character).

For this multi-label classification task, we use a coupled network as described above. The first network has the trigram cues as input, and the lexomes as output. A given set of input cues produces a vector of activations over the lexomes. When presented with the sentence *John kicked the bucket*, a network trained on the mini-corpus summarized in Table 1 incorrectly assigns a higher activation to the grammatical lexome DEF than to the lexome DIE (see Figure 2, left upper panel, and related discussion below). A language model bringing in (often implicitly) sophisticated, high-level 'knowledge about the world', could help alleviating this kind of problem for words in utterances, by providing 'hints' to desired outcomes. However, any such language model would give its contribution "for free", as nothing would be revealed about how this knowledge was acquired in the first place.

Classification accuracy is improved by taking the vector of activations produced by the first network, and giving the second network the task of discriminating between the lexomes encoded in the utterance and those that are not part of the message. This second network is a lexome-to-lexome network, but the inputs are no longer dichotomous (1 or 0, depending on whether the lexome is present in the input) but real-valued (see left panels in Figure 2). As a consequence, the Rescorla-Wagner equations cannot be used. Instead, the closely related learning rule of Widrow and Hoff (1960), identical to the Rescorla-Wagner rule under proper parameter selection, can be used for incremental updating of the weights, learning event by learning event. Instead of the Widrow-Hoff learning rule, the weights of the second network can also be estimated by

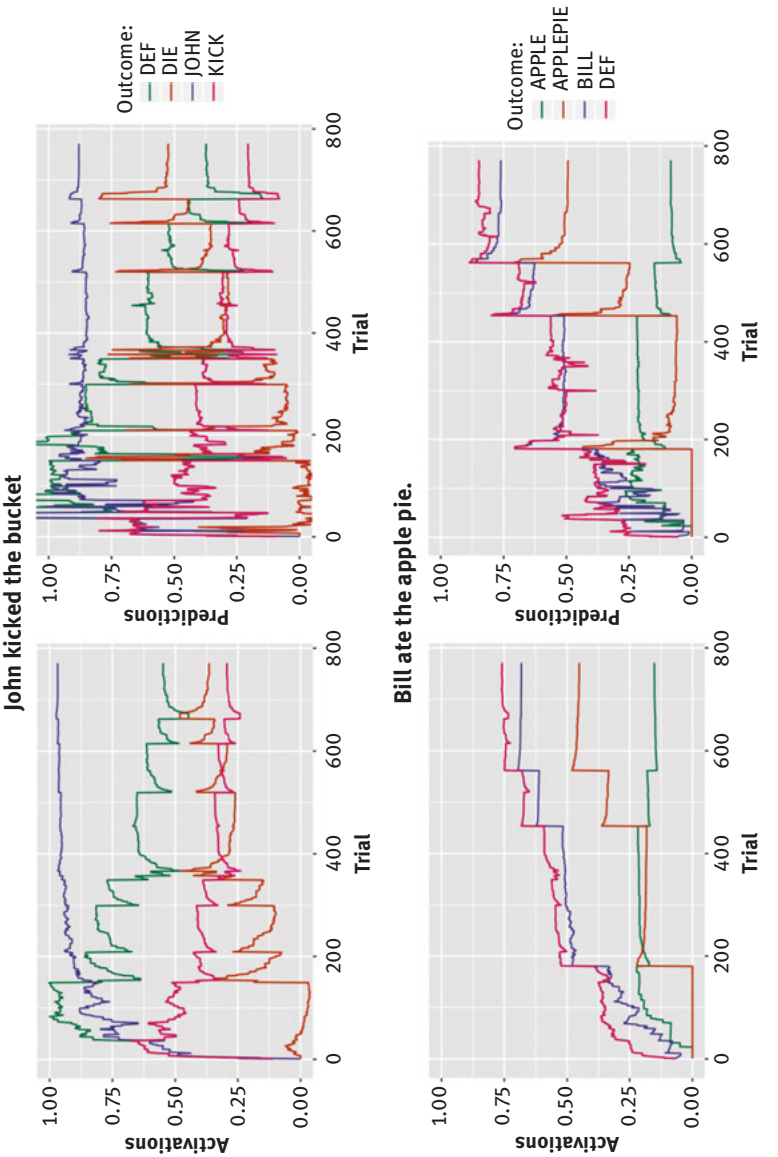

Figure 2: Activations and prediction strengths for selected lexomes in the learning events of sentences 8 and 9 in Table 1, using incremental coupled Rescorla-Wagner and Kalman Filtering. Left panels: activations from the first network; right panel: predictions from the second network. For clarity we are not presenting all of the outcomes.

means of Kalman Filtering (KF: Kalman 1960). The Kalman filter improves on Widrow-Hoff learning by taking the cues' uncertainty (i.e. variance-covariance) into account.[7]

For the mini-corpus presented in Table 1, we estimated the weights for the two networks. For each learning event, we first updated the weights of the first network, then calculated the vector of activations over the lexomes, and subsequently used this as input for the second network; we used the Rescorla-Wagner learning rule for the first network, and the Kalman filter for the second network. By setting all relevant parameters of the two networks to compatible values (for both networks, the learning rate (η) was set to 0.01 and for the second network initial variances – input variance (i.e. cue uncertainty), and output variance (i.e. noise) were all set to 1.0), we can inspect the details of an incremental training regime when the networks are trained in parallel.

Figure 2 shows how the performance of the model develops for selected lexomes in sentences 8 and 9 (see Table 1), *John kicked the bucket* and *Bill ate the apple pie*. For training, the 771 sentence tokens, each constituting one learning event, were randomly ordered. To avoid clogging up the figure, only lexomes of interest are graphed. The upper left panel presents the activations of the lexomes DEF, DIE, JOHN, and KICK. Initially, the network assigns a high activation to KICK and a low activation to DIE. As training proceeds, the activation of the unintended lexome KICK decreases while the activation of DIE increases. The jagged pattern in the learning curves reflects that weights are strengthened only when a given lexome is present in the learning trial, while they are weakened whenever cues supporting e.g. DIE in a sentence with *kick the bucket* are used in sentences that do not contain DIE. Thus, the weight on the connection from the trigram ed# to DIE will be weakened whenever the sentence *Ann tasted an apple* is encountered. The upper left panel also illustrates that the lexome DEF has an inappropriately high activation even at the end of training. The upper right panel shows the activations produced by the second network. By the end of training, the lexomes DEF and KICK are properly downgraded, and the lexomes actually encoded in the input, JOHN and DIE, correctly appear with the highest activations.

7 A computationally efficient implementation of both WH and KF is currently under development by the last author (P. Milin) and his research group (https://outofourminds.bham.ac.uk/). Alternatively, given a set of learning events and the vectors of activations over the outcomes for these learning events, finding the weights of the second network amounts to solving a set of equations, which can be accomplished mathematically with the generalized inverse. In current implementations, this second method is much faster, but, unfortunately, it misses out on the consequences of incremental learning.

The bottom panels present the development of activations for *Bill ate the apple pie*. Here, the relative activations of APPLE and APPLEPIE are of interest. Note that in the initial stages of learning, APPLE receives a higher activation than APPLEPIE. By the end of learning, the first network already succeeds in discriminating apple pies from apples, and the second network enhances the difference in activation even further. The fact that APPLE has not been completely suppressed is, in our opinion, an asset of the model. In a multi-label classification problem, a winner-takes-all set-up, as commonly found in interactive activation models, cannot work. In fact, we think that semantic percepts are co-determined by all lexomes in the system, proportional to their activation. (In the semantic vector space, this hypothesis translates into all lexomes having vectors the length and prominence of which is modulated by their activation.) Thus, according to the present example model, there is an *apple* in *apple pie*, but the model also knows very well that Bill ate an APPLEPIE and not an APPLE. This highlights that in the present approach, the semantics of complex words are not derived from the semantics of their parts by some combinatorial operation.

Comparing the panels in the left and right columns of Figure 2 reveals that the first network (the Rescorla-Wagner network) shows a more stable behavior, which means that it 'learns' faster than the second network trained with the Kalman Filter.[8] Nevertheless, by the end of the learning sequence, only the second network succeeds in giving the intended lexomes higher scores.

3.1.4 Advantages of NDL

An important design property of NDL is that 'lexical access' is defined as a multi-label classification problem driven by low-level, sublexical features. A hierarchy of units, such as letter features, letters, morphemes and words for reading, and phonemes, syllables, morphemes, and words for auditory comprehension, is not part of the model. In fact, such a hierarchy of units is viewed as disadvantageous, because low-level co-occurrence information is a-priori ruled out to influence comprehension. For instance, fine phonetic detail below the phoneme that is present across (co-articulated) syllables is lost when comprehension is filtered first through abstract phonemes and then through abstract syllables. Baayen, Shaoul et al. (2016) show how the word segmentation problem, which is computationally

8 However, the Kalman Filter network learns much faster than a network trained with the Widrow-Hoff learning rule, as can be seen by comparing the present results with those reported in Sering et al. (2018) using a variant of WH.

hard, is no longer an issue in a discrimination-driven approach. Arnold et al. (2017), furthermore, show that an NDL network trained on cues derived from the speech signal achieves an identification accuracy that is within the range of human identification performance.[9]

Inspired by Word and Paradigm Morphology (Matthews 1974; Blevins 2016), NDL likewise avoids the popular idea that the morpheme is a linguistic sign, which goes back to post-Bloomfieldian American structuralism. This does not imply that NDL denies the relevance of all linguistic variables such as tense, aspect, person, or number. In fact, the approach implements such variables through 'experiential' lexomes, as illustrated above in Figure 2. However, form units for morphemes are not part of the model (cf. Milin, Feldman et al. 2017; and also consult Schmidtke et al. 2017). Finally, the discriminative perspective also sheds light on why – often fairly idiosyncratic – allomorphy is widespread in morphological systems. Such allomorphy requires complex adjustment rules (or extensive listing) in classic decompositional approaches, while from the discrimination stance allomorphy renders the base word and the complex word less similar in form, which consequently makes the two easier to distinguish (see also Blevins, Milin, and Ramscar 2017).

In the discriminative framework, NDL is a computational implementation of implicit learning, i.e. the learning that goes on without conscious reflection. This kind of learning is not unique to language. For instance, Marsolek (2008) discusses how error-driven updating of visual features affects cognition. Implicit learning is likely the dominant form of learning in young children, whose cognitive control systems are not well-developed. As prefrontal systems mature, it becomes possible to consider multiple sources of information simultaneously, leading to markedly different performance on a variety of tasks (Ramscar and Gitcho 2007). Indeed, Ramscar, Dye, and Klein (2013) provide an example of the very different performance, on the same novel-object labelling task, of young children on the one hand and adults on the other, with the children following discriminative informativity, and the adults applying logical reasoning. As a

9 The auditory model also takes acoustic reductions in its stride. Standard computational models of auditory comprehension are challenged by strongly reduced forms, which are ubiquitous in spontaneous speech. When reduced forms are added to the inventory of word forms, recognition systems tend not to improve. Although some words may be recognized better, the addition of many short, reduced forms typically increases problems elsewhere (Johnson 2004). From a discriminative perspective, reduced forms simply have different acoustic features, and as the requirement is dropped that comprehension must proceed through an abstract standardized form representation, the acoustic features that are highly specific for the reduced form can straightforwardly support the intended lexomes.

consequence, NDL networks will often not be sufficient for predicting adult behavior in experiments addressing morphological learning using stimuli constructed according to some artificial grammar. For such data, NDL can still be useful for clarifying where human performance deviates from what one would expect if learning were restricted to implicit learning, which in turn is informative about where additional processes of cognitive control addressing response competition are at work. If the goal is to clarify implicit learning in adults, which we think takes place continuously (but not exclusively), great care is required to ensure that participants do not have time to think about the task they are performing or to develop response strategies.

NDL networks provide functional models for tracing the consequences of discriminative learning for lexical processing. Although there is ample neurobiological evidence for error-driven learning (e.g. Schultz 1998), actual neural computation is much more complex than suggested by the architecture of a two-layer artificial neural network. Because of this, the NDL model remains agnostic about possible spatial clustering of cues and outcomes in neural tissue.

Published work using NDL addresses primarily aspects of language comprehension. Much less work has been done on speech production. Ramscar, Dye, and McCauley (2013) show how discrimination learning predicts the U-shaped learning curve often observed for the acquisition of irregular morphology. Hendrix (2015) developed a computational model for word naming that is built on two discrimination networks. Recent studies (Tucker et al. 2017; Lensink et al. 2017) suggest that specifically the activation diversity measure helps predict the acoustic durations with which segments or utterances are realized in speech. Whether a computational model of speech production that eschews representations for phonemes and morphemes can be made to work is currently under investigation.

3.2 Temporal Self-Organizing Maps

Although most recent work in discriminative word learning has primarily focused on form-meaning relationships based on highly-distributed a-morphous representations, a recurrent network variant of discriminative learning has recently been used in one-level self-organizing grids of processing nodes known as Temporal Self-Organizing Maps (TSOMs, Ferro et al. 2011; Marzi et al. 2014; Pirrelli et al. 2015). TSOMs develop Markov-like chains of memory nodes that can mirror effects of gradient morphological structure and emergent paradigmatic organization upon exposure to simple inflected forms. By developing specialized patterns of input receptors through recurrent connections, TSOMs recode *one-level stimuli auto-associatively*, thereby exploiting the formal redundancy of

temporal series of symbols. From this perspective, discriminative learning proves to be a powerful strategy for scaffolding the input stream into internalized structured representations, which turn out to be useful for efficient word recognition and production. Here we will show how TSOMs can be used as lexical memories.

3.2.1 Architecture outline

The core of a TSOM consists of an array of nodes with two weighted layers of synaptic connectivity (Figure 3). Input connections link each node to the current

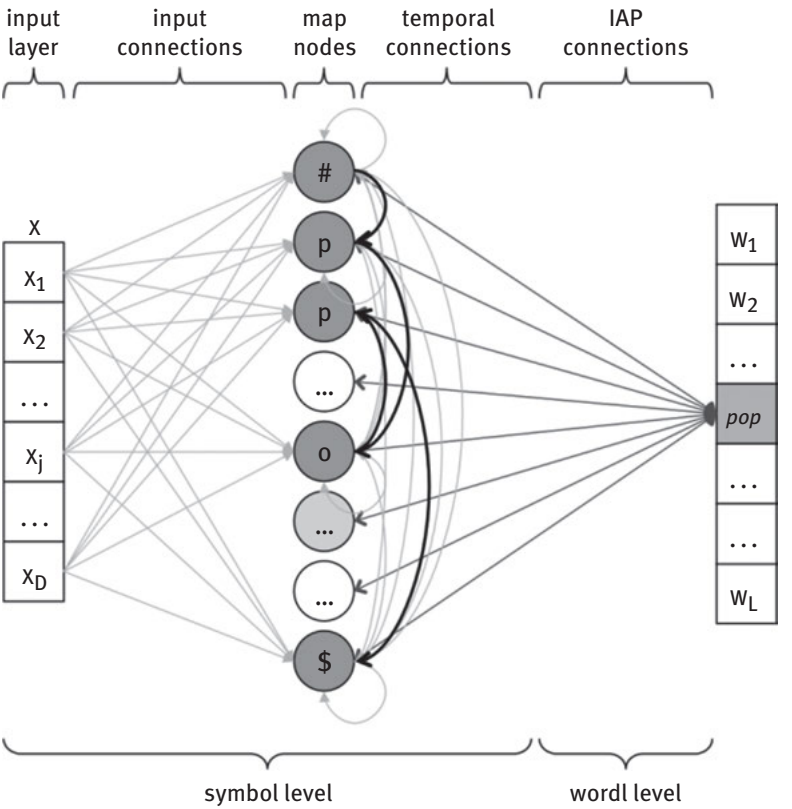

Figure 3: Functional architecture of a Temporal Self-Organizing Map (TSOM). Shades of grey represent levels of activation of map nodes, from low (light grey) to high (dark grey). The figure depicts the integrated level of activation of the map after the word *pop* ('#pop$') is shown in input.

input stimulus (e.g. a letter or a sound), a one-hot vector presented on the input layer at a discrete time tick. Temporal connections link each map node to the pattern of node activation of the same map at the immediately preceding time tick. In Figure 3, these connections are depicted as re-entrant directed arcs, leaving from and to map nodes. Nodes are labelled by the input characters that fire them most strongly. '#' and '$' are special characters, marking the beginning and the end of an input word respectively.

3.2.2 Processing and storage

Storage and processing are traditionally seen as independent, non-interactive functions, carried out by distinct computer components, with data representations defined prior to processing, and processing applied independently of input data. Conversely, in a TSOM storage and processing are two different time-scales of the same underlying process, defined by a unique pool of principles: (i) long-term storage depends on processing, as it consists in routinized time-bound chains of sequentially activated nodes; (ii) processing is memory-based since it consists in the short-term reactivation of node chains that successfully responded to past input. As a result of this mutual interaction, weights on input and temporal connections are adaptively adjusted as a continuous function of the distributional patterns of input data.

Algorithmically, when an input vector $x(t)$ (say the letter *o* in Figure 3) is input to the map at time *t*, activation propagates to all map nodes through both input and temporal connections. The most highly activated node at time *t* is termed Best Matching Unit ($BMU(t)$ for short), and represents the processing response of the map to the current input.

Following this short-term processing step, both input and temporal connections are updated incrementally, for map nodes to be made more sensitive to the current input. In particular, for each j^{th} input value $x_j(t)$ in the input vector, its connection weight $w_{i,j}$ to the i^{th} map node is incremented by *equation 3*:

$$\Delta w_{i,j}(t) = \gamma_I(E) \cdot G_I(d_i(t)) \cdot \left[x_j(t) - w_{i,j}(t)\right] \tag{3}$$

Likewise, the temporal connections of the i^{th} node are synchronized to the activation state of the map at time *t-1*, by increasing the weight $m_{i,BMU(t-1)}$ on the connection from $BMU(t-1)$ to the i^{th} node (*equation 4*), and by decreasing all other temporal connections to the i^{th} node (*equation 5*).

$$\Delta m_{i,h}(t) = \gamma_T(E) \cdot G_T(d_i(t)) \cdot [1 - m_{i,h}(t)]; \quad h = BMU(t-1). \tag{4}$$

$$\Delta m_{i,h}(t) = \gamma_T(E) \cdot G_T(d_i(t)) \cdot [0 - m_{i,h}(t)]; \quad h \neq BMU(t-1). \tag{5}$$

Note that for both input and temporal connections, the resulting long-term increment (respectively $\Delta w_{i,j}(t)$ and $\Delta m_{i,h}(t)$) is an inverse function (respectively $G_I(\cdot)$ and $G_T(\cdot)$) of the topological distance $d_i(t)$ between the i^{th} node and the current $BMU(t)$, and a direct function (respectively $\gamma_I(\cdot)$ and $\gamma_T(\cdot)$) of the map's learning rate at epoch E.[10]

Because of this dynamic, $BMU(t)$ will benefit most from weight adjustment at time t, but information will nonetheless spread radially from $BMU(t)$ to topologically neighbouring nodes. In the end, the map develops a topological organization where nodes responding to the same symbol tend to cluster in a connected area. Figure 4 shows a map trained on German verb inflected forms: each map node is labelled with the input letter it responds most highly to. A node *N* gets the label *L*, if the *L* input vector is at a minimal distance from *N*'s vector of spatial weights. Nodes that are labelled with the same symbol are specialized for responding to that symbol in different temporal contexts. Intuitively, they store long-term information about the typical contexts where the symbol happened to be found in input. Notably, the node that stores specialized information about the *L* symbol in a specific context is the same node that responds most highly to *L* when *L* happens to be input in that particular context.

3.2.3 Information of 'what' and information of 'when'

Input connections store information about the nature of the current input (or 'what' information). The layer of temporal connections encodes the expectation for the current state of map activation given the activation of the map at the previous time tick (or 'when' information). *Equation 4* and *equation 5*, by which 'when' connections are dynamically trained in TSOMs, are strongly reminiscent of Rescorla-Wagner's *equation 2*. Given the input bigram '*ab*', the connection strength between BMU('*a*') at time *t-1* and BMU('*b*') at time *t* will

(i) increase every time '*a*' precedes '*b*' in training (entrenchment)
(ii) decrease every time '*b*' is preceded by a symbol other than '*a*' (competition and inhibition).

10 Intuitively the two functions define the degree of *plasticity* of the map, i.e. how readily the map adjusts itself to the current input stimuli. Hence, they are inverse functions of the map's learning epoch *E*, i.e. their impact decreases as learning progresses.

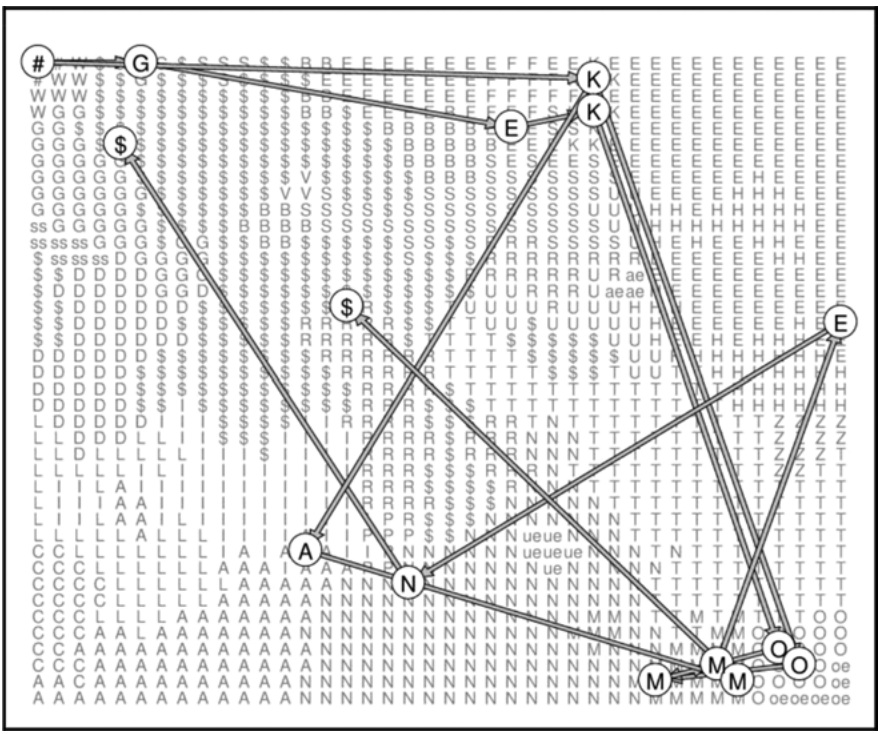

Figure 4: A labelled TSOM trained on German verb inflected forms. Highlighted nodes depict the BMUs activated by the forms *kommen* 'come' (infinitive/1P-3P present indicative), gekommen 'come' (past participle) and *kam* 'came' (1S-3S past tense), with directed arrows representing their activation timeline.

Note, however, that *equation 4* and *equation 5* apparently reverse the cue-outcome relationship of *equation 2*: $BMU(t)$ acts as a cue to $BMU(t-1)$ and strengthens the temporal connection from $BMU(t-1)$ to $BMU(t)$ accordingly (entrenchment). At the same time, all the temporal connections to $BMU(t)$ emanating from nodes other than $BMU(t-1)$ are depressed (competition). To understand this apparent reversal, it is useful to bear in mind that the output of a TSOM is an optimal self-organization of map nodes, based on past stimuli. This is done incrementally, by adjusting the weights on temporal connections to optimize processing of the current input string. Ultimately, *equation 4* and *equation 5* concur to develop the most *discriminative* chains of *BMUs* given a set of training data. This means that $BMU(t)$ is not the map's outcome, but the internally encoded cue to the map's optimal self-organization. By differentially adjusting the incoming temporal connections that emanate from $BMU(t-1)$ and

non-$BMU(t-1)$, the current $BMU(t)$ is in fact specializing a chain of *BMUs* for them to keep in memory, at each step, as many previous processing steps as possible. The *outcome* of $BMU(t)$ is thus the incremental step in building such maximally discriminative chain.

The interaction between entrenchment and competition accounts for effects of context-sensitive specialization of map nodes. If the bigram '*ab*' is repeatedly input, a TSOM tends to develop a dedicated node for '*b*' in '*ab*'. Since node specialization propagates with time, if '*c*' is a frequent follower of '*ab*', the map will strengthen a temporal connection to another dedicated *BMU* responding to '*c*' preceded by '*b*' when preceded by '*a*'. Ultimately, the TSOM is biased towards memorizing input strings through *BMUs* structured in a word tree (Figure 5). As we shall see later in the section on serial word processing, a tree-like memory structure favors word recognition by looking for word uniqueness points as early as possible in the input string.

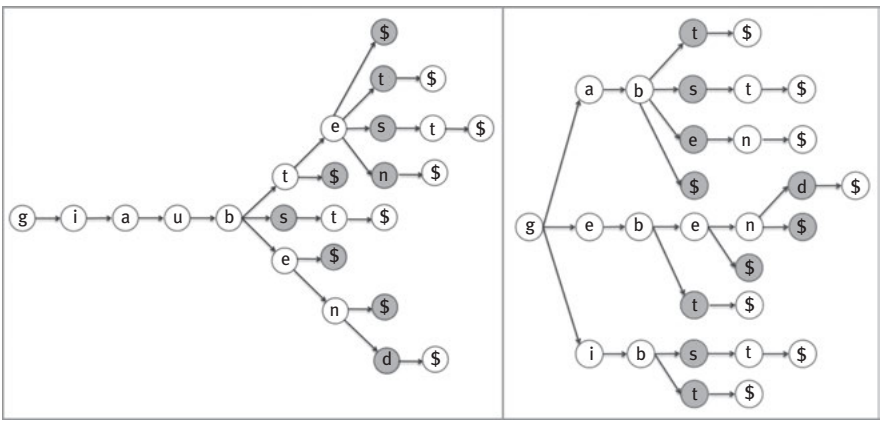

Figure 5: Word-tree representations of sub-paradigms of German *glauben* ('believe') and *geben* ('give'). Shaded nodes represent word Complex Uniqueness Points (see note 2, and the section below on serial word processing).

Figure 6 shows the scatter plot of the number of *BMUs* responding to input symbols in a 40×40 node TSOM trained on 750 German verb forms, regressed on the number of distinct nodes required to represent the same symbols in a word-tree (Pearson's $r = .95$, $p < .00001$). On average, the more contexts a symbol is found in during training (accurately approximated by the number of distinct tree nodes associated with the symbol), the more map nodes will be specialized for that symbol. Context-sensitive specialization of *BMUs* allows a TSOM to

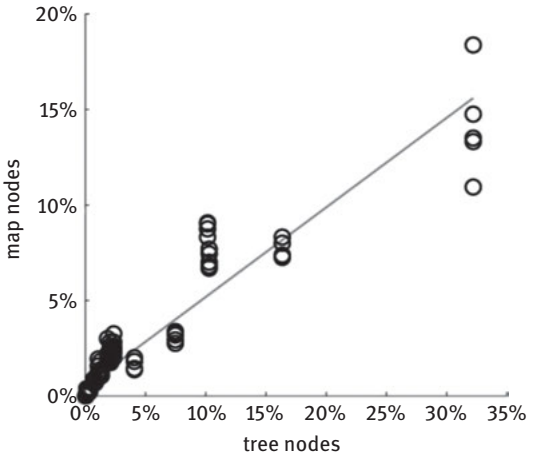

Figure 6: Scatter plot of per-symbol nodes allocated in a map trained on German verb forms. Data are regressed on the number of nodes in a word-tree representing the training data.

allocate specific resources to input symbols that occur at specific points in time. A TSOM develops a growing sensitivity to surface distributional properties of input data (e.g. language-specific constraints on admissible symbol arrangements, as well as probabilistic expectations of their occurrence), turning chains of randomly connected, general-purpose nodes into specialized sub-chains of *BMUs* that respond to specific letter strings in specific contexts. This ability is fundamental to storing symbolic time-series like words.

3.2.4 Using TSOMs as lexical memories

In showing a word like *#pop$* one symbol at a time on the input layer (Figure 3), the activation pattern produced on the map by each symbol in the string is incrementally overlaid with all patterns generated by all other symbols making up the same string. The resulting *Integrated Activation Pattern* (IAP) is shown in Figure 3 by levels of node activation represented as shaded nodes. IAP activation levels are calculated according to the following equation:

$$\hat{y}_i = max_{t=1,\ldots,k}\{y_i(t)\}; i = 1, \ldots, N \tag{6}$$

where i ranges over the number of nodes in the map, and t ranges over symbol positions in the input string. Intuitively, each node in the *IAP* is associated with the maximum activation reached by the node in processing the whole input word. Note that, in Figure 3, the same symbol 'p', occurring twice in *#pop$*,

activates two different BMUs depending on its position in the string. After presentation of *#pop$*, integrated levels of node activation are stored in the weights of a third level of *IAP* connectivity, linking the map nodes to the lexical map proper (rightmost vector structure in Figure 3). The resulting *IAP* is not only the short-term processing response of a map to *#pop$*. The long-term knowledge sitting in the lexical connections makes the current *IAP* a routinized memory trace of the map processing response. Given an *IAP* and the temporal connections between BMUs, a TSOM can thus use this knowledge to predict, for any currently activated BMU in the *IAP*, the most likely upcoming BMU. This makes it possible to test the behavior of a TSOM on two classical lexical tasks: immediate word recall and serial word processing.

3.2.4.1 Word recall

Word recall refers to the process of retrieving lexical information from the long-term word store. We can test the accuracy of the IAPs as long-term lexical representations by simulating a process of recall of a target word from its own IAP. Since an IAP is a synchronous pattern of activated nodes, the task tests how accurately levels of node activation in the IAP encode information about the timing of the symbols that make up the target word. The process of recall consists in the following steps:

(i) initialize:
   a) reinstate the word IAP on the map
   b) prompt the map with the start-of-word symbol '#'
   c) integrate the word IAP with the temporal expectations of '#'
(ii) calculate the current BMU and output its associated label
(iii) if the output label is NOT symbol '$':
   a) integrate the word IAP with the temporal expectations of the current BMU
   b) go back to step (ii)
(iv) stop

A word is recalled correctly from its IAP if all its symbols are output correctly in the appropriate left-to-right order.

There are a number of features that make IAPs interesting correlates of lexical long-term memory traces. First, activation of an IAP makes all its BMUs simultaneously available. This accounts for "buffering effects" (Goldrick and Rapp 2007; Goldrick et al. 2010), where the idea that symbol representations are concurrently maintained while being manipulated for recall explains the distribution of substitution, deletion and transposition errors. Secondly, IAPs

encode word letters in a context-sensitive way, allowing for representation of multiple occurrences of one letter type in the same word. In addition, they rely on a predictive bias, capturing facilitative effects of probabilistic expectation on word processing. Finally, they may contain highly activated nodes that are BMUs of other non-target IAPs, causing strong co-activation (and possible interference) of the latter. To illustrate, if two input strings present some symbols in common (e.g. English *write* and *written*, or German *macht* '(s)he makes' and *gemacht* 'made', past participle), they will tend to activate largely overlapping patterns of nodes.

A TSOM can be said to have acquired a new word form when the word form is accurately recalled from its own IAP. Accordingly, the time of acquisition of a word can be defined as the earliest learning epoch since the word is always recalled accurately. Monitoring the pace of acquisition of words through learning epochs thus allows us to observe which factors affect word acquisition. Concurrent memorization of morphologically redundant forms in inflectional paradigms prompts competition for the same memory resources (processing nodes and temporal connections). Due to *equation 4*, at each processing step, weights on the temporal connection between $BMU(t)$ and $BMU(t+1)$ are reinforced (entrenchment). At the same time, *equation 5* depresses presynaptic connections to $BMU(t+1)$ from any other node than $BMU(t)$(competition). This simple per-node dynamic has far-reaching consequences on the global self-organization of the map at the word level.

First, the number of nodes responding to a specific input symbol is directly proportional to the token frequency of that symbol. As a result of this correlation (Pearson's r = .95, p < .00001), at early learning epochs, high-frequency words are assigned a larger pool of processing resources than low-frequency words are. In addition, entrenchment makes the time taken for a form to develop strong temporal connections an inverse function of the token frequency of the form. The large availability of processing nodes and dedicated connections causes high-frequency words to be acquired (i.e. accurately recalled from their own IAPs) at earlier learning epochs than low-frequency words (Figure 7, right panel).

Figure 7, left panel, shows the pace of acquisition for regular and irregular verb forms in German, focusing on the interaction between word length and inflectional regularity.

Together with word frequency, word length appears to be a major factor delaying the time of acquisition. Longer words are more difficult to recall since more, concurrently-activated BMUs in an IAP are easier to be confused, missed or jumbled than fewer BMUs are. When word length and word frequency are controlled, regularly inflected forms are recalled at earlier stages than irregulars. The evidence

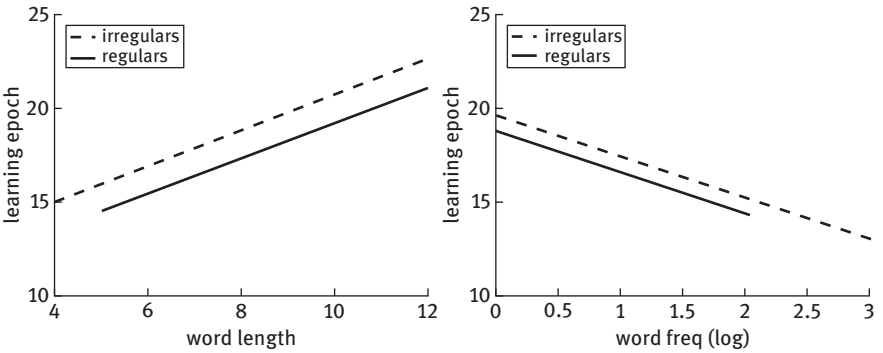

Figure 7: Marginal plots of interaction effects between word length (left panel), word frequency (right panel), and inflectional regularity (solid lines = regulars, dotted lines = irregulars) in an LME model fitting word learning epochs in German.

is in line with the observation that speakers produce words that belong to bigger neighbor families more quickly than isolated words (Chen and Mirman 2012).

3.2.4.2 Serial word processing

Serial word processing involves the processing of an input signal unfolding with time, as is the case with auditory word recognition. Serial lexical access and competition are based on the incremental activation of onset-sharing items, forming a cohort-like set of concurrently activated lexical competitors (Marslen-Wilson 1984; Marslen-Wilson and Welsh 1978). The so-called Uniqueness Point (UP) defines the position in the input string where the cohort of competitors winnows down to unity, meaning that there is only one possible lexical continuation of the currently activated node chain. Figure 5 provides a few examples of Complex Uniqueness Point (or CUP: Balling and Baayen 2008, 2012) for trees of inflectionally related lexical items. Unlike Marslen-Wilson's original definition of UP, which is meant to mark the point in time at which morphologically unrelated words are teased apart, at CUP a target input word is distinguished from the set of its paradigmatically-related companions.

To analyze serial word processing with TSOMs, we monitor the activation state of a map incrementally presented with an input word. Upon each symbol presentation on the input layer at time t, a TSOM is prompted to complete the current input by predicting its most likely lexical continuation. The map propagates the activation of the current BMU(t) through its forward temporal connections, and outputs the label $L_{BMU(t+1)}$ of the most strongly (pre)activated node $BMU(t+1)$:

$$BMU(t+1) = argmax_{i=1,\ldots,N}\{m_{i,h}\}; \quad h = BMU(t), \tag{7}$$

where $m_{i,h}$ is the weight on the forward temporal connection from node h to node i, and N the overall number of map nodes. Prediction accuracy across the input word is calculated by assigning each correctly anticipated symbol in the input word a 1-point score. Otherwise, the symbol receives a 0-point score. We can then sum up the per-symbol prediction scores in an input word and average the sum by the input word length, to obtain a per-word prediction score; the higher the score, the easier for the map to process the input word.

