Abstraction, storage and naive discriminative learning

Harald Baayen and Michael Ramscar

keywords: rules, schemata, exemplars, analogy, hybrid models, discrimination learning

1 Introduction

The English sentence *you want milk* can be uttered in a variety of circumstances, such as a mother about to feed her baby (answer: *bweeeh*), a father asking a toddler whether she would like a glass of milk (answer: *yes please*), or an air hostess serving black tea in economy class (answer: *sure*). Furthermore, similar sentences (*you want coffee, you want water, would you like coffee, would you like a cup of coffee*) are also produced and understood in a wide variety of contexts. What are the cognitive principles that allow us to produce and understand many different sentences across an even greater kaleidoscope of contexts and situations?

In this chapter, we discuss three very different approaches that seek to answer this fundamental question about how language works. We begin with the oldest one, the structuralist tradition and its formalist offshoots, which posits that rules obtained by a process of *abstraction* are essential to understanding language. The second approach argues that generalizations are achieved not through abstraction, but by analogical reasoning over large numbers of instances of language use stored in memory. The third approach takes the perspective that to understand language and productivity in language, it is essential to take into account well-established basic principles of discrimination learning.

2 Abstraction

In traditional abstractionist approaches to language, it is assumed that the contexts in which a question such as *you want milk* is uttered are so varied, that the properties characterizing these contexts must be powerless as predictors of a given utterance. What the child has to learn is to abstract away from all the irrelevant contextual information, and identify a level of elemental representations that capture abstract commonalities in instances of useage. The common core of all utterances of *you want milk* is thus identified as roughly an abstract tri-partite knowledge structure comprising the phonological elements ([(ju)$_w$(wɒnt)$_w$(mɪlk)$_s$]), a syntactic structure comprising the elements ([$_{\text{NP}}$ you [$_{\text{VP}}$ want [$_{\text{NP}}$ milk]]]), and a semantic structure comprising the elements DESIRE(YOU, MILK). Rules link the volitional agent element in the semantic structure to the subject element of the syntactic structure, and yet other rules spell out the pronoun element *you* as a strings of phonemic elements [ju]. Typically, the knowledge base is kept as lean as possible, by storing in memory only the most elementary units (phonemes, morphemes, semantic primitives) and the rules

for combining these units into well-formed sequences. Thus, the semantic structure DESIRE(YOU, MILK) would not be available in memory as such. Instead, a more abstract structure, DESIRE(X, Y) would be stored in memory, where X is a symbolic placeholder for any volitional agent able or imagined to be able to have desires, and Y any object, person, state, or event that is desired, or can be imagined to be desirable.

Furthermore, in order to cut down on memory requirements, and to make relations between words and utterances as transparent as possible, the formalism of inheritence hierarchies as developed in the context of object-oriented programming languages has been found useful (see, e.g., Steels and De Beule, 2006, for fluid construction grammar). Thus, instead of having to store different kinds of milk (*cow milk, goat's milk,sheep milk, mother milk, camel milk, coffee milk, coconut milk, ...*) and all their properties in separate lexical entries, one can set up one entry for the most typical kind of milk (e.g., the cow milk as bought in the supermarket),

$$\text{MILK} \quad \left[\begin{array}{ll} \text{type:} & \text{thing} \\ \text{properties:} & \text{concrete, inanimate, imageable, fluid, } \ldots \\ \text{function:} & \text{to be consumed by drinking} \\ \text{color:} & \text{white} \\ \text{source:} & \text{cows} \end{array}\right],$$

and keep the entries for the other kinds of milk lean by having them inherit all specifications from the entry for milk except where specified otherwise:

$$\text{CAMEL MILK}: \quad \text{MILK} \quad \left[\ \text{source:} \quad \text{female camels}\ \right].$$

When a mother offers milk to her child, while uttering *you want milk*, the semantic structure of the utterance might be characterized by lexical conceptual structures (Jackendoff, 1990) such as

$$\text{OFFER(MOTHER, CHILD, MILK)}$$
$$\text{ASK(MOTHER, CHILD, IS-TRUE(DESIRE(CHILD, MILK)))}\ .$$

However, these structures are themselves the outcome of the application of more abstract semantic structures

$$\text{OFFER(X, Y, Z)}$$
$$\text{ASK(X, Y, IS-TRUE((DESIRE, Y, Z)))}$$

which also cover utterances such as *you want to play* and *you want to sleep*.

Several proposals have been made as to how such abstract structures (and the elements that they combine) might be identified or acquired. One class of theories holds that the language learner is genetically endowed with a set of abstract rules, constraints or primitives. This innate knowledge of an underlying universal abstract grammar would then relieve the learner of having to figure out the basic principles of human grammars, since these basics can be assumed to be already given. The learner's task is then reduced to solving the simpler problems such as figuring out the proper word order in English for three-argument verbs, given the innate knowledge that verbs can have three arguments, that word order can be fixed, etc. However, innate rules and constraints by themselves have no explanatory value, and half a century of research has not lead to any solid and generally accepted results confirming that the basic principles of formal (computer) languages as developed in the second half of the twentieth century are part of the human race's genetic endowment.

It should be noted, however, that in constraint-based approaches (see, e.g., Dressler, 1985; Prince and Smolensky, 2008), constraints can be argued to have functional motivations (see, e.g.

Boersma, 1998; Boersma and Hayes, 2001). In phonology, for instance, voiceless realizations might be dispreferred due to voiced segments, as voiced segments require more articulatory effort, and hence more energy, than voiceless segments. In syntax, constraints might also be functionally grounded. For the dative alternation, for instance, a functional rationale motivating the observed preferences for particular constituent orders would be to provide a consistent and predictable flow of information, with given referents preceding non-given referents, pronouns preceding non-pronouns, definites preceding indefinites, and shorter constituents preceeding longer constituents (Bresnan et al., 2007). However, even for constraints with reasonably plausible functional motivations, it is unclear how these constraints are learned. The problem here is that what is a hard constraint in one language, can be a soft constraint in another, and not a constraint at all in yet a third language. Sceptics of functional explanations will argue that functionally motivated constraints are unhelpful because it is not clear under what circumstances they are more, or less, in force.

Would it be possible to induce rules without invoking innate principles or supposed functional constraints? The minimum generalization learning algorithm proposed by Albright and Hayes (2003) seeks to do exactly this in the domain of morphology. This algorithm gradually learns more abstract rules by iteratively comparing pairs of forms. Each comparison identifies what a pair of forms have in common, and wherever possible creates a more abstract rule on the basis of shared features. For instance, transposed to syntax, the minimum generalization learning algorithm would, given the utterances *you want milk* and *you want juice*, derive the structure

OFFER(MOTHER, CHILD, Z)
ASK(MOTHER, CHILD, IS-TRUE(DESIRE(CHILD, Z)))

$$\text{Z} \begin{bmatrix} \text{type:} & \text{thing} \\ \text{properties:} & \text{concrete, inanimate, imageable, fluid, } \ldots \\ \text{function:} & \text{to be consumed by drinking} \end{bmatrix}$$

by deletion of the feature-value pairs [source:cow] and [source:fruit] in the respective semantic structures of the individual sentences. For the pair of utterances *you want to play* and *you want to eat*, the shared abstract structure would be

OFFER(MOTHER, CHILD, Z)
ASK(MOTHER, CHILD, IS-TRUE(DESIRE(CHILD, Z)))

$$\text{Z} \begin{bmatrix} \text{type:} & \text{event} \\ \text{properties:} & \text{volitional agent, social activity, } \ldots \\ \text{agent:} & \text{the child} \end{bmatrix}.$$

When in turn these structures are compared for further abstraction, all that remains is

OFFER(MOTHER, CHILD, Z)
ASK(MOTHER, CHILD, IS-TRUE(DESIRE(CHILD, Z)))

which in turn, when the utterances are used with different interlocutors, will undergo further abstraction to

OFFER(X, Y, Z)
ASK(X, Y, IS-TRUE((DESIRE, Y, Z))).

A salient property of abstractionist theories is that although the rules and constructions are deduced from a systematic and comprehensive scan of all pairs of utterances, the utterances themselves are discarded once the rules and constructions have been properly inferred. From the perspective of language processing, however, this raises several questions. First, if utterances are required for

rule deduction, and hence have to be available in memory, why would they be discarded once the rules have been discovered?

Second, rule deduction requires a comprehensive set of utterances, but in real life, utterances become available one by one over time. Do we have to assume that at some point in late childhood, rule deduction is completed, the language has been learned, and that therefore the traces of past experience with the language can be erased from memory? Such a fundamental discontinuity in the learning process seems at odds with recent evidence that language learning is a process that continues throughout one's lifetime (see, e.g., Ramscar et al., 2014, 2013d). Third, the number of utterances to be stored in memory for rule deduction may be prohibitively large. Corpus surveys show that in English there are hundreds of millions of sequences of just four words. Some studies have reported frequency effects for sequences of words (Bannard and Matthews, 2008; Arnon and Snider, 2010; Tremblay and Baayen, 2010). These frequency effects have been argued to support the existence of representations of multi-word sequences in the mental lexicon (or mental constructicon). However, as pointed out by Shaoul et al. (2013), knowledge about word sequences appears to be restricted to sequences no longer than four, perhaps five, words. It is therefore unlikely that syntactic rules, especially those for complex sentences with main and subordinate clauses, could arise by a process of abstraction from a large set of stored full sentences, as the current evidence suggests that the brain doesn't retain memory traces of long complex sentences but only of short sequences of words.

The main strength of abstractionist approaches — thanks to the presupposition that at its heart language is best understood as a formal calculus — is that these approaches have at their disposal all the technology developed over many decades in computer science. It is worth noting that, in fact, most computationally implemented theories of different aspects of linguistic cognition, whatever the very different schools of thought they come from, make use of abstractionist decompositional frameworks. Although the lexical conceptual structures of (Jackendoff, 1990) and Lieber (2004) look very different from the schemata of Langacker (1987) and Dabrowska (2004a), differences concern what aspects of human experience are found worthy of being formalized and how they should be formalized, whereas both approaches share the conviction that abstraction is at the heart of the language engine.

This can also be seen in the treatment of conceptual blending (for details on blending, see the chapter on blending) by Veale et al. (2000). Consider the production of metaphorical expressions suchs as *elephants were the tanks of Hannibal's army*. Veale et al. (2000) propose a computationally implemented model that generates such conceptual blends from knowledge structures for elephants, tanks, classical and modern warfare, and Hannibal, in conjunction with an abstract rule that searches for n-tuples of knowledge structures in one domain (e.g., Roman warfare) that match, on the basis of their features, n-tuples of knowledge structures in another domain (e.g., modern warfare). Given matching features (such as elephants being the strongest and most dangerous units in ancient warfare, and tanks being the strongest and most dangerous units in modern warfare), the algorithm can blend *elephants were the strongest units of Hannibal's army* with *tanks are the strongest units of a modern army* to create *elephants were the tanks of Hannibal's army*. To do so, the algorithm abstracts away from concrete examples, and searches for correspondences across knowledge domains.

The tools of computer science provide the language engineer with valuable control over how a given computational operationalization will function. A further advantage is that, in principle, computational implementations can be evaluated precisely against empirical data. However, this technology also has its share of disadvantages. First, both representations and rules typically require extensive labor-intensive hand-crafting.

Second, and more importantly, language is fundamentally contextual. A sentence such as *She cut her finger with a knife* typically suggests that the finger was not completely severed from the hand,

phrase	freq.	p	prep.	freq.	q
with the onion	8,867	0.305	*with*	2,171,020	0.074
in the onion	7,058	0.243	*in*	10,212,008	0.347
to the onion	5,734	0.197	*to*	4,148,449	0.141
from the onion	2,213	0.076	*from*	2,150,946	0.073
on the onion	1,922	0.066	*on*	4,010,429	0.136
into the onion	1,337	0.046	*into*	1,296,889	0.044
up the onion	1,091	0.038	*up*	403,114	0.014
...	...	...	...	...	...
over the onion	826	0.028	*over*	269,847	0.009
total	29,048	1.000		29,401,403	1.000

Table 1: Frequencies and relative frequencies of English prepositional phrases (left: specific to *onion*, right: summed across all nouns) that enter into the calculation of prepositional relative entropy.

whereas the sentence *the lumberjacks cut trees for a living* typically means that trees were cut down and severed from their roots. The interpretation of the verb in *Outlines of animals were cut out of paper* is different yet again. Here, the verb indicates creation by means of cutting. Interestingly, the context in which a word such as *cut* is used generates expectations that arise surprisingly early in the comprehension processing record (see, e.g., Elman, 2009, for a review), much earlier than one would expect given theories that assume an initial stage of purely form-based processing. Within the abstractionist enterprise, one can of course distinguish between different senses of *cut* (WordNet distinguishes 41), each with its own semantic structure, with sufficiently narrowly defined features to make a sense fit only in very specific contexts (see also the chapter on polysemy). But the problem here is that the expectations that readers form about how to understand *cut* depend on subjects such as *she*, *lumberjacks*, and *outlines of animals*. While one might consider specifying in the lexical representation for *lumberjack* that this is a person whose profession it is to cut down trees, it stretches belief that *outlines of animals* (a lexical entry used by Google as a caption for images of outlines of animals[1]) would have an entry in the mental lexicon specifying that these are cutable.

