Précis of *Bridging levels of analysis: Learning, information theory, and the lexicon*

Melody Dye

1. Overview

Natural languages are hybrid systems, products of both a common biological endowment (shared across languages) and a particular ecological niche (specific to a particular language). That shared endowment – the architecture of the human nervous system – serves as a powerful constraint on how languages vary and evolve. Nevertheless, the world's languages exhibit remarkable diversity in sound, meaning, and structural organization. Such diversity complicates the search for the invariant universal properties that underpin the uniquely human capacity for language (Christiansen & Chater, 2008; Evans & Levinson, 2009).

In my dissertation work, I take the view that human languages are the end point of complex processes of cultural evolution, occurring over generations, and that their features can thus be analyzed as adaptive solutions to a complex constraint-satisfaction problem. This framing guides my investigation of the challenges posed by building and maintaining a functional lexicon. In particular, the projects detailed here seek to understand:

(1) how cognitive principles of learning and memory serve to constrain cross-linguistic variation,

(2) how social and historical contingencies select for certain designs, and

(3) how different 'design' choices can incur trade-offs between early acquisition and adult processing.

The dissertation is highly interdisciplinary in both its topical coverage and its techniques. It explores questions in language processing, typology, and evolution, and capitalizes on large-scale corpus analyses, behavioral experimentation, and computational modeling as the varied means of

investigation. Information theory, the theoretical lens under which these investigations are conducted, serves as a central, organizing principle.

1.1 Information Theory as Rational Analysis

On the cultural transmission model (Becker et al. 2009; Tomasello, 2003), a language's structural features are subject to selection pressures, and variation among languages results both from random drift and selective adaptation to variable circumstances. Within this framework, a critical question is how to establish the 'fitness' of a particular linguistic feature. Information theory (Shannon, 1948) supplies a functional answer: Its theorems, and their correlates, specify how to construct a maximally efficient code (such that communication proceeds as rapidly and reliably as possible) and how to quantify the extent to which a given code deviates from this theoretical maximum. One means of measuring a feature's fitness is thus in terms of its communicative efficiency.

This approach makes the assumption of ***rationality***—that languages behave as optimal solutions to the communication problem speakers face (Anderson & Schooler, 1991). However, this is not to imply that the solutions that different languages converge on are 'equally' optimal to some pre-specified degree: Evolutionary processes achieve local (rather than global) optima, and are chained to their particular historical lineage (Simon, 1989). Rather, the idea is to provide an overarching framework in which the host of interacting variables may be arrayed, so as to better understand how the system maintains and restores a functional equilibrium. In particular, it allows us to ask: How are the perturbations in one part of the system balanced by compensating forces in another? This mode of inquiry can help uncover how languages use different means to nevertheless

achieve similar functional ends, and the potential trade-offs—in terms of complexity and efficiency—that these different strategies may incur (Pellegrino, Coupé, & Marsico, 2011).

Importantly, information theory is a 'computational-level' theory (Marr, 1982), which functions primarily as a descriptive tool—useful for characterizing the properties of language as a communicative code, rather than pointing to the underlying cognitive mechanisms that might generate or interact with such a system. Its principles are not meant to supplant mechanistic frameworks, but rather to show why the design they implement is rational.

To date, the majority of research adopting this approach has focused on cataloging how speakers—in their utterances—and languages—in their design—conform to information theoretic principles. By contrast, comparatively little work has been done to explicate how fundamental cognitive mechanisms, like learning and memory, give rise to such rational behavior. Showing *why* speakers conform to such principles is an important correlate to showing *that* they do. A communication system, no matter how efficiently coded, must also be possible for humans to learn and to use. The studies detailed in this dissertation were conceived and analyzed with an aim to begin to bridge the gap between these levels of explanation.

1.2 Information Theory & Communication Systems

Information theory presents a set of solutions for engineering digital communication systems (Shannon, 1948). Within the framework it provides, the fundamental problem of communication concerns how to code a message generated at one point (the source) into a physical signal that can be transmitted to another point (the destination) as efficiently and reliably as possible. Successful communication relies on both the sender, at the source, and the receiver, at the destination, sharing sufficiently similar codes, such that the receiver can reconstruct the

original message from the received signal.