The panel in Figure 8 shows how prediction scores vary, on average, in 750 German verb forms, as a function of the incremental left-to-right processing of input symbols. Input symbols are plotted by their distance from the word stem-ending boundary (x = 0 denotes the first position in the input string after the base stem). Training forms are selected from the 50 top-ranked German verb paradigms by their cumulative frequency in Celex (Baayen et al. 1995), and classified as either regular or irregular.[11]

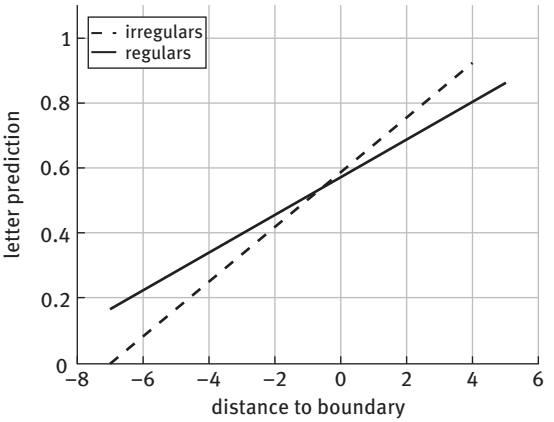

Figure 8: Marginal plot of interaction effects between letter distance to the stem-inflection boundary (x axis, with $x = 0$ marking the first letter in the inflectional ending) and inflectional regularity (regular = solid line vs. irregular = dashed line) in an LME model fitting letter prediction (y axis) in a TSOM trained on German verbs.

11 Following a paradigm-based approach to inflection (Aronoff 1994; Blevins 2016; Matthews 1991), all inflected forms belonging to regular paradigms share an invariant base stem (e.g. *walk*, *walk-s*, *walk-ed*, *walk-ing*), whereas irregular paradigms exhibit a more or less wide variety of phonologically unpredictable stems (*sing*, *sing-s*, *sang*, *sung*, *sing-ing*). Paradigms can thus be classified according to the number of base stems they select, and individual forms are more or less regular depending on the number of their stem-sharing neighbors.

In Figure 8, prediction scores are found to get higher while the end of the form is approached. This is an expected consequence of the reduction in uncertainty for possible lexical continuations at lower nodes in a word-tree. However, the rate of increase follows significantly different slopes in regulars and irregulars.

The evidence is accounted for by the way regularly and irregularly inflected forms are structured in a word-tree (Figure 5). German irregular paradigms (e.g. *geben*) typically present vowel-alternating stems (e.g. *geb-*, *gib-*, *gab-*), which cause their tree-like representation to branch out at higher nodes in the hierarchy (Figure 5). Stems in regular verbs, on the other hand, do not suffer from the competition of other stem alternants within the same paradigm.[12] The general pattern is plotted in Figure 9, depicting the branching-out factor (or node "arity") in the word-tree representation of German verb forms by inflectional regularity and letter distance from the morpheme boundary. Irregulars appear to show a higher branching-out factor at early nodes in the word-tree representation. This factor, however, shrinks further down in the hierarchy more quickly in irregulars than in regulars. This means that processing decisions made on early nodes in the tree-structure reduce the level of processing uncertainty downstream in the lexical tree. Intuitively, once a specific stem alternant is

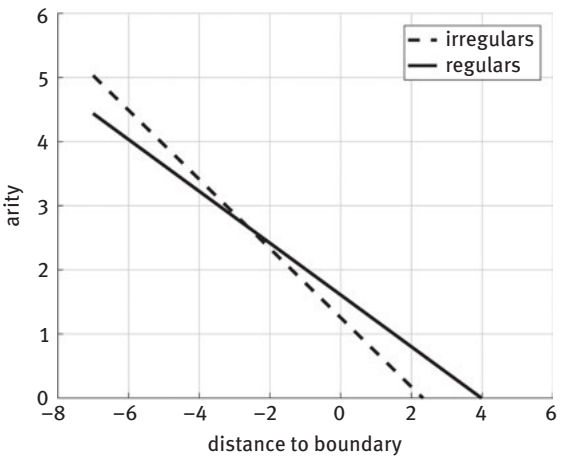

Figure 9: Marginal plot of interaction effects between distance to stem-inflection boundary (x axis) and inflectional regularity (regular = solid line vs. irregular = dashed line) in an LME model fitting node arity (y axis) in a word-tree of German verbs.

12 Clearly, both regular and irregular stems can be onset-aligned with other paradigmatically-unrelated stems. Our evidence shows that this extra-paradigmatic "competition" affects both regulars and irregulars to approximately the same extent.

found at the beginning of an irregularly inflected form (e.g. *gab-* in *gaben*), the number of admissible paths branching out at the end of the selected stem goes down dramatically.

These structural properties accord well with evidence that time latencies in processing words out of context are a function of lexical uniqueness points, i.e. the word-internal positions where the human processor can uniquely identify an input word. Balling and Baayen (2008, 2012) show that, in morphologically complex words, lexical processing is paced by two disambiguation points: (i) the uniqueness point distinguishing the input stem form other morphologically-unrelated onset-overlapping stems (or UP1), and (ii) the complex uniqueness point distinguishing the input form from other morphologically-related forms sharing the same stem (or CUP). To illustrate (see Figure 5), in a toy German lexicon containing two paradigms only, namely *geben* ('give') and *glauben* ('believe'), UP1 for *gebt* ('you give', second person plural) is the leftmost letter telling *gebt* from all forms of *glauben*: namely, *e* in second position. Its CUP is the leftmost letter that distinguishes *gebt* from all other forms of *geben*: *i.e. t* in fourth position.

Balling and Baayen show that late UP1s are inhibitory and elicit prolonged reaction times in acoustic word recognition. The evidence challenges the Bayesian decision framework of Shortlist B (Norris and McQueen 2008), where intermediate points of disambiguation play no role in predicting response latencies in auditory comprehension. Balling and Baayen's evidence is nonetheless modelled by a quantitative analysis of the TSOM processing response.

Figure 10 (top panel) depicts average prediction scores in a TSOM processing input symbols in German verb stems, plotted by increasing position values of UP1 in the word form, measured as a distance from the start of the word. Late UP1s slow down processing by decreasing prediction scores. The bottom panel of Figure 10 shows a similar pattern. As expected, late CUPs elicit lower suffix prediction scores than early CUPs.

Finally, when the influence of both UP1 and CUP is taken into account, their joint effect on processing is additive: for any two words with the same CUP position, the word with a later UP1 is processed more slowly by a TSOM than the word with an earlier UP1, in keeping with evidence of human processing (Balling and Baayen 2012).

3.2.5 Competition and entropy

There is a clear connection linking competition among members of a morphological family, and the entropy of the frequency distribution of family members (Baayen et al. 2006; Moscoso del Prado Martín et al. 2004). Milin and colleagues

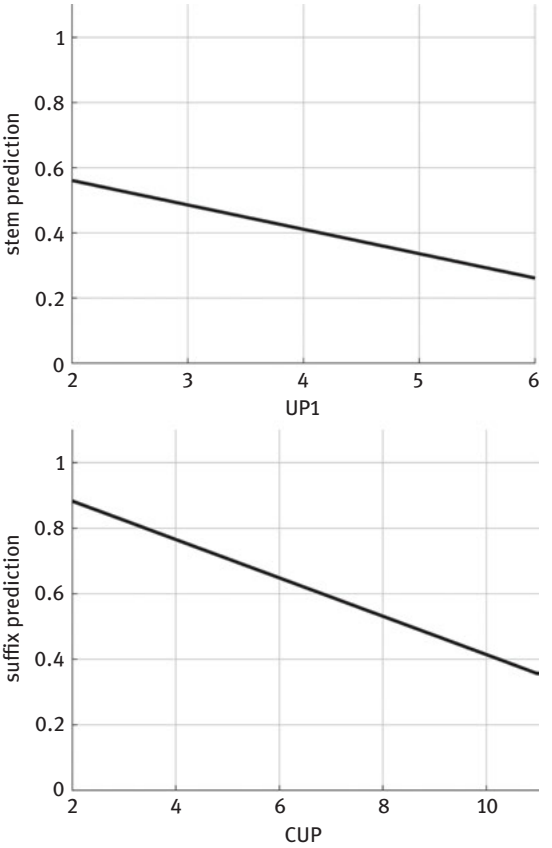

Figure 10: Top panel – marginal plot of interaction effects between UP1 position (x axis) and stem length in an LME model fitting letter prediction in verb stems (y axis) by a TSOM trained on German verbs. Bottom panel – marginal plot of interaction effects between CUP position (x axis) and length of inflectional endings in an LME model fitting letter prediction in verb endings (y axis) by a TSOM trained on German verbs.

(Milin, Filipović Đurđević et al. 2009, Milin, Kuperman et al. 2009) put considerable emphasis on the interactive role of intra-paradigmatic and inter-paradigmatic distributions in accounting for differential effects on visual lexical recognition. In particular, they focus on the divergence between the distribution of inflectional endings within each single paradigm (measured as the entropy of the distribution of paradigmatically-related forms, or Paradigm Entropy), and the distribution of the same endings within their broader inflectional class (measured as the entropy of the distributions of inflectional endings across all paradigms, or Inflectional Entropy). Both entropic scores are known

to facilitate visual lexical recognition. If the two distributions differ, however, a conflict may arise, resulting in slower recognition of the words. These effects are the by-product of a model of the lexicon offering more or less explicit mechanisms dealing with the simultaneous existence of potentially competing paradigmatically related forms, and with the simultaneous existence of multiple paradigms. Similar results are reported by Kuperman et al. (2010) on reading times for Dutch derived words, and are interpreted as reflecting an information imbalance between the family of the base word (e.g. *plaats* 'place' in *plaatsing* 'placement') and the family of the suffix (*-ing*).

The difference between Paradigm Entropy and Inflectional Entropy can be expressed in terms of Relative Entropy, or Kullback-Leibler divergence (D_{KL}, Kullback 1987), as follows:

$$D_{KL}(p(e|s)||p(e)) = \sum_e p(e|s) log \frac{p(e|s)}{p(e)}, \tag{8}$$

where $p(e|s)$ represents the probability of having a specific inflected form (an ending e) given a stem s, and $p(e)$ the probability of encountering e. For any specific paradigm being selected, the larger D_{KL}, the more difficult is, on average, the visual recognition of members of that paradigm.

The relatively simple learning dynamic of TSOMs, expressed by rules (i) and (ii) above, accounts for facilitative effects of paradigm entropy and inflectional entropy on word learning.

To illustrate, we trained a TSOM on three mini-paradigms, whose forms are obtained by combining three stems ('A', 'B' and 'C') with two endings (symbols 'X' and 'Y'). Mini-paradigms were administered to the map on six training regimes (R1-R6, see Table 2), whose distribution was intended to control the comparative probability distribution of 'X' and 'Y', and the comparative probability distribution of the stems 'A', 'B' and 'C' relative to each ending. Across regimes 1–3, we kept the frequency distribution of X constant (but we made it vary across paradigms), while increasing the distribution of Y both within each paradigm (R2), and across paradigms (R3). Across regimes 4-5, the frequency of Y was held constant, while X frequencies were made vary. Finally in R6 all word frequencies were set to 100. Note that in R3 and R6 $p(e|s) = p(e)$: *i.e.*, the distribution of each inflected form within a paradigm equals the distribution of its ending (given its inflection class).

Results of the different training regimes are shown in Figure 11, where we plotted weights on the connection between stems ('A', 'B' and 'C') and endings ('X' and 'Y') by learning epochs, averaged over 100 repetitions of the same experiment on each regime. Results were analyzed with linear mixed-effects models, with stem-ending connection weights as our dependent variable and the

Table 2: Frequency distribution of 3 mini-paradigms (rows) in 6 training regimes (columns).

		Frequency					
paradigm id	**items**	**regime 1**	**regime 2**	**regime 3**	**regime 4**	**regime 5**	**regime 6**
A	#,A,X,$	5	5	5	5	5	100
A	#,A,Y,$	5	50	50	333	333	100
B	#,B,X,$	10	10	10	10	100	100
B	#,B,Y,$	10	100	100	333	333	100
C	#,C,X,$	85	85	85	85	850	100
C	#,C,Y,$	10	100	850	333	333	100

following three fixed effects: (i) the word probability $p(s, e)$, expressed as a stem-ending combination, (ii) the probability $p(e \mid s)$ of a stem selecting a specific ending (or intra-paradigmatic competition), and (iii) the conditional probability $p(s \mid e)$ of a given ending being selected by a specific stem (inter-paradigmatic competition). Experiment repetitions were used as random effects. We refer the interested reader to Ferro et al. (2018) for a thorough analysis of the effects. Here, we shortly summarize the main results observed, and provide an analytical interpretation of this evidence.