A further challenge for traditional abstractionist theories comes from paradigmatic effects in language processing. The paradigmatic dimension of language is difficult to capture in abstractionist frameworks. Consider prepositional phrases in English, such as *with the onion, over the onion, in the onion, ...* When abstraction is taken as the basis of generalization, then a structure such as [$_{\text{PP}}$P [$_{\text{NP}}$ the [$_{\text{N}}$ N]] captures crucial aspects of the abstract knowledge of prepositional phrases, in conjunction with the set of prepositions and the set of nouns in the lexicon. All prior experiences with actual prepositional phrases (*with the onion, over the onion, in the onion, ...*) are lost from memory. The abstractionist grammar thus reduces a rich slice of experience to a prepositional symbol, freely replaceable without reference to context by a single instance from the set of prepositions, followed by a definite determiner, in turn is followed by a noun symbol that is again selected without reference to context, from the set of nouns.

However, native speakers of English know, albeit implicitly, much more about how prepositions are actually used in English. Speakers of English know, without being aware of this at conscious levels of reflection, that some prepositions are quite atypical for *onion*, whereas other prepositions are rather popular with *onion*. To see this, consider Table 1. The second column of this table lists the frequencies with which prepositions occur with the noun *onion* in the British National

[1] As of October 20, 2014.

Corpus (Burnard, 1995). The third column lists the corresponding relative frequencies (or sample probabilities), obtained by dividing the counts in the second column by the total of the counts in that column. The fifth column lists the counts of occurrences of these prepositions with any noun, and the final column lists the corresponding probabilities. What this table shows is that *with the onion* is used much more frequently with *onion* compared to nouns in general, with relative frequencies of 0.305 verses 0.074 respectively. By contrast, *on the onion* is used somewhat less frequently than *on* followed by an arbitrary noun. Does this matter? To judge from both behavioral (Baayen et al., 2011) and electrophysiological (Hendrix and Baayen, 2014) evidence, it does. Nouns that make use of prepositions in a way that is very different from how an average noun uses its prepositions, show very different processing profiles. A measure capturing how well the use of prepositions by a specific noun corresponds to how prepositions are used in general is the Kulback-Leibler divergence, also known as *relative entropy*:

$$\text{relative entropy}(p, q) = \sum_{i=1} (p_i * \log_2 (p_i/q_i)), \tag{1}$$

where p and q refer to the probability distributions in columns 3 and 6 of Table 1. It turns out that when the relative entropy for a noun is large, i.e., when the noun makes atypical use of prepositions, response latencies to the noun, even when presented in isolation in the visual lexical decision task, are longer. Furthermore, in speech production, as gauged by a picture naming paradigm, relative entropy is an effective statistical predictor of the brain's electrophysiological response (Hendrix and Baayen, 2014). Crucially, the effect of relative entropy arises irrespective of whether nouns are presented in isolation, or whether nouns are presented in the context of a particular preposition. What matters is how different a noun's use of prepositions is from prototypical prepositional use in English. This paradigmatic effect poses a fundamental challenge to abstractionist theories, precisely because an abstract representation of "the" prepositional phrase has been crafted to have amnesia about how a noun is actually used.

3 Analogy

In traditional grammar, analogy is used to denote an incidental similarity-based extension of a pattern that is not supported by a general rule. In more recent theories, however, analogy is seen as a much more foundational process of which rules are a special, typically more productive, case (see, e.g., Langacker, 1987; Pothos, 2005).

In morphology, Matthews (1974) and Blevins (2003) developed a framework, known as Word and Paradigm Morphology, in which words, rather than morphemes and exponents, are the basic units in the lexicon. Proportional analogy (*hand: hands = tree: trees*) is posited as driving production and comprehension of novel forms. Explicit algorithms for capturing the core idea of analogy-driven prediction have been developed within the context of a class of computational approaches commonly referred to as exemplar models.

Exemplar models start off with the assumption of extensive storage in memory of instances of language use, typically referred to as *exemplars*. Instead of seeking to account for the productivity of language through abstract rules operating over hand-tailored representations, exemplar models base their predictions about novel forms on the exemplars in memory, in combination with a general, domain a-specific similarity-driven algorithm. One of the earliest linguistic exemplar models was developed by Skousen (1989), who grounded his approach, analogical modeling of language (AML), in probability theory (Skousen, 2002, 2000). Skousen's algorithm searches for sets of exemplars with characteristics that consistently support a particular outcome, where an outcome can be a

construction, a phonetic feature (such as voicing alternation Ernestus and Baayen, 2003), or the choice between rival affixes (Arndt-Lappe, 2014). Given the resulting subset of consistent exemplars, the *analogical set*, the different outcomes are ranked by the number of exemplars supporting these outcomes. The best-supported, highest-ranked outcome is selected as the most likely outcome.

Skousen's AML model is computationally expensive for data with many features. The *memory based learning* (MBL) framework of Daelemans and Van den Bosch (2005) sidesteps this computational problem. Just as AML, it searches for the set of nearest neighbors, and selects as its choice the outcome with the best support in the nearest neighbor set. In the very simplest set-up, the nearest neighbors are those instances in memory that share most features with a given case for which the appropriate outcome class has to be determined. This simplest set-up is not very useful, however, because in the presence of many irrelevant predictors, classification accuracy can plummet. By weighting features for their relevance for a given choice problem, accuracy can be improved dramatically while keeping computational costs down. By way of example, consider the choice of the plural allomorph in English, which is [iz] following sibilants, [s] following voiceless consonants, and [z] elsewhere. Knowledge of a word's final consonant nearly eliminates uncertainty about the appropriate allomorph, whereas knowledge of the initial consonant of the word is completely uninformative. Since manner of articulation and voicing of the final consonant are informative features, they can be assigned large weights, whereas manner and voicing for initial consonants can be assigned low weights. The values of these weights can be estimated straightforwardly from the data, for instance, by considering to what extent knowledge of the value of a feature reduces one's uncertainty about the class outcome. The extent to which uncertainty is reduced then becomes the weight for the importance of that feature.

An important observation coming from the literature on memory based learning is that forgetting is harmful (Daelemans et al., 1999). The larger the set of exemplars is, the better MBL is able to approximate human performance. The message here is exactly the opposite of that of abstractionist models, which seek to keep the knowledge base as lean as possible. However, differences are not so large as they would seem. As pointed out by Keuleers (2008), minimum generalization learning and memory based learning (under certain parameter configurations) are mathematically nearly indistinguishable. But whereas minimum generalization learning first deduces rules, then forgets about exemplars, and uses rules at run-time (greedy learning), memory-based learning simply stores exemplars, and runs its similarity-based algorithm at runtime (lazy learning).

