Learning the Pieces of Language

Timothy J. O'Donnell,[1]

McGill University

[1]Corresponding author. Address: 1085 Dr. Penfield Rm. 304, Montréal, QC H3A 1A7. Phone: +1 (617) 413-1697. Email: timothy.odonnell@mcgill.ca

Learning the Pieces of Language

1. Introduction

It has been understood since ancient times that the amazing expressive capacity of language is made possible by *compositionality*—utterance are built by combining and recombining an inventory of units like words and morphemes to express a vast array of possible meanings. It is also a truism that the set of basic units such as words and morphemes is different for each language, dialect, and even speaker and thus myst be learned. Determining the set of units for a given a language is one of the fundamental challenges of language learning. Because the exact form, meaning, and combinatorial properties of linguistics units cannot be known in advance, learners must solve this this problem using phonological, semantic, and morphosyntactic regularities.

What is less commonly appreciated is the difficulty of this task, especially at the outset of learning. The learner faces pervasive ambiguity and uncertainty and there will often be hundreds of thousands, millions, or even billions candidate sets of units consistent with any input. This remains the case, even when phonology, semantics, or morphosyntax provide strong constraints on the form and meaning of possible units.

To illustrate the problem, consider the three English suffixes in Table 1, *-ness*, *-ity*, and *-th*. All three of these suffixes are relatively common in English (representative type and token counts appear in the table).[2] All three suffixes are resonably phonologically regular and have similar morphological properties—they typically attach to adjectives to form nouns. The suffixes mean approximately the same thing as well, denoting the abstract state or property indicated by the adjective to which they attach.

However, these three suffixes differ in a crucial way: while *-ness* can freely combining with adjectival stems to give rise to new forms (e.g., *pine-scented*/*pine-scentedness*), *-th* cannot be used to generate new forms, even with stems satisfying its selectional restrictions such as the

[2]The token and word type counts in the second and third column of the table are taken from the corpus we discuss below. For comparison, the entire corpus contained over $25,000$ word types distributed over 7.2 word tokens.

monosyllabic, Germanic, obstruent-final stem *cool* (i.e., **coolth*). It seems clears that *-ness* is an lexical unit for most English speakers while *-th* is not. The suffix *-ity* is more complex—in general, it does not freely combine with bases—however, there are morphological contexts in which it it can be used to form novel words, for example, after the suffix *-able* (e.g., *modelable*/*modelability*).

Suffix	Word Tokens	Word Types	Examples
-ness	15568	1024	*goodness*, *cheapness*,*forgiveness*, *circuitousness*,*grandness*, *orderliness*, *business*, *goodness*,...
-th	12562	16	*truth*, *warmth*, *width*, *depth*, *filth*,*sloth*, *strength*, *death*, