Journal of Child Language (2021), **48**, 984–1022
doi:10.1017/S0305000921000544

CAMBRIDGE UNIVERSITY PRESS

ARTICLE

How children learn to communicate discriminatively

Michael RAMSCAR

Department of Linguistics, University of Tübingen, Germany
Address for correspondence: Michael Ramscar, Department of Linguistics, University of Tübingen, Germany. E-mail: michael.ramscar@uni-tuebingen.de

(Received 29 January 2020; revised 7 February 2021; accepted 16 June 2021)

Abstract

How do children learn to communicate, and what do they learn? Traditionally, most theories have taken an associative, compositional approach to these questions, supposing children acquire an inventory of form-meaning associations, and procedures for composing / decomposing them; into / from messages in production and comprehension. This paper presents an alternative account of human communication and its acquisition based on the systematic, discriminative approach embodied in psychological and computational models of learning, and formally described by communication theory. It describes how discriminative learning theory offers an alternative perspective on the way that systems of semantic cues are conditioned onto communicative codes, while information theory provides a very different view of the nature of the codes themselves. It shows how the distributional properties of languages satisfy the communicative requirements described in information theory, enabling language learners to align their expectations despite the vastly different levels of experience among language users, and to master communication systems far more abstract than linguistic intuitions traditionally assume. Topics reviewed include morphological development, the acquisition of verb argument structures, and the functions of linguistic systems that have proven to be stumbling blocks for compositional theories: grammatical gender and personal names.

Keywords: language learning; usage based model; information theory; human communication

What do children learn when they learn to communicate?

Most theoretical accounts of human communication assume that languages comprise an inventory of elements and some procedures for combining them into messages. These elements are conceived of at various levels of description: PHONEMES, the acoustic/psychological equivalent of letters; MORPHEMES, basic units of meaning that cannot be further decomposed (such that the word *uni-corn-s* comprises three morphemes); WORDS, which can be either mono- or multi-morphemic; and SENTENCES which comprise more complex combinations of words / morphemes. It is further

assumed that the way meanings combine in these more complex structures follows the principle of compositionality, which holds that "the meaning of an expression is a function of the meanings of its parts and the way they are syntactically combined" (Partee, 1984).

Yet although human communication is almost universally assumed to be compositional, attempts to cash out this assumption have inevitably proven unsuccessful (Ramscar & Port, 2016). Phonemes fail to capture many of the acoustic properties that are essential to spoken communication, and although phonemes are assumed to 'spell out' speech signals, they are often impossible to identify from acoustic information alone (Port & Leary, 2005; Samuel, 2020). The idea that morphemes comprise basic sound-meaning units has been undermined by analyses of their functions (which are often meaningless), and the discovery of context effects that contradict the idea of them being elemental units of meaning (Blevins, 2016; Lieber, 2019). Meanwhile, centuries of dedicated research has failed to make theoretical sense of meaning units (with philosophers such as Wittgenstein, 1953, and Quine, 1960, concluding that meanings cannot possibly be atomic), while efforts to find psychological evidence for their existence have produced more questions than answers (Ramscar & Port, 2015). Since similar problems have arisen when theories have attempted to explain how smaller 'elements of meaning' are combined to form larger compositional elements (Culicover, 1999), it seems that viewed dispassionately, the results of the massive body of research seeking to explain language processing in terms of the composition and decomposition of basic elements of form and meaning do not so much support this view as indicate that human communication does not work that way at all.

This paper describes an alternative theory of language that begins with learning, a subject that ought to lie at the heart of any theory of language acquisition, and ends with an account of human communication that is in many ways the opposite of received wisdom. It describes how research on 'associative' learning has resulted in theories that are DISCRIMINATIVE in processing terms, and explains how this further undermines the idea that languages comprise inventories of discrete form-meaning associations, or that human communication is a compositional process. Instead, because the best models of learning and communication are fundamentally systematic, fundamentally probabilistic, and fundamentally discriminative (such that the processes that underlie the use of language are also likely to be systematic, probabilistic, and discriminative), this account reframes human communication and its learning in these terms – so as to make it compatible with these processes. It also results in a number of concrete predictions about the kinds of problems that children will encounter in learning a language, and the kinds of properties we should expect human communicative codes to possess for children to learn them. In keeping with the spirit of this special issue, these predictions will be highlighted as this review seeks to explain why (theoretically) communication and children's learning to communicate are best characterized in discriminative terms.

How do children learn?

Accounts of communicative learning seek to answer two questions: WHAT do children learn, and HOW do they learn it? Since starting from HOW they learn allows for the establishment of useful constraints on WHAT children might be able to learn, it seems reasonable to begin by offering an account of learning. Beginning with learning also

https://doi.org/10.1017/S0305000921000544 Published online by Cambridge University Press

benefits from the fact that considerable agreement exists about the nature of many basic learning processes, which tend to be specified in greater detail than is the case in linguistics. There are two main reasons for this: first, humans share their basic learning mechanisms with other animals, allowing animal models to offer insight into the neural processes of learning (O'Doherty, Dayan, Friston, Critchley & Dolan, 2003; Schultz, 2006); second, human and animal learning mechanisms appear to be error-driven, a learning method that has been subject to a large amount of computational research that provides considerable insight into the capabilities of – and constraints on – this kind of learning in real-world situations (Hinton, McClelland & Rumelhart, 1986; Hinton, Deng, Yu, Dahl, Mohamed, Jaitly, Vanhoucke, Nguyen, Kingsbury & Sainath, 2012; Hannun, Case, Casper, Catanzaro, Diamos, Elsen, Prenger, Satheesh, Sengupta, Coates & Ng, 2014; LeCun, Bengio & Hinton, 2015).

With regards to animal learning, it has long been known that simple association rates (the frequency at which a 'stimulus' – say the German article *das* – is associated with a 'response,' the noun *Mädchen*) are incapable of explaining basic conditioning (Ramscar, Dye & McCauley, 2013a). Two further factors are critical to predicting and explaining learning the predictive relationships between cues and events: cue BACKGROUND RATES (Rescorla, 1968; Ramscar, Dye & Klein, 2013b; how often *das* occurs absent *Mädchen*), and BLOCKING (Kamin, 1969; Arnon & Ramscar, 2012; the prior predictability of *Mädchen* in context). Learning is then a product of the interactions between these factors in experience, with association rates tending to promote learning, and blocking and background rates tending to inhibit it.

Formal models embodying these principles are adept at fitting and predicting learning effects. A classic example, the Rescorla and Wagner (1972) learning rule specifies how learning can be described in terms of the computation of discrepancies between a learner's expectations and reality, with the difference between the two serving to modify the values of a set of predictive cues in relation to a set of expected outcomes in trial-by-trial learning. Although it was originally proposed as part of an elemental model of learning couched in associative terms (see also Miller, Barnet & Grahame, 1995; Siegel & Allan, 1996; Ellis, 2006), the error-driven learning mechanism described by the rule is best understood by re-conceptualizing learning as a discriminative process that reduces uncertainty about events in the world, such that learning only occurs when uncertainty is present (Ramscar, Yarlett, Dye, Denny & Thorpe, 2010; Hoppe, Hendriks, Ramscar & van Rij, in press). If an event (outcome) whose likelihood is underestimated occurs, the values of cues to it are strengthened, whereas if its likelihood is overestimated, the values of these cues are weakened. Because uncertainty is finite (learning can often result in certainty), cues compete for predictive value, leading to the discovery of reliable cues through competition, and the discriminatory weakening and elimination of others.

The Rescorla-Wagner learning rule is relatively simple, allowing processes such as error-driven learning and cue competition to be explained in relatively intuitive terms. However the algorithm it describes is simply the linear form of an earlier rule proposed by Widrow and Hoff (1960; see Stone, 1986), that is in turn formally equivalent to the delta-rule used in connectionist networks (Sutton & Barto, 1981). In all of these variants, error-driven learning is a systematic process that produces a mapping that best discriminates the informative relationships in a set of inputs and a set of outputs given a training schedule. Because of this, Ramscar et al. (2010) suggest that from a computational perspective it is best understood as describing a discriminative learning mechanism (this point also applies to the error-driven learning

algorithms found at the heart of most connectionist / neural network model; Ng & Jordan, 2002, and Bayesian models of learning, e.g., Daw, Courville & Dayan, 2008). Conceptually, the logic of discrimination enshrined in these models is far removed from the "blooming, buzzing confusion" assumed in many developmental theories.

For the sake of clarity, it is worth noting that 'discrimination learning' has been used in at least three ways in the literature (see Hoppe, van Rij, Hendriks & Ramscar, 2021 for discussion). The term DISCRIMINATION LEARNING was widely used in the animal learning literature in the early twentieth century, and, consistent with the behaviourist principles that dominated theory at this time, it was (and still is) used in a mechanism-neutral way to describe the fact that, objectively, both animals and humans are capable of learning different responses to different stimuli (Rescorla & Wagner, 1972). The second area to use the term discrimination learning is machine learning, where the concept of a discriminative model was introduced to provide a mechanism neutral contrast to GENERATIVE MODELS. Whereas the latter learn the data that generates a set of labels, DISCRIMINATIVE MODELS simply learn to maximize the conditional probabilities of labels for a set of labels given an input representation (Ng & Jordan, 2002). Finally, DISCRIMINATIVE LEARNING can be used to describe the mechanism implemented in the error-driven learning models. Because in most learning situations this mechanism enforces cue competition – which serves to discriminate against or in favor of the units that serve as inputs by re-weighting the influence of individual units – it serves to re-represent an input representation so as to maximize its informativity about a set of outputs (Ramscar et al., 2010).

From this latter perspective, the mind of a newborn learner can be thought of as an undifferentiated set of inputs that are connected to an undifferentiated set of output states. Starting from what is, in functional terms, a system containing a single entity, the learner's representation of the world will grow into a larger set of (more or less individuated) entities as the error resulting from perceptible variances and invariances in the environment produces differentiation, increasing the degree of discrimination between inputs and outputs (Ramscar et al., 2010, 2013a; see also James, 1890).

Learning and morphology – where connectionism went wrong

To turn the second question posed above – WHAT do children learn about language? – I will initially consider it in relation to linguistic morphology, and, in particular, English inflectional morphology. Although this is a relatively simple morphological system, its properties embody many characteristics of language as a whole: it is SYSTEMATIC (the formation of most but not all English plurals can be described as adding an allomorph of the morpheme *–s* to a singular noun form); it is PRODUCTIVE (children can readily generate past tenses for novel forms such as *rick-ricked*); and yet the system is QUASIREGULAR in that irregular forms deviate by degrees from the regular pattern (e.g., *knife-knives, mouse-mice, child-children*; Seidenberg & Plaut, 2014; Ramscar, in press).

These properties have led to morphological development being used as a test domain for theories of language acquisition, with the question of WHAT children learn in acquiring a morphological system being the subject of considerable debate. For example, Pinker (1998) claimed that COMPOSITIONAL RULES are explicitly represented components of MORPHOLOGICAL knowledge, and argued that the processes of inflectional morphology, which seemingly distinguish productive regular items from

https://doi.org/10.1017/S0305000921000544 Published online by Cambridge University Press

unproductive irregulars, provide evidence for this claim. This was in contrast to the classic connectionist model presented by Rumelhart and McClelland (1986), which took the phonological form of a verb's present tense as input, and generated the phonological form of its past tense as output, using a uniform procedure for the generation of both regular and irregular (and novel) forms (see Ramscar, in press, for a review).

For current purposes, what was notable about the famous 'past tense' debate is that both sides agreed that WHAT children learn in the course of morphological development are ways of composing and decomposing morphemes (Ramscar, in press). In the case of English plurals, it is assumed that a child learns a morpheme that associates the concept *mouse* with the word 'mouse,' an association between the concept *mice* and 'mice,' an association between the concept *rat* and 'rat,' and an association between the concept for PLURALITY (or sets of objects, excluding multiple mouses etc.) and a morpheme *+s*, etc. Yet the discrete system of associations between forms and meanings envisaged here is difficult to reconcile with the highly interconnected systems that associative learning actually produces. Moreover, this neat picture of form-meaning mapping is also inconsistent with the results of research into human categorization, which show that human performance on categorization tasks is best accounted for by models that treat categorization as a process in which discrete outcomes such as labels, responses, etc., are discriminated in context from a more continuous system of inputs (Love, Medin & Gureckis, 2004).

If we allow that morphological systems are not sets of discrete mappings between units of meaning and lexical forms, some of the limitations in the assumptions shared by both sides of this debate become apparent. The Rumelhart and McClelland (1986) model assumed children learn transformational rules that add a discrete English past tense morpheme *+ed* to a verb stem to generate a past tense form, or a discrete plural morpheme *+s* to a singular noun stem to generate a plural. In keeping with this, the model's training set is a list of uninflected stems that are transformed into past tense forms, as if the learning environment contained speakers producing repetitive bursts of singular-plural forms. Yet in reality, adults do not go around repeating, "*go-went*, *walk-walks*." Instead children learn in context, from hearing sentences like, "*shall we go walk the dog?*" (Gleitman, 1965). Thus not only is the learning scenario assumed implausible, but critically, Rumelhart and McClelland's theoretical account of inflection learning appears to be compositional, even though it is implemented in a discriminative learning model (in which compositionality seems to make little sense, Lake & Baroni, 2018). All of which raises a question: what might a discriminative theoretical account of morphological development actually look like?

A discriminative model of morphological development and processing

If we accept that children encounter morphological variations in contexts that offer no obvious evidence for transformation (they don't actually here numerous repetitions of "*go-went*, *dog-dogs*), the WHAT of learning in morphological development can be straightforwardly recast in terms of their learning WHAT it is about the environment that warrants the use of a given linguistic form in a given context (Ramscar, in press). To illustrate how this works, I will briefly describe how a discriminative model learning English noun morphology accounts for the patterns of over-regularization often observed in development (Seidenberg & Plaut, 2014). The

https://doi.org/10.1017/S0305000921000544 Published online by Cambridge University Press

model uses essentially the same learning rule as the Rumelhart and McClelland (1986) model. Where it differs is how it represents the learning task, and the nature of linguistic knowledge itself (representations critical to the performance of learning models, Bröker & Ramscar, 2020).

Because of the nature of the input, English noun inflection is difficult to learn even in comparison to verb inflection. Whereas children encounter more inflected than uninflected verb forms, and more of these forms (by token) are irregular than regular, plurals are different. Children mainly encounter singular nouns, and most plural noun types *and tokens* are regular. Yet as with the past tense, children's irregular plural production follows a 'U-shaped' developmental trajectory. Children who have produced 'mice' in one context will still produce over-regularized forms like 'mouses' in others (Ramscar & Yarlett, 2007).

The discriminative model of morphological development described here (Ramscar & Yarlett, 2007; Ramscar & Dye, 2009; Ramscar et al., 2013a) is based on results indicating that lexical learning involves discriminating the cues to word use (Ramscar, Thorpe & Denny, 2007; Ramscar et al., 2010; Ramscar et al., 2013b), and that patterns of morphological variation reflect similar semantic and contextual factors (Ramscar, 2002; Ramscar & Dye, 2011; Ramscar, Dye & Hübner, 2013c). Accordingly, the model assumes that children encounter words and morphological variations in context, and are faced with the task of discriminating the cues that are informative about their use. Since initially any kind of *stuff* in the world is potentially informative about any lexical contrast, an important aspect of learning to discriminate the cues to noun forms involves discriminating the more specific cue dimensions in the objects associated with them from the other, less specific dimensions they also comprise (that *mousiness* is a better cue to '*mice*' than *stuff*). Similarly, learning to discriminate singulars from plurals requires learning the dimensions of NUMEROSITY that best predict different forms (that *multiple mouse objects* best predicts '*mice*').