Communication thus entails both that there be a set of ***alternatives*** to select among to communicate, and that there be a ***code*** systematically relating these alternatives to a physical signal. This framework can be readily extended to natural language: The store of concepts known to a community of speakers provides the space of communicable alternatives, which are, in turn, mapped to a shared linguistic code. The code consists in a finite stock of sounds, gestures, and written symbols, and internalized norms for their selection, ordering, and combination. Human communication can thus be seen as a probabilistic enterprise, in which speakers and listeners cooperate in order to discriminate the intended message from possible alternatives. Much like the forking branches of a ***decision tree***, each communicative act serves to iteratively reduce uncertainty, further pruning the space of possibilities (Ramscar & Baayen, 2013).

Given a sequence of speech or text, information theory specifies how to identify the uncertainty at a given point about what will come next, and the extent to which that uncertainty is subsequently reduced by what follows. The more freedom the speaker has in selecting among alternatives, the greater the uncertainty—formally known as entropy. Once a choice has been made, the information in the signal represents the amount of uncertainty that has been reduced. Information can also be understood as the predictability of a particular choice in context, given the range of available possibilities. The less predictable the choice, the more informative it is.

Formally, given a particular time point in a sequence, with a set of n equally likely continuations, each with probability $p_x = 1/n$ of being selected, the self-information (or surprisal) of a specific outcome x, can be written as:

$$I(x) = log_2(n) = log_2\frac{1}{p_x} \tag{1.1}$$

When the outcome probability is used, the equation generalizes to non-uniform distributions.

The entropy *H* captures the expected value of information over all outcomes:

$$H(p) = H(X) = -\sum_{x \epsilon X} p(x) log_2 p(x) \tag{1.2}$$

Identifying ***information*** with ***predictability***, and ***entropy*** with ***uncertainty***, establishes a clear bridge between psycholinguistics and information theory. Like other aspects of human cognition, language comprehension and production are incremental, predictive processes. In making predictive inferences about upcoming speech or text, communicators draw on multiple sources of linguistic information, including lexical, semantic, and discourse-level (Pickering & Garrod, 2007). Communicators appear sensitive to the likelihood both of individual words in context (Conway et al., 2010), and of larger units, such as multi-word phrases (Bannard & Matthews, 2008), syntactic clauses (Reali & Christiansen, 2007), and constructions (Goldberg, 2006). Likewise, points of greater uncertainty correlate with difficulties in comprehension and production (van Rooij & Plomp, 1991).

Information theory offers psychological research a model-independent means of characterizing predictability and uncertainty mathematically. Measures of information can thus be used to compare the predictions of alternative models (e.g., comparing probability estimations derived from a recurrent neural network vs. a probabilistic context-free grammar; Frank & Bod, 2011). However, information theory is more than simply a unifying mathematical framework. It can also be used to derive empirical predictions about how natural languages should rationally be organized. In particular, it makes concrete predictions about (1) the distribution of uncertainty over elements in a sequence, and (2) the shape of the likelihood distribution at a particular point in the sequence.

2. Rational Hypotheses

This dissertation takes a three-pronged approach to understanding language from an information theoretic perspective. The first line of investigation tests novel predictions suggested by this model of communication, assessing the fit between its rational principles and the design of various morphological and syntactic features. A second and complementary line of inquiry draws on behavioral research to bridge between these design principles and behavior. The third strand deals with limits on the application of information theoretic principles, and identifies where we might expect to find them.

2.1 The Entropy Rate Principle

One of the central empirical predictions made by an information theoretic approach concerns how uncertainty should vary over the speech stream. Formally, the most efficient means of transmitting information across a channel is at a constant rate at (or approaching) the channel's capacity (Shannon, 1948). Thus, if human speakers structure their utterances optimally, they should distribute uncertainty evenly across discourse—a rational principle known as the ***entropy rate constancy*** hypothesis (Genzel & Charniak, 2002). Speakers appear to conform to this principle in their utterances, and languages in their design, and selectively increasing uncertainty over a given element predictably disrupts processing and production (see Jaeger & Tily, 2011 for a review).