Another similarity between MGL and MBL is that a new model is required for each individual problem set within a domain of inquiry. For instance, when modeling phonological form, one model will handle past tenses, another model the choice between the allomorphy of nominalizations in *-ion*, and yet a third model the allomorphy of the plural suffix. Thus, both approaches work with different rules (or schemas) for different phenomena, and differ only as to how these rules/schemas are implemented under the hood.

Exemplar models such as AML and MBL offer several advantages. First, because analogical rules are executed at run-time, new exemplars in the instance base will automatically lead to updated prediction performance. In MGL, by contrast, once the rule system has been deduced, it remains fixed and cannot be updated for principled reasons. (Technically, of course, the rules can be recalculated for an updated set of exemplars, but doing so implies that the exemplars are held in reserve, and are not erased from memory.)

Another important advantage of AML and MBL is that getting these algorithms to work for a given data set requires very little hand-crafting. These algorithms discover themselves which features are important.

Of course, these models also have disadvantages. First, compared to handcrafted abstractionist systems developed over many years and fine-tuned to all kinds of exceptions, AML and MBL may

show lack of precision. Second, it remains to be seen how plausible the assumption is of storing each and any exemplar in memory. Above, the hundreds of millions of four-word sequences were already mentioned that would have to be stored in an English mental lexicon. For languages with highly productive inflectional systems, millions of forms are in use just at the word level. Furthermore, the rampant variability in the speech signal makes it highly unlikely that each pronunciation variant of every word ever heard would be stored in memory.

4 Hybrid models

Hybrid models hold that schemata (or rules) and exemplars exist side by side. For instance, Langacker (2010) argues for a hybrid approach when he states that "structure emerges from usage, is immanent in usage, and is influenced by usage on an ongoing basis". The co-existence of rules and exemplars (see also Langacker, 1987; Dabrowska, 2004b) implies a system with redundancy, such that, for instance, in comprehension, an interpretation can be arrived at either by retrieving the appropriate holistic exemplar, or by application of a rule or schema to the relevant exemplars of smaller units. For morphological processing, Baayen et al. (1997) similarly argued for the existence of whole-word representations for complex words, side by side with a parsing mechanism operating on the morphemic constituents of these words.

The redundancy offered by hybrid models is generally taken to make the processing system more robust. For instance, when one processing route fails to complete, another processing route may still be effective. In horse race models, which make the assumption that processing routes run independently and in parallel, statistical facilitation can take place. If processing time is determined by the first route to win the race, and if the distributions of the completion times of the different routes overlap, then, across many trials, the average processing time of the combined routes will be shorter than the average processing time of the fastest route by itself (Baayen et al., 1997).

However, hybrid models also comes with several disadvantages. From exemplar models, hybrid models inherit the problem of a high-entropy exemplar space. It might be argued that not *all* exemplars are stored, but only large numbers of exemplars. However, this raises the question of under what circumstances exemplars are, or are not, stored. Positing a frequency threshold for storage runs into logical difficulties, because any new exemplar will start with an initial frequency of 1, far below the threshold, and hence will never be stored.

From abstractionist models, hybrid models inherit the problem of selecting the correct analysis from the multitude of possible analyses (Bod, 1998, 2006; Baayen and Schreuder, 2000). When schemata are assumed to be in operation at multiple levels of abstraction, how does the system know which level of abstraction is the appropriate one? How is competition between more concrete and more abstract schemata resolved?

5 Discrimination

We agree with Langacker (2010) that usage shapes the grammar on an ongoing basis. But we believe that in order to justice to the insights driving abstractionist approaches, exemplar models, and hybrid models, while avoiding their weak points, it is essential to turn to learning theory.

Modern learning theory begins with Ivan Pavlov and his famous observations about bells and dog-food. Pavlov first noticed that his dogs salivated in the presence of the technician who usually fed them. He then devised an experiment in which he rang a bell before he presented the dogs with food. After a few repetitions, the dogs started to salivate in response to the bell, anticipating the food they expected to see (Pavlov, 1927). Pavlov's initial results led to a straightforward theory of

learning that seems obvious and feels intuitively right: If a cue is present, and an outcome follows, an animal notices the co-occurrence and subsequently learns to associate the two.

It turns out, however, that this simple associative view of learning provides a one-sided and misleading perspective on the actual learning process and its consequences. For example, a dog trained to expect food when a bell is rung, can later be given training in which a light is flashed simultaneously with the bell. After repeated exposure to bell and light, followed by food, only a light is flashed. Will the dog drool? Surprisingly, the answer is no: the dog doesn't drool. Even though the light consistently co-occurred with the food in training, the dog does not learn to associate it with the food, a phenomenon known as blocking.

The problem for, e.g., memory-based learning is that it would pick out the light as an informative cue for food. After all, whenever the light is present, food is present. Since there is no uncertainty about the food given the light, the model predicts that the light should be an excellent cue, and that this cue should build strong expectations for food, contrary to fact.

The learning equations that Rescorla developed together with Wagner (Wagner and Rescorla, 1972), however, perfectly capture this finding. The reason that the light never becomes an effective cue for food is that the bell is already a perfectly predictive cue for the food. Because there are no situations in which the light predicts food but the bell does not, the light does not add any new information: it is not predictive of the food over and above the bell. As this and many similar experiments have revealed, associative learning is sensitive to the *informativity* of co-occurrences, rather than their mere existence.

The learning theory of Rescorla (1988) not only predicts a substantial body of findings in the animal literature, but has recently also been found to predict aspects of first language acquisition as well as implicit learning in adults (see, e.g., Ramscar and Yarlett, 2007; Ramscar and Gitcho, 2007; Ramscar et al., 2010, 2013c). This learning theory specifies how the association weights from the cues in the environment (such as a bell and a flashing light in the case of Pavlov's dog) to an outcome (e.g., food) should be modified over time. The basic insights are, first, that if a cue is not present, association weights from that cue to outcomes are left untouched. For instance, whiskers are visual cues to various animals, such as cats, rabbits, rats, and mice. If there are no whiskers to be seen, then the weights on the links between whiskers and cats, rabbits, rats, and mice, are left unchanged, even though these animals might be present (as when they are observed from the back). When whiskers are seen, and a cat is present but no rabbits, rats, or mice, then the weight from whiskers to cat is increased. At the same time, the weights from whiskers to rabbits, rats, and mice are decreased, even though these animals have whiskers. This is a crucial element of Rescorla's theory, that sets it apart from its associationist, even behaviorist, predecessors. Learning is sensitive not only to associations forming when cues and outcomes co-occur. Learning is also sensitive to the success and failure of the implicit predictions that prior experiences relating cues to outcomes generate. Whiskers do not only predict cats, but also rabbits and other rodents. When these predictions turn out to be false, the weights that connect whiskers to the animals that were mispredicted to be present will be tuned down. As a result of this, outcomes (cats, rabbits, mice, rats) *compete* for the cues, while at the same time, cues compete for outcomes.