Figure 1a depicts some of the environmental dimensions that reliably covary with the irregular plural '*mice*.' Critically, while ALL these dimensions co-occur with '*mice*' at the same rate, their covariance with other nouns differs, resulting in cue competition. Because generic cues like *stuff* are reinforced when '*mice*' is encountered, learners will expect mice to occur whenever stuff is present, resulting in prediction-error in the contexts where stuff occurs and '*mice*' is not heard. This process will cause the value of these generic cues to weaken over time, such that *multiple mouse-items* will be learned as the best cue to '*mice*.'

Accordingly, the actual pattern of reinforcement and unlearning of the environmental cues to different forms will depend on their distribution in the learning environment. Figure 1b shows how the various potential semantic cues to '*mice*' overlap relative to a simple set of idealized cues to different singular and plural forms – irregulars, regular stems, and the regular plural contrast *+s* – in learning. Broadly speaking, the plural forms classed as 'regular' in English are similar in that they all end in a final sibilant that discriminates plural from singular forms (they also differ slightly in that different forms use different allomorphs of this final sibilant). By contrast, irregulars discriminate singular from plural forms in various ways. It is thus important to note that in this model, 'rats' is not conceptualized as comprising a stem that is inflected for plurality by adding +s, but rather, 'rat' and 'rats' are different word forms. Children must learn to discriminate one from the other, and the contextual cues appropriate to the usage of each (see Hoppe et al., in press, for a comprehensive tutorial on this kind of learning process).

https://doi.org/10.1017/S0305000921000544 Published online by Cambridge University Press

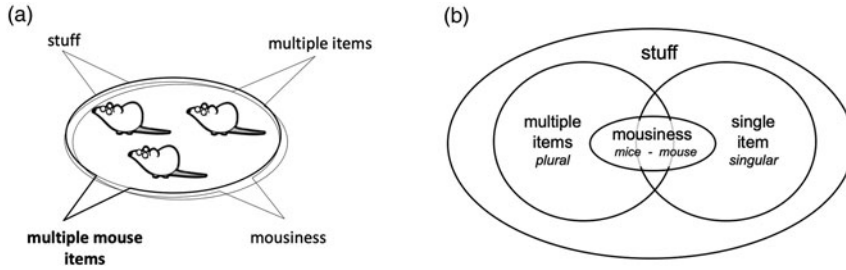

Figure 1. A: Some of the semantic dimensions that will be present whenever a child is exposed to the word 'mice' in the context of mice. **B:** A more abstract representation of the relative specificity of these dimensions as cues to plural forms. Although the less specific cues (*stuff* and *mousiness*) will be reinforced during early in learning, their ubiquity will ultimately cause them to produce more errors than the uniquely informative cues. As a result, the influence of these less specific cues will wane as experience grows.

At the same time, because 'rat' and 'rats' are very similar forms that appear in very similar ratty contexts, the specific cues to them are difficult to discriminate. Similarly, the fact that a final sibilant discriminates the plural and singular forms of a great many words means that it is difficult to discriminate the specific semantic cues to the final sibilant in 'rats' 'cats' and 'bats'. Given the distribution of cues and forms – in which regulars are by far more frequent – a learner's language model will initially come to expect plural forms that end in sibilants whenever sets of objects are described. This over-general expectation causes interference when irregular plurals are produced, causing the production of over-regularized forms.

However, further exposure to the same distribution serves to eliminate this interference. This is because the same generic cues that lead to over-regularization must inevitably also produce expectations for irregular forms ('*mice*') in contexts where REGULAR forms will be used. The prediction errors that result from this will cause the unlearning of these generic dimensions as cues to *mice*, increasing the relative strength of more specific cues, and reducing the likelihood of future over-regularization. Accordingly, this model makes a clear, unambiguous prediction: at an appropriate stage of development, exposing children to REGULAR FORMS ALONE ought to serve to reduce over-regularization, in any morphological paradigm.[1] To perform a test of this hypothesis in relation to English plurals, regular and irregular plural and singular forms were elicited from two groups of children, after which one performed a control task and the other a task that required them to access the regular forms from the elicitation test in memory. Whereas children in the control

[1]For example, although the semantic structure of verb arguments is more subtle than that of noun phrases (a point I shall return to later), the contextual cues to past tense forms appear to be sufficiently straightforward to support a simple adaptation of the plural model above. In this model, contextual features would serve to cue irregular past / present tense forms that will be suppletive, whereas for regulars, although +ed serves to discriminate tense at a form level, their morphological features otherwise offer little information in this regard. The approach to learning implemented in the plural models described earlier employed a set of forms sharing a common context – form mapping (in this case, would be past-context–+ed) to generate error in the representations of a set of 'exceptions,' thereby reducing the noise in the system that led to over- regularization. Applying the same logic here predicts that getting children of an appropriate age to repeatedly produce past tense forms in context ought to result in the same kind of reduction in over- regularization for verbs as was described for nouns above.

https://doi.org/10.1017/S0305000921000544 Published online by Cambridge University Press

task over-regularized at the same rate in a posttest, the experimental manipulation – which involved processing ONLY REGULAR FORMS – produced a significant DECREASE in the rate at which IRREGULAR FORMS were over-regularized (Ramscar et al., 2013a).

Regularity, information and coding

The plural learning model described above offers an interesting perspective on the idea of regularity in language. From a generative/compositional perspective, regularity is often assumed to be a desirable or normative goal for morphological systems, such that irregular paradigms represent deviations from the norm. However this assumption is at odds with phenomena like suppletion, where the exponents of inflectional paradigms are more or less related allomorphs (e.g., '*go*'-'*went*' /'*mouse*'-'*mice*') that serve to obscure the form-meaning relationship in the paradigm as a whole. Critically, although suppletive forms are often type infrequent, they tend to be HIGHLY FREQUENT as tokens, raising the question of what 'normative' means when languages are viewed statistically. It is thus interesting to note that in the model described above, suppletive forms serve to ACCELERATE the discrimination of the specific cues to individual forms in learning. From this perspective, suppletive irregular forms like '*went*' and '*feet*' can be seen to encode strong examples of the FORM CONTRASTS that learning and communication ultimately rely upon. Given that these processes will be facilitated by strong contrasts – '*one*,' '*two*,' '*three*' – and impaired by weak contrasts – '*one*,' '*john*,' '*gone*' – when seen from the perspective of learning and communication it is not suppletion that needs explaining so much as regularity (Ramscar, Dye, Blevins & Baayen, 2018): why do languages tolerate the less discriminative form contrasts that lead to over-regularization?

In answering to this question, it will help to imagine a world whose language is defined in an enormous codebook. The code comprises a set of sound-symbols (codewords), each of which serves to uniquely discriminate one of the many, specific messages a speaker might possibly wish to send from the other messages they might have sent, such that the task facing a learner is that of memorizing every codeword / message of what is a fully suppletive language. Using this language will prove problematic, and considering why can help highlight some problems – especially with regards to learning – that any ACTUAL language must solve.

First, given that people will want to communicate many different messages, the codebook will be have to be enormous (in fact, boundless). Second, in order to generate the enormous number of codewords this will require, many codewords will have to be long and/or complex. Third, since memorizing the whole book will be impossible, each individual will have learned only a fraction of it, restricting communication to only those codewords speakers share knowledge of. These factors will make the language massively inefficient, if not unusable.

How might one offset these problems? One way to simplify the code would be to employ combinatorics: codewords that discriminate different KINDS of messages – questions, declarations, etc. – could be defined, and combined with other codewords in SIGNALS, enabling the number of unique words required to be reduced. The code could be further improved by analyzing the RATE at which people use each message. SHORTER SIGNALS could then be assigned to MORE FREQUENT MESSAGES, making it easier for people to learn the signals they are more likely to use before less generally useful signals.

These latter modifications presuppose there is some consistency in the likelihood with which every person wants to send any given message. Yet how likely is it that

https://doi.org/10.1017/S0305000921000544 Published online by Cambridge University Press

this will always be the case? To address this, a further, less obvious modification could be made: an analysis could be made of the various kinds of signals people send, along with their probabilities. The code could then be redesigned so as to distribute the discriminatory power of different kinds of codewords more efficiently and more abstractly across signals. The idea would be to strike a balance, such that shared needs use less coding resources, and those that vary a lot require more, so that the code's overall structure was matched to the collective needs of the community.

Finally, the code could be modified so as to tolerate AMBIGUITY. If the same codewords/signals could be used to communicate different (yet similar) messages, and if speakers were to use context to resolve any uncertainty about the different meanings intended, the code's impossibly vast vocabulary could be reduced dramatically (Piantadosi, Tily & Gibson, 2012).

However, whereas learning the original code simply involved memorizing form-meaning pairings one-by-one, it is far from clear this strategy will be compatible with all of these modifications. While combinatorics is compatible with the memorization of form-meaning pairs, this strategy would become less plausible as the code began to get more complex, and as its level of abstraction grew. Finally, when ambiguity – and the requirement that context be used to resolve uncertainty about intended meanings – are introduced, the codebook and the memorization of form-meaning pairs become unworkable. If context is necessary to learn how signals are used, then either the codebook will have to describe context, which is impossible (the book would have to resemble the fabulous cartographers' 1:1 scale map of the empire, Borges, 1946), or else the code will have to be learned in context, and its design will somehow have to facilitate this process.

Given this imaginary language was initially defined in a book, it made sense to talk about it through a process of top down analysis and modification modification. However, if we allow that human communicative codes have evolved in response to the selection pressures that arise out of people's desire to communicate more or less important things more or less successfully, we might allow that the ambiguous, highly abstract structure of the modified code might arise naturally, as codes are used, reused and changed by successive generations. In what follows, I will assume that it is exactly this kind of socially evolved code that children must learn in order to enter into a linguistic community. I will seek to elucidate how these codes work, describe in detail how some aspects of them are learned, and illustrate how in fact, rather their being unlearnable (as other characterizations of language suggest), it appears that the structures of human communicative codes have evolved (culturally) to support their learning.

Before doing so, it is worth noting that thinking in theoretical linguistics (Goldberg, 2003; Partee, 2007; Ellis & Ferreira-Junior, 2009) and computational linguistics (Bender & Koller, 2020; Lake & Baroni, 2018) is still dominated by the assumption that languages revolve around inventories of form-meaning parings, such that language acquisition is usually thought to involve learning these pairings (Pinker, 1998; Bohn & Frank, 2019). This is despite the fact that, as noted earlier, centuries of study have failed to provide a detailed account of what form-meaning parings are, or how children learn them. By contrast, the plural model described above did not learn a set of individual form-meaning pairings. Rather, it learned the systematic relationships between a set of environmental cues and a set of forms based on their distribution in training. In this model (and in more sophisticated versions of it), prior learning serves to eliminate the majority of potential relationships between a

given set of inputs and the totality of possible outputs, such that more or less 'discrete' form-meaning 'pairings' emerge dynamically (and probabilistically) as the model is run (as it processes the inputs present in a given context). Thus while the model learns comparatively discrete mapping between the form *mice* and its contextual cues, many of the cues it learns to regular plural forms (*cats, rats*) are generic, such that in most cases the model DOES NOT learn a simple pairing between a set of meaning cues and a form, but rather it learns to 'pair' various aspects of meaning with various sets of forms.

These generic mappings – which underpin the model's ability to infer the form of novel plurals – are often described as 'distributed representations' in the connectionist literature (Rumelhart, Hinton & McClelland, 1986). However, the use of 'representation' here is somewhat misleading, because the outputs of error-driven learning models cannot be read from the 'representations' they learn (the weights between their input and output layers), because as the generic cues in the plural model described above serve to highlight, error-driven learning models do not learn discrete pathways from inputs to outputs. Rather it is the interaction between a specific set of inputs and the learned state of a model (that occurs as it is run) that serves to discriminate an output state (such that similar inputs can result in the same or different outputs, depending on training).

In practice – given an appropriately structured environment and an appropriate amount of training – these models settle into learned states which represent distributed, somewhat abstract relationships between their sets of inputs and outputs that are stable and predictable even in the absence of a discrete pathway from a given input to an output. This points to an interesting state of affairs: although these models don't learn form-meaning pairings, they are able to simulate combinatoric generalization WITHOUT ever implementing a compositional system (as when the plural inflection model described above infers that a novel noun should end in a sibilant in a context where plurality is implied). Moreover, these models can also help explain the intuitions that people have about form-meaning pairings and compositionality, and why these intuitions are ultimately misleading, because they can explain how people can learn to discriminate specific form-meaning relationships in USE (Wittgenstein, 1953) WITHOUT ever learning inventories of 'form-meaning pairings' (see also Ramscar, in press).

These considerations point in turn to a clear difference in the predictions that compositional accounts of language make about the way that linguistic forms are distributed, and the kind of distributive predictions made by the highly abstract communicative systems described here. If languages are compositional, we should expect the distribution of form-meaning pairings to reflect human habit and interest. That is, if people prefer red to green, then this preference, and its degree, ought to be reflected in the distribution of the words 'red' and 'green,' such that patterns of usage should reflect patterns of relevance.

By contrast, if human communicative codes are akin to the final iteration of the imaginary language described above – such that meanings and messages are distributed abstractly across the code – this predicts that any relationships between patterns of usage and patterns of human interest ought to be FAR more abstract. If languages are communication systems, THE USE OF INDIVIDUAL WORDS AND CONSTRUCTIONS OUGHT TO BE DETERMINED BY COMMUNICATIVE CONSTRAINTS, AND WE SHOULD EXPECT THAT PATTERNS OF USAGE REFLECT THESE CONSTRAINTS.

https://doi.org/10.1017/S0305000921000544 Published online by Cambridge University Press

To offer a first test of these different hypotheses, Figure 2a plots the usage frequencies (in the Corpus of Contemporary American English, Davies, 2009) for a set of common colour words taken from Wikipedia (*red, orange, yellow, green, blue, purple, brown, magenta, tan, cyan, olive, maroon, navy, aquamarine, turquoise, silver, lime, teal, indigo, violet, pink, black, white, grey/gray*). Figure 2b then shows how the distribution of these frequencies is geometric (the geometric being the discrete form of the exponential distribution).

Empirically, it appears when people talk about colour in English, they use *white* exponentially more frequently than *black, black* exponentially more frequently than *red,* and *red* exponentially more frequently than *green* etc. Figure 3a then plots the COCA frequencies of a set of common kinship terms (*mother, father, son, daughter, brother, sister, uncle, grandmother, aunt, grandfather, cousin, grandson, nephew, niece, granddaughter*; taken from Kemp & Regier, 2012), and Figure 3b shows that the distribution of these words is also geometric.

Figure 4 plots the correlation between the distributions of colour and kin word (along with a replication using the 100 million-word British National Corpus), showing how the usage of what are supposed to be compositional items in these completely different domains follows exactly the same – highly abstract – pattern. Given sufficient time and effort, it is clear one could come up with plausible-sounding stories to account for these patterns of use. One might argue that colour word usage reflects biases on people's colour concepts, and that kinship usage reflects other biases, and that the suspicious similarities between their usage patters are coincidental, a consideration that points to a problem facing scientific studies of human communication: 'language' is such a complex and poorly defined construct that it is unclear that any of the many theories that compete to explain it can be falsified on the basis of a few failed predictions or contrary findings.