Chapter 2 illustrates how two closely related languages comply with its maxims by smoothing the peaks in entropy that occur over nouns. **Chapter 3** shows how deviating from this information structure impairs the learning and recall of verbal sequences. **Chapter 5** reframes the principle in mechanistic terms and finds that "there is no free lunch": While rate-constancy

facilitates efficient processing and precise recall, it hinders semantic learning of the elements that comprise it.

2.2 Distributional Efficiency

According to the entropy rate principle, uncertainty should remain relatively constant across speech and text. A closely related question then, is how the probability distribution of each element should be coded. Empirically, this question is not difficult to answer: the universal scaling law for word frequencies, commonly known as Zipf's Law (1949), is one of the most striking and robust regularities that language exhibits.

The law states that frequency distribution of words in a given context will approximate an inverse power law, in which a steep decline in frequency over the highest ranked items eases off as rank grows, producing a long tail of words with similarly low frequencies. To state this formally, if r is rank, $p(w_r)$ is the probability of a word of rank r, and α is the law's exponent, varying between 0 and 1, then:

$$p(w_r) \propto r^{-\alpha} \tag{1.3}$$

A more difficult question is why lexical frequency distributions exhibit this particular property. This has been the subject of considerable debate (see Piantadosi, 2014 for a review). However, a general point of agreement is that highly skewed distributions, like power laws and exponentials, are more efficient at minimizing uncertainty than those that approach uniformity, where uncertainty is maximal. For the purposes of this dissertation, the critical lesson is simply that given a pair of distributions over the same set of items, (1) their comparative efficiency can be established empirically, and (2) that items from the more efficient distributions should be easier to process and to retrieve.

Chapter 3 investigates how varying the efficiency of a particular distribution affects memory for the items that comprise it, an idea inverted in **Chapter 4**, which examines how memory for an item depends on the distribution of contexts in which it occurs.

2.3 Coding a Lexicon

How might a lexicon with these properties emerge? Zipf argued that language's characteristic statistical structure reflects a compromise that balances the desire for a many-to-one code (in which there is a single, maximally frequent word) against the desire for one-to-one code (in which there are a vast number of low-frequency words). In the terms of optimal coding theory, these balancing forces of unification and diversification can be framed as a compromise between 'word-by-word' coding and 'large-block' coding (Mandelbrot, 1953). Thus, the problem of language design is one of how to distribute the information necessary to discriminate the repertoire of possible messages across acoustic signals (Baayen & Ramscar, 2015; Piantadosi, Tily, & Gibson, 2012).

In the limit, a one-to-one code would contain distinct sounds to express every possible thought. However, in practice, no human language assigns a sound to every arbitrary semantic possibility. Instead, languages are combinatorial, with complex ideas communicated iteratively over a sequence of elements (Hockett, 1960). Critically, the manner in which they are communicated does not have a single, universal solution (Slobin, 2003; Boroditsky, 2006). Rather, what a given language offers is a particular method of partitioning reality that results in a broadly similar source domain among its speakers, in which the habitual modes of distinction are distributed over varying numbers of units, including sounds (Monaghan et al., 2014), words, constructions (Goldberg, 2006), and contexts (Jones, Johns, & Recchia, 2010). The problem that

each language solves is how to strike this distributional balance over its elements. Qualitatively, what Zipf's law suggests is that a highly structured core vocabulary provides the scaffolding—the branches in the decision tree—on which the rest of the lexicon hangs (Baayen, 2009).

In recent years, attempts have been made to formalize Zipf's proposal, in hopes of better understanding how combinatorial solutions emerge in natural communication systems (see Ferrer-i-Cancho, 2006 for a review). These models reveal that given certain (modest) assumptions about our perceptual and physiological capabilities, a scale-free combinatorial system becomes necessary beyond a small signal repertoire.

The model systems described in **Chapters 2** and **3** serve to illustrate how information can be effectively distributed over the signal in a combinatorial system, showing how and why the burden of discriminating the identity of a specific entity must be shouldered by multiple elements. **Chapter 3** specifically examines the challenges associated with a rapidly expanding lexicon.

3. Investigations

3.1 Chapter 2: Model System – Nouns

Chapter 2 reports a pair of large-scale corpus analyses of two closely-related Germanic tongues. The aim is to compare the alternative approaches to the problem of managing nominal entropy adopted by English and German, showing that while both languages have adopted solutions that strike a balance between efficiency and learnability, the precise nature of this compromise reflects the particular demands of different populations of speakers. In particular, while German relies on a deterministic system to facilitate noun selection (grammatical gender), English employs a probabilistic one (prenominal adjectives). Despite these differences, both systems act to efficiently smooth information over discourse, making nouns more equally

predictable in context. Our investigations reveal that this facilitates processing in multiple ways: (1) by helping speakers avoid the peaks in uncertainty that would otherwise occur over nouns, (2) by reducing competition between nouns that are highly confusable in context (**Figures 1, 2**), and (3) by facilitating the use of a richer array of lexical items. The 'choice' of solution cross-linguistically appears to reflects certain social and historical forces at work. In particular, we examine the proposal that the structural form of a language is coupled to its population (and history) of adult learners (Johnson & Newport, 1989; Lupyan & Dale, 2010). On this account, English has traded efficiency—in communicative terms—for error tolerance, making it more amenable to later learning.

Figure 1. In German, semantic dispersion across genders is commonly seen among high frequency items, while semantic clustering is more common for low frequency items.

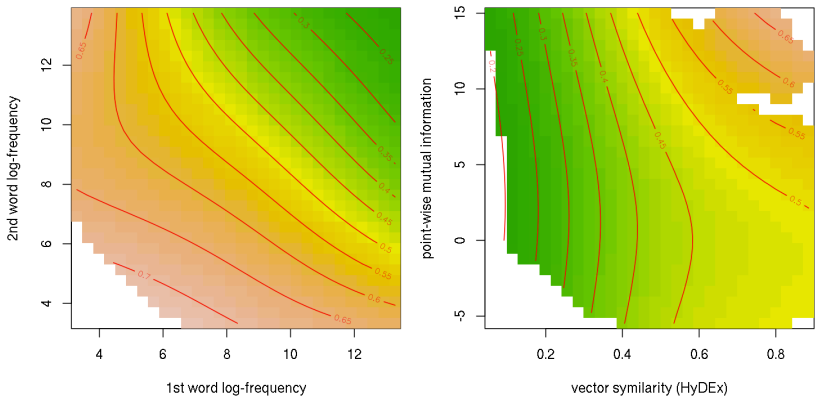

Figure 2. A Generalized Additive Model with binomial link function revealed that among German noun pairs,

assignment to the same gender class was predicted by two composite factors: (1) the frequency of the words in the pairing; and (2) the semantic similarity of the pair modulated by their co-occurrence likelihood.

3.2 Chapter 3: Model System – Name Grammars

Chapter 3 uses a related model system, name grammars, to investigate how the rational principles of efficient coding and rate constancy are realized in an ever-expanding lexicon, bounded by cognitive and physiological constraints, and susceptible to external interference. The chapter's theoretical analysis explores why combinatorial systems are a necessary solution to the problem of naming and illustrates how identifying information can be effectively distributed over multiple elements (**Figures 3, 4**).

In previous work, we identified an information structure common to the world's naming systems, and examined how the concomitant forces of social legislation and rapid population growth have altered these structures in Western naming practices. Modeling simulations predict that these deviations from communicative efficiency should render names increasingly difficult to process and remember. Here, we test this thesis across three classic memory paradigms—fluency, recognition, and recall. Our results lend support to the thesis that naming systems have evolved in line with communicative principles that optimize for ease of processing and memorability, and that external interference has been predictably damaging.

In particular, we find that: (1) Items drawn from more efficient distributions are easier to retrieve from memory, even when their frequency is held constant. (2) Entropy rate constancy supports precise sequence recall. Specifically: (1) Individual names of the same frequency are more fluently recalled, and more quickly processed in reading, when they are drawn from a more efficient distribution. (2) In an artificial name-grammar experiment, full name sequences are significantly better recalled when identifying information is evenly smoothed across elements,

rather than concentrated over a single element. Taken together, these results suggest that information-based measures have much to contribute to the study of verbal learning and memory.

