Research Statement – Michael Ramscar

My research seeks to understand how the unique human capacity for thought and language emerge out of the processes of human learning. Whereas cognitive psychology traditionally sees language and thought as being combinatoric mental processes, my work takes the discriminative nature of learning processes as its starting point, and focuses on understanding how thought and language can be understood in discriminative terms. This has led me to exploring the way that these learning processes can inform a structural relationship between form and meaning, to research that focus on uncovering the culturally evolved information structures on which our linguistic and cognitive capacities are founded, and to work that is focused on understanding how the developments of the brain's learning capacities have adapted to take advantage of these structures.

The discriminative perspective that forms the basis of this work offers several benefits: It is compatible with the brain's basic learning mechanisms; It is easily and plausibly formalized; It integrates well with information theory (which, as Claude Shannon [1] emphasized, is also based on a discriminative model); And it continually generates successful predictions about previously unobserved behavior.

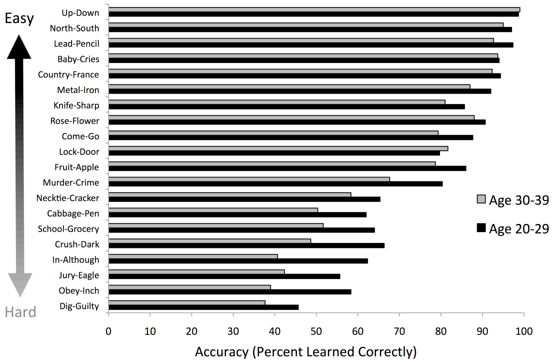
Although discriminativism has a long history in experimental psychology, it presents a theoretical alternative to combinatoric cognitive theories when applied to high-level cognition, and challenges historical ideas about the mind typically adopted in cognitive science. Accordingly, I have sought to apply this approach broadly in my research, relating it to lifespan development, formal and computational models of learning and communication, the brain bases of behavior, the synchronic and diachronic evolution of language, and the world outside of my lab. My purpose in doing so is *not* to show that discriminative models of cognition are 'right.' As George Box noted, all models are wrong, and their value is in their utility. Rather, I have sought to illustrate the *value* of discriminativism to research psychology and linguistics, its practical relevance to everyday life, and its theoretical utility to other areas of science.

This brief description of this work is organized around the broad axes that comprise this approach. First, I describe discriminative learning, and why it is a good starting point for understanding human cognition. I illustrate this by explaining how it provide a good account of what appears to be a canonical combinatoric process, paired associate learning, and how it yields surprising and successful predictions in this domain. Second, I describe how it accounts for the way children learn various aspects of language. Third, since all animals are capable of discriminative learning, I describe how the development of learning processes can explain why children learn language while their pets don't. Finally, I describe the distributional structures embodied in human communication systems, and describe how a discriminative approach can help account for aspects of language that have traditionally been ignored by researchers in the combinatoric tradition, including the semantics of person names, and the function of grammatical gender.

1. Discrimination learning

Associative learning is one of the best-understood processes in psychology and neuroscience. A critical finding in this vast body of research is that learning an association between a behavioral cue (e.g., hearing a *bell*) and a behavioral outcome (getting *food*) is not a simple process of mentally counting how often hearing the *bell* is followed by *food* (the association rate). Two other quantitative factors, the background rate of the cue (how often the *bell* is heard without *food*) and blocking (the predictability of the *food* given the *bell*) have been found to be critical to determining how informative cues are learned. While high association rates tend to promote learning—which is why this process has historically been called 'associative learning'—blocking and background rates serve to inhibit and shape the formation of associations through a process called cue-competition. The way a dog learns to associate a bell with food is best explained as a process in which the dog's mind implicitly assigns predictive value to the bell. And because the predictive value of the bell is affected by everything else in the dog's environment [2], computationally, what we traditionally think of as an 'associative' process has turned out to be better described in discriminative terms.