This should not surprise us, since this problem is not unique to the study of language. Indeed, despite the importance of falsification to science (Popper, 1958), it is far from clear that theories themselves are actually falsifiable. Rather, it seems that whereas falsification allows for the rejection of specific hypotheses, it is explanatory adequacy that ultimately determines the fate of theories (Kuhn, 1962). Thus, although the account of learning and communication presented here predicts these exact patterns, AND PREDICTS THAT THEY SHOULD BE UBIQUITOUS IN COMMUNICATIVE CODES ACROSS DIFFERENT LEVELS OF DESCRIPTION (RAMSCAR, 2019, 2020; LINKE & RAMSCAR, 2020) – SUCH THAT THE THEORY WOULD BE FALSIFIED IF COMMUNICATIVE CODES WERE NOT STRUCTURED IN THIS WAY[2], the exact prediction and the exact findings that would falsify it are ultimately determined by the specific details of the various mechanisms proposed by different parts of the theory. Just like any other theory, the accuracy of

[2]The prediction is that the statistical structure of codes should respect communicative constraints at every level of analysis, such that in context, their structure should be better described by communicative (and related combinatoric) factors than, say, compositionality. For example, at a lexical level, the theory predicts that in comparisons of languages like English and German, whose articles and pronominal adjectives convey different amounts of information about upcoming nouns, the extra reliance of English on adjectives as compared to German (Dye, Milin, Futrell & Ramscar, 2017; Dye, Milin, Futrell & Ramscar, 2018) will be reflected in the frequency distributions of the words in the two languages. Specifically, to return to the example of color words above, it predicts that color words will be much more frequent in English than German. It further predicts that this difference will simply reflect the different information structures of these two languages, and not that English speakers are somehow more 'interested' in color than German speakers.

https://doi.org/10.1017/S0305000921000544 Published online by Cambridge University Press

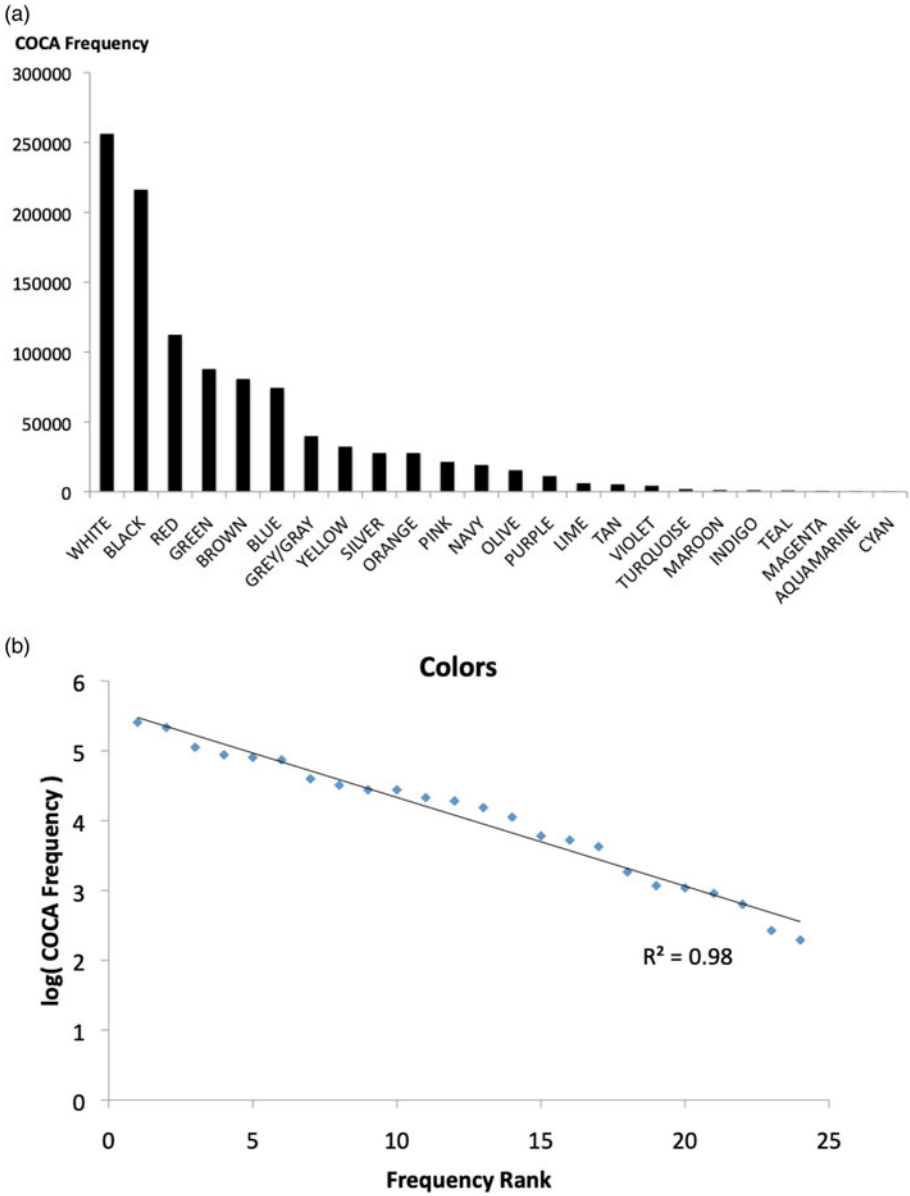

Figure 2. A: Corpus of Contemporary American English (COCA) frequencies of 24 color common English color words **B:** Log frequency x frequency rank plot of the same words (linear = exponential, R^2=.98).

the communicative account of language learning / use presented here will ultimately stand (or fall) on its ability to predict and explain linguistic phenomena more adequately and more accurately than other theories (Kuhn, 1962; Cohen, 2017). Accordingly, the remainder of this article describes the mechanisms that predict

https://doi.org/10.1017/S0305000921000544 Published online by Cambridge University Press

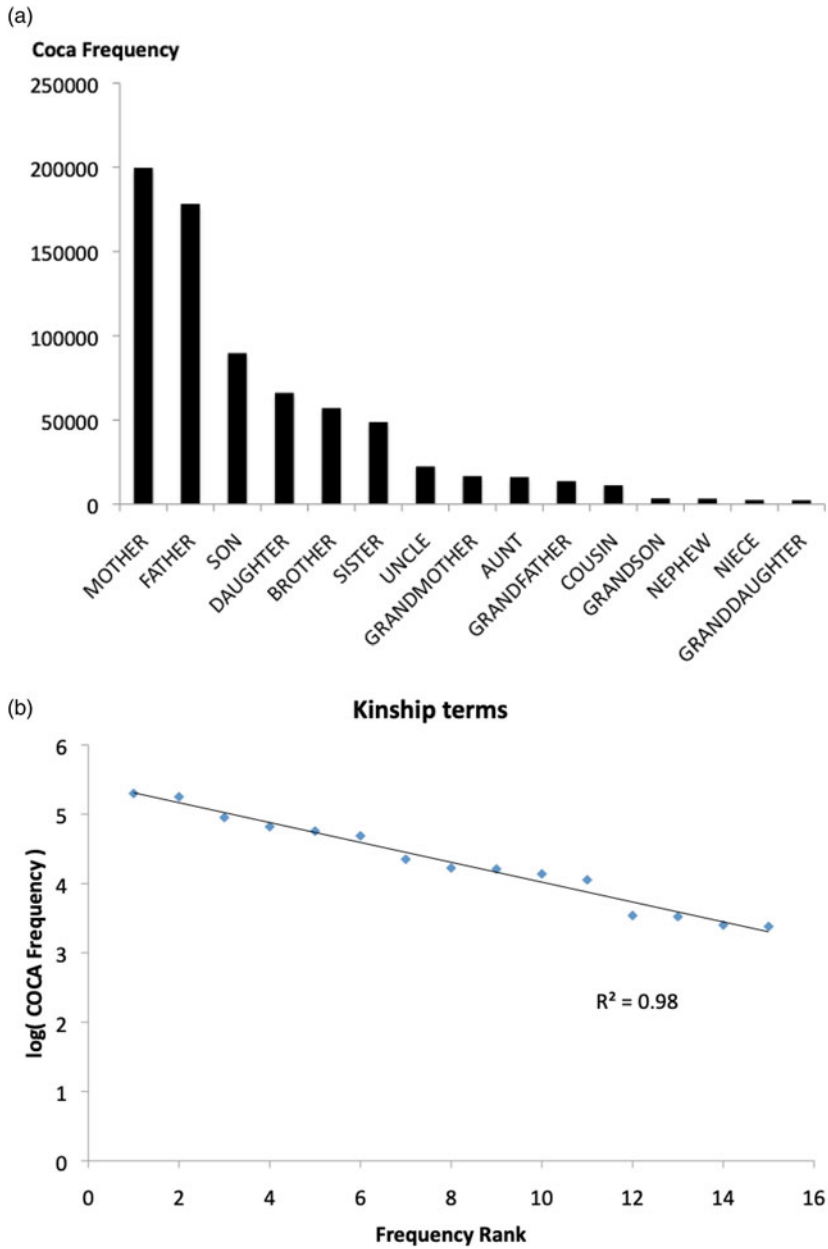

Figure 3. A: COCA frequencies of 15 common English kinship terms (Kemp & Regier, 2012) B: Log frequency x frequency rank plot of the same word frequencies (R^2=.98).

these usage patterns in more detail, along with other predictions and explanations that can be derived from them, and findings in other domains that further support these analyses.

https://doi.org/10.1017/S0305000921000544 Published online by Cambridge University Press

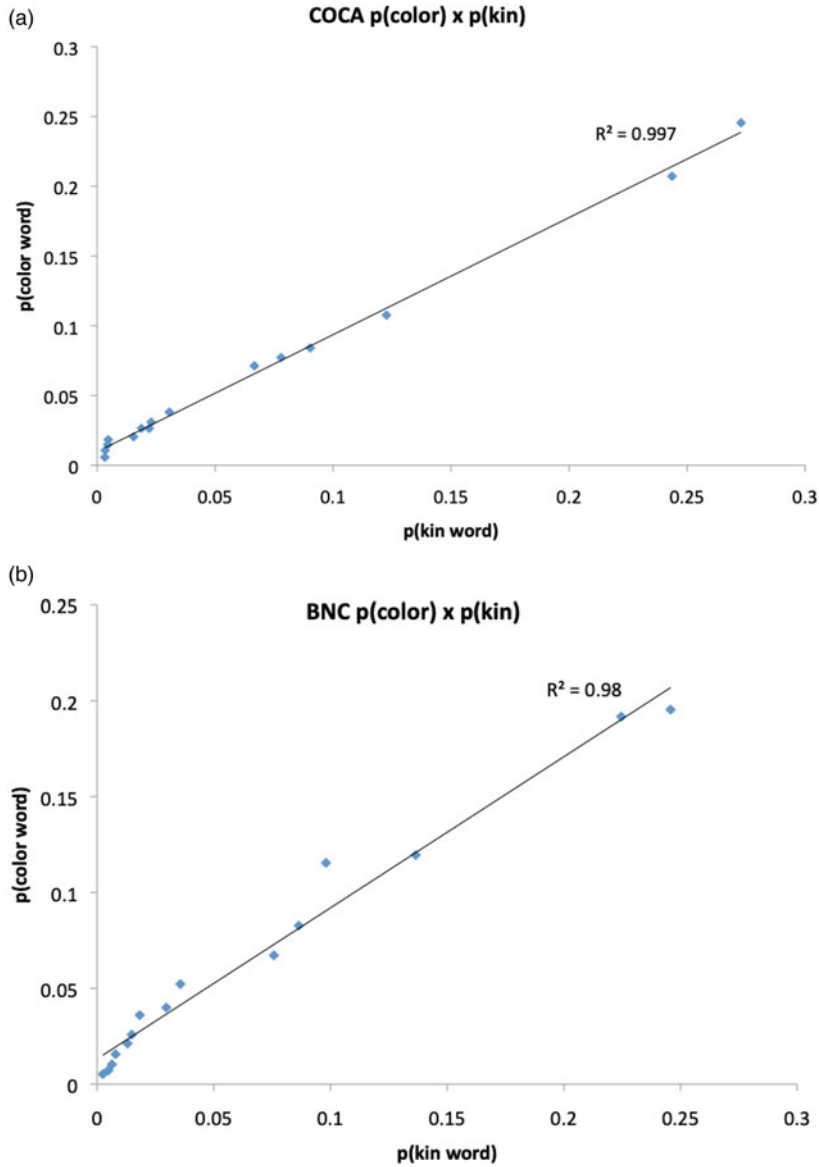

Figure 4. Point-wise comparisons of (A) the COCA probabilities of the 15 most frequent English colour words in Figure 1 (H=3.4 bits, calculated over all 24 items) and the probabilities of the English kin terms (3 bits) in Figure 3 (R^2=.997), and (B) the same probabilities in the BNC (R^2=.97; colour 3.4 bits; kin, 3 bits).

Information theory and human communication

One advantage of framing human communication in discriminative terms is that it enables linguistic theory to benefit from the insights gained from the study of discriminative codes in information theory. This work has shown, for example, how

https://doi.org/10.1017/S0305000921000544 Published online by Cambridge University Press

variable-length codes (in which codewords comprise a variable number of bits) enable the efficiency of communication to be maximized close to its theoretical limit (Shannon, 1948). The benefits of variable-length coding can be explained in intuitive terms: if *one* is easy to articulate, whereas *five* and *twenty* require successively more effort, then making *one* highly frequent and *five* and *twenty* successively less frequent will reduce the effort speakers must expend on number articulation. As the distribution and articulation of *one*, *five* and *twenty* suggests, the organization of natural languages do appear to reduce articulatory effort in this way (Lindblom, 1990): word lengths and frequencies are reliably associated, with more frequent words (*one*) being shorter than less frequent words (*nineteen*, Piantadosi, Tily & Gibson, 2011).

In recent years the development of massive speech and text corpora, and the tools to analyze and mine them has served to highlight these kinds of statistical regularities in linguistic codes. Studies of speech production have shown how words that are more predictable in context are more likely to undergo articulatory reduction (Bell, Brenier, Gregory, Girand & Jurafsky, 2009; Seyfarth, 2014; Linke & Ramscar, 2020), whereas disfluencies and other difficulties in language processing are strongly associated with increases in lexical information (Howes, 1957; van Rooij & Plomp, 1991; Westbury, Shaoul, Moroschan & Ramscar, 2016). The frequency distributions of words in natural languages have been shown to resemble variable length codes in that they are systematically skewed (Estoup, 1916), so that half the words in a corpus will reliably comprise tokens of around a hundred high-frequency types ('*and*', '*the*'), while the other half comprises very large numbers of low-frequency types ('*comprise*', '*corpus*'). The fact that word frequency distributions appear to have power law distributions has been taken as evidence that they have evolved (culturally) to optimize communication (Zipf, 1935, 1949), and many theories seek to explain how they help optimize linguistic communication (Mandelbrot, 1966; Manin, 2009; Piantadosi, 2014). These methods have also led to a resurgence of interest in the application of information theory to linguistics with many similarities between human and digital communication systems (and human behaviour and information theoretic predictions) being highlighted (Gibson et al., 2019).

At the same time, however, some important properties of the communicative model described by information theory are difficult to reconcile with what is known about natural languages, and these differences become particularly salient when we consider how children learn to communicate, and what they must learn in doing so:

1. Shannon's (1948) theory of communication describes a SYSTEM solution to the problem of signaling over a noisy channel (MacKay, 2003), based on predefined SOURCE and CHANNEL CODES. Like the codebook of the imaginary language, SOURCE CODES map messages onto source symbols (which are configured to optimize the use of bandwidth), while CHANNEL CODING increases the reliability of transmission by adding REDUNDANCY to coded vectors of source symbols. These codes define a probabilistic model SHARED by every sender/receiver in a system. However the distributions of words in languages, and the fact that people LEARN them, guarantees NO speaker / hearer ever learns an entire code (Ramscar, Hendrix, Love & Baayen, 2013d).
2. Individual samples of linguistic codes vary enormously (Ramscar et al., 2014; Ramscar, Sun, Hendrix & Baayen, 2017) suggesting that the probability models individuals learn may also vary considerably. By contrast, in information

https://doi.org/10.1017/S0305000921000544 Published online by Cambridge University Press

systems the provision of common codes ensures there is NO DIFFERENCE between the individual probability models of each sender / receiver.

3. The power law distributions observed in natural languages are NOT the most efficient distributions for variable length codes (Huffman, 1952; Gallager & Van Voorhis, 1975).