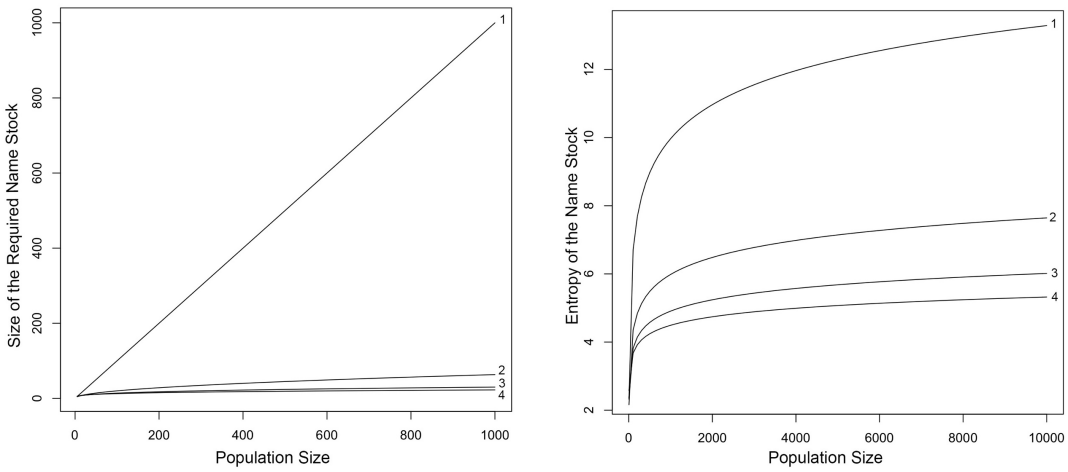

Figure 3. The requisite size and complexity of a name stock varies as a function of population size and the number of elements that comprise each name.

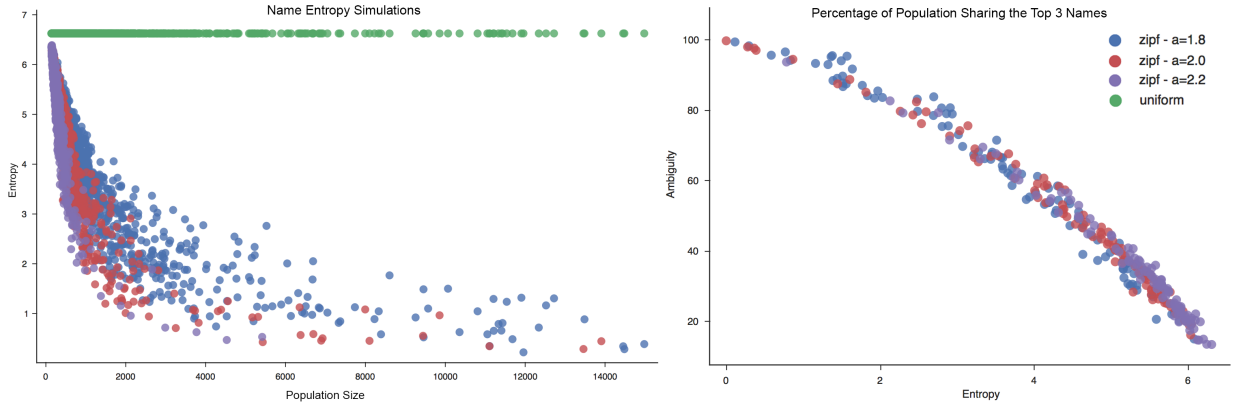

Figure 4. (***a***): Simulations of name entropy (y-axis) for a name stock of n=100 types, a population of up to 15,000 individuals (x-axis), and three Zipf distributions over that population, with varying choice of exponent. A matched uniform distribution serves as a baseline, depicting the least efficient coding strategy. (***b***): The percentage of the population that shares one of the top three name types as a function of name entropy.

3.3 Chapter 4: The Organization of the Lexicon

Chapter 4 affirms this conclusion in its examination of the principles that underpin the organization of the lexicon in memory. The chapter reviews a current debate in the literature over the contributions of context and repetition to lexical memory (Adelman, Brown, & Quesada, 2006; McDonald & Shillcock, 2003), presenting evidence that measures of a word's occurrence that are weighted by the informational redundancy of its contexts significantly outperform raw frequency measures, a result predicted by both learning and information theoretic accounts.