Baayen et al. (2011) used the Rescorla-Wagner equations to build a computational model for the reading of words, as gauged by the visual lexical decision task. The basic structure of the model is very simple, and is exemplified by Figure 1. The bottom layer of the network has nodes representing letter pairs (digraphs). The top layer of the network specifies lexemes, in the sense of Aronoff (1994), that is, as lexical nodes that are the symbols linking up to rich form information (such as letter digraphs) on the one hand, and rich world knowledge (not shown in Figure 1) on the other hand. Thus, lexemes are symbolic focal points mediating between linguistic form and experience of the world, that themselves are neither form nor meaning.

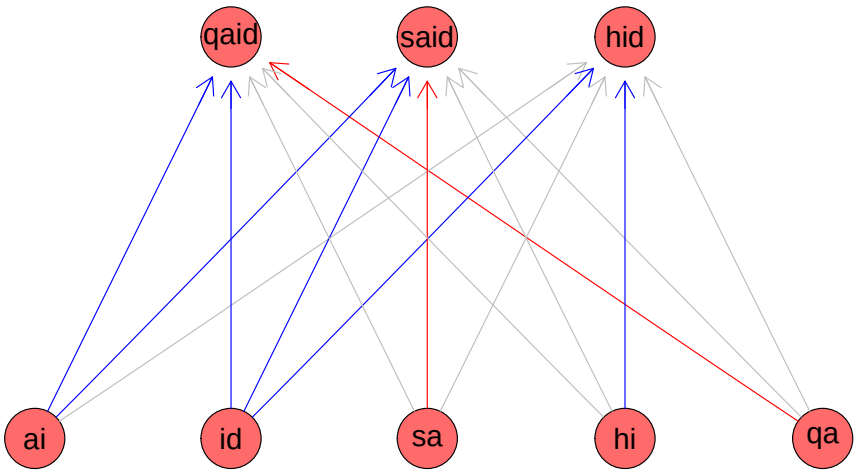


Figure 1: A Rescorla-Wagner network with five digraphs as cues, and three lexemes as outcomes.

Of course, this raises the question how the elements of form (n-graphs, n-phones) and the elements of experience (the lexemes) themselves are learned. Here, we assume that these units are simply available to the learner. Any computational implementation has to work with units that are primitives to that implementation, but that themselves have arisen as the outcome of other classification processes. One kind of learning process that might give rise to these units is unsupervised category induction (see, e.g., Love et al., 2004, for a computational implementation, and also the chapter on categorization).

The first word, the legal scrabble word *qaid* ('tribal chieftain'), has one letter pair, *qa*, that uniquely distinguishes it from the two other lexemes. The Rescorla-Wagner equations predict that this cue is strongly associated with *qaid*, and negatively associated with *said* and *hid*. Conversely, the letter pair *id* occurs in all three words. Hence it is not very useful for discriminating between the three lexemes. As a consequence, the weights on its connections are all small. The total support that cues in the input provide for a lexeme, its *activation*, is obtained by summation over the weights on the connections from these cues (for *qaid*, the cues *qa*, *ai*, and *id*) to the outcome (the lexeme of *qaid*). This activation represents the learnability of the lexemes given the cues.

The naive discriminative learner model of Baayen et al. (2011) takes this simple network architecture and applies it rigorously to word triplets in the British National Corpus. For each word triplet, all the letter digraphs in the three words were collected. These served as cues. From the same words, all "content" lexemes and "grammatical" lexemes (number, tense, person, etc.) were collected and served as outcomes. The Rescorla-Wagner equations were then used to adjust the weights from the digraph cues to the lexeme outcomes.[2] For any given word in the corpus, its activation was obtained by summing the weights from its orthographic cues to its lexemes. For words with multiple lexemes, such as a plural or a compound, the activations of its lexemes were summed. It turns out that these activation weights are excellent predictors of lexical decision laten-

[2]In the actual implementation, a mathematical shortcut, due to Danks (2003), was used for estimating the weights.

cies: words with longer responses are the words with lower activations, i.e., the words that cannot be learned that well given their orthographic properties. The activation weights turn out to mirror a wide range of effects reported in the experimental literature, such as the word frequency effect, orthographic neighborhood effects, morphological family size effects, constituent frequency effects, and paradigmatic entropy effects (including the abovementioned prepositional relative entropy effect). What is especially interesting is that the model covers the full range of morphological effects, without having any representations for words, morphemes, exponents, or allomorphs.

In this approach, the morphology and syntax is implicit in the distribution of cues and outcomes, which jointly shape a network that is continuously updated with usage. Since morphology and syntax are implicit in the usage, we refer to this approach as *implicit morphology* and *implicit grammar*. Interestingly, this approach to language dovetails well with the mathematical theory of communication developed by Shannon (1948).

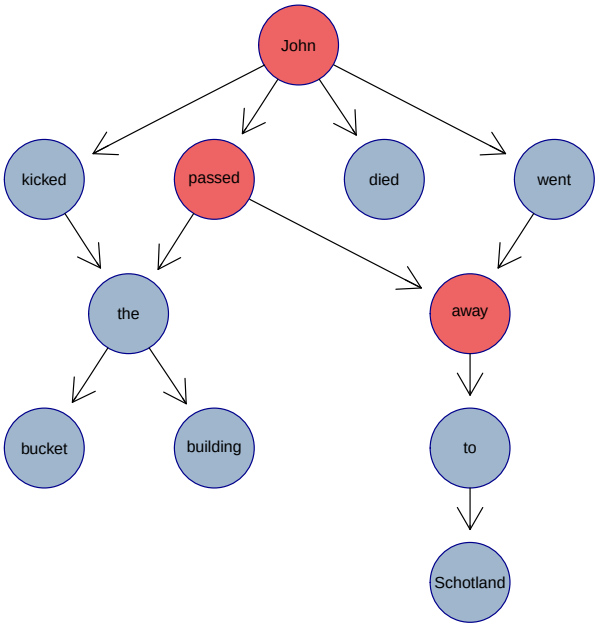
When a photograph is sent over a cable from a camera to a laptop, it is not the case that the objects in the photograph (say a rose on a table, next to which is a chair), are sent down the wire one by one (first the chair, and than the rose plus table). To the contrary, the picture is transformed into a binary stream that is optimized for the transmission channel as well as protected against data loss by error-correcting code. The laptop is able to reconstruct the picture, not by applying a grammar to reconstruct the picture from the signal, but by making use of the same coding scheme that the camera used, to select the appropriate distribution of pixel colors over the canvas from the possible distributions of pixel colors that coding schemes allow for.