In other words, information theoretic models of communication are predicated on SHARED CODES, whereas human communicative codes do not seem to be shared in anything like the same way. Further, while information theory describes a number of specific features that enable the development of optimized codes, these features do not appear to be present in linguistic codes. Finally, information theory describes a deductive process (Shannon, 1956) that serves to eliminate a receiver's uncertainty about the identity of a sender's message, using common codes that maximize the discriminability of messages while minimizing the cost of signaling. By contrast, most linguistic theories, even when they claim inspiration from information theory, adhere to compositional principles that make inductive assumptions about the way meanings are communicated.

These concerns differ in the degree to which they undermine the application of information theoretic concepts to human communication: however, they highlight two problems that appear to be critical. First, how do natural languages provide a degree of SYSTEMATICITY that is at least SUFFICIENT to support communication (how can they allow users to communicate despite their having partial access to a code)? Second, how do users of natural languages learn probability models that converge SUFFICIENTLY for languages to be PROBABILISTIC SYSTEMS (how can the common model that defines the INFORMATION in a communication system be acquired and transmitted)? Unless these questions are answered, it would appear that the analogies researchers draw between language use and information theory, as well as any talk about languages as 'probabilistic codes,' can only be speculative at best

Partial attestation, regularity and the poverty of the stimulus

The productive nature of the morphological neighbourhoods found across languages suggests at least a partial solution to the first of these problems, in that they provide a means for CONSISTENTLY filling the inevitable gaps in individual language users' experience (Ramscar et al., 2018). Formally, this solution emerges as a virtuous outcome of learning from the same distribution of forms that led to OVER-REGULARIZATION in the discriminative model of plural morphology described above (see also Marzi, Ferro & Pirrelli, 2019). In that model, the distribution of regular plural forms inhibited the unlearning of generic meaning cues (such that the cues to regular plurals typically comprised a mix of generic and specific cues). One consequence of learning this pattern of input-output relationships is that the representations the model learned serve to implicitly encode the forms of regular noun plurals BEFORE they are encountered.

This result suggests an alternative perspective on the coexistence of regular and irregular patterns in morphology, since it suggests that the existence of regular and irregular forms represents a trade-off that balances the opposing communicative pressures of DISCRIMINABILITY and LEARNABILITY in the evolution of communicative codes. From this perspective, the existence of frequent, well-discriminated irregular forms serves to make important communicative contrasts more discriminable and

https://doi.org/10.1017/S0305000921000544 Published online by Cambridge University Press

thus also more learnable. By contrast, BECAUSE regularity entails less discriminability, learners' representations of lexico-morphological neighbourhoods will tend to be more generic, which causes the forms of large numbers of less frequent items to be learned IMPLICITLY, compensating for the incompleteness of individual experience.

It follows that if natural languages are communicative systems in the probabilistic, discriminative way learning and information theory suggest, we should expect linguistic codes to be structured along these lines at every level of description. The structure of forms known to all speakers should tend to maximize discriminability in learning and communication, while the distributional structure of less frequent forms should support productive regularities of the kind that maintain systematicity. With this in mind, I next consider the form and function of a linguistic subsystem that will be noticeable for its absence in the other articles in this special issue: systems of personal names. It turns out that, as a subsystem, names provide a helpful introduction to reasons why, theoretically, probabilistic theories of communication predict the occurrence of specific kinds of distributional structure EVERYWHERE, while also offering an intuitive introduction to what it is that 'meaningful communication' actually involves from a discriminative, non-compositional perspective.

Learnability and the cost of naming

Personal names serve to discriminate individuals from their peers. Given that an obvious way of achieving this would be to give each individual a unique name, it seems that if any subsystem of language is going to be suppletive, it is names. It is thus notable that the world's languages DO NOT take a suppletive approach to naming. Rather, somewhat strikingly, all languages adopt the same combinatoric solution to the problem of discriminating individuals, forming personal names from sequences of hierarchically structured name tokens that allow huge sets of identifiers to be constructed from relatively small sets of name types (Ramscar, Smith, Dye, Futrell, Hendrix, Baayen & Starr, 2013e).

Name grammars

Formally, the information provided by a set of discrete codewords depends on their distribution (Shannon, 1948; see also Hartley, 1928; Nyquist, 1924). Perhaps surprisingly, names offer a very intuitive explanation of this idea. Imagine that 33% of males are called *John*, and only 1% *Cornelius*. In this scenario, learning someone is named *Cornelius* is more informative than learning their name is *John* (*Corneliuses* are better discriminated by their names than *Johns*). On the other hand, *Johns* will be easier to remember (guessing '*John*' will be correct 1/3 of the time). Further, although the memory advantage of *John* relies on its frequency, the memorability of *Cornelius* also benefits from this: *Cornelius* is easier to remember if the system contains fewer names (also, as discussed earlier, if *John* is easier to say than *Cornelius*, this will reduce the average effort of name articulation).

Finally, since *John* is so easy to remember and articulate, it may be possible to figure out a way of distributing names so that combining *John* with a high frequency word that serves as a name in context (*Green, Smith* etc.) can allow us to produce a combinatoric name that requires little/no more processing effort than *Cornelius*. That is, if the names *John* and *Cornelius* are distributed along the lines just described, one might possibly devise a system that balances the competing demands of discriminating individuals

https://doi.org/10.1017/S0305000921000544 Published online by Cambridge University Press

for communication purposes, while managing the processing and memory demands of using a code, in a near optimal way. Information theory offers formal proofs of these points, and shows that the most efficient way of distributing codewords (*John, Cornelius*) is by distributing their probabilities geometrically (Gallager & Van Voorhis, 1975; the geometric distribution being the discrete form of the exponential).[3]

The actual name grammars of languages have been heavily impacted by name laws in the modern era, and the formal fixing of patronymic 'family names' for legal purposes in particular (in the Sinosphere, given names were legislated into patronyms, whereas in the West, bynames have become patronyms). While this has resulted in significant changes to the traditional, vernacular naming systems that existed before these laws were imposed (Ramscar, 2019), it is still possible to discern a common structure in the name grammars of many languages. Perhaps surprisingly, the form of this structure is remarkably close to the *John/Cornelius* example above.

Names in modern Chinese (a family of Sino-Tibetan languages, Handel, 2008) and modern Korean, a language isolate (Song, 2006) typically comprise two or three elements (Kiet, Baek, Kim & Jeong, 2007). As a Korean name is encountered in speech, these comprise sequentially: first, one of a small number of family names (patronyms), second, a clan/generational name, and third, a given name. The size of the set of name tokens each is drawn from increases as names unfold, such that these names have a hierarchical structure in which each element simultaneously increases the degree to which an individual is identified, and reduces the number of alternatives that need to be discriminated between at each step.

Historically Sinosphere first names were drawn from a small set comprising around 100 or so name tokens (Baek, Kiet & Kim, 2007), and unlike family names in most of the world's languages (whose distributions appear to follow power laws) Chinese and Korean family names have been shown to be geometrically distributed (Yuan, 2002; Kiet et al., 2007; Guo, Chen & Wang, 2011). Since this suggests that the name grammars of Korean and Chinese may provide an optimal solution to some of the communicative problems posed by naming, Ramscar (2019) reconstructed a partial Vietnamese family name distribution from US census data (most Vietnamese Americans in the 2000 US census were named in Vietnam), and compared this to data from the 2000 South Korean census to examine whether this finding generalized to another, unrelated language (Vietnamese, which also employs patronyms as first names, is an Austroasiatic language, albeit with many lexical borrowings from Chinese; Sidwell & Blench, 2011). As Figure 5 shows, the distribution of US Vietnamese and Korean first names is essentially identical.

The distribution of English first names

The organization of Western and Sinosphere names suggests that Western given names and Sinosphere family names share a similar communicative function: given names come first as Western names are encountered in speech, and are drawn from a far smaller set than family names, which come last). However, analyses of U.S. census data show contemporary English first names to be Zipf distributed (Ramscar, 2019). This is hardly surprising, first because the US population is over 325 million people (making it empirically impossible for individuals to sample the entire name

[3]In the LEAST efficient distribution, names are equiprobable.

https://doi.org/10.1017/S0305000921000544 Published online by Cambridge University Press

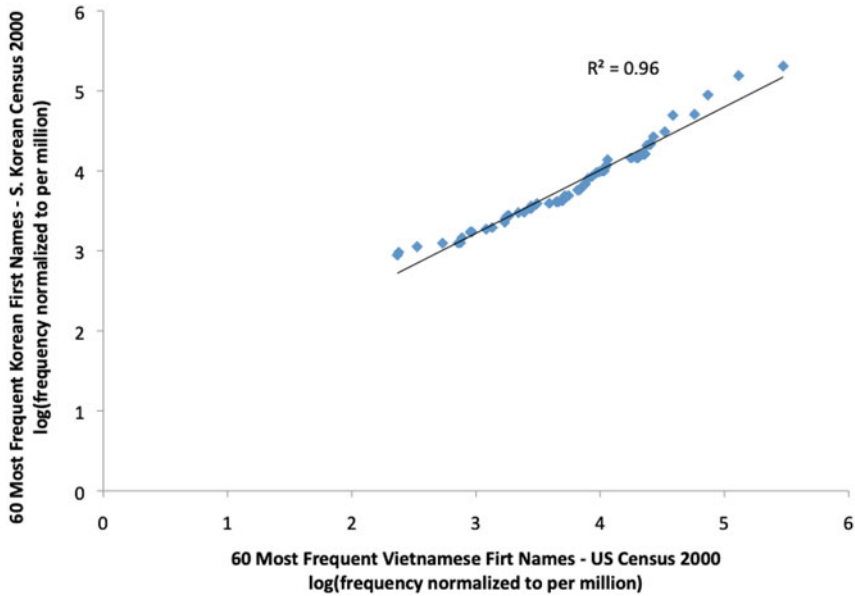

Figure 5. A frequency normalized comparison of the distribution of the 60 most frequent first names in the 2000 South Korean Census to the 60 most frequent first names in a Vietnamese first name distribution constructed from the 2000 US Census (R^2=.96; data from Ramscar, 2019).

distribution), and second because mixtures of exponentials often form produce power law distributions (Newman, 2005). What is more relevant to actual communication – and critically, its development – are the distributions learners actually experience. In this regard, it is notable that analyses of the US social security records for the 50 US states and the District of Columbia across the 20th Century (Ramscar, 2020), show the average cumulative distributions of first names from 1910-2010 across the States to be geometric.

However, perhaps the most striking evidence for the communicative function of first names comes from comparing first name distributions in 18th Century Scotland and England (Ramscar, 2019) with those of China and Korea. Historically, first names in the latter comprised a stock of around 100 names (the colloquial Chinese expression for the common people – '*Bǎijiāxìng*' – means 'the hundred names') and local name stocks for England and Scotland also comprised around 100 names, such that the distribution (r^2=.96) and information entropy (Korea=4.7 bits; Scotland=4.8 bits) of historical Scottish / modern Korean first names share remarkably similar information structures (Ramscar, 2019; Figure 6e)

How do name grammars work?

If we assume names encode identities (something neurotypical human brains seem to be adapted to discriminate, Kanwisher, McDermott & Chun, 1997), it seems that the codes people use to communicate them employ exactly the kind of structures information theoretic accounts of human communication predict. Importantly, what should also be clear from the foregoing is that the function served by the name

https://doi.org/10.1017/S0305000921000544 Published online by Cambridge University Press

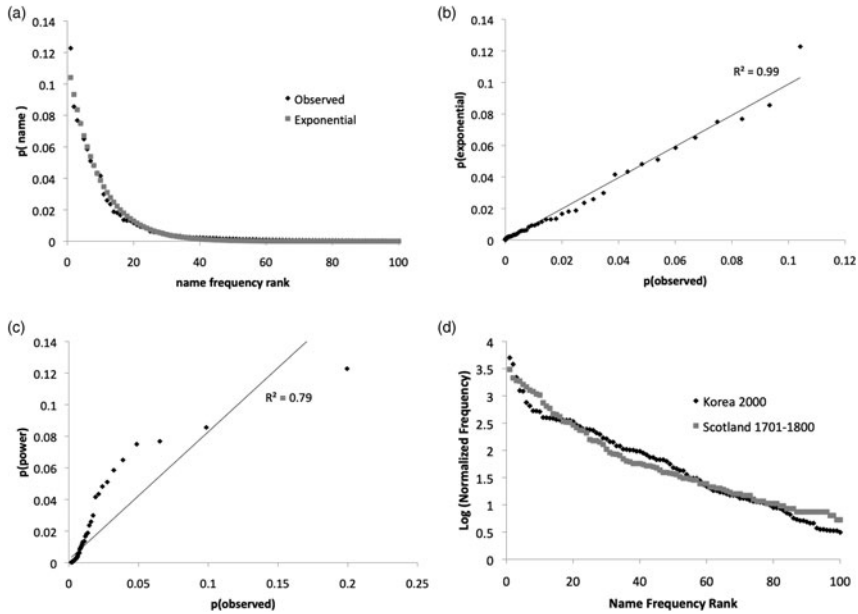

Figure 6. A: Probabilities of the 100 most frequent given names (98% of population) by frequency rank in 4 Scottish parishes 1701–1800 plotted against an idealized exponential distribution. B: Pointwise comparison of the observed distribution to idealized exponential distribution. C: Pointwise comparison of the observed distribution to an idealized power-law distribution. D: Log (normalized) frequency x frequency rank plot comparing the distribution of first names in South Korea 2000 to that in Scotland 1701–1800 (Ramscar, 2019).

David is simply that of maximizing the likelihood that *Davids* (identities conditioned on *David*) will be discriminated from *Marys* etc. in communication (Figure 7).

In some contexts, comprehenders may be able to discriminate a communicated message (identity) from a first name alone; in others, more information will have to be signaled. If we assume the contextual distributions of surnames are similar to those of first names, then empirical name grammars will tend to generate signals that minimize the average cost of name processing (Meylan & Griffiths, 2017), smooth the information communicated across signals (Aylett & Turk, 2004), and make names easier to access, use and recall (Dye, Johns, Jones & Ramscar, 2016). It is thus important to note that from this perspective the name *David Bowie* is discriminative encoding, and its function can be explained WITHOUT assuming that it is composed from the concepts DAVID and BOWIE.

Memorylessness and the alignment of communicative models

Thus far our discussion of geometric distributions has focused on their contribution to the efficiency of communication. However, these distributions have a further property that is particularly important to communicative learning. The geometric distribution is unique in being the only discrete MEMORYLESS distribution (just as the exponential distribution is unique in being the only continuous MEMORYLESS distribution). This property is important because it suggests a solution to a problem, raised above, of explaining how learners with very different experiences of a probability distribution

https://doi.org/10.1017/S0305000921000544 Published online by Cambridge University Press

Figure 7. Pictures discriminated by the search terms "David", "David Bowie", and "David Bowie Ziggy period" by Google image search (13/2/2019). "David" eliminates pictures not related to David, and "David Bowie" and "David Bowie Ziggy period" refine this discriminative process.

nevertheless learn the same model of it, so that their communicative expectations are actually shared (a problem ALL probabilistic models of communication must face).

The memoryless property is best explained in terms of waiting times: if the probabilities of encountering people at any given point in time are distributed exponentially, then (because of the way these distributions interact with the laws of conditional probability) it can be proven that the probability of encountering someone at any specific point in time t_n is independent of the time that has elapsed since t_1, the time a person was last encountered. A counterintuitive result of this proof is that when the periods between encounters are exponentially distributed, the likelihood of encountering another individual in a minute is independent of the time

https://doi.org/10.1017/S0305000921000544 Published online by Cambridge University Press

that has elapsed since the last person was encountered, such that the likelihood is the same 30 minutes after the last encounter as it was 2 minutes after; and it will remain the same 2 hours later.