Due to entrenchment (*equation 4*), the strength of each connection at the morpheme boundary tends to be a direct function of the probability of each word form, or $p(s, e)$ (see panel R3 in Figure 11). However, other factors interact with word frequency: connection strengths are affected by the probability of each ending $p(e)$, with low-frequency words that contain high-frequency endings (e.g. "AX" in panel R1) showing a stronger boundary connection than low-frequency words that contain less frequent endings ("AY" in panel R1). This boosting effect is modulated by two further interactions: the conditional probability distribution $p(e|s)$, with connections to 'X' suffering from an increase in the probability mass of 'Y' (panels R2 and R4), and the competition between words selecting the same ending (rule ii), modulated by the entropy of the conditional probability distribution $p(s|e)$, or $H(s|e)$ (panels R4 and R5). In particular, if we control $H(s)$, i.e. the distribution of paradigms in the input data, the entropy $H(s|e)$ is expressed analytically by the following equation:

$$H(s|e) = H(s) - \sum_{s,e} p(s,e) log \frac{p(s,e)}{p(s)p(e)}. \quad (9)$$

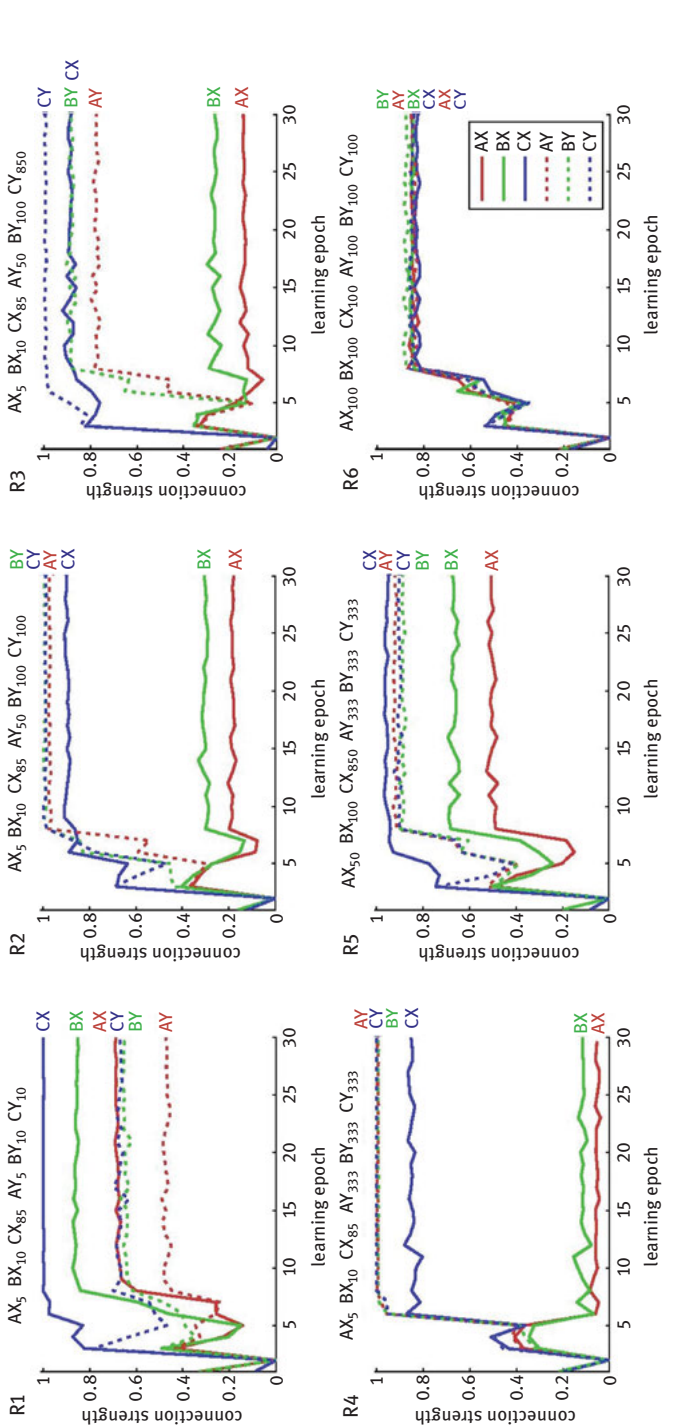

Figure 11: Developmental trends of connection strength at the stem-ending boundary under different training regimes with three mini-paradigms (R1-R6, see Table 2). Weights are plotted against the first 30 learning epochs.

Where $\sum_{s,e} p(s,e) log \frac{p(s,e)}{p(s)p(e)}$ is known as *Mutual Information*, a measure of the mutual dependence between stems and endings, defined as the divergence of the distribution $p(s,e)$ of the verb forms in our training set from the hypothesis that $p(s,e) = p(s)p(e)$, or independence hypothesis (Manning and Schütze 1999). Using the Bayesian equality $p(s,e) = p(s)p(e|s)$, we can rewrite *equation 9* above as follows:

$$H(s|e) = H(s) - \sum_{s} p(s) \sum_{e} p(e|s) log \frac{p(e|s)}{p(e)}, \tag{10}$$

where $\sum_s p(s) \sum_e p(e|s) log \frac{p(e|s)}{p(e)}$ is the Kullback-Leibler divergence between $p(e|s)$ and $p(e)$ in *equation 8* above. *Equation 10* shows that, when $H(s)$ is kept fixed, $H(s|e)$ is maximized by minimizing the average divergence between the intra-paradigmatic distribution $p(e|s)$ of the endings given a stem, and the marginal distribution $p(e)$ of the endings (see Table 3). In other words, verb paradigms are learned more accurately by a TSOM when, on average, the distribution $p(e|s)$ of the forms within each paradigm approximates the marginal distribution of each ending in the corresponding conjugation class (compare R4 and R6). This behavior, accounted for by the interaction of entrenchment and competition/inhibition in discriminative learning, is in line with the facilitation effects reported for visual lexical recognition of inflected words and reading times of derived words (Milin, Filipović Đurđević et al. 2009, Milin, Kuperman et al. 2009; Kuperman et al. 2010). Besides, the evidence is compatible with more extensive experiments on German and Italian verbs (Marzi et al. 2014), showing that, for comparable cumulative frequencies, uniform distributions in training data (R6) facilitate paradigm acquisition (see also Marzi et al. 2020, this volume).

Table 3: Different intra-paradigmatic frequency distributions obtained by keeping marginal distributions fixed. The right-hand distribution is obtained with $p(s,e) = p(s) \cdot p(e)$, to make $D_{KL}(p(e|s)||p(e)) = 0$. For the distribution on the left, $D_{KL}(p(e|s)||p(e)) > 0$.

$p(s,e)$	X	Y	$p(s)$		$p(s,e)$	X	Y	$p(s)$
A	0.04	0.04	0.08		A	0.064	0.016	0.08
B	0.08	0.08	0.16	>	B	0.128	0.032	0.16
C	0.68	0.08	0.76		C	0.608	0.152	0.76
$p(e)$	0.8	0.2	1.00		$p(e)$	0.8	0.2	1.00

Ferro et al. (2018) report comparable results with TSOMs trained on real inflection systems. In two experiments, a TSOM is trained on the same 50 German

verb paradigms by varying their frequency distributions: in the first experiment, forms are presented with a uniform frequency distribution (Kullback-Leibler divergence = 0); in the second experiment, the same set of forms was presented with realistic frequency distributions (Kullback-Leibler divergence > 0). For each map, the number of BMUs recruited for the recognition of inflectional ending was counted. The two experiments were repeated 5 times, and results were averaged across repetitions. As shown in Figure 12, in the training regime with uniform distributions, inflectional endings recruit a larger number of *BMUs* than in the realistic training regime.

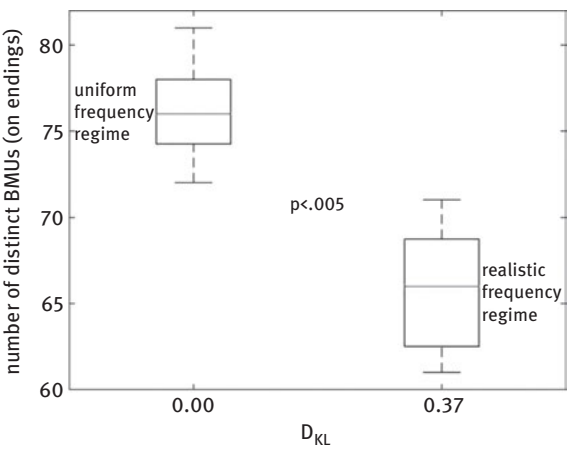

Figure 12: Number of BMUs recruited by inflectional endings, in two experiments where a TSOM is trained with German verb forms with uniform (left) and realistic (right) frequency distributions.

4 Concluding remarks

This chapter provided (i) a selective overview of mathematical and computational approaches to the mental lexicon with a view to prospective unification, (ii) a reappraisal of traditional issues of word storage and processing, and (iii) a novel perspective on these issues from a discriminative learning perspective. In principle, a learning perspective on matters of lexical content, organization and processing crucially can part ways with two alternative views: (i) that lexical representations and word processing strategies are completely predetermined by nature and structure of input data; (ii) that they are completely predetermined by the blueprints of the human word processor. Computer simulations of lexical

modelling tell us that both word representations and processing strategies are shaped up by a complex, dynamic interaction between the innate, domain general principles governing the way humans encode and manage input stimuli, and the structure, paradigmatic organization and frequency distribution of the language input. Different global effects in the operation of a pool of low-level interactive processes are the by-products of domain-specific levels of input representations giving rise to a relatively autonomous organization. Likewise, in a neuro-anatomical perspective, words can be investigated as emergent properties of the functional interaction of different brain areas, each participating in multiple functions (Price 2012, 2017).

We believe that the question of how much of a speaker's internalized word knowledge is determined and accounted for by the informativeness of the language input, and how much is due to the operation of innate principles of serial processing and storage is entirely empirical and, according to our current understanding, not yet amenable to a unifying theory. This is why any strongly dualistic view on lexical matters, sharply separating lexicon from rules, storage from processing, exceptions from regularities, declarative from procedural knowledge strikes us as premature if not unwarranted. A more sensible way to make progress in this area is to focus on some basic cognitive operations and their interaction, and investigate how higher-level language functions and operations emerge from them. From this perspective, learning is not only central to language inquiry as such, but it is also a fundamental key to methodological unification between psycholinguistic and cognitive evidence on the one hand, constraining important aspects of algorithmic modelling, and computer simulations on the other hand. In this chapter, we showed that very simple principles of discriminative learning can go a long way to accounting for complex behavioral evidence. Future work will tell us if these accounts are entirely correct, or should be refined or rejected altogether. Nonetheless, we see no serious alternative to a minimalist, bottom-up approach, whereby innatistic assumptions and ad hoc language principles are introduced as cautiously as possible.

This approach shifts the research focus from a "modular" view of lexical storage, segregated and fundamentally independent from processing, to a radically "integrative view", where storage and processing are in fact two different dynamics of the same underlying process. We provide here a list of some criterial features of such an integrative storage-processing framework (adapted from Marzi and Pirrelli 2015):

- **non-enumerative**: there is no such thing as a finite list of stored items in the human brain; there are many more (potential) pathways in our network of partially overlapping lexical items, than those attested in the input; as a result, the notion of "wordlikeness" (or "lexicality") is a gradient one (a

lexical entry can be perceived as more or less "typical"), and is not co-extensive with the linguistic notion of "listedness" (Di Sciullo and Williams 1987);
- **parallel**: lexical items are activated simultaneously and accessed globally, through resolution of highly distributed, shared sublexical relations;
- **dynamic**: information is never stable; every time a lexical representation is successfully accessed, its content changes accordingly (e.g. through consolidation of connection strengths); moreover, access of any lexical representation affects, more or less deeply, the activation state of all other representations in the same lexicon;
- **processing-dependent**: a lexical representation is fundamentally grounded in processing principles; in fact, it may consist in the same processing units that are fired by the input word associated with the lexical representation;
- **redundant**: lexical representations consist of highly redundant, distributed relations, subsuming both lexical and sublexical structures;
- **emergent/abstractive**: word structure is not a prior, but the perceived by-product of stored, unsegmented input stimuli (full forms or units larger than full forms); perception of structure eventually feeds back on processing;
- **multidimensional**: the lexicon develops structural units defined over many hierarchically arranged levels of representation, ranging from sounds, syllables and morphemes, to words, phrases and sentences; nonetheless, the hypothesis that complex units are processed through a staged sequence of steps going from irreducible primitives to the whole input, is questioned by the highly interactive nature of representation levels, showing pervasive top-down effects on the processing of lower level units;
- **two-way interaction**: lexical representations affect processing, and are crucially affected by processing.