To make this more concrete, consider a coding scheme devised to transmit for experiences: the experience of a fountain, the experience of a fountain pen, the experience of an orange, and the experience of orange juice. Assume a code, shared by encoder and decoder, specifying that the four experiences are signalled by the digit strings 00, 01, 10, and 11 respectively. When seeking to communicate the experience of a fountain pen, the speaker will encode 01, and thanks to the shared code, the listener will decode 01 into the experience of a fountain pen. There is no need whatsoever to consider whether the individual ones and zeroes compositionally contribute to the experiences transmitted.

Thus, we can view language-as-form (ink on paper, pixels on a computer screen, the speech signal, gestures) as a signal that serves to discriminate between complex experiences of the world. The success of the signal hinges on the interlocutors sharing the code for encoding and decoding the signal (Wieling et al., 2014). The same code that allows the speaker to discriminate between past experiences in memory and encode a discriminated experience in the language signal, is then used by the listener to discriminate between her past experiences. Discrimination is important here, as speakers will seldom share the same experiences. Consider, for example, a speaker mentioning a larch tree. The interlocutor may not know what exactly a larch tree is, because she never realized the differences between larches, spruces, and pine trees. Nevertheless, the communicative event may be relatively successful in the sense that the listener was able to reduce the set of potential past experiences to her experiences of trees. She might request further clarification of what a larch tree is, or, not having any interest in biology, she might just be satisfied that some (to her irrelevant) subspecies of trees is at issue. Thus implicit grammar views the language signal as separating encoded relevant experiences from the larger set of a listener's irrelevant experiences.

Thus far, we have discussed comprehension. What about speech production? The model we are developing (see also Baayen and Blevins, 2014), proposes a two-layered knowledge structure, consisting of a directed graph specifying the order between production outcomes on the one hand, and of Recorla-Wagner networks associated with the vertices in the network on the other hand. Figure 2 presents such a knowledge structure for the sentences *John passed away, John kicked the bucket, John died*, *John passed the building*, and *John went away to Scotland*. The left panel presents

the directed graph specifying the potential paths defined by these sentences, and the right panel summarizes the connection strengths between lexemic cues (rows) and word outcomes (columns) in tabular form. These connection strengths are obtained with the Rescorla-Wagner equations (for detailed discussion of these equations, see Ramscar et al. (2010) and Baayen et al. (2011)) applied to all sentences containing John.



	died	kicked	passed	went
BUILDING	-0.273	0.045	0.455	-0.091
CONSIDERATE	-0.364	-0.273	0.773	0.045
DIE	0.455	0.091	-0.091	-0.182
GO	-0.273	0.045	-0.045	0.409
JOHN	-0.091	0.182	0.318	0.136
NEUTRAL	0.636	-0.273	-0.227	0.045
PASS	-0.273	0.045	0.455	-0.091
RUDE	-0.364	0.727	-0.227	0.045
SCHOTLAN	-0.273	0.045	-0.045	0.409

Figure 2: An example of a directed word graph, with the path for *John passed away* highlighted, and the Rescorla-Wagner control network at the node *John*.

All sentences in this simple example begin with `John`, hence this is the top node. Given `John`, the possible continuations are `kicked, passed, died`, and `went`. When the speaker has the intention of communicating in a considerate way that John died (indicated by the lexemes JOHN, DIE, CONSIDERATE, highlighted in the table of weights), then the word `passed` has a total activation of 1 (the sum of the highlighted weights in the `passed` column), whereas the other continuations have activations of zero. Thus, sentences emerge as paths through the directed graph, where each choice where to go next is governed by the accumulated knowledge discriminating between the different options, guided by past experience of which lexemes predict which word outcomes.

Knowledge structures such as illustrated in Figure 2 can be formulated for sequences of words, but also for sequences of diphones or demi-syllables. It is currently an open question whether separate structures above and below the word are really necessary. What is important is that the digraphs provide a very economical storage format. In a word graph, any word form is represented by a single vertex. In a diphone graph, any diphone is present only once. This is a large step away from standard conceptions of the mental lexicon informed by the dictionary metaphor, in which a letter or diphone pair is represented many times, at least once for each entry. The directed graph also sidesteps the problem of having to assume distinct exemplars for sequences of demi-syllables or sequences of words. In the present example, for instance, an opaque idiom (*kick the bucket*), a semi-transparent idiom (*to pass away*), and a literal expression (*die*) are represented economically

with dynamical control from the Rescorla-Wagner networks.

From the discriminative perspective, questions as to how opaque and semi-transparent idioms are "stored" in the mental dictionary, decomposed, or not decomposed, simply do not arise because words are now part of a signal for which traditional questions of compositionality are simply not relevant. Thus, in implicit grammar, rules, schemata, constructions, inheritence hierarchies, multiple entries of homonyms in dictionary lists, and all other constructs based on formal grammars are unnecessary. These constructs may provide high-level descriptions of aspects of language which may be insightful for the analyst reflecting on language, but in the discriminative approach, they have no cognitive reality.

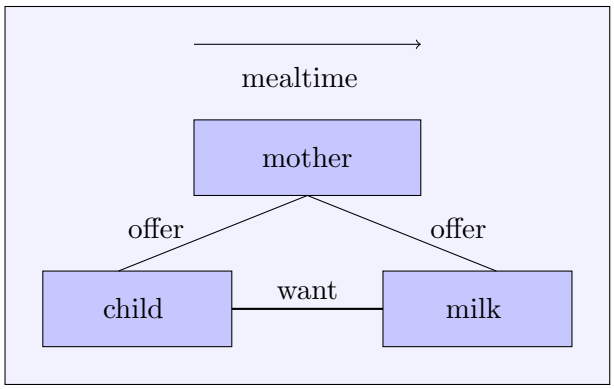
The knowledge structures of implicit grammar do not permit redundancy, in the sense that different sets of representations, and different rules for achieving the same result, would co-exist. The theory acknowledges that the linguistic signal is rich, and that the experiences we encode in the signal are richer by many orders of magnitude. But redundancy in the sense of having multiple ways in which to achieve exactly the same goal is ruled out. The directed graph and the Rescorla-Wagner networks define one unique most-probable path for the expression of a given message.

Research on child language acquisition (e.g., Bannard and Matthews, 2008; Tomasello and Tomasello, 2009) has shown that children are conservative learners who stay very close to known exemplars, and initially do not use constructions productively. One explanation holds that initially, children work with large unanalyzed holistic chunks, which they learn, over time, to break down into smaller chunks, with as end product the abstract schemata of the adult speaker (Dabrowska, 2004b; Dabrowska and Lieven, 2005; Borensztajn et al., 2009; Beekhuizen et al., 2014). Implicit grammar offers a very different — and currently still speculative — perspective on the acquisition process.

Consider a child inquiring about what activity her interlocutor is engaged in. Typically, an English-speaking child in North America or the U.K. will have ample experience with such questions, which often arise in the context of reading a picture book ("What's the bear doing? It's eating honey!"). However, with very little command over her vocal apparatus, in the initial stage of speech production, the full message (a question about the event an actor is engaged in) has to be expressed by the child in a single word, e.g., "Mommy?". However, single-word expressions will often not be effective, as "Mommy?" could also be short-hand for what adults would express as "Mommy, where are you?" or "Mommy, I'm hungry". From a learning perspective, the word uttered (*Mommy*), and the lexemes in the message (QUESTION, EVENT, MOMMY) constitute the cues in a learning event with the success of the communicative event as outcome. Over the course of learning during the one-word stage, the lexemes QUESTION, EVENT, AGENT will acquire low or even negative weights to communicative success. Only *Mommy* will acquire substantial positive weights, thanks to the single-word utterances being successful for attracting attention.

By the end of the one-word stage, the child has a production graph with only vertices and no edges. Once the child succeeds in uttering sentences with more than one word (*What's Mommy doing*), thanks to increasing motor control over the articulators, the chances of successful communication rise dramatically. This will prompt the reuse of multi-word sequences, and the construction of edges between the vertices in the graph, together with the Rescorla-Wagner networks that discriminate between where to go next in the graph given the child's communicative intentions. The first path in the graph will be re-used often, consolidating both the edges between the vertices in the directed graph, as well as the associated Rescorla-Wagner control networks, which, in terms of what the child actually produces, will enable the child to demonstrate increasing fluency with multiword productions.

In this approach to learning, the empirical phenomenon of children proceeding in their production from a prefab such as "What's Mommy doing?" to utterances of the form "What's X V-ing",



OFFER(MOTHER, CHILD, Z)
ASK(MOTHER, CHILD, IS-TRUE(CHILD(DESIRE, Z)))

$$Z \begin{bmatrix} \text{type:} & \text{thing} \\ \text{properties:} & \text{concrete, inanimate, imageable, fluid, } \ldots \\ \text{function:} & \text{to be consumed by drinking} \end{bmatrix}$$

Figure 3: Semantic representations in the style of cognitive grammar (after Dabrowska (2004, page 221) and Jackendoff's lexical conceptual structures.

analysed in cognitive grammar as *schematization*, in implicit grammar does not involve any abstraction. What is at stake, instead, is learning to think for speaking (Slobin, 1996). During the one-word stage, children gradually learn that many aspects of the experiences they want to express cannot be packed into a single word. Once they have accumulated enough articulatory experience to launch word sequences, they can develop their production graph and the associated control networks. As this graph is expanded, syntactic productivity, which is already nascent in small worlds such as shown in Figure 2, will increase exponentially.

It is worth noting that the process of chunking in acquisition, with the child as a miniature linguist trying to find units at varies hierarchical levels in the speech signal, is also is at odds with the ACT-R theory of cognition, according to which chunking evolves in the opposite direction, starting with the small chunks that are all that can be handled initially, and that only with experience over time can be aggregated into the greater chunks representing the automatization of cognitive skills (Anderson, 2007).

Theoretical frameworks have developed different notational schemes for describing the semantics of utterences such as *you want milk*, as illustrated in Figure 3 for cognitive grammar (top) and lexical conceptual structures in the style of Jackendoff (1990). From the perspective of implicit grammar, the knowledge summarized in such representations is valueable and insightful, but too dependent on a multitude of interpretational conventions to be immediately implementable in a discriminative learning model. What needs to be done is to unpack such descriptions into a set of basic descriptors that can function as lexemes in comprehension and production models. For instance, OFFER(MOTHER, CHILD, MILK) has to be unpacked into lexemes not only for offer, mother, child, and milk, but also for the mother as the initiator of the offering, the milk as the thing offered, etc. In other words, the insights expressed by the different frameworks can and should be made available to the learning algorithms in the form of lexemic units. How exactly these units conspire within the memory system defined by the directed graph and its control networks is determined by

how they are used in the language community and the learning algorithms of the brain.

Implicit grammar is a new computational theory which is still under development. We have illustrated that this theory makes it possible to reflect on language and cognition from a very different perspective. Computational simulations for comprehension indicate that the model scales up to corpora with many billions of words. For speech production, simulations of the production of complex words promise low error rates (Baayen and Blevins, 2014), but whether the same holds for sentence and discourse production remains to be shown.

Implicit grammar is a theory that grounds language in discrimination learning. There is, of course, much more to language and cognition than implicit discriminative learning. For discussion of the role of higher-order cognitive processes in resolving processing conflicts and integrating implicit learning with speakers' goals, and also the importance of the late development of these higher-order processes, see Ramscar and Gitcho (2007); Ramscar et al. (2013a,b). A further complication is that with the advent of the cultural technology of writing, literate speakers bring extensive meta-linguistic skills into the arena of language use and language processing. How exactly the many multimodal experiences of language use at both implicit and conscious levels shape how a given speaker processes language is a serious computational challenge for future research, not only for implicit grammar, but also for abstractionist and exemplar approaches, as well as hybrid models such as cognitive grammar.

6 Concluding remarks

When comparing different algorithms, it is important to keep in mind, irrespective of whether they come from abstractionist, exemplar-based, or discriminative theories, that they tend to perform with similar precision. For instance, Ernestus and Baayen (2003) compared AML, stochastic optimality theory, and two classifiers from the statistical literature, among others, and observed very similar performance. Keuleers (2008) showed equivalent performance for memory-based learning and minimum generalization learning for past-tense formation in English. Baayen et al. (2013) compared two statistical techniques with naive discrimination learning, and again observed similar performance. This state of affairs indicates that the typical data sets that have fuelled debates over rules, schemas, and analogy, tend to have a quantitative structure that can be well-approximated from very different theoretical perspectives. Therefore, the value of different approaches to language, language use, and language processes will have to be evaluated by means of the simplicity of computational implementations, the neuro-biological support for these implementations, and the extent to which the models generate concrete, falsifiable predictions regarding unseen data.