If we assume these probability laws apply to learning from lexical distributions, it follows that when words are geometrically distributed individuals will learn similar models of the words' underlying probabilities, even if the size of two samples varies considerably (Townsend & Ashby, 1983, pp. 38–45; Linke & Ramscar, 2020). Accordingly, it is further worth noting that the human frontal cortex develops over a prolonged period, such that the control mechanisms that allow mature learners to more flexibly sample their environments take two to three decades to develop. From a communicative perspective this a hugely beneficial developmental trait, since restricting learners to a naïve sampling strategy will further increase the likelihood that typically developing children learn the same communicative models as their peers (Ramscar & Gitcho, 2007; Thompson-Schill, Ramscar & Chrysikou, 2009; Ramscar et al., 2013b; Ramscar, Dye, Gustafson & Klein, 2013f).

Whether these mathematical points about sampling and learning actually apply to human learners are empirical questions. This account makes clear predictions in regard to them: if learners are exposed to sets of geometrically distributed forms, they should acquire models of their probabilities that better approximate one another than when learning from other distributions. Conversely, if learning from geometric distributions does not produce convergence, it would suggest the probabilistic account of communication described here (indeed, any probabilistic account of communication) is false.[4]

Semantics and the distributional structure of codes

As noted above, while almost all linguistic theories assume compositionality, no adequate account of what compositional meanings are, or how they are learned, actually exists (Ramscar & Port, 2015). While the many difficulties involved in defining meanings are acknowledged, the consensus is that these problems MUST be solvable. If children learn to use words like dog or red, this MUST BE because they learn – or innately have – the concepts DOG and RED. The communicative account of names presented above offers a way out of this circular thinking. If we accept that the information contributed by *Mary* is not derived from the concept MARY, but rather is a function of *Mary's* role in a discriminative system of names, then

[4]As a reviewer helpfully pointed out, because of the recurrent levels of structure in natural languages (Ramscar, 2019, Linke & Ramscar, 2020), any empirical test of these predictions must involve some analysis of the information structure of the forms in these distributions as well. Whereas word learning studies typically involve arbitrary forms, and 'control for frequency' by employing uniform frequency distributions in training, natural lexical distributions are inevitably highly skewed, and the discriminability of the forms they comprise seems to change as frequency ranks descend (compare one and two with fourteen and fifteen). This indicates that training on new (i.e., unknown) natural languages might produce different results to training on completely novel forms. Similarly, while uniform frequency distributions offer an obvious control when it comes to testing the kind of input that does or does not lead to the alignment of learners expectations, further questions arise as to the degree to which different distributions – e.g., power law versus geometric – do or do not result in alignment. While this suggests that such testing must necessarily be somewhat exhaustive, it is interesting to note that insofar as artificial language learning studies ignore these matters – as most currently do – they raise the question of how much these studies actually have to tell us about language learning.

https://doi.org/10.1017/S0305000921000544 Published online by Cambridge University Press

providing a non-compositional account of how names are learned and used in context is a straightforward task. This line of reasoning points to an interesting conclusion: although the theoretical problems posed by compositionality seem particularly acute with regards to names (Gray, 2014; Fara, 2015), a series of analyses by Wittgenstein (1953) indicate that the problems involved in explaining the meanings of common nouns and names are ultimately the same. If Wittgenstein's analyses are right, then the use of nouns, verbs and adjectives etc., should be amenable to the same functional analysis.

A source of support for this suggestion comes from studies of colour and number words. Although infants can distinguish basic colour categories (Bornstein, Kessen & Weiskopf, 1976), and despite their high frequencies in the input, children's use of colour words is haphazard for a surprisingly long period (Sandhofer & Smith, 1999). Children's learning of number words shows a very similar pattern of problems, and again, these do not stem from an inability to discriminate along the appropriate dimension (Ramscar, Dye, Popick & O'Donnell-McCarthy, 2011). Discriminatively, the obvious problem here is that while children might encounter 'three apples' or 'red apples,' three and red are never encountered independently. Rather, since these words inevitably occur in ambiguous contexts (Figure 8) children must learn to discriminate the cues to their use in context. As with the learning of the cues to *mouse* and *mice* described earlier, if language is used informatively, children will be able to solve this problem by discriminating a distributed representation of the environmental features that predict the use of various lexical contrasts in context in the code. However, because DISCRIMINATIVE learning relies on cue competition and prediction error, the temporal structure of information is a critical factor in it.

This point is best illustrated by comparing the effects of learning in contexts when complex (multi-feature) stimuli predict discrete linguistic forms (Labels), to its inverse. In the examples described so far in this paper, FEATURES in the world have served as cues to LABELS (FL-learning; Ramscar et al., 2010), an information structure that naturally allows for features to compete as cues to labels. When this relationship is reversed (see Figure 9), such that labels serve as cues (LF-learning), cue competition becomes problematic, because the serial nature of speech means that only one label cue is present at any time. Since a single cue cannot compete with itself, learning ceases to be discriminative, and produces a representation of the probability of each feature given the label instead (Ramscar, 2013; Hoppe et al., 2021, Vujović, Ramscar & Wonnacott, 2021; see also Rische & Komarova, 2016; Ma & Komarova, 2017).

With regards to colours and numbers, although discourse factors may make these temporal relationships rather more complicated in the real world, this analysis predicts that post-nominal constructions will be more likely to facilitate the discrimination of the appropriate system of cues to set of these words than pre-nominal constructions. Empirical results support this prediction, showing that training with post-nominal constructions significantly improves the accuracy and consistency of two-year olds' number and colour word use, whereas pre-nominal training has no effect on their performance. These results also help explain why children struggle to learn colour and number words despite their frequency in the input: in English, where these problems have mostly been studied, children overwhelmingly encounter colour and number words in pre-nominal constructions (Ramscar et al., 2010, 2011).

https://doi.org/10.1017/S0305000921000544 Published online by Cambridge University Press

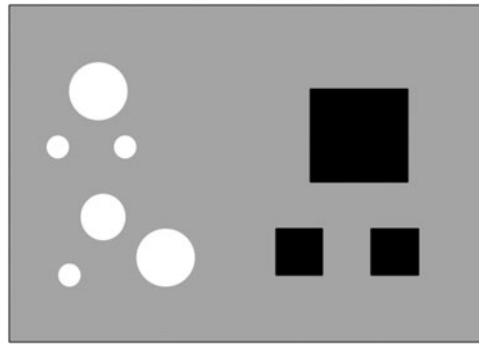

Figure 8. An illustration of the challenge presented by colour and number learning. This picture contains: **six** circles, and **three** squares; **white** circles and **black** squares; and **more** circles than squares / **less** squares than circles; some of the circles and squares are **larger** and some are **smaller**. Somehow children must learn the cues that discriminate between the appropriate and inappropriate use of these words in context.

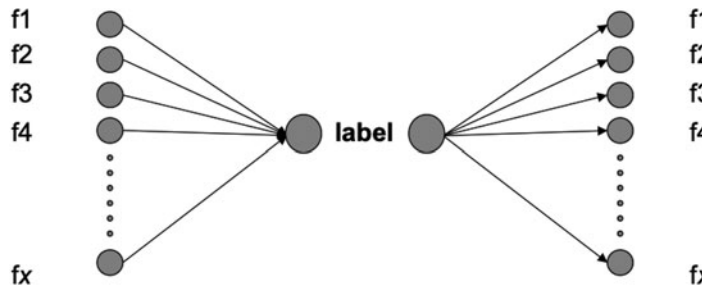

Figure 9. The possible predictive relationships labels (words or affixes) can enter into with the other features of the world (or a code). A feature-to-label relationship (left) will tend to facilitate cue competition between features, and the abstraction of the informative dimensions that predict labels in learning, whereas a label-tofeature relationship (right) will facilitate learning of the probabilities of the features given the label.

However, this analysis raises a question: if sentences like "*red ball*" are unhelpful to learners, why do people use them so often? The answer appears to lie in the specific problems nouns pose in communication. Because humankind has a propensity for inventing things that need names, in most languages nouns comprise a very large set of types. Analyses show that, by splitting them into classes, the German grammatical gender system serves to reduce the uncertainty associated with nouns in context, a function that English (a largely non-gendered language), achieves through its distribution of pre-nominal adjectives (which are more frequent, and more informative about nouns than German pre-nominal adjectives; Dye, Milin, Futrell & Ramscar, 2017; Dye, Milin, Futrell & Ramscar, 2018).

These findings indicate that gender systems may communicate more information about content than is often supposed (and that English pre-nominal adjectives communicate more grammatical information than is often supposed), but more importantly, they also indicate that content information may be far more DISTRIBUTED than compositional theories assume.

https://doi.org/10.1017/S0305000921000544 Published online by Cambridge University Press

Meaning, function, and the distributional structure of codes

The distinction between function and content words that is assumed in many linguistic theories appears to be further complicated by findings showing that measures derived from the distributional patterns of words can accurately predict human behaviour in tasks normally associated with 'semantic processing' (Landauer & Dumais, 1997; McDonald & Ramscar, 2001; Ramscar, 2002; Ramscar & Yarlett, 2003; Johns & Jones, 2010). These models are typically described at a lexical level, as capturing distributional aspects of the 'meaning' of individual lexical items. However, for present purposes, what is important to understand about these models is that they simply measure the conditioning history of a word in relation to the other words in a sample. When two words have the same conditioning histories (if their co-occurrence patterns are identical), then although learning will discriminate them from words that don't share their history, it will not result in their being discriminated from one another. Further, because learning is probabilistic, when a subset of words share conditioning histories that only slightly vary from one another, a learner's expectations about the behaviour of the words within this subset will be far less discriminated from each other than they will be from the rest of the lexical system. The members of this subset will thus form a probabilistic cluster within this system, and a learner's expectations will tend to relate as much to this subset as any individual item within it.

These considerations suggest that, because semantically similar words covary systematically in the lexicon, they will cluster together when they are learned in context. From this perspective, co-occurrence patterns can be seen to discriminate a level of coarse semantic similarities that is at a higher level than content words, and yet informative about them (again, blurring distinctions between 'content' and 'function' words). Which further suggests that all meaningful linguistic communication – including communication about seemingly concrete things like nouns – might be discriminative, based on the same process of incremental uncertainty reduction as names. This hypothesis makes clear, falsifiable predictions*:* If grammar works discriminatively, then for the same functional reasons that name distributions are geometrically distributed, it predicts any other class of words that is systematically encoded in distributional patterns should also be geometrically distributed. Moreover, the theory further predicts that children should be able to learn these classes along with their native language.[5] Initial support for this prediction comes from analyses showing how a set of relatively unambiguous semantic clusters that can be reliably extracted from child-directed speech using a range of co-occurrence model (Asr, Willits & Jones, 2016; see **Table 1**) are geometrically distributed (Ramscar, 2109; **Figure 10**). Assuming these findings generalize to the other lexical subcategories discriminated by co-occurrence patterns in speech (see Figure 5), they can begin to explain how adult and child speakers are able to align their communicative expectations.

[5]This prediction comprises two parts, one relating to the distribution of forms in context, and the other to the abilities of young learners, and whether they are able to learn and use contextual information in the way that adult learners seem to do (MacDonald & Ramscar, 2001). Accordingly, while testing the first part of this prediction will involve the analysis of corpus data (and possibly the development of appropriate corpora), the second part will involve empirical analyses of what children can learn from distributional information, and analysis of the degree to which these capacities support the kind of processing envisaged here.

https://doi.org/10.1017/S0305000921000544 Published online by Cambridge University Press

Table 1: Nouns in two categories defined by context in CHILDES (Asr et al., 2016).

CLOTHING
Clothes, Dress, Suit, Shirt, Coat, Hat, Tie, Jacket, Cap, Belt, Uniform, Hood, Shoe, Skirt, Cape, Purse, Boot, T-Shirt, Shorts, Helmet, Outfit, Sweater, Glove, Gown, Underwear, Robe, Sunglasses, Scarf, Blouse, Vest, Bra, Apron, Buckle, Sock, Diaper, Slacks, Sweatshirt, Nightgown, Tights, Bathrobe, Pant, Bonnet, Sneaker, Slipper, Pajama, Sandal, Undershirt, Bib, Mitten, Snowsuit, Shoelace
BODY
Back, Head, Hand, Hair, Nose, Mouth, Face, Foot, Mind, Bottom, Behind, Finger, Tummy, Eye, Leg, Arm, Ear, Knee, Neck, Tongue, Toe, Heart, Body, Belly, Lap, Bone, Thumb, Chin, Skin, Tooth, Stomach, Nail, Blood, Mustache, Beard, Throat, Chest, Shoulder, Calf, Elbow, Mood, Cheek, Lip, Butt, Wrist, Memory, Brain, Hip, Weener, Ponytail, Forehead, Fingernail, Bruise, Penis, Breast, Ankle, Muscle, Eyebrow, Waist, Eyeball, Liver, Braid, Nipple, Ass, Skull, Toenail, Lungs, Jaw, Vagina, Rib

Learning verbs and arguments

So far this article focused on the learning of systems of lexical items and morphological contrasts, albeit that it has sought to emphasize the importance of context in defining these classes, and the way that the learning of classes of items depends on shared patterns of variance and invariance in relation to other items. From this perspective, a long-studied feature of verbs is particularly notable: namely, that verbs inevitably take arguments, the structure of which often differs systematically across verb classes. Thus unlike nouns and names, which can be characterized as being learned as classes of lexical items, verbs are different. Verb learning seems to be best characterized in terms of the acquisition of classes of arguments (patterns of relations between lexical items), because an aspect of verbs that is relatively invariant across contexts is their relationship to the arguments they occur in.

The idea that distributional patterns might systematically discriminate verbs into coherent subcategories has a long history (Levin, 1993), and numerous theories have been put forward to explain the relationship between the semantic properties of verb subcategories and their different argument structures (Fillmore, 1968; Jackendoff, 1972; Goldberg, 1995), and the way children learn corrects patterns of generalization within them (Gropen, Pinker, Hollander & Goldberg, 1991; Brooks & Zizak, 2002; Ambridge et al., 2013, 2014). For example, constructionist accounts of argument learning propose that a child's knowledge of a language is initially a set of initial fixed patterns, which then develop into semi-productive item-based constructions, before adult competence (in which the scope of some constructions remains limited, while the scope of others seems more open-ended) is achieved.

At a broad level these theories account well for the patterns of behaviour associated with verb argument learning (Goldberg, 1995; Cameron-Faulkner, Lieven & Tomasello, 2003; Tomasello, 2006; Ambridge & Lieven, 2011; Ambridge et al., 2014). However, the mechanisms that they use to explain the developmental progression described above – schematization and analogy – tend to be poorly specified (Beekhuizen, Bod & Verhagen, 2014), as is the relationship between traditional and constructionist ideas about the basic functions of language, such as compositionality (Kay & Michaelis, 2012). By contrast, the processes of 'schematization' and 'analogy' (or at least their discriminative analogues) are clearly specified in learning models. Cue competition

https://doi.org/10.1017/S0305000921000544 Published online by Cambridge University Press

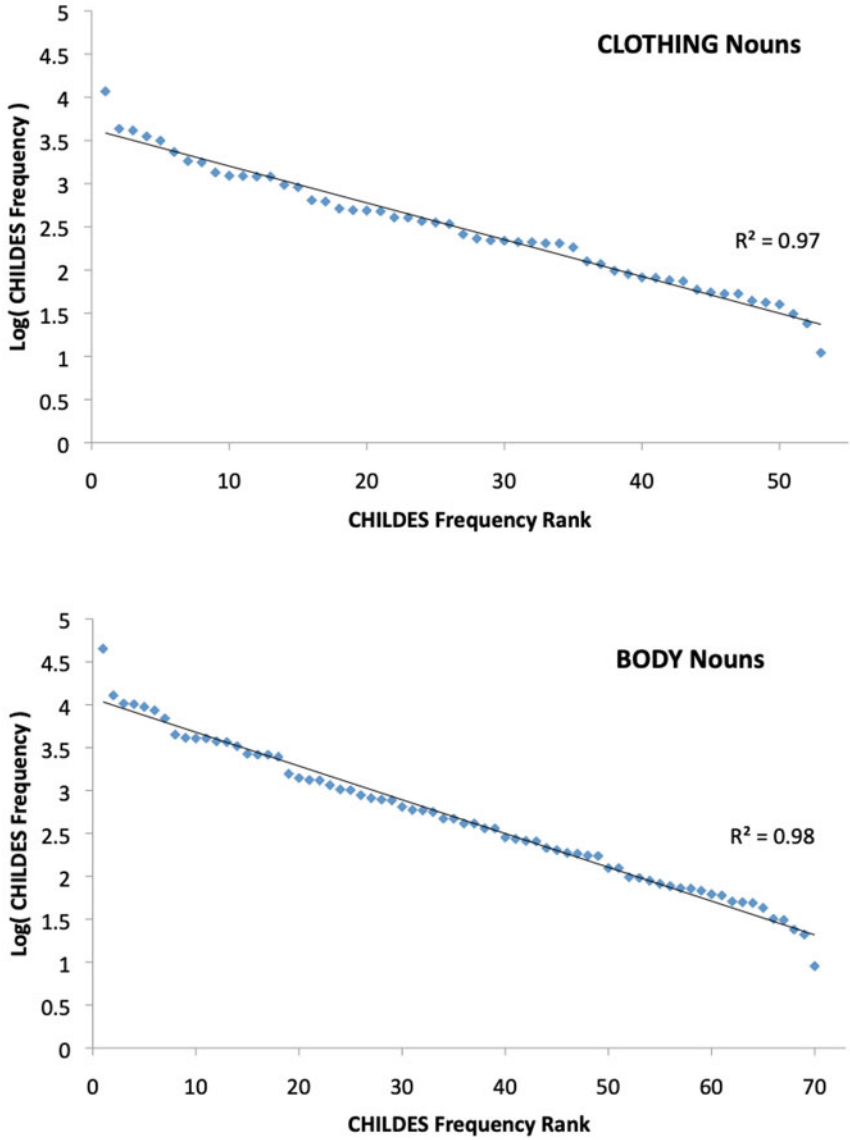

Figure 10. Log frequency x frequency rank plots of the two noun categories extracted from by CHILDES (Asr et al, 2016). As can be seen, both of these categories, which are discriminated by the contexts in which they occur, have a geometric distribution.

leads to abstraction, while overgeneralization / analogy simply represent the probabilistic output of a system given its current state of learning (Ramscar et al., 2013a).

To date, however, there have been very few investigations of verb argument learning from an explicitly discriminative perspective, such that when it comes to accounting for their acquisition, only promissory notes and predictions can be offered in this article (though see Bidgood et al., 2021; Ambridge et al., 2020, for encouraging signs in this

https://doi.org/10.1017/S0305000921000544 Published online by Cambridge University Press

direction). Accordingly, what is notable for current purposes is the degree of compatibility between constructionist accounts of verb argument acquisition and a discriminative learning perspective. One aspect of children's early grammars emphasized by the former is that they appear to comprise a mixture of words, holophrases and 'unanalyzed expressions' (Pine & Lieven, 1997; Tomasello, 2003), such that when language use initially becomes productive, utterances tend to be organized around particular words (Tomasello, 2000). These early constructions typically comprise what Braine (1976) describes as 'pivot words' (specific relational terms, which are typically verbs), and 'open slots,' that are filled in turn by a wider range of words or expressions (typically semantically appropriate nouns). Initially, only the items that are related by pivot words tend to vary, with variance in other words only emerging later, as children's experience grows.

This pattern of learning is highly compatible with the behaviour of discriminative learning models, in which the initial process of association is always maximal: when a child hears a string comprised of a series of novel acoustic contrasts for the very first time, all and any available contextual / semantic cues in the environment will be associated to that string (Arnon & Ramscar, 2012). The discrimination of the more or less discrete 'components' in a system then depends on experience and error; such that the degree to which different aspects of form and meaning are associated in context will be a function of the distribution of the forms and a learner's experience of it (Ramscar et al., 2013a). Moreover, explicitly treating the learning of argument structure as a discriminative process also highlights the way that communicative systems have evolved to support learning: analyses of the empirical distribution of a set of English verb subcategories defined by the shared alternation patterns of their members (Levin, 1993) reveal that like nouns and names, verb arguments are geometrically distributed (Ramscar, 2019, 2020; see Figure 11 for an example).

In the light of this discussion of the discriminability and learnability of forms in morphological paradigms, it is worth highlighting a further aspect of the distribution of verb alternation subcategories: that the most frequent arguments in each distribution tend to involve irregular verbs (strongly marking at least some of the semantic contrasts in the argument), whereas the less frequent arguments are regular, a pattern that will appear to support the appropriate generalization of lower frequency items across the full set of alternations warranted by each subcategory. Given that the account of verb learning put forward here assumes that learning to produce a plural or use an argument differ only in their degrees of complexity, it follows that this discriminative account of the acquisition of verb arguments predicts that the incorrect use of arguments (**don't giggle me*) is not merely the product of a child's failure to adequately learn the specific cues to a particular argument, but also their failure to UNLEARN the generic cues that lead to over-generalization. The theory thus predicts that, at an appropriate stage of development, training children on correct arguments (don't tickle me) should result in a DECREASE in over-generalization errors (*don't giggle me). Failure to find evidence for this would not only falsify this prediction (and the assumptions that inform it), it would also raise serious questions about the account of morphological development described earlier.[6]

[6]Children's production of verb arguments also appears to be U-shaped. While younger children produce sentences that violate normal argument conventions – e.g., saying, "don't giggle me" – they eventually converge on the same model as adults – preferring, "don't make me giggle" – in much the same way as

https://doi.org/10.1017/S0305000921000544 Published online by Cambridge University Press

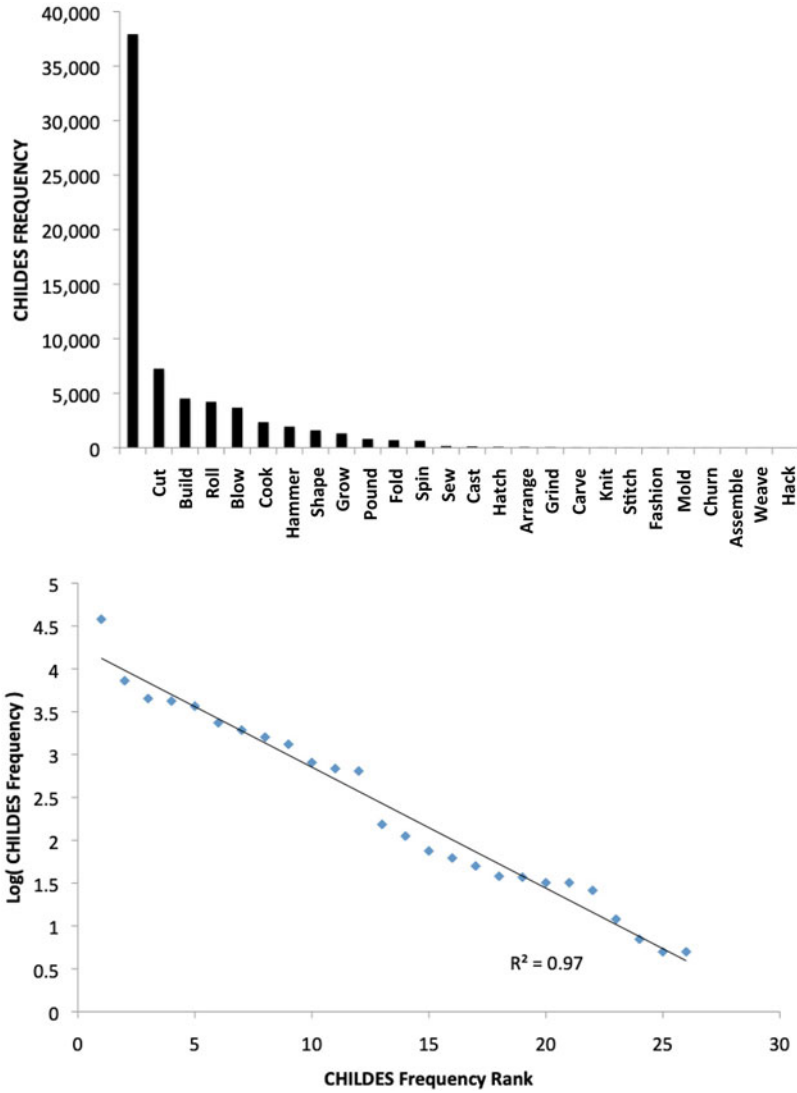

Figure 11. Table 2 shows the verbs in the Build subcategory (Levin, 1993). The top panel plots their frequencies in CHILDES, and the bottom panel shows the fit of these frequencies to a geometric distribution. A comparison of 40 sets of verb alternation patterns to the sets of verbs beginning with the 20 most frequent English letters showed that although the frequency distributions of verbs following letters are Zipf distributed, the frequency distributions of the verbs defined by their alternation patterns are all geometric (Ramscar, 2020).

they come to produce 'mice' rather than 'mouses' (Ambridge et al., 2013). Because the semantic structure of verbs appears to be more abstract than that of noun phrases, and because verbs always tend to appear in arguments, this makes their structure more complex in modeling terms (a point discussed further below): however, because it seems that the same learning principles underlie this pattern of behavior as well, it follows that to be consistent with the approach to learning argued for above, one would have to predict that given an appropriate analysis, an intervention based on the same logic of using a set of dominant

https://doi.org/10.1017/S0305000921000544 Published online by Cambridge University Press

Table 2: The 'Build' verb alternation class (Levin, 1993).

Make, Cut, Build, Roll, Blow, Cook, Hammer, Shape, Grow, Pound, Fold, Spin, Sew, Cast, Hatch, Arrange, Grind, Carve, Knit, Stitch, Fashion, Mold, Churn, Assemble, Weave, Hack

Models of learning and representation

It is clear that modeling the acquisition and processing of verbs is a more complicated task than modeling inflectional morphology or number and colour word learning. While the semantic dimensions of the morphological models described earlier were crude, they succeeded in capturing many of the more important features of the discriminative puzzle solved by a child learning to use plural forms in a useful, informative way (and as Box, 1979, pointed out, usefulness is all scientific models can aspire to). Moreover, the fact that the representations employed in the various models described earlier were actually sufficient for their purpose highlights a commonality in all of them: that seemingly different learning tasks such as inflectional morphology and colour word learning both involve the discrimination of relationships between relatively straightforward perceptual/contextual dimensions and relatively simple forms. The simplicity of this approach came with clear theoretical benefits, in that it is amenable to modeling using simple two-layer networks. In these, the discriminative nature of error-driven learning, the contribution of the input and output representation, and any theoretical insights that might be gleaned from interactions between them, can be made fairly transparent.

By contrast, given what they communicate – causality, agency, manner, etc., often all at once – learning to use verb arguments involves the task of discriminating semantic relationships that are multidimensional, that are distributed across different items that arguments comprise, and that are likely far more semantically abstract than anything considered earlier. Meanwhile, the distributions described above guarantee that learners encounter some argument structures at very high rates, and others hardly at all. It thus follows that in any individual, the degree to which learning will have discriminated any given argument structure along its various semantic / form dimensions is likely to vary considerably at any given time. For example, in the 460+ million word COCA corpus, *make*, the most frequent of the *build* verbs described above, occurs 130,000 times; *embroider*, the least frequent, occurs 5 times. Learners will thus encounter *make* across a wide range of contexts, in which a wide variety of forms will fill its arguments. This will lead them to discriminate very abstract representations of the relationship between *make* arguments and the cues to the forms they comprise. By contrast, *embroider* will be encountered in a far more restrictive set of contexts in which a small set of forms comprise its arguments. Accordingly, learners will acquire representations that are less abstract, such that at the form level *embroider* will be associated more with specific items and less with abstract classes, while, at the semantic level, the cues to *embroider* arguments will be more associated with the whole structure, and less with specific parts of it (in both

forms to generate error in the representations of a set of 'exceptions' ought to result in the same pattern of reduction in over generalization.

https://doi.org/10.1017/S0305000921000544 Published online by Cambridge University Press

cases, this is a consequence of the fact that learners will have had less opportunity to discriminate more abstract relationships at the form and semantic level; Ramscar et al., 2013a). In other words, not only must learners discriminate their representations of the cues to *make* and *embroider* arguments across multiple dimensions (e.g., between forms and forms, and semantics and forms) at multiple levels of abstraction at the same time, but since their experiences of the two different arguments will differ greatly, their representations of them will likely differ greatly as well.

Since the two-layer networks described earlier are ill-suited to modeling interactions at multiple layers of abstraction in learning, it follows that accounting for verb argument learning in children will involve the addressing of some difficult methodological and theoretical challenges. For example, at first blush, multi-layer, deep learning networks appear ideally suited to capturing this kind of complex multidimensional learning (LeCun et al., 2015). They develop representations at multiple layers of abstraction (Kozma, Ilin & Siegelmann, 2018), and appear – at least in principle – to be capable of learning many of the complex relational patterns that characterize human communicative codes (Graves, 2012; Hannun et al., 2014; Sutskever, Vinyals & Le, 2014; Jozefowicz, Vinyals, Schuster, Shazeer & Wu, 2016). However, although multi-layer networks are theoretically capable of the kind of complex, multi-level learning that appears to characterize this domain, this kind of modeling often leads to Bonini's problem (Bonini, 1963), in that understanding exactly how these models actually learn their functions can be as challenging as understanding children's learning itself.

Given this problem, recent attempts to understand the performance of multi-layer networks in language processing tasks by treating them as experimental subjects (McCloskey, 1991) are interesting (see e.g., Futrell, Wilcox, Morita, Qian, Ballesteros & Levy, 2019; Linzen, Dupoux & Goldberg, 2016; Wilcox, Levy, Morita & Futrell, 2018), first because these approaches underline the difficulties involved in actually translating the complexities of learning at multiple levels of abstraction into theoretical insight (Bonini's paradox), and secondly because they suggest that the study of communicative development may be maturing away from straw-man arguments about learnability, and towards the development of accounts of what human communication actually comprises and what learning to communicate actually entails.

As well as requiring the kind of reappraisal of the processes involved in human communication outlined here, it seems inevitable that explaining how children learn to communicate will also involve a reappraisal of what we might expect from theories and models themselves. For example, just as there seems little use in asking whether multi-layer network models of language processing are capable of compositionality if it turns out that human communication is not compositional, the complex architecture of the human brain seems to rule out the idea that human learning can be reduced to a single set of representations that are processed in a uniform fashion within a single model.

Whereas for the purposes of exposition this article has treated 'error-driven learning' as a somewhat monolithic, abstract concept, it seems clear that the brain contains a range of different circuits that are capable of learning from prediction error depending upon the complexity of the stimuli and the temporal dynamics of the learning 'episode'[7] involved (Freedberg, Toader, Wassermann & Voss, 2020). For example, dissociable neurobiological circuits implementing error sensitive learning mechanisms have been identified in the striatum (Schultz, Dayan & Montague, 1997;

[7]The models described earlier treat time as a discrete sequence of events, yet the temporal dynamics of even simple learning scenarios are far more subtle and complex than this (Apfelbaum & McMurray, 2017).

https://doi.org/10.1017/S0305000921000544 Published online by Cambridge University Press

Daw, Niv & Dayan, 2005) and the medial temporal lobe (MTL; Delgado & Dickerson, 2012; Shohamy & Daw, 2014), yet they appear to differ in their temporal sensitivity to prediction error. In tasks that tap people's ability to learn probabilistic associations, the performance of amnesic patients with MTL damage is impaired when response-contingent feedback is delayed, but not when feedback is provided immediately. By contrast, patients with Parkinson's disease (which involves the progressive degeneration of substantia nigra dopaminergic neurons and their projections into the striatum) show the opposite pattern (Foerde, Race, Verfaellie & Shohamy, 2013). These findings suggest that changes in trial durations on the 'same' task can elicit responses from different learning circuits in the brain, resulting in concomitant differences in what is learned.

In a similar vein, the fact that learners simultaneously appear sensitive to both the specifics of individual experiences and abstractions from them indicates that the brain's multiple learning mechanisms produce multiple representations at multiple levels of abstraction (Squire, 2004), suggesting that no single 'language learning model' will suffice to account for the full range of linguistic behaviour observed in individuals. Thus although many theories have suggested that the 'end state' of language learning can be characterized by the acquisition of a particular set of abstract linguistic representations (Gold, 1967; Pinker, 1998) or that abstractions are barely learned at all (Ambridge, 2020), it seems far more likely to be the case that language learning has no end state, and that explaining an individual's communicative capacities requires more than a single model operating on one level of representation.

Rather, in the same way that linguistic units are better thought of as descriptive idealizations as opposed to psychological elements (Ramscar & Port, 2016; Samuel, 2020), we should accept that the linguistic representations posited by all theories of communicative development are wrong, yet some will prove more theoretically useful than others. Accordingly, it is to be hoped the focus of the field of language development can move away from vague questions about 'learnability' to more specific questions about what children learn and how they learn it. To return to an earlier example, the WHAT and HOW of Rumelhart and McClelland's (1986) past tense model are easily stated: it assumed that morphological learning involves mastering the transformation of root forms into past tense forms using phonetic information alone. Given a set of phonological cues representing root forms, it used error-driven learning in order to try to discriminate the set of values that best predicted a corresponding set of phonological representations representing past tense forms. From the perspective of the discriminative models of inflection described earlier, the Rumelhart and McClelland model used the wrong input-output representations (sound-to-sound) and attempted to learn the wrong function (transforming root forms into past tense forms, rather than trying to use context to predict the forms of words, Ramscar, in press). However, the important point here is not so much which of these models is 'right' and which is 'wrong' (though there are numerous reasons to believe that Rumelhart & McClelland's approach was misguided from the outset, Ramscar, in press), but rather what is important is that although these models all use the same basic learning algorithm, they embody very different theoretical models, and these differences matter (Bröker & Ramscar, 2020).

From the perspective outlined here, a child learning to communicate must master a systematic set of mappings between their experiences of 'the world' and a highly structured, resolutely probabilistic system of forms. Since any model of this process is likely to be complex and 'incomplete,' understanding how a specific model

https://doi.org/10.1017/S0305000921000544 Published online by Cambridge University Press

contributes to theory must necessarily involve a discussion of its limitations as well as its capabilities. For example, when it comes to word learning and what it means for a child to 'know a word,' there is a vast difference between learning about *yellow* from lexical co-occurrence data and learning about *yellow* as most children eventually do (as 'embodied' sensory agents, in contexts where objects of various hues are experienced along with various form contrasts). However, it is likely that, for a child, learning that *yellow* serves as a cue to just a subset of words can still be useful in language processing even before the child knows what yellow 'means', because pronominal adjectives can provide statistical information about upcoming nouns even before one understands their semantic relationship to the world (Dye et al., 2017). However to conflate knowledge of the former with knowledge of the latter is a theoretical failure. Rather, when it comes to understanding how language is used and processed in communication, it seems clear that what is required are more subtle and more detailed theoretical approaches that break these different senses of a child's 'knowing' a word down into their component parts. In this review, I have tried to show how information theory and learning theory can offer important tools for the development of these kinds of more subtle and detailed kinds of theoretical description.

Conclusion: learning, communication and discrimination

One of the most important ideas contributed by information theory is the proposal that codewords do not contribute information in isolation, but rather as part of a SYSTEM. The information value of any given codeword is a function of the SET of codewords it belongs to – i.e., the codewords that might be expected to occur in a given context – and it contributes information as a function of the expectancies provided by the system, in an ELIMINATIVE rather than compositional manner. Similarly, one of the most important ideas contributed by learning theory is the proposal that what is learned is not just a function of events that occur together, but also of events that might have been expected to occur but do not. And again, this process works in an ELIMINATIVE manner, discriminating against and eliminating potential associations that result in error so as to positively weight informative cues and negatively weight uninformative cues. In reality, learning is as much about learning to ignore as it is about learning to associate.

To date it seems fair to say that the literature on human communication and its development has not covered itself in glory when it comes to grasping either of these ideas. However there seems to be no principled reason to suppose that progress towards better models of these processes is impossible. What I have sought to outline here is how, when seen through the lens of learning and information theory, much of the structure of natural communications systems begin to make sense, and many of the mysteries of communicative development begin to seem a lot less puzzling. Clearly a more complete discriminative theory of human communication will require more flesh on its bones than this short review can provide. A better alignment between linguistic theory and the appropriate formal models of communication, computation and learning can only help in this regard.

References

Ambridge, B. (2020). Against stored abstractions: A radical exemplar model of language acquisition. *First Language*, *40*(5–6), 509–559.

Ambridge, B., & Lieven, E. V. (2011). *Child language acquisition: Contrasting theoretical approaches*. Cambridge University Press.

https://doi.org/10.1017/S0305000921000544 Published online by Cambridge University Press

Ambridge, B., Pine, J. M., Rowland, C. F., Chang, F., & Bidgood, A. (2013). The retreat from overgeneralization in child language acquisition: Word learning, morphology, and verb argument structure. *Wiley Interdisciplinary Reviews: Cognitive Science*, *4*(1), 47–62.

Ambridge, B., Pine, J. M., Rowland, C. F., Freudenthal, D., & Chang, F. (2014). Avoiding dative overgeneralisation errors: semantics, statistics or both? *Language, Cognition and Neuroscience*, *29*(2), 218–243.

Ambridge, B., Maitreyee, R., Tatsumi, T., Doherty, L., Zicherman, S., Pedro, P.M., Bannard, C., Samanta, S., McCauley, S., Arnon, I., & Bekman, D. (2020). The crosslinguistic acquisition of sentence structure: Computational modeling and grammaticality judgments from adult and child speakers of English, Japanese, Hindi, Hebrew and K'iche'. Cognition, 202, 104310.

Arnon, I., & Ramscar, M. (2012). Granularity and the acquisition of grammatical gender: How order of acquisition affects what gets learned. *Cognition*, *122*, 292–305.

Asr, F. T., Willits, J., & Jones, M. (2016). Comparing Predictive and Co-occurrence Based Models of Lexical Semantics Trained on Child-directed Speech. In *Proceedings of the 38th Annual Conference of the Cognitive Science Society*.

Aylett, M., & Turk, A. (2004). The smooth signal redundancy hypothesis: A functional explanation for relationships between redundancy, prosodic prominence, and duration in spontaneous speech. *Language and Speech*, *47*(1), 31–56.

Baek, S.K., Kiet, H.A.T, & Kim, B.J. (2007). Family name distributions: Master equation approach, *Physical Review E*, *76*, 046113.

Beekhuizen, B., Bod, R., & Verhagen, A. (2014). The linking problem is a special case of a general problem none of us has solved: Commentary on Ambridge, Pine, and Lieven. *Language*, *90*(3), e91-e96.

Bell, A., Brenier, J. M., Gregory, M., Girand, C., & Jurafsky, D. (2009). Predictability effects on durations of content and function words in conversational English. *Journal of Memory and Language*, *60*(1), 92–111.

Bender, E. M., & Koller, A. (2020). Climbing towards NLU: On meaning, form, and understanding in the age of data. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics* (pp. 5185–5198).

Bidgood, A., Pine, J., Rowland, C., Sala, G., Freudenthal, D., & Ambridge, B. (2021). Verb argument structure overgeneralisations for the English intransitive and transitive constructions: grammaticality judgments and production priming. *Language and Cognition*, 1–41.

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

Bohn, M., & Frank, M. C. (2019). The pervasive role of pragmatics in early language. *Annual Review of Developmental Psychology*, *1*, 223–249.

Bonini, C. P. (1963). *Simulation of information and decision systems in the firm*, Englewood Cliffs, N. J.: Prentice-Hall

Borges, J. L. (1946). On exactitude in science. In Borges J. L. (1998, Collected Fictions, Translated by A. Hurley, Penguin Books: New York

Bornstein, M. H., Kessen, W., & Weiskopf, S. (1976). Color vision and hue categorization in young human infants. *Journal of Experimental Psychology: Human Perception and Performance*, *2*(1), 115.

Box, G. E. (1979). Robustness in the strategy of scientific model building. In *Robustness in statistics* (pp. 201–236). Academic Press.

Braine, M. D. (1976). Children's first word combinations. *Monographs of the society for research in child development*, 1–104.

Bröker, F., & Ramscar, M. (2020). Representing absence of evidence: why algorithms and representations matter in models of language and cognition. *Language, Cognition and Neuroscience*, DOI: 10.1080/23273798.2020.1862257.

Brooks, P. J., & Zizak, O. (2002). Does preemption help children learn verb transitivity?. *Journal of Child Language*, *29*(4), 759–781.

Cameron-Faulkner, T., Lieven, E., & Tomasello, M. (2003). A construction based analysis of child directed speech. *Cognitive Science*, *27*(6), 843–873.

Cohen, B. A. (2017). Point of View: How should novelty be valued in science? *Elife*, *6*, e28699.

Culicover, P. W. (1999). *Syntactic Nuts: Hard Cases in Syntax*. Oxford University Press, Oxford.

Davies, M. (2009). The 385+ million word Corpus of Contemporary American English (1990–2008+): Design, architecture, and linguistic insights. *International Journal of Corpus Linguistics*, *14*(2), 159–190.2008)

Daw, N. D., Courville, A. C., & Dayan, P. (2008). Semi-rational models of conditioning: The case of trial order. *The probabilistic mind*, 431–452.

Daw, N. D., Niv, Y., & Dayan, P. (2005). Uncertainty-based competition between prefrontal and dorsolateral striatal systems for behavioral control. *Nature neuroscience*, *8*(12), 1704–1711.

Delgado, M. R., & Dickerson, K. C. (2012). Reward-related learning via multiple memory systems. *Biological psychiatry*, *72*(2), 134–141.

Dye, M., Johns, B. T., Jones, M. N., & Ramscar, M. (2016). The Structure of Names in Memory: Deviations from Uniform Entropy Impair Memory for Linguistic Sequences. Proceedings of the 38th Annual Conference of the Cognitive Science Society, Philadelphia, PA.

Dye, M., Milin, P., Futrell, R., & Ramscar, M. (2017). A functional theory of gender paradigms. In F. Kiefer, J. P. Blevins, & H. Bartos (eds.), *Perspectives on morphological organization: Data and analyses* (pp. 212–239). Leiden: Brill

Dye, M., Milin, P., Futrell, R., & Ramscar, M. (2018). Alternative solutions to a language design problem: The role of adjectives and gender marking in efficient communication. *Topics in Cognitive Science*, *10*(1), 209–224.

Ellis, N. C. (2006). Language acquisition as rational contingency learning. *Applied Linguistics*, *27*, 1–24. doi:10.1093/applin/ami038

Ellis, N. C., & Ferreira-Junior, F. (2009). Constructions and their acquisition: Islands and the distinctiveness of their occupancy. *Annual Review of Cognitive Linguistics*, *7*(1), 188–221

Estoup, J. B. (1916). *Gammes Stenographiques*. Institut Stenographique de France, Paris

Fara, D. G. (2015). Names are predicates. *Philosophical Review*, *124*(1), 59–117.

Fillmore, C. (1968). "The Case for Case," in E. Bach and R. T. Harms (eds.), *Universals in Linguistic Theory*, Holt, Rinehart and Winston, New York, 1–90

Foerde, K., Race, E., Verfaellie, M., & Shohamy, D. (2013). A role for the medial temporal lobe in feedback-driven learning: evidence from amnesia. *Journal of Neuroscience*, *33*(13), 5698–5704.

Freedberg, M., Toader, A. C., Wassermann, E. M., & Voss, J. L. (2020). Competitive and cooperative interactions between medial temporal and striatal learning systems. *Neuropsychologia*, *136*, 107257.

Futrell, R., Wilcox, E., Morita, T., Qian, P., Ballesteros, M., & Levy, R. (2019). Neural language models as psycholinguistic subjects: Representations of syntactic state. *arXiv preprint arXiv:1903*.03260.

Gallager, R., & Van Voorhis, D. (1975). Optimal source codes for geometrically distributed integer alphabets (corresp.). *IEEE Transactions on Information theory*, *21*(2), 228–230.

Gibson, E., Futrell, R., Piantadosi, S. P., Dautriche, I., Mahowald, K., Bergen, L., & Levy, R. (2019). How efficiency shapes human language. *Trends in Cognitive Sciences*, *23*(5), 389–407.

Gleitman, L. R. (1965). Coordinating conjunctions in English. *Language*, *41*(2), 260–293.

Gold, E. M. (1967). Language Identification in the Limit, *Information and Control 10*: 447–474.

Goldberg, A. E. (1995). *Constructions: A construction grammar approach to argument structure*. University of Chicago Press.

Goldberg, A. E. (2003). Constructions: A new theoretical approach to language, *Trends in Cognitive Science*, *7*.5, 219–224.

Graves, A. (2012). *Supervised Sequence Labelling with Recurrent Neural Networks*. Springer, Berlin.

Gray, A. (2014). Name-bearing, reference, and circularity. *Philosophical Studies*, *171*(2), 207–231.

Gropen, J., Pinker, S., Hollander, M., & Goldberg, R. (1991). Affectedness and direct objects: The role of lexical semantics in the acquisition of verb argument structure. *Cognition*, *41*(1–3), 153–195.

Guo, J-Z., Chen, Q-H., & Wang, Y-G. (2011). Statistical distribution of Chinese names. *Chinese Physics B*, *20*.11 118901

Handel, Z. (2008). What is Sino-Tibetan? Snapshot of a Field and a Language Family in Flux. *Language and Linguistics Compass*, *2*(3), 422–441.

Hartley, R. (1928). Transmission of Information, *Bell System Technical Journal 7*, no. 3: 535–63

Hannun, A., Case, C., Casper, J., Catanzaro, B., Diamos, G., Elsen, E., Prenger, R., Satheesh, S., Sengupta, S., Coates, A., & Ng, A.Y. (2014). Deep speech: Scaling up end-to-end speech recognition. *arXiv preprint arXiv:1412*.5567.

Hinton, G. E., McClelland, J. L., & Rumelhart, D. E. (1986). Distributed representations. In D. E. Rumelhart, J. L. McClelland, & PDP Research Group (eds.), *Parallel distributed processing. explorations in the microstructure of cognition*. Vol. 1. Foundations Cambridge, MA: MIT Press.

Hinton, G., Deng, L., Yu, D., Dahl, G., Mohamed, A. R., Jaitly, N., Vanhoucke, V., Nguyen, P., Kingsbury., B., & Sainath, T. (2012). Deep neural networks for acoustic modeling in speech recognition. *IEEE Signal processing magazine*, *29*.

Hoppe, D., van Rij, J., Hendriks, P., & Ramscar, M. (2021). *Order Matters! Influences of Linear Order on Linguistic Category Learning Cognitive Science*, Volume 44, Issue 11, https://doi.org/10.1111/cogs.12910

Hoppe, D. B., Hendriks, P., Ramscar, M., & van Rij, J. (in press). An exploration of error-driven learning in simple two-layer networks from a discriminative learning perspective. *Behavior Research Methods*.

Howes, D. (1957). On the relation between the intelligibility and frequency of occurrence of English words. *The Journal of the Acoustical Society of America*, *29*(2), 296–305.

Huffman, D. A. (1952). A method for the construction of minimum-redundancy codes. *Proceedings of the IRE*, *40*(9), 1098–1101.

Jackendoff, R.S. (1972). *Semantic Interpretation in Generative Grammar*, MIT Press, Cambridge, MA.

Johns, B. T., & Jones, M. N. (2010). Evaluating the random representation assumption of lexical semantics in cognitive models. *Psychonomic Bulletin & Review*, *17*(5), 662–672.

Jozefowicz, R., Vinyals, O., Schuster, M., Shazeer, N., & Wu, Y. (2016). Exploring the limits of language modeling. *arXiv*, *1602*.02410.

Kamin, L. J. (1969). Predictability, surprise, attention, and conditioning. In B. Campbell & R. Church (eds.), *Punishment and aversive behavior* (pp. 279–296). New York, NY: Appleton-Century-Crofts.

Kanwisher, N., McDermott, J., & Chun, M. M. (1997). The fusiform face area: a module in human extrastriate cortex specialized for face perception. *Journal of Neuroscience*, *17*(11), 4302–4311.

Kay, P., & Michaelis, L. A. (2012). Constructional Meaning and Compositionality. In C. Maienborn, K. von Heusinger & P. Portner (eds), *Semantics: An International Handbook of Natural Language Meaning*. Berlin: Mouton de Gruyter.

Kemp, C., & Regier, T. (2012). Kinship categories across languages reflect general communicative principles. *Science*, *336*(6084), 1049–1054.

Kiet, H. A. T., Baek, S. K., Kim, B. J., & Jeong, H. (2007). Korean Family Name Distribution in the Past. *Journal of the Korean Physical Society*, *51*(5), 1812–1816

Kozma, R., Ilin, R., & Siegelmann, H. T. (2018). Evolution of Abstraction Across Layers in Deep Learning Neural Networks. *Procedia Computer Science*, *144*, 203–213.

Kuhn, T. (1962). *The structure of scientific revolutions* (2nd ed.). Chicago, IL: University of Chicago Press.

Lake, B., & Baroni, M. (2018). Generalization without systematicity: On the compositional skills of sequence-to-sequence recurrent networks. In *International Conference on Machine Learning* (pp. 2873–2882). PMLR.

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

LeCun, Y., Bengio, Y., & Hinton, G. (2015). Deep learning. *Nature*, *521*(7553), 436–444.

Levin, B. (1993). *English verb classes and alternations: A preliminary investigation*. University of Chicago Press.

Lieber, R. (2019). 3 Semantics of derivational morphology. *Semantics-Interfaces*, *75*.

Lindblom, B. (1990). Explaining phonetic variation: A sketch of the H&H theory, in W. Hardcastle and A. Marchal (eds.), *Speech Production and Speech Modeling*, Kluwer, Dordrecht, pp. 403–439.

Linke, M., & Ramscar, M. (2020). How the Probabilistic Structure of Grammatical Context Shapes Speech. *Entropy*, *22*(1), 90.

Linzen, T., Dupoux, E., & Goldberg, Y. (2016). Assessing the ability of LSTMs to learn syntax-sensitive dependencies. *Transactions of the Association for Computational Linguistics*, *4*, 521–535.

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

Ma, T., & Komarova, N. L. (2017). Mathematical modeling of learning from an inconsistent source: A nonlinear approach. *Bulletin of mathematical biology*, *79*(3), 635–661.

MacKay, D. J. (2003). *Information theory, inference and learning algorith*ms. Cambridge: Cambridge University Press.

Mandelbrot, B. (1966). Information theory and psycholinguistics: A theory of word frequencies. In P. Lazarsfield & N. Henry (eds.), *Readings in mathematical social sciences*. Cambridge: MIT Press.

Manin, D. (2009). Mandelbrot's Model for Zipf's Law: Can Mandelbrot's Model Explain Zipf's Law for Language? *Journal of Quantitative Linguistics*, *16*(3), 274–285.

Marzi, C., Ferro, M., & Pirrelli, V. (2019). A processing-oriented investigation of inflectional complexity. *Frontiers in Communication*, *4*, 48.

McCloskey, M. (1991). Networks and theories: the place of connectionism in cognitive science. *Psychological Science*, *2*, 387–395.

McDonald, S., & Ramscar, M. (2001). Testing the Distributioanl Hypothesis: The influence of Context on Judgements of Semantic Similarity. In *Proceedings of the Annual Meeting of the Cognitive Science Society* (Vol. *23*, No. 23).

Meylan, S. C., & Griffiths, T. L. (2017). Word forms-not just their lengths-are optimized for efficient communication. *arXiv preprint arXiv:1703*.01694.

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

Newman, M. E. (2005). Power laws, Pareto distributions and Zipf's law. *Contemporary Physics*, *46*(5), 323–351.

Ng, A. Y., & Jordan, M. I. (2002). On discriminative vs. generative classifiers: A comparison of logistic regression and naive bayes. In *Advances in Neural Information Processing Systems* (pp. 841–848).

Nyquist, H. (1924). Certain factors affecting telegraph speed. *Transactions of the American Institute of Electrical Engineers*, *43*, 412–422.

O'Doherty, J. P., Dayan, P., Friston, K., Critchley, H., & Dolan, R. J. (2003). Temporal difference models and reward-related learning in the human brain. *Neuron*, *38*(2), 329–337.

Partee, B. (1984). Compositionality. In F. Landman & F. Veltman (eds.), *Varieties of formal semantics*. Dordrecht: Foris.

Partee, B. H. (2007). Compositionality and coercion in semantics: The dynamics of adjective meaning. In G. Bouma, I. Krämer & J. Zwarts (eds.), *Cognitive Foundations of Interpretation*, 145–161. Amsterdam: Royal Netherlands Academy of Arts and Sciences.

Piantadosi, S. T. (2014). Zipf's word frequency law in natural language: A critical review and future directions. *Psychonomic Bulletin & Review*, *21*(5), 1112–1130.

Piantadosi, S. T., Tily, H., & Gibson, E. (2011). Word lengths are optimized for efficient communication. *Proceedings of the National Academy of Sciences*, *108*(9), 3526.

Piantadosi, S. T., Tily, H., & Gibson, E. (2012). The communicative function of ambiguity in language. *Cognition*, *122*(3), 280–291.

Pine, J. M., & Lieven, E. V. (1997). Slot and frame patterns and the development of the determiner category. *Applied Psycholinguistics*, *18*(2), 123–138.

Pinker, S. (1998). Words and rules. *Lingua*, *106*(1–4), 219–242.

Popper, K. R. (1958). *The Logic of Scientific Discovery*, Hutchinson, London.

Port, R. F., & Leary, A. P. (2005). Against formal phonology. *Language*, *81*(4), 927–964.

Quine, W. V. O. (1960). *Word and object*. MIT press.

Ramscar, M. (2002). The role of meaning in inflection: Why the past tense does not require a rule. *Cognitive Psychology*, *45*(1), 45–94.

Ramscar, M., & Yarlett, D. (2003). Semantic grounding in models of analogy: an environmental approach. *Cognitive Science*, *27*(1), 41–71.

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

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

Ramscar, M., Thorpe, K., & Denny, K. (2007). Surprise in the learning of color words *Proceedings of the 29th Meeting of the Cognitive Science Society*, Nashville, TN.

Ramscar, M., & Dye, M. (2009). Error and expectation in language learning: An inquiry into the many curious incidences of 'mouses' in adult speech. In *Proceedings of the 31st Meeting of the Cognitive Science Society*, Amsterdam, NE.

Ramscar, M., Yarlett, D., Dye, M., Denny, K., & Thorpe, K. (2010). The Effects of Feature-Label-Order and their implications for symbolic learning. *Cognitive Science*, *34*(6), 909–957.

Ramscar, M., & Dye, M. (2011). Learning language from the input: Why innate constraints can't explain noun compounding. *Cognitive Psychology*, *62*(1), 1–40.

Ramscar, M., Dye, M., Popick, H. M., & O'Donnell-McCarthy, F. (2011). The enigma of number: Why children find the meanings of even small number words hard to learn and how we can help them do better. *PloS one*, *6*(7), e22501.

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

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

Ramscar, M., Dye, M., & Hübner, M. (2013c). When the fly flied and when the fly flew: How semantics affect the processing of inflected verbs. *Language and Cognitive Processes*, *28*(4), 468–497.

Ramscar, M., Hendrix, P., Love, B., & Baayen, R. H. (2013d). Learning is not decline. *The Mental Lexicon*, *8*(3), 450–481.

Ramscar, M., Smith, A. H., Dye, M., Futrell, R., Hendrix, P., Baayen, R. H., & Starr, R. (2013e). The 'universal'structure of name grammars and the impact of social engineering on the evolution of natural information systems. In M. Knauff, M. Pauen, N. Sebanz & I. Wachsmuth (eds.), *Proceedings of the 35th Meeting of the Cognitive Science Society*, Berlin, Germany.

Ramscar, M., Dye, M., Gustafson, J. W., & Klein, J. (2013f). Dual routes to cognitive flexibility: Learning and response-conflict resolution in the Dimensional Change Card Sort task. *Child Development*, *84*(4), 1308–1323.

Ramscar, M. (2013). Suffixing, prefixing, and the functional order of regularities in meaningful strings. *Psihologija*, *46*(4), 377–396.

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

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

Ramscar, M., & Port, R. F. (2016). How spoken languages work in the absence of an inventory of discrete units. *Language Sciences*, *53*, 58–74.

Ramscar, M., Sun, C. C., Hendrix, P., & Baayen, H. (2017). The mismeasurement of mind: Life-span changes in paired-associate-learning scores reflect the "cost" of learning, not cognitive decline. *Psychological science*, *28*(8), 1171–1179.

Ramscar, M., Dye, M., Blevins, J., & Baayen, H. (2018). Morphological development. In A. Bar On & D. Ravit (eds.), *Handbook of communication disorders*, Berlin, DE: Mouton de Gruyter, pp 181–202.

Ramscar, M. (2019). Source codes in human communication. arXiv preprint arXiv:1904.03991.

Ramscar, M. (2020). "The empirical structure of word frequency distributions." arXiv preprint arXiv:2001.05292

Ramscar, M. (in press). A discriminative account of the learning, representation and processing of inflection systems. *Language, Cognition & Neuroscience*.

Rescorla, R. A. (1968). Probability of shock in the presence and absence of CS in fear conditioning. *Journal of Comparative and Physiological Psychology*, *66*, 1–5.

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

Rische, J. L., & Komarova, N. L. (2016). Regularization of languages by adults and children: A mathematical framework. *Cognitive Psychology*, *84*, 1–30.

Rumelhart, D. E., & McClelland, J. L. (1986). On learning the past tense of English verbs. In J. L. McClelland & D. E. Rumelhart (eds.), *Parallel Distributed Processing: Explorations in the Microstructure of Cognition*, Cambridge, MA: MIT Press, 216–271.

Rumelhart, D. E., Hinton, G. E., & McClelland, J. L. (1986). A general framework for parallel distributed processing. in D. E. Rumelhart, G. E. Hinton & J. L McClelland (eds.), *Parallel distributed processing: Explorations in the microstructure of cognition*, 45–76.

Samuel, A. G. (2020). Psycholinguists should resist the allure of linguistic units as perceptual units. *Journal of Memory and Language*, *111*, 104070.

Sandhofer, C. M., & Smith, L. B. (1999). Learning color words involves learning a system of mappings. *Developmental Psychology*, *35*(3), 668.

Schultz, W. (2006). Behavioral theories and the neurophysiology of reward. *Annual Review of Psychology*, *57*, 87–115.

Schultz, W., Dayan, P., & Montague, P. R. (1997). A neural substrate of prediction and reward. *Science*, *275*(5306), 1593–1599.

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

Seyfarth, S. (2014). Word informativity influences acoustic duration: Effects of contextual predictability on lexical representation. *Cognition*, *133*(1), 140–155.

Shannon, C. E. (1948). A Mathematical Theory of Communication, *Bell System Technical Journal*, *27*, pp. 379–423 & 623–656.

Shannon, C. E. (1956). The bandwagon. *IRE Transactions on Information Theory*, *2*(1), 3.

Shohamy, D., & Daw, N. D. (2014). Habits and reinforcement learning. In M. S. Gazzaniga & G. R. Mangun (eds.), *The cognitive neurosciences*, pp. 591–603, MIT Press.

Sidwell, P., & Blench, R. (2011). 14 the austroasiatic urheimat: the southeastern riverine hypothesis. *Dynamics of Human Diversity*, *315*.

Siegel, S., & Allan, L. G. (1996). The widespread influence of the Rescorla-Wagner model. *Psychonomic Bulletin & Review*, *3*(3), 314–321.

Song, J. J. (2006). *The Korean language: Structure, use and context*. Routledge.

Squire, L. R. (2004). Memory systems of the brain: a brief history and current perspective. *Neurobiology of Learning and Memory*, *82*(3), 171–177.

Stone, G. O. (1986). An analysis of the delta rule and the learning of statistical associations. In D. Rumelhart, J. McClelland, & the PDP Research Group (eds.), Parallel distributed processing: Explorations in the microstructure of cognition, pp. 444–459, Cambridge, MA: MIT Press.

Sutskever, I., Vinyals, O., & Le, Q. V. (2014). Sequence to sequence learning with neural networks. Adv*ances in Neural Information Processing Systems*. 3104–3112.

Sutton, R. S., & Barto, A. G. (1981). Toward a modern theory of adaptive networks: expectation and prediction. *Psychological Review*, *88*(2), 135.

Thompson-Schill, S. L., Ramscar, M., & Chrysikou, E. G. (2009). Cognition without control: When a little frontal lobe goes a long way. *Current Directions in Psychological Science*, *18*(5), 259–263.

Tomasello, M. (2000). Do young children have adult syntactic competence? *Cognition*, *74*, 209–253.

Tomasello, M. (2003). *Constructing a language: A usage-based theory of language acquisition*. Harvard University Press.

Tomasello, M. (2006). Construction grammer for kids. *Constructions*, *1*(11), 3–23.

Townsend, J. T., & Ashby, E. G. (1983). *Stochastic modeling of elementary psychological processes*. Cambridge, MA: Cambridge University Press.

van Rooij, J. C., & Plomp, R. (1991). The effect of linguistic entropy on speech perception in noise in young and elderly listeners. *The Journal of the Acoustical Society of America*, *90*(6), 2985–2991.

Vujović, M., Ramscar, M., & Wonnacott, E. (2021). Language learning as uncertainty reduction: The role of prediction error in linguistic generalization and item-learning, *Journal of Memory and Language*, *119*, 104231.

Westbury, C., Shaoul, C., Moroschan, G., & Ramscar, M. (2016). Telling the world's least funny jokes: On the quantification of humor as entropy. *Journal of Memory and Language*, *86*, 141–156.

Widrow, B., & Hoff, M. E. (1960). *Adaptive switching circuits (No. TR-1553-1)*. Stanford Electronics Labs: Stanford CA.

Wilcox, E., Levy, R., Morita, T., & Futrell, R. (2018). What do RNN language models learn about filler-gap dependencies? In T. Linzen, G. Chrupała & A. Alishahi (eds.), *Proceedings of the 2018 EMNLP Workshop BlackboxNLP: Analyzing and Interpreting Neural Networks for NLP*, *Brussels*, Belgium, November 2018 pp. 211–221. Association for Computational Linguistics.

Wittgenstein, L. (1953). *Philosophical Investigations*. Oxford: Blackwell.

Yuan, Y. (2002). *Chinese Surnames*. East China Normal University Press, Sanghai, 21–57.

Zipf, G. K. (1935). *The Psychobiology of Language*. Boston: Houghton-Mifflin.

Zipf, G. K. (1949). *Human Behavior and the Principle of Least-Effort*. Cambridge, MA: Addison-Wesley.

Cite this article: Ramscar M (2021). How children learn to communicate discriminatively. *Journal of Child Language* **48**, 984–1022. https://doi.org/10.1017/S0305000921000544