An important cross-linguistic implication of this view is that not all morphologies are processed equally. They do not give rise to homogenous effects of global self-organization. Differences may depend on differences in morphological structure and degrees of predictability (Bompolas et al. 2017; Marzi et al. 2018; and Marzi et al. 2020, this volume). In turn, perception of morphological structure may vary as a function of word length, frequency, perceptual salience, size of lexical neighborhood, distribution of neighborhood members, valence, age of acquisition, embedding context and yet other factors. Computer simulations have so far only scratched the surface of such a multifaceted dynamic interaction. An important emerging trend in the recent literature is that a comparatively small pool of basic, language-independent principles can

account for a number of differential effects that were commonly understood to require different modules and functionally specialized processing routes.

Competition among multiple lexical cues for their discriminative value is key to understanding fundamental aspects of the word learning dynamic (Baayen et al. 2011; Ramscar and Yarlett 2007; Ramscar, Dye, and McCauley 2013; Milin, Feldman et al. 2017). In most discriminative approaches to language learning reviewed here, units defined on one level of representation are understood and modelled to cue units on a different level. For example, forms are cues to either lexical or morpho-syntactic content. Although this is the most intuitive way to conceptualize a cue-outcome relationship in language learning, we saw here that discriminative equations can be used to develop maximally efficient processing structures for symbolic series defined on one representation level only: e.g. sequences of letters/sounds in TSOMs, and lexome-to-lexome discriminative networks for word recognition. One-level, re-entrant discriminative networks prove to be effective in a number of tasks, from prediction-driven processing of upcoming symbols, to context-sensitive filtering of irrelevant units in context. The most efficient way to learn these tasks is to build a maximally discriminative network given the input context. We showed that this straightforward principle can account for complex effects of relative entropy on human processing of verb paradigms.

Finally, in spite of the wide variety of attested self-organizing systems, there seems to be an upper limit on the level of structural complexity they can exhibit, measured as the speaker's uncertainty in making processing predictions about an unknown inflected form in word production (or cell filling problem). Ackerman and Malouf (2013) use Shannon's information entropy to quantify the average conditional entropy of predicting each form in a paradigm on the basis of any other form in the same paradigm, to conjecture that inflectional systems tend to minimize such figure of merit for inflectional complexity. In a discriminative learning framework, Ackerman and Malouf's conjecture can naturally be interpreted in terms of the average degree of predictability of word forms in either recognition or recall. Based on evidence from German and Italian, we showed that processing uncertainty is differently apportioned, depending on the nature of the processing task (Marzi et al. 2016). While irregulars can hardly be predicted when they are unknown because they typically have fewer neighbors than regulars have, irregulars are readily accessed once they are acquired, for exactly the same reason. Thus, existence of irregulars is not dysfunctional, but instrumental to the need to balance processing costs in the two tasks. Similarly, in a typological perspective, non-concatenative morphologies make stems harder to process, due to the variety of their allomorphs, but easier to be completed with their appropriate inflectional endings. Conversely, concatenative morphologies tend to make stems easier to process, but increase processing uncertainty in the selection of the inflectional

ending at the morpheme boundary (Marzi et al. 2018; Ferro et al. 2018; Marzi et al. 2019; Marzi et al. 2020, this volume).

Of late, the advent and exponential growth of neuroimaging technology has allowed in-vivo investigation of the connection between brain data and psychological evidence, establishing a level of material continuity between observations and hypotheses in the domains of neuroscience and cognitive psychology. In the near future, further technological progress will be able to improve the spatial and temporal resolution with which functional regions are located anatomically, to provide novel evidence and constraints on computations and word representations in the brain. Nonetheless, the greatest challenge ahead of us is probably to understand "how" processing takes place in each region and how it interacts with information processed in other regions recruited for the same linguistic task. In this connection, computational and mathematical models of behavioral evidence and functionally related anatomic data have a great potential in bridging the persisting gap between low-level, interactive brain processes and high-level, cognitive models of language knowledge and language behavior. We believe that such integrative, multi-scale, performance-based models of word knowledge will provide an important contribution to a deeper understanding of how language works and is implemented in the brain.

References

Ackerman, Farrell, James P. Blevins & Robert Malouf. 2009. Parts and wholes: implicative patterns in inflectional paradigms. In James P. Blevins & Juliette Blevins (eds.), *Analogy in Grammar*, 54–82. Oxford: Oxford University Press.

Acquaviva, Paolo, Alessandro Lenci, Carita Paradis & Ida Raffaelli. 2020. In Vito Pirrelli, Ingo Plag & Wolfgang U. Dressler (eds.), *Word knowledge and word usage: a cross-disciplinary guide to the mental lexicon*, 354–404. De Gruyter.

Ackerman, Farrell & Robert Malouf. 2013. Morphological organization: The low conditional entropy conjecture. *Language* 89 (3). 429–464.

Albright, Adam. 2002. Islands of reliability for regular morphology: Evidence from Italian. *Language* 78 (4). 684–709.

Alvargonzález, David. 2011. Multidisciplinarity, interdisciplinarity, transdisciplinarity, and the sciences. *International Studies in the Philosophy of Science* 25 (4). 387–403.

Anderson, Stephen R. 1992. *A-Morphous Morphology*.Cambridge (UK): Cambridge University Press.

Arnold, Denis, Fabian Tomaschek, Konstantin Sering, Florence Lopez & R. Harald Baayen. 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. 1–16.

Aronoff, Mark (1976). *Word Formation in Generative Grammar*. Cambridge (Mass.): MIT Press.
Aronoff, Mark (1994). *Morphology by Itself: Stems and Inflectional Classes*. Cambridge (Mass.): MIT Press.
Arppe, Antti, Peter Hendrix, Petar Milin, R. Harald Baayen, Tino Sering & Cyrus Shaoul. 2015. ndl: Naive Discriminative Learning. R package. Retrieved from https://CRAN.R-project.org/package=ndl
Baayen R. Harald (2007). Storage and computation in the mental lexicon. In Gonia Jarema & Gary Libben (eds.), *The Mental Lexicon: Core Perspectives*, 81–104. Elsevier.
Baayen, R. Harald, Laurie B. Feldman & Robert Schreuder 2006. Morphological influences on the recognition of monosyllabic monomorphemic words. *Journal of Memory and Language* 55 (2). 290–313.
Baayen, R. Harald, Petar Milin, Dušica Filipović Đurđević, Peter Hendrix, Marco Marelli. 2011. An amorphous model for morphological processing in visual comprehension based on naive discriminative learning. *Psychological Review* 118 (3). 438–481.
Baayen, R. Harald, Petar Milin & Michael Ramscar. 2016. Frequency in lexical processing. *Aphasiology 30* (11). 1174–1220.
Baayen, R. Harald, Richard Piepenbrock & Leon Gulikers. 1995. The CELEX Lexical Database (CD-ROM). Philadelphia: Linguistic Data Consortium.
Baayen, R. Harald, Tino Sering, Cyrus Shaoul & Petar Milin. 2017. Language comprehension as a multiple label classification problem. In Marco Grzegorczyk & Giacomo Ceoldo (eds.), *Proceedings of the 32nd International Workshop on Statistical Modelling (IWSM)*, 21–34. Groningen, University of Groningen.
Baayen, R. Harald, Cyrus Shaoul, Jon Willits & Michael Ramscar. 2016. Comprehension without segmentation: A proof of concept with naive discriminative learning. *Language, Cognition and Neuroscience* 31 (1). 106–128.
Baayen, R. Harald, Lee H. Wurm & Joanna Aycock. 2007. Lexical dynamics for low-frequency complex words. A regression study across tasks and modalities. *The Mental Lexicon* 2. 419–463.
Baddeley Alan D. (1986). *Working Memory*. Oxford, Oxford University Press.
Balling, Laura W. & Baayen, R. Harald. 2008. Morphological effects in auditory word recognition: Evidence from Danish. *Language and Cognitive Processes* 23 (7–8). 1159–1190.
Balling, Laura W. & Baayen, R. Harald. 2012. Probability and surprisal in auditory comprehension of morphologically complex words. *Cognition* 125(1). 80–106.
Becker, Curtis A. 1980. Semantic context effects in visual word recognition: An analysis of semantic strategies. *Memory and Cognition* 8. 493–512.
Beckner, Clay, Richard Blythe, Joan Bybee, Morten H. Christiansen, William Croft, Nick C. Ellis, John Holland, Jinyun Ke, Diane Larsen-Freeman & Tom Schoenemann. 2009. Language is a complex adaptive system: Position paper. *Language learning* 59. 1–26.
Beard, Robert. 1995. *Lexeme-Morpheme Base Morphology: A General Theory of Inflection and Word Formation*. SUNY Press.
Berg, Thomas. 2020. Morphological slips of the tongue. In Vito Pirrelli, Ingo Plag & Wolfgang U. Dressler (eds.), *Word knowledge and word usage: a cross-disciplinary guide to the mental lexicon*, 635–679. De Gruyter.
Blevins, James P. 2006. Word-based morphology. *Journal of Linguistics* 42. 531–573.
Blevins, James P. 2016. *Word and paradigm morphology*. Oxford: Oxford University Press.

Blevins, James P., Farrell Ackerman, Robert Malouf & Michael Ramscar. 2016. Morphology as an adaptive discriminative system. In Daniel Siddiqi & Heidi Harley (eds.), *Morphological metatheory*, 271–302. John Benjamins.

Blevins, James P., Petar Milin & Michael Ramscar. 2017. The Zipfian paradigm cell filling problem. In Ferenc Kiefer, James P. Blevins & Huba Bartos (eds.), *Perspectives on Morphological Organization: Data and Analyses*, 139–158. Leiden: Brill.

Bloch, Bernard. 1947. English verb inflection. *Language* 23. 399–418.

Bloomfield, Leonard. 1933. *Language*. New York, Henry Holt.

Blough, Donald S. 1975. Steady state data and a quantitative model of operant generalization and discrimination. *Journal of Experimental Psychology: Animal Behavior Processes* 1 (1). 3.

Bompolas Stavros, Marcello Ferro, Claudia Marzi, Franco Alberto Cardillo & Vito Pirrelli. 2017. For a performance-oriented notion of regularity in inflection: the case of Modern Greek conjugation. *Italian Journal of Computational Linguistics* 3 (1). 77–92.

Booij, Geert. 2010. *Construction Morphology*. Oxford University Press.

Bresnan, Joan, Anna Cueni, Tatiana Nikitina, & R. Harald Baayen. 2005. Predicting the dative alternation. *Proceeding of the (KNAW) Academy Colloquium: Cognitive foundations of interpretation*, Amsterdam.

Burzio, Luigi. 1998. Multiple Correspondence, *Lingua* 104. 79–109.

Bybee Joan. 1995. Regular Morphology and the Lexicon. *Language and Cognitive Processes* 10 (5). 425–455.

Bybee, Joan. 2002. Sequentiality as the basis of constituent structure. In Thomas Givón & Bertram F. Malle (eds.), *The Evolution of Language out of Pre-Language*, 107–132. John Benjamins.

Bybee, Joan & Paul J. Hopper (eds.). 2001. *Frequency and the Emergence of Linguistic Structure*. John Benjamin Publishing Company.

Bybee, Joan & James L. McClelland. 2005. Alternatives to the combinatorial paradigm of linguistic theory based on domain general principles of human cognition. *The Linguistic Review* 22 (2–4). 381–410.

Caramazza, Alfonso, Alessandro Laudanna & Cristina Romani. 1988. Lexical access and inflectional morphology. *Cognition*, 28, 297–332.

Cardillo, Franco Alberto, Marcello Ferro, Claudia Marzi & Vito Pirrelli. 2018. Deep learning of inflection and the Cell-filling problem. *Italian Journal of Computational Linguistics* 4(1). 57–75.