References

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

Anderson, J. R. (2007). *How can the human mind occur in the physical universe?* Oxford University Press.

Arndt-Lappe, S. (2014). Analogy in suffix rivalry: the case of English ity and ness. *English Language and Linguistics*, in press.

Arnon, I. and Snider, N. (2010). More than words: Frequency effects for multi-word phrases. *Journal of Memory and Language*, 62(1):67–82.

Aronoff, M. (1994). *Morphology by itself: Stems and inflectional classes*. The MIT Press, Cambridge, Mass.

Baayen, R. H. and Blevins, J. P. (2014). Implicit morphology. *Manuscript, University of Tübingen*.

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

Baayen, R. H., Janda, L. A., Nesset, T., Endresen, A., and Makarova, A. (2013). Making choices in Russian: Pros and cons of statistical methods for rival forms. *Russian Linguistics*, 37:253–291.

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

Baayen, R. H. and Schreuder, R. (2000). Towards a psycholinguistic computational model for morphological parsing. *Philosophical Transactions of the Royal Society (Series A: Mathematical, Physical and Engineering Sciences)*, 358:1–13.

Bannard, C. and Matthews, D. (2008). Stored word sequences in language learning: The effect of familiarity on children's repetition of four-word combinations. *Psychological Science*, 19:241–248.

Beekhuizen, B., Bod, R., Fazly, A., Stevenson, S., and Verhagen, A. (2014). A usage-based model of early grammatical development. In *Proceedings of the 2014 ACL workshop on cognitive modeling and computational linguistics*, 46–54. Association for computational linguistics.

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

Bod, R. (1998). *Beyond Grammar: An Experience-based Theory of Language*. CSLI publications, Stanford, CA.

Bod, R. (2006). Exemplar-based syntax: How to get productivity from examples. *The Linguistic Review*, 23(3):291–320.

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

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

Borensztajn, G., Zuidema, W., and Bod, R. (2009). Children's grammars grow more abstract with age — evidence from an automatic procedure for identifying the productive units of language. *Topics in Cognitive Science*, 1(1):175–188.

Bresnan, J., Cueni, A., Nikitina, T., and Baayen, R. H. (2007). Predicting the dative alternation. In Bouma, G., Kraemer, I., and Zwarts, J., editors, *Cognitive Foundations of Interpretation*, 69–94, Amsterdam. Royal Netherlands Academy of Arts and Sciences.

Burnard, L. (1995). *Users guide for the British National Corpus*. British National Corpus consortium, Oxford university computing service.

Dabrowska, E. (2004a). *Language, mind and brain. Some psychological and neurological constraints on theories of grammar*. Edinburgh University Press, Edinburgh.

Dabrowska, E. (2004b). Rules or schemas? Evidence from Polish. *Language and Cognitive Processes*, 19:225–271.

Dabrowska, E. and Lieven, E. (2005). Towards a lexically specific grammar of children's question constructions. *Cognitive Linguistics*, 16(3):437–474.

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

Daelemans, W., Van den Bosch, A., and Zavrel, J. (1999). Forgetting exceptions is harmful in language learning. *Machine learning, special issue on natural language learning*, 34:11–41.

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

Dressler, W. (1985). On the predictiveness of natural morphology. *Journal of Linguistics*, 21:321–337.

Elman, J. (2009). On the meaning of words and dinosaur bones: Lexical knowledge without a lexicon. *Cognitive Science*, 33:1–36.

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

Hendrix, P. and Baayen, R. (2014). Distinct erp signatures of word frequency, phrase frequency, and prototypicality in speech production. To appear in *Journal of Experimental Psychology: Learning, Memory and Cognition*.

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

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

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

Langacker, R. W. (2010). How not to disagree: The emergence of structure from usage. *Language usage and language structure*, 213:107.

Lieber, R. (2004). *Morphology and Lexical Semantics*. Cambridge University Press.

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

Matthews, P. H. (1974). *Morphology. An introduction to the theory of word structure*. Cambridge University Press, London.

Pavlov, I. P. (1927). *Conditioned reflexes: An investigation of the physiological activity of the cerebral cortex (translated by G. V. Anrep)*. Oxford University Press, London.

Pothos, E. M. (2005). The rules versus similarity distinction. *Behavioral and Brain Sciences*, 28(01):1–14.

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

Ramscar, M., Dye, M., Gustafson, J., and Klein, J. (2013a). Dual routes to cognitive flexibility: Learning and response conflict resolution in the dimensional change card sort task. *Child Development*, doi 10.1111/cdev.12044.

Ramscar, M., Dye, M., and Klein, J. (2013b). Children value informativity over logic in word learning. *Psychological science*, 24(6):1017–1023.

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

Ramscar, M. and Gitcho, N. (2007). Developmental change and the nature of learning in childhood. *Trends In Cognitive Science*, 11(7):274–279.

Ramscar, M., Hendrix, P., Love, B., and Baayen, R. (2013d). Learning is not decline: The mental lexicon as a window into cognition across the lifespan. *The Mental Lexicon*, 8:450–481.

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

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

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

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

Shannon, C. (1948). A mathematical theory of communication. *Bell System Technical Journal*, 27:379–423.

Shaoul, C., Westbury, C. F., and Baayen, H. R. (2013). The subjective frequency of word n-grams. *Psihologija*, 46(4):497–537.

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

Skousen, R. (2000). Analogical modeling and quantum computing. Los Alamos National Laboratory <http://arXiv.org>.

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

Slobin, D. (1996). From 'thought to language' to 'thinking for speaking'. In Gumperz, J. and Levinson, S., editors, *Rethinking linguistic relativity*, 70–96. CUP, Cambridge.

Steels, L. and De Beule, J. (2006). A (very) brief introduction to fluid construction grammar. In *Proceedings of the Third Workshop on Scalable Natural Language Understanding*, 73–80. Association for Computational Linguistics.

Tomasello, M. (2009). *Constructing a language: A usage-based theory of language acquisition*. Harvard University Press.

Tremblay, A. and Baayen, R. H. (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 Wood, D., editor, *Perspectives on Formulaic Language: Acquisition and communication*, 151–173. The Continuum International Publishing Group, London.

Veale, T., O Donoghue, D., and Keane, M. T. (2000). Computation and blending. *Cognitive Linguistics*, 11(3/4):253–282.

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

Wieling, M., Nerbonne, J., Bloem, J., Gooskens, C., Heeringa, W., and Baayen, R. H. (2014). A cognitively grounded measure of pronunciation distance. *PloS One*, 9(1):e75734.

R. Harald Baayen and Michael Ramscar, Eberhard Karls University, Tübingen, Germany