To further explore this idea, the chapter presents a distributional model of learning and semantic memory that relies on an expectancy-congruency mechanism to update its lexical representations, and can be used to generate predictions about both lexical access and similarity. The model predicts that while experiencing a novel word across semantically variable contexts should facilitate subsequent lexical access, more consistent contexts of occurrence should support the development of a superior semantic representation—a trade-off between ease of access and ease of acquisition. These predictions find support in the results of a novel word learning experiment (**Figure 5**).

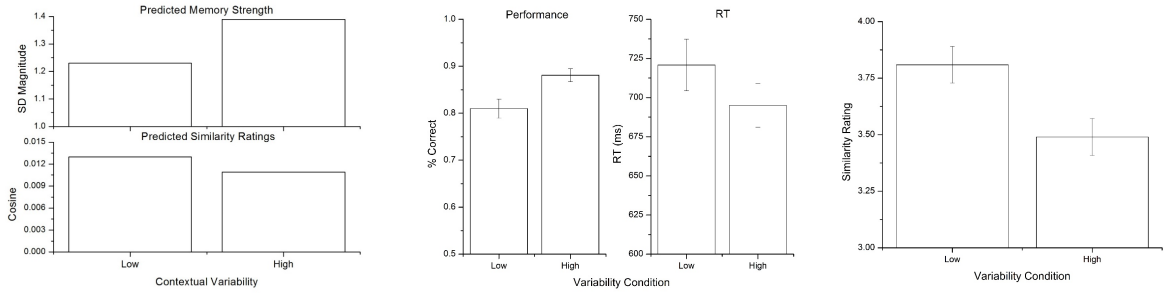

Figure 5. (***a***): Predictions from the distributional learning model after training on the same materials as our subjects. (***b***): Performance and RT results from the lexical decision task. (***c***): Mean similarity ratings of studied items and target associates by training type.

3.4 Chapter 5: The Information Structure of Verbal Sequences

Chapter 5 examines how the information structure of verbal sequences that comply with the rate constancy hypothesis differ from those that deviate. Rate constancy implies that rationally-designed verbal sequences should follow a characteristic tree-branching structure, in which early occurring elements act to reduce the entropy of later elements. Inverting this structure has the consequence of concentrating uncertainty over the first element. Structures of the 'optimal' type are observable in (e.g.) many of the world's naming systems (**Chapter 3**) and in languages in which nominal modifiers are placed prenominally (**Chapter 2**). However, inversions of this structure also exist, both within and across languages. Cross-linguistically, postnominal modifier bias is actually significantly more common than its prenominal counterpart (Culbertson, Smolensky, & Legendre, 2012). The question is: Why do languages systematically deviate from this rational principle?

The central insight of our analysis is that the structure of these alternative sequences can be mapped onto distinct paradigms from associative learning (Osgood, 1949). In particular, 'optimal' and 'suboptimal' codings map neatly onto 'convergent' and 'divergent' learning schemas, which have been shown to produce markedly different behavioral outcomes in tasks like categorization and causal reasoning (Yamauchi & Markman, 1988; Ramscar et al., 2010). Across three experiments, we investigate how information structure relates to different learning outcomes in an implicit word learning task, comparing learning from a divergent information structure (which resembles a decision tree, and is well designed to keep entropy constant), with learning from a convergent structure (in which the tree is inverted).

These results reveal a tradeoff between learnability and efficiency in the structure of verbal sequences. While 'optimally' coded sequences are more efficiently processed and better recalled (**Chapter 3**), 'suboptimal' sequences are better structured for semantic learning. Since languages

are designed to be both useable and learnable, their grammars must reflect a compromise between these desiderata. This raises the interesting theoretical possibility that linguistic regularities may play different functional roles depending on their relative temporal order.

3.5 Chapter 6: Limitations and Future Directions

> "If the human being acts in some situations like an ideal decoder, this is an experimental and not a mathematical fact, and as such must be tested under a wide variety of experimental situations."
>
> –Claude Shannon in "The Bandwagon" (1956)

Chapter 6 reviews the projects in the preceding chapters and explores the future applications of information theory to natural language. Broadly, information theoretic measures can be used to redescribe extant findings with greater precision and greater clarity, discriminate competing accounts, and suggest more nuanced empirical inroads. At the same time, the overarching theory of communication supplied by information theory invites a fundamental rethinking of the lexicon, which in turn motivates a host of new investigative projects.

Yet an important caveat is in order. Our results in **Chapter 5** reveal that the structure of verbal sequences incurs a tradeoff between processing efficiency and learnability. This suggests that the applicability of information theoretic principles depends closely on the communicative goal. Shannon's communicative framework was devised for systems in which the sender and receiver already share both a common source domain and a mutually agreed upon procedure for encoding and decoding messages into signals. Natural languages, by contrast, must be ***learnable***, such that children can acquire them, and ***adaptable***, such that speakers can flexibly update their representations in response to other speakers, or contexts. Thus, while languages may appear to

comply with the rate constancy principle at a macroscopic level, they should also systematically deviate from it in ways that support semantic learning.

The picture of natural language provided by information theory is thus incomplete, recalling George Box's injunction that "all models are wrong, but some are useful". Notably missing from this account is how the system is first acquired in childhood, how it can continue to flexibly develop and adapt over the lifespan, and how interlocutors with different priors coordinate their communicative efforts to converge on the same message. Computational models of learning and memory are thus a necessary complement to the picture supplied by Shannon. Future work should bear these lessons in mind.

4. Contributions

This dissertation makes several novel theoretical contributions to the literature:

(1) **Chapters 2** and **3** show how the rational analysis supplied by information theory can help disclose the functional 'design' properties of the lexicon. Social and historical analyses can then shed further light on alternative design choices.

(2) **Chapters 4** and **5** offer a thoroughgoing review of the conceptual links between learning and information theory, as well as an analysis of the utility of information-based measures over and above traditional frequency-based measures.

(3) **Chapters 3** and **5** reveal definite limits on the applicability of information theoretic principles to natural language, underscoring the importance of complementing rational analysis with mechanistic models.

Thesis Publications

Chapter 2

Dye, M., Milin, P., Futrell, R., & Ramscar, M. (*in press*). Alternative solutions to a language design problem: The role of adjectives and gender marking in efficient communication. *Topics in Cognitive Science*. doi: 10.1111/tops.12316

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*. Brill: Leiden.

Chapter 3

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.

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

Chapter 4

Jones, M. N., **Dye, M.**, & Johns, B. T. (2017). Context as an organizing principle of the lexicon. In B. Ross (Ed.), *The Psychology of Learning and Motivation* (pp. 239–283).

Johns, B.T., **Dye, M.**, & Jones, M.N. (2016). The influence of contextual diversity on word learning. *Psychonomic Bulletin and Review* 23: 1214–1220.

Chapter 5

Dye, M., Jones, M., Yarlett, D., & Ramscar, M. (2017). Refining the distributional hypothesis: A role for time and context in semantic representation. *Proceedings of the 39th Annual Conference of the Cognitive Science Society*, London, UK.

References

Adelman, J. S., Brown, G. D. A., & Quesada, J. F. (2006). Contextual diversity, not word frequency, determines word-naming and lexical decision times. *Psychological Science*, 17, 814-823.

Anderson, J. R., & Schooler, L. J. (1991). Reflections of the Environment in Memory. *Psychological Science*, *2*(6), 396–408.

Baayen, R. H. (2009) Corpus linguistics in morphology: morphological productivity. In A. Luedeling, and M. Kyto (eds.), *Corpus Linguistics. An international handbook* (pp 900-919). Mouton De Gruyter: Berlin.

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. Berlin: De Gruyter Mouton.

Bannard, C., & Matthews, D. (2008). Stored Word Sequences in Language Learning: The Effect of Familiarity on Children's Repetition of Four-Word Combinations. *Psychological Science*, *19*(3), 241–248.

Beckner, C.A. et al. (2009). Language is a complex adaptive system: Position paper. *Language Learning*, 59(1), 1-26.

Boroditsky, L. (2006). Linguistic Relativity. *Encyclopedia of Cognitive Science*.

Christiansen, M. H., & Chater, N. (2008). Language as shaped by the brain. *Behavioral and Brain Sciences*, 31(05), 489–509.

Conway, C. M., Bauernschmidt, A., Huang, S. S., & Pisoni, D. B. (2010). Implicit statistical learning in language processing: Word predictability is the key. *Cognition*, 114(3), 356–371.

Culbertson, J., Smolensky, P., & Legendre, G. (2012). Learning biases predict a word order universal. *Cognition*, 122(3), 306-329.

Evans, N. & Levinson, S.C. (2009). The myth of language universals: Language diversity and its importance for cognitive science. *Behavioral and Brain Sciences*, 32(5), 429-448.

Ferrer i Cancho, R. (2006). When language breaks into pieces. A conflict between communication through isolated signals and language. *BioSystems*, 84, 242-253.

Frank, S. L., & Bod, R. (2011). Insensitivity of the Human Sentence-Processing System to Hierarchical Structure. *Psychological Science,* 22(6), 829–834.

Genzel, D., & Charniak, E. (2002). Entropy rate constancy in text (pp. 199–206). In *Proceedings of the 40th Annual Meeting on Association for Computational Linguistics*. Association for Computational Linguistics: Morristown, NJ.

Goldberg, A.E. (2006). *Constructions at work: The nature of generalization in language*. Oxford University Press.

Hockett, C. F. (1960). The origin of speech. *Scientific American*, 203, 89–96.

Jaeger, T. F., & Tily, H. (2011). On language "utility": processing complexity and communicative efficiency. *Wiley Interdisciplinary Reviews: Cognitive Science*, 2(3), 323–335.

Johnson, J. S., & Newport, E. L. (1989). Critical period effects in second language learning: The influence of maturational state on the acquisition of English as a second language. *Cognitive Psychology*, 21(1), 60-99.

Jones, M.N., Johns, B.T., & Recchia, G. (2012). The role of semantic diversity in lexical organization. *Canadian Journal of Experimental Psychology*, *66*(2), 115–124.

Labov, S. (1972). *Sociolinguistic Patterns*. University of Pennsylvania Press.

Lupyan, G., & Dale, R. (2010). Language Structure Is Partly Determined by Social Structure. *PLoS ONE*, 5(1), e8559.

Mandelbrot, B. (1953). An informational theory of the statistical structure of language. *Communication theory*, 84, 486-502.

Marr, D. (1982). *Vision: A computational investigation into the human representation and processing of visual information*. MIT Press.

McDonald, S. A., & Shillcock, R. C. Rethinking the word frequency effect: The neglected role of

distributional information in lexical processing. *Language and Speech*, 44(3), 295-322.

Monaghan, P., Shillcock, R.C., Christiansen, M.H., & Kirby, S. (2014). How arbitrary is language? *Philosophical Transactions of the Royal Society B*, 369 20130299.

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

Pellegrino, F., Coupé, C., & Marsico, E. (2011). A cross-language perspective on speech information rate. *Language*, *87*(3), 539–558.

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. (2012). The communicative function of ambiguity in language. *Cognition*, 1–12.

Pickering, M. J., & Garrod, S. (2007). Do people use language production to make predictions during comprehension? *Trends in Cognitive Sciences*, *11*(3), 105–110.

Ramscar, M. & Baayen, R.H. (2013). Production, comprehension and synthesis: A communicative perspective on language. *Frontiers in Language Sciences*. 4:233. doi: 10.3389/fpsyg.2013.00233

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.

Reali, F., & Christiansen, M. H. (2007). Processing of relative clauses is made easier by frequency of occurrence. *Journal of Memory and Language*, *57*(1), 1–23.

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

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

Simon, H.A. (1989). Cognitive Architectures and Rational Analysis: Comment. *Technical Report AIP*, 58, 1–25.

Slobin, D. I. (2003). From "thought and language" to "thinking for speaking." 1–14.

Tomasello, M. (2003). *Constructing a Language: A Usage-Based Theory of Language Acquisition*. Harvard University Press.

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

Yamauchi, T., Love, B.C., & Markman, A. B. (2002). Learning nonlinearly separable categories by inference and classification. *Journal of Experimental Psychology: Learning, Memory, and Cognition*, 28(3), 585-593.

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