Chen, Qi & Daniel Mirman. 2012. Competition and cooperation among similar representations: toward a unified account of facilitative and inhibitory effects of lexical neighbors. *Psychological review* 119 (2). 417–430.

Chomsky, Noam & Morris Halle. 1968. *The sound pattern of English*. New York, Harper and Row.

Clark, Alexander, Chris Fox & Shalom Lapping (eds.). 2010. *The Handbook of Computational Linguistics and Natural Language Processing*. Wiley Blackwell.

Clark, Eve V. 1987. The Principle of Contrast: A Constraint on Language Acquisition. In Brian MacWhinney (ed.), *Mechanisms of Language Acquisition*, 1–33. Lawrence Elbaum.

Clark, Eve V. 1990. On the pragmatics of contrast. *Journal of child language* 17 (2). 417–431.

Corbett, Greville G. & Norman M. Fraser. 1993. Network Morphology: A DATR account of Russian nominal inflection. *Journal of Linguistics* 29. 113–142.

Dahan, Delphine & James S. Magnuson. 2006. Spoken-word recognition. In Matthew J. Traxler& Morton A. Gernsbacher (eds.), *Handbook of Psycholinguistics*, 249–283. Elsevier.

Danks, David. 2003. Equilibria of the Rescorla–Wagner model. *Journal of Mathematical Psychology* 47 (2). 109–121.

Davis, Colin J. 2010. The Spatial Coding Model of Visual Word Identification, *Psychological Review* 117 (3). 713–758.

Di Sciullo, Anna Maria & Edwin Williams. 1987. *On the Definition of Word*. Cambridge: MIT Press.

Divjak, Dagmar & Stephan T. Gries. 2006. Ways of trying in Russian: Clustering behavioral profiles. *Corpus Linguistics and Linguistic Theory* 2 (1). 23–60.

Ellis, Nick C. 2006a. Language acquisition as rational contingency learning. *Applied linguistics 27* (1). 1–24.

Ellis, Nick C. 2006b. Selective attention and transfer phenomena in L2 acquisition: Contingency, cue competition, salience, interference, overshadowing, blocking, and perceptual learning. *Applied linguistics 27* (2). 164–194.

Ellis, Nick C. & Diane Larsen-Freeman. 2006. Language energence: Implications for applied linguistics – Introduction to the special issue. *Applied Linguistics* 27. 558–589.

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

Fábregas, Antonio and Martina Penke. 2020. Word storage and computation. In Vito Pirrelli, Ingo Plag & Wolfgang U. Dressler (eds.), *Word knowledge and word usage: a cross-disciplinary guide to the mental lexicon*, 455–505. De Gruyter.

Ferro, Marcello, Claudia Marzi & Vito Pirrelli. 2011. A Self-Organizing Model of Word Storage and Processing: Implications for Morphology Learning. *Lingue e Linguaggio* X (2). 209–226.

Ferro, Marcello, Claudia Marzi & Vito Pirrelli. 2018. Discriminative word learning is sensitive to inflectional entropy. *Lingue e Linguaggio* XVII (2). 307–327.

Finkel, Raphael & Gregory Stump. 2007. Principal parts and morphological typology. *Morphology* 17. 39–75.

Gaskell, M. Gareth & William D. Marslen-Wilson. 2002. Representation and competition in the perception of spoken words. *Cognitive psychology* 45 (2). 220–266.

Geeraert, Kristina, John Newman & R. Harald Baayen. 2017. Idiom variation: Experimental data and a blueprint of a computational model. In Morten Christiansen & Inbal Arnon (eds.), *More than Words: The Role of Multiword Sequences in Language Learning and Use*. Special issue of *Topics in Cognitive Science* 9 (3). 653–669.

Ghirlanda, Stefano (2005) Retrospective revaluation as simple associative learning. *Journal of Experimental Psychology: Animal Behavior Processes*, 31(1), 107.

Ghirlanda, Stefano, Johan Lind & Magnus Enquist. 2017. Memory for stimulus sequences: a divide between humans and other animals? *Royal Society Open Science* 4. 161011. http://dx.doi.org/10.1098/rsos.161011

Giraudo, Hélène & Jonathan Grainger. 2000. Effects of prime word frequency and cumulative root frequency in masked morphological priming. *Language and Cognitive Processes* 15 (4/5). 421–444.

Goldberg, Adele. 2006. *Constructions at work. the nature of generalization in language*. Oxford University Press.

Goldrick, Matthew, Jocelyn R. Folk & Brenda Rapp. 2010. Mrs. Malaprop's neighborhood: Using word errors to reveal neighborhood structure. *Journal of Memory and Language* 62 (2). 113–134.
Goldrick, Matthew & Brenda Rapp. 2007. Lexical and post-lexical phonological representations in spoken production. *Cognition* 102 (2). 219–260.
Grainger, Jonathan, Pascale Colé & Juan Segui, J. (1991). Masked morphological priming in visual word recognition. *Journal of Memory and Language*, 30, 370–384.
Halle, Morris (1973). Prolegomena to a theory of word formation. *Linguistic Inquiry* 4. 451–464.
Hay, Jennifer B. & R. Harald Baayen. 2005. Shifting Paradigms: Gradient structure in morphology. *Trends in Cognitive Sciences* 9 (7). 342–348.
Hendrix, Peter. 2015. Experimental explorations of a discrimination learning approach to language processing. Doctoral dissertation, University of Tuebingen.
Hendrix, Peter, Patrick Bolger & R. Harald Baayen. 2017. Distinct ERP signatures of word frequency, phrase frequency, and prototypicality in speech production. *Journal of Experimental Psychology: Learning, Memory, and Cognition 43* (1). 128.
Henson, Richard N. 1998. Short-term memory for serial order: The start-end model, *Cognitive Psychology* 36. 73–137.
Hockett, Charles F. 1954. Two models of grammatical description. *Word* 10. 210–231.
Jackendoff, Ray. 1975. Morphological and semantic regularities in the lexicon. *Language* 51. 639–671.
Jackendoff, Ray. 2002. *Foundations of Language. Brain, Meaning, Grammar, Evolution*. Oxford University Press.
James, William. 1890. *Principles of Psychology*. New York: Henry Holt & Company.
James, William. 1907. *Pragmatism: A New Name for Some Old Ways of Thinking*. New York: Longmans, Green, & Co.
James, William. 1909. *The Meaning of Truth: A Sequel to "Pragmatism"*. New York: Longmans, Green, & Co.
Johnson, Keith. 2004. Massive reduction in conversational American English. In *Spontaneous speech: Data and analysis. Proceedings of the 1st session of the 10th international symposium*, 29–54. Tokyo, Japan: The National International Institute for Japanese Language.
Kalman, Rudolph E. 1960. A new approach to linear filtering and prediction problems. *Journal of basic Engineering 82* (1). 35–45.
Karttunen, Lauri. 2003. Computing with realizational morphology. In Alexander Gelbukh (ed.), *Computational Linguistics and Intelligent Text Processing, Proceedings of CICLing 2003*, 203–214. Springer Verlag, Berlin.
Kullback, Solomon. 1987. Letter to the editor: The Kullback-Leibler distance. *The American Statistician* 41 (4), 340–341.
Kuperman, Victor, Raymond Bertram & R. Harald Baayen. 2010. Processing trade-offs in the reading of Dutch derived words. *Journal of Memory and Language* 62(2). 83–97.
Kuperman, Victor, Robert Schreuder, Raymond Bertram & R. Harald Baayen. 2009. Reading of multimorphemic Dutch compounds: towards a multiple route model of lexical processing. *Journal of Experimental Psychology: HPP* 35. 876–895.
Landauer, Thomas K., Peter W. Foltz & Darrell Laham. 1998. An introduction to latent semantic analysis. *Discourse Processes* 25 (2–3). 259–284.

Larsen-Freeman, Diane & Lynne Cameron. 2008. *Complex Systems and Applied Linguistics*. Oxford University Press.
Lensink, S.E., Verhagen, A., Schiller, N. & R. Harald Baayen. 2017. Keeping them apart: on using a discriminative approach to study the nature and processing of multi-word units. Manuscript, University of Leiden.
Levelt, Willem J., Ardi Roelofs & Antje S. Meyer. 1999. A theory of lexical access in speech production. *Behavioral and Brain Sciences* 22. 1–38.
Libben Gary. 2005. Everything is psycholinguistics: Material and methodological considerations in the study of compound processing. *Canadian Journal of Linguistics* 50. 267–283.
Libben, Gary. 2016. The quantum metaphor and the organization of words in the mind. *Journal of Culture Cognitive Science* 1. 49–55.
Lieber, Rochelle 1980. *On the organization of the lexicon*. PhD thesis. Cambridge, MIT.
Linke, Maja, Franziska Bröker, Michael Ramscar & R. Harald Baayen. 2017. Are baboons learning "orthographic" representations? Probably not. *PloS one* 12 (8). e0183876.
Luce, Paul A. 1986. A computational analysis of uniqueness points in auditory word recognition. *Perception & Psychophysics* 39. 155–158.
Luce, Paul A. & David B. Pisoni. 1998. Recognizing spoken words: The neighborhood activation model. *Ear and Hearing* 19. 1–36.
MacKay, Donald G. 1982. The problems of flexibility, fluency, and speed-accuracy trade-off in skilled behavior. *Psychological Review* 89. 483–506.
MacWhinney, Brian (ed.). 1999. *The emergence of language*. Lawrence Erlbaum Associates.
MacWhinney, Brian & William O'Grady. (eds.). 2015. *The Handbook of Language Emergence*. Wiley Blackwell.
Manning, Christopher D. & Hinrich Schütze. 1999. *Foundations of statistical natural language processing*. Cambridge: MIT press.
Marangolo, Paola & Costanza Papagno. 2020. Neuroscientific protocols for exploring the mental lexicon. In Vito Pirrelli, Ingo Plag & Wolfgang U. Dressler (eds.), *Word knowledge and word usage: a cross-disciplinary guide to the mental lexicon*, 127–166. De Gruyter.
Marelli, Marco & Marco Baroni. 2015. Affixation in semantic space: Modeling morpheme meanings with compositional distributional semantics. *Psychological review* 122 (3). 485–515.
Marr, David. 1982. *Vision*. San Francisco: W.H. Freeman.
Marslen-Wilson, William D. 1984. Function and process in spoken word recognition: A tutorial overview. In Herman Bouma & Don G. Bouwhuis (eds.), *Attention and performance X: Control of language processes*, 125–150. Hillsdale: Erlbaum.
Marslen-Wilson, Wiliam D. & Alan Welsh. 1978. Processing interactions and lexical access during word recognition in continuous speech. *Cognitive Psychology* 10. 29–63.
Marsolek, Chad J. 2008. What antipriming reveals about priming. *Trends in Cognitive Sciences* 12 (5). 176–181.
Marzi Claudia, James P. Blevins, Geert Booij & Vito Pirrelli. 2020. Inflection at the morphology-syntax interface. In Vito Pirrelli, Ingo Plag & Wolfgang U. Dressler (eds.), *Word Knowledge and Word Usage: a Cross-disciplinary Guide to the Mental Lexicon*, 228–294. De Gruyter.
Marzi, Claudia, Marcello Ferro, Franco Alberto Cardillo & Vito Pirrelli. 2016. Effects of frequency and regularity in an integrative model of word storage and processing. *Italian Journal of Linguistics* 28 (1). 79–114.

Marzi, Claudia, Marcello Ferro, Ouafae Nahli, Patrizia Belik, Stavros Bompolas & Vito Pirrelli. 2018. Evaluating Inflectional Complexity Crosslinguistically: a Processing Perspective. In *Proceedings of 11th LREC 2018*, Miyazaki, Japan. 3860–3866.

Marzi, Claudia, Marcello Ferro & Vito Pirrelli. 2014. Morphological structure through lexical parsability. *Lingue e Linguaggio* XIII (2). 263–290.

Marzi, Claudia, Marcello Ferro & Vito Pirrelli. 2019. A processing-oriented investigation of inflectional complexity. Frontiers in Communication 4. 48, 1–23. https://doi.org/10.3389/fcomm.2019.00048

Marzi, Claudia & Vito Pirrelli. 2015. A neuro-Computational Approach to Understanding the Mental Lexicon. *Journal of Cognitive Science* 16 (4). 491–533.

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

Matthews Peter H. 1991. *Morphology*. Cambridge, Cambridge University Press.

McClelland, James L. & Elman, Jeffrey L. 1986. The TRACE model of speech perception. *Cognitive Psychology* 18. 1–86.

Milin, Petar, Dagmar Divjak & R. Harald Baayen. 2017. A learning perspective on individual differences in skilled reading: Exploring and exploiting orthographic and semantic discrimination cues. *Journal of Experimental Psychology: Learning, Memory, and Cognition* 43 (11). 1730–1751.

Milin, Petar, Laurie B. Feldman, Michael Ramscar, Peter Hendrix, R. Harald Baayen. 2017. Discrimination in lexical decision. *PloS one* 12 (2). e0171935.

Milin, Petar, Dušica Filipović Đurđević & Fermín Moscoso del Prado Martín. 2009. The simultaneous effects of inflectional paradigms and classes on lexical recognition: Evidence from Serbian. *Journal of Memory and Language* 60 (1). 50–64.

Milin, Petar, Victor Kuperman, Aleksandar Kostić & R. Harald Baayen. 2009. Words and paradigms bit by bit: An information theoretic approach to the processing of paradigmatic structure in inflection and derivation. In James P. Blevins & Juliette Blevins (eds.), *Analogy in grammar: Form and acquisition*, 214–252. Oxford University Press.

Morton, John. 1969. Interaction of information in word recognition. *Psychological review* 76 (2). 165–178.

Morton, John. 1970. A functional model for memory. *Models of human memory*. 203–254.

Morton, John. 1979. Facilitation in word recognition: Experiments causing change in the logogen model. In *Processing of visible language*, Springer, 259–268.

Moscoso del Prado Martín, Fermín, Aleksandar Kostić, & R. Harald Baayen. 2004. Putting the bits together: An information theoretical perspective on morphological processing. *Cognition* 94 (1). 1–18.

Norris, Dennis. 1994. Shortlist: A connectionist model of continuous speech recognition. *Cognition* 52. 189–234.

Norris, Dennis. 2006. The Bayesian reader: explaining word recognition as an optimal Bayesian decision process. *Psychological review* 113 (2). 327.

Norris, Dennis & James M. McQueen. 2008. Shortlist B: a Bayesian model of continuous speech recognition. *Psychological review* 115 (2). 357–395.

Norris, Dennis, James M. McQueen & Anne Cutler. 1995. Competition and segmentation in spoken-word recognition. *Journal of Experimental Psychology: Learning, Memory, and Cognition* 21 (5). 1209–1228.

Osgood, Charles E. 1946. Meaningful similarity and interference in learning. *Journal of Experimental Psychology* 36 (4). 277–301.

Osgood, Charles E. 1949. The similarity paradox in human learning: A resolution. *Psychological Review* 56 (3). 132–143.

Osgood, Charles E. 1966. Meaning cannot be r_m?. *Journal of Verbal Learning and Verbal Behavior* 5 (4). 402–407.

Pham, Hien & R. Harald Baayen. 2015. Vietnamese compounds show an anti-frequency effect in visual lexical decision. *Language, Cognition and Neuroscience* 30. 1077–1095.

Pinker, Steven & Alan Prince. 1988. On language and connectionism: Analysis of a parallel distributed processing model of language acquisition. *Cognition* 29, 195–247.

Pinker, Steven & Alan Prince. 1994. Regular, and Irregular Morphology, and the Psychological Status of Rules of Grammar. In Susan D. Lima, Roberta Corrigan & Gregory K. Iverson (eds.), *The Reality of Linguistic Rules*, 321–351. John Benjamins.

Pinker, Stevan & Michael T. Ullman. 2002. The past and future of the past tense. *Trends in Cognitive Science* 6. 456–463.

Pirrelli, Vito. 2018. Morphological Theory And Computational Linguistics. In Jenny Audring & Francesca Masini (eds.), *The Oxford Handbook of Morphological Theory*, 573–593. Oxford University Press.

Pirrelli, Vito & Marco Battista. 2000. The paradigmatic dimension of stem allomorphy in Italian verb inflection. *Italian Journal of Linguistics* 12. 307–380.

Pirrelli, Vito, Marcello Ferro & Claudia Marzi. 2015. Computational complexity of abstractive morphology. In Matthew Baerman, Dunstan Brown & Greville Corbett (eds.), *Understanding and Measuring Morphological Complexity*, 141–166. Oxford University Press.

Pirrelli, Vito & François Yvon. 1999. The hidden dimension: a paradigmatic view of data-driven NLP. *Journal of Experimental & Theoretical Artificial Intelligence* 11 (3). 391–408.

Plaut, David C. & Laura M. Gonnerman. 2000. Are nonsemantic morphological effects incompatible with a distributed connectionist approach to lexical processing? *Language and Cognitive Processes* 15 (4/5). 445–485.

Poggio, Tomaso. 2010. Afterword. Marr's Vision and Computational Neuroscience. In David Marr, *Vision*, 362–367. The MIT Press.

Poggio, Tomaso. 2012. The levels of understanding framework, revised. *Perception*, 41(9), 1017–1023.

Price, Cathy J. 2012. A review and synthesis of the first 20 years of PET and fMRI studies of heard speech, spoken language and reading. *Neuroimage* 62 (2). 816–847.

Price, Cathy J. 2017. The evolution of cognitive models: From neuropsychology to neuroimaging and back. *Cortex* 107. 37–49.

Quine, Willard V.O. (1960) *Word and object*. Cambridge: MIT Press.

Ramscar, Michael, Melody Dye & Joseph Klein. 2013. Children value informativity over logic in word learning. *Psychological Science* 24 (6). 1017–1023.

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

Ramscar, Michael & Nicole Gitcho. 2007. Developmental change and the nature of learning in childhood. *Trends in cognitive sciences*, 11 (7). 274–279.

Ramscar, Michael, Peter Hendrix, Cyrus Shaoul, Petar Milin & R. Harald Baayen. 2014. The myth of cognitive decline: Non-linear dynamics of lifelong learning. *Topics in cognitive science* 6 (1). 5–42.

Ramscar, Michael, Ching C. Sun, Peter Hendrix & R. Harald Baayen. 2017. The Mismeasurement of Mind: Lifespan Changes in Paired Associate Test Scores Reflect The 'Cost' of Learning, Not Cognitive Decline. *Psychological Science*. 1171–1179.

Ramscar, Michael & Daniel Yarlett. 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, Michael, Daniel Yarlett, Melody Dye, Katie Denny & Kirsten Thorpe. 2010. The Effects of Feature-Label-Order and their Implications for Symbolic Learning. *Cognitive Science* 34 (6). 909–957.

Rescorla, Robert A. 1988. Behavioral Studies of Pavlovian conditioning. *Annual review of neuroscience* 11 (1). 329–352.

Rescorla, Robert A. & Allan R. Wagner. 1972. A theory of Pavlovian conditioning: Variations in the effectiveness of reinforcement and nonreinforcement. *Classical conditioning II: Current research and theory* 2. 64–99.

Rubenstein, Herbert, Lonnie Garfield & Jane A. Millikan. 1970. Homographic entries in the internal lexicon. *Journal of verbal learning and verbal behavior* 9 (5). 487–494.

Rubenstein, Herbert, Spafford S. Lewis & Mollie A. Rubenstein. 1971. Homographic entries in the internal lexicon: Effects of systematicity and relative frequency of meanings. *Journal of Memory and Language* 10 (1). 57.

Rueckl, Jay G. & Michal Raveh. 1999. The influence of morphological regularities on the dynamics of a connectionist network. *Brain and Language* 68. 110–117.

Rumelhart David & James McClelland. 1986. On learning the past tense of English verbs. In David Rumelhart, James McClelland & PDP Research Group, *Parallel distributed processing: Explanations in the microstructure of cognition*, vol. I. 216–271. The MIT Press.

Rumelhart David, James McClelland & PDP Research Group. 1986. *Parallel distributed processing: Explanations in the microstructure of cognition*. Voll. 1 & 2. The MIT Press.

Scalise, Sergio. 1984. *Generative Morphology*. Dordrecht: Foris.

Schmidtke, Daniel, Kazunaga Matsuki & Victor Kuperman. 2017. Surviving blind decomposition: a distributional analysis of the time-course of complex word recognition. *Journal of experimental psychology. Learning, memory, and cognition* 43(11). 1793–1820.

Schreuder, Robert & R. Harald Baayen. 1995. Modeling morphological processing. In Laurie B. Feldman (ed.), *Morphological aspects of language processing*, 131–56. Hillsdale, NJ: Erlbaum.

Schultz, Wolfram. 1998. Predictive reward signal of dopamine neurons. *Journal of neurophysiology* 80 (1). 1–27.

Seidenberg, Mark S. & Maryellen C. MacDonald. 1999. A probabilistic constraints approach to language acquisition and processing. *Cognitive science* 23(4). 569–588.

Selkirk, Elisabeth. 1984. *Phonology and Syntax*. The MIT Press.

Selfridge, Oliver G. 1959. Pandemonium: A paradigm for learning. In D. V. Blake & A. M. Uttley (eds.), *Proceedings of the Symposium on Mechanisation of Thought Processes*, 511–529.

Sering, Konstantin, Petar Milin & R. Harald Baayen. 2018. Language comprehension as a multiple label classification problem. *Statistica Neerlandica* 72 (3). 339–353.

Shannon, Claude. 1948. A mathematical theory of communication. *The Bell System Technical Journal* 27. 379–423, 623–656.

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

Skinner, Burrhus F. 1953. *Science and Human Behavior*. Simon and Schuster.
Skinner, Burrhus F. 1957. *Verbal Behavior*. Copley Publishing Group.
Snodgrass, Joan G. & Robert Jarvella. 1972. Some linguistic determinants of word classification times. *Pyschonomic Science* 27 (4). 220–222.
Spivey, Michael J, Ken McRae & Marc F. Joanisse. 2012. *The Cambridge Handbook of Psycholinguistics*. Cambridge University Press.
Taft, Marcus. 1994. Interactive-activation as a framework for understanding morphological processing. *Language and cognitive processes* 9 (3). 271–294.
Taft, Marcus. 2004. Morphological decomposition and the reverse base frequency effect. *Quarterly Journal of Experimental Psychology Section A* 57 (4). 745–765.
Taft, Marcus & Kenneth I. Forster. 1975. Lexical storage and retrieval of prefixed words. *Journal of Verbal Learning and Verbal Behavior* 14. 638–647.
Tiedemann, Jörg. 2012. *Parallel Data, Tools and Interfaces in OPUS. Proceedings of the Eighth International Conference on Language Resources and Evaluation* (LREC). 2214–2218. Istanbul, Turkey.
Tolman, Edvard C. 1932. *Purposive behavior in animals and men*. New York: Century.
Tolman, Edvard C. 1951. *Behavior and psychological man: essays in motivation and learning*. Berkeley: University of California Press.
Tremblay, Antoine & R. Harald Baayen. 2010. Holistic processing of regular four-word sequences: A behavioral and erp study of the effects of structure, frequency, and probability on immediate free recall. In David Wood (ed.), *Perspectives on formulaic language: Acquisition and communication*, 151–173. Bloomsbury Publishing.
Tucker, Benjamin V., Michelle Sims & R. Harald Baayen. 2017. *Opposing forces on acoustic duration*. Manuscript, University of Alberta and University of Tübingen.
Vulchanova, Mila, David Saldaña and Giosué Baggio. 2020. Word structure and word processing in developmental disorders. In Vito Pirrelli, Ingo Plag & Wolfgang U. Dressler (eds.), *Word knowledge and word usage: a cross-disciplinary guide to the mental lexicon*, 680–707. De Gruyter.
Weitz, Marc, Konstantin Sering, David-Elias Künstle & Lennard Schneider. 2017. pyndl: Naive Discriminative Learning. Python3 package. Retrieved from https://github.com/quantling/pyndl.
Widrow, Bernard & Marcian E-Hoff. 1960. *Adaptive switching circuits* (No. TR-1553-1). Stanford University, CA Stanford Electronics labs.
Wittgenstein, Ludwig. 1953. *Philosophical Investigations*. Gertrude E. M. Anscombe (*Trans*.), Blackwell Publishing.