To illustrate how this process works, the figure below plots data from a paired-associate learning (PAL) test [3], in which people hear arbitrary word pairs like *north–dog,* and later have to recall *dog* given *north.* The mean recall accuracy of 20-29 year-olds here is significantly better than that of 30-39 year-olds:



Note that the old-young shift in the plot is not linear. As compared to easy PAL items (*up-down*), hard items (*dig-guilty*) get harder with experience. While this is not easily explained by simple ideas about cognitive decline, it is actually predicted by discriminative learning models. These models assume that knowledge is represented as a fully connected graph, with every cue linked to every outcome, and that learning serves to minimize the mismatch between what experience teaches a learner to expect, and the events that actually occur. This means that each time a learner hears *jury-duty* it will slightly diminish her expectation for *jury-eagle.* Depending on the structure of the environment, these experiences may even lead a learner to negatively expect *eagle* given *jury* (i.e., learn a negative association between the two). This predicts experience will make learning hard pairs harder – because the negative expectations formed for hard pairs will have to be unlearned before the pairing can be learned – and consistent with this, a regression analysis of the data in the plot above using corpus derived parameters for the association rates, background rate and blocking factors for these words pairs accounts for over 85% of the variance in both groups, suggesting both are similarly sensitive to the same underlying relations in the items [5]. Performance in other age ranges can be explained similarly [6], and analyses show that the critical change in performance across the lifespan is the increased influence of the parameters associated with learning negative expectations. This suggests that these data do not reveal a decline in learning performance, but rather that older adults' greater experience of English is making their understanding of the English words that do – and critically – do *not* go together much better discriminated. Collaborative work with Ching Chu Sun and Peter Hendrix (Tübingen) further supports this suggestion: when tested in German, the PAL performance of older bilingual Chinese speakers—who have far less experience of German—is far *better* than age-matched native German speakers [5].

2. Learning about the World

In another line of work, I have shown how conceptualizing language learning in discriminative terms can account for a range of puzzling findings. For example, Darwin first noted the specific problem children have in mastering the meanings of color words. Seen discriminatively, this problem stems from the ubiquity of color, which renders color words heard in isolation uninformative—in any given multi-hued context, what does "green" denote? I have shown that once children learn nouns, the properties of the objects referenced by nouns can serve as cues to subsequent color words, enabling children to make use of the systematic covariance of colors across objects and scenes to discriminate which hue goes with which color word. My model of the problem successfully predicts that color word learning is facilitated by post-nominal ("the dog is brown") but not pre-nominal constructions ("the brown dog") [7]. It also accounts for why the distribution of post-nominal color constructions in English and Spanish is identical, despite there being very large differences in the overall distribution of color words in the two languages [8].

A similar analysis accounts for the development of a core mathematical ability, subitization—our ability to 'see' sets of, say, three objects without counting [9]. The analysis makes use of two basic facts: a) that sets become harder to discriminate as they get larger (proportionally, 1 is easier to discriminate from 2 than 11 is from 12), and b) that the distribution of numbers in language is highly skewed: 1 is far more frequent that 2 which is more frequent than 3, and so on. If we take linguistic mentions of set size as a rough measure of its behavioral relevance, this suggests that even in the absence of language, learners will encounter far more contexts in which small set sizes are relevant than is the case for larger sets. In simulations in which the confusability problem is represented in this way, and training is distributed as it is in language (i.e., in which the easiest discriminations are trained the most), sets 1 through 4 are learned relatively easily, after which performance begins to degrade, and the discrimination of sets greater than 7 is highly unreliable—a pattern which closely approximates adult subitization performance. The model also successfully predicts the training contexts that can best support the use of language in *improving* children's subitization skills, which appear to be an important precursor to later mathematical ability.

Finally, modeling language learning as a process in which the semantic dimensions of experience are discriminated by (and in turn serve to discriminate) predictable linguistic regularities provides an alternative account of plural over-regularization (children saying *mouses* rather than *mice*). The model accounts both for why over-regularizations happen and why they stop, and also critically predicts the kinds of experimental interventions that can influence this process. In the model, an irregular form like *mice* is more semantically specific than a regular form like *cats*. As a result, the production of *mice* will suffer interference from erroneous expectations that are initially supported by the overall cue structure. (Such as, e.g., 'whenever multiple objects are present, expect a sibilant to follow the word form associated with the particular semantics.') At the same time, the model is faced with the challenge of discriminating when *mouse* and *mice* are contextually appropriate (because clearly, they share semantic dimensions). Compare this to a child learning the correct use of *cat* and *cats*. About 70% of the noun tokens a child hears are singular, which will support the learning of *cat*; and even if the child fails to discriminate the exact semantic cues to *cats* (multiple cat things), the likely over-hypothesis is going to support the production of the correct form, *cats*. Meanwhile, the statistics of English will render errors like *catses* highly unlikely. Formally modeled, this analysis makes a surprising prediction: For much the same reasons that learning about *jury-duty* serves to make *jury-eagle* increasingly less likely, at a certain point in the model's development, hearing *cats* generates implicit error that makes *mouses* less likely, thereby decreasing the tendency to over-regularize. Empirically, this prediction appears to be right: a study I conducted with Melody Dye (Indiana University) and Stewart McCauley (Cornell) found that training older children on regular plurals did indeed bring about a reduction in the rate at which they over-regularized irregular plurals [10]. Importantly, while this model makes use of exactly the same learning rule as Rumelhart & McClelland's 1986 past tense model, it differs dramatically in its framing of the problem. It thus both provides a very different insight into phenomena at stake, and highlights an important distinction—between algorithms and the way they are conceptualized in models—that is often overlooked in the psychology literature.

3. Language and the Development of Learning

This line of work has examined the differences in the ways children and adults learn. A recent experiment is illustrative of the basic approach [11]: In it, children were trained on novel word meanings while the background rates of the objects paired with the accompanying labels were manipulated. Children first saw two different novel objects 'A' and 'B' together, and heard them labeled ambiguously as a "DAX." Subsequently, 'B' was presented with a new object, 'C,' and another ambiguous label – "PID." This training was repeated, and then the children were presented with all three objects, and asked to identify the "DAX," the "PID," or a "WUG" (which they hadn't heard before). Because 'B' occurs with "PID" and "DAX," it has a higher background rate than 'A,' making 'A' more informative about "DAX" than 'B'. For the same reason, 'C's informativity about "PID" is greater than 'B's. And importantly, B's higher background rate makes it less informative about "WUG" than either 'A' or 'C' (which are equally informative about "WUG"). In informational terms, 'A' is DAX, 'C' is PID, and 'A' or 'C' are WUG. The two-year-olds we tested agreed.

While the children matched the objects and labels based on informativity, adults did not. They agreed about 'A' and 'C' but chose 'B' as the WUG—presumably because A and C were taken. However, while the adults here seem to have behaved "logically," the model we used to predict the children's behavior is the exact same one that proved so successful at accounting for adult PAL learning. Why the difference? The obvious answer is that in this small, constrained task, adults think hard about their answers, and in the course of doing so, over-ride whatever they learned about the informativity of the objects in training. This is especially interesting because, when considered from a discriminative perspective, communication relies on speakers and listeners sharing a common code. A listener uses a speaker's words, intonation, gestures, etc. as cues to help discriminate what a speaker intends to mean from what might have been meant. Although the codes shared by speakers need not be identical—not least because in most, if not all, speech acts a sender intended to bring about learning in a receiver—it does require them to be proximate. This raises an important question. In the models described in the previous section, I assumed that learners faithfully estimate the informativity of their environment. Yet the adults in the labeling task actively ignored the informativity of the cues they were trained on. Given that this is the case, how do language users manage to develop proximate models in the first place?

The answer, I suggest, lies in the fact that the ability to select amongst specific responses—and ignore informativity—is a function of the brain's frontal regions; and of course, in humans, these regions are very slow to develop. In a complementary line of work [12,13], I consider how the nature of prefrontal development serves an adaptive function in the acquisition of language and culture. Because the development of the top-down learning of the kind exhibited by adults in this experiment is delayed, all normally developing children initially sample their linguistic and cultural environments in much the same naïve bottom-up way. As a result, when the structure of those environments is regular—as linguistic data tends to be—they develop very similar models of their environment. Empirically, the PAL data I described above lends support to this idea: the fact that corpus data can predict averaged adult behavior so well across the lifespan indicates that linguistic data is indeed highly structured, and that typically developing adults converge on similar language models. My suggestion that this is facilitated by early sampling constraints on learning offers a mechanistic recasting of the widely accepted relationship between language and theory of mind. From this perspective, theory of mind is not what drives language development. Rather, naïve sampling causes children to develop the highly similar linguistic and cultural models critical to language acquisition, and by dint of making the actions and intentions of others more predictable, helps develop theory of mind. This idea may also help explain why, in disorders where prefrontal development is accelerated—such as autism [14]—language and theory of mind development are both impaired.

4. The discriminatory nature of human communication

The three lines of work above laid the foundations for my most recent work, which is best described as uncovering the nature and structure of the cultural software that feeds into the functional system of learning and processing examined above, and which serves to facilitate human communication and thought.

Beginning with recent work from Korea and China showing that Sinosphere family names are exponentially distributed [15], I have shown how the naming systems of all the world's languages appear to share an interesting property: In Korean, family names – which come first in speech – convey the least information (i.e., *Kim Yong-il*, *Kim Il-sung*), whereas in English it is given names – *Michael* – that are uninformative. The sets of first name tokens are always vastly smaller than the sets of later names, and they are exponentially distributed (indeed, historically, the distribution of English given names and Korean family names was identical). This is notable because although lexical distributions have long been thought to follow power laws, and many suggestions have been made to explain this [e.g., 16], there are many reasons to believe that exponential distributions are preferable to Zipfian power distributions in communication, not least because: 1. they are optimal for the purposes coding and 2. they are memory-less [1]. That is, 1., it can be proven that the average message length encoding a set of items – in terms of communicated bits – can be minimized if they are distributed exponentially; and 2., the exponential distribution is unique in being the only continuous distribution to have the memory-less property. This property is usually explained in terms of waiting times: If wait times between encounters with other individuals are distributed exponentially, then because of the way exponential distributions interact with the laws of conditional probability, we can prove that the probability of encountering someone after any waiting period t_n is independent of the time that has elapsed since t_1, the time a last person was encountered. This means that similarly, in a situation where names are distributed exponentially, the chances of randomly being introduced to someone named *John* at any given point in time is independent of whether or not the last person you were introduced to was called *John*. In other words, empirical distributions of names both maximize coding economy, and if we assume that names implement a Markovian process that iteratively serve to discriminate identities, then their memory-less properties maximize the likelihood that people sampling from name distributions will all learn the same models of them.

For names at least, this offers a solution to a key question in psychology and linguistics: how do learners ever manage to converge on the grammar that makes spoken communication possible? Further, empirically, names are a relevant linguistic distribution: Although any name can appear after *I'd like you to meet...!,* most other nouns will <u>not</u> occur here, which means a hearer has partial semantic prior to a name (a name is coming…), and the parts of the name then serve to reduce semantic uncertainty about a specific identity. I then examined whether other, similarly relevant distributions share this property. This work has shown that the distributions of the sets of items in English verb alternations [17] and the distributions of nouns that condition on the variance in lexical collocations in child directed speech (in CHILDES) are all exponentials. Taken together with the results I described above, these findings support a very different model of language to that normally assumed in cognitive science, since they further suggest that all of human communication involves the same kind of discriminative processing as names, in which linguistic signals serve to reduce semantic uncertainty, instead of "conveying" meanings. Thus while linguistic signaling can be thought of in compositional terms from this perspective (in that signaling makes use of sequences of discrete form contrasts), the process that hearers use to comprehend linguistic messages (i.e., meanings) is discriminative. Hearers do not 'extract' meanings that are combinatorially 'encoded' in signals, as linguists have long supposed, but rather it appears that hearers use signals to help them discriminate the semantic message speakers intend them to receive from the messages they might have sent (which is exactly how, Shannon [defines 'communication' 1]). What makes this communicative process possible is the fact that the speaker and listener both share the richly structured common codes that systematically patterns forms with meanings that my recent work describes.

5. <u>Further applications of this approach</u>

5.1 Grammatical gender: historically linguists have considered grammatical gender systems—which assign nouns to seemingly arbitrary classes—to be useless linguistic artifacts. However, from a discriminative perspective of communication the role of gender immediately becomes apparent: It helps to reduce uncertainty about upcoming nouns. Artificial language learning studies with Inbal Arnon (Hebrew University) [18] have shown how the gender of nouns facilitates their later retrieval, and a series of corpus studies with Melody Dye, Richard Futrell (MIT) and Petar Milin (Sheffield) indicate that the gender of most German nouns is predictable from their acoustic and semantic properties. What makes German gender seem perplexing is that this is *not* true of high-frequency nouns. Yet their gender assignment is neither random, nor arbitrary; rather, the likelihood of two high frequency nouns appearing in similar lexical contexts is predictive of their being in <u>different</u> gender classes. Seen from this perspective, German gender looks like any other grammatical system: high frequency types forming an exception class which maximizes discriminability, whereas simply because of their lower frequencies, types which cannot be lexicalized behave paradigmatically.

5.2 Adult lexical processing: Work with Harald Baayen and our group in Tübingen has extended these ideas to adult language processing. In so doing, we have developed the notion of a "lexome" to deal with some of the problems that simple ideas about "words" present when considered across the lifespan. The theoretical basis of the lexomes is best illustrated by considering an example: As young children, we have only a rudimentary grasp of the various regular plural allomorphs (e.g., the different realization of the final sibilant in *cats* and *dogs*). However, adults not only discriminate a range of plural allomorphs (*dog/z/* versus *cat/s/*), but when played recordings of plural noun from which the plural morpheme segment are deleted (i.e. *cat/s/* with the */s/* deleted), an adult will hear a weird plural, <u>not</u> a singular. This is because, in normal speech, adults articulate the 'common' phonetic features of *dog* and *dogs* differently. Given that linguistic knowledge seems to be continuous, lexomes serve to discretize useful—if artificial—'points' in linguistic systems for descriptive and modeling purposes. For example, our models of reading assume that orthographic cues (and context) serve to discriminate lexomes, and when trained on large samples of corpus data, they have proven adept at predicting and explaining a broad range of psycholinguistic data [15,16].

References

[1] Shannon, C. (1956). The Bandwagon. *IRE Transactions on Information Theory*, 1(2), 3.

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

[3] des Rosiers, G., & Ivison, D. (1986). Paired-associate learning: Normative data for differences between high and low associate word pairs. *Journal of Clinical Experimental Neuropsychology*, 8, 637.

[4] Danks, D. (2003). Equilibria of the rescorla-wagner model. *Journal of Mathematical Psychology*, 47(2), 109

[5] Ramscar, M., Sun, C.C., Hendrix, P. & Baayen, H. (2015). The mismeasurement of mind: Why neuropsychological test results overestimate cognitive decline.

[6] Ramscar, M., Hendrix, P., Love, B. & Baayen, H. (2013) Learning is not decline: The mental lexicon as a window into cognition across the lifespan. *The Mental Lexicon*, 8(3), 450

[7] 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

[8] Ramscar, M. (2014), June 23). The errors in my answer to Darwin [Web log post]. Retrieved October 4, 2015 from https://ramscar.wordpress.com/2014/06/23/the-errors-in-my-answer-to-darwin/

[9] 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. doi:10.1371/journal.pone.0022501

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

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

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

[13] Thompson-Schill, S., Ramscar, M., & Chrysikou, M. (2009). Cognition without control: When a little frontal lobe goes a long way. *Current Directions in Psychological Science*, 8(5), 259

[14] Courchesne, E., Mouton, P. R., Calhoun, M. E., Semendeferi, K., Ahrens-Barbeau, C., Hallet, M. J., & Pierce, K. (2011). Neuron number and size in prefrontal cortex of children with autism. *Jama*, 306(18), 2001-2010.

[15] Kim, B. J., & Park, S. M. (2005). Distribution of Korean family names. *Physica A: Statistical Mechanics and its Applications*, 347, 683-694

[16] Zipf, G. K. (1949). *Human behavior and the principle of least-effort*. Philadelphia, PA: Addison-Wesley

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

[18] Arnon, I. & Ramscar, M. (2012). Granularity and the acquisition of grammatical gender: How order of acquisition affects what gets learned. *Cognition*, 122(3), 292-305

[19] Baayen, R.H., Hendrix, P., & Ramscar, M. (2013). Sidestepping the combinatorial explosion: Towards a processing model based on discriminative learning. *Language and Speech*, 56(3), 329

[20] Baayen, R.H. & Ramscar, M. (2015). Abstraction, storage, and naive discriminative learning. In Dabrowska, E. and Divjak, D. (eds.), *Handbook of Cognitive Linguistics*, 99-120. De Gruyter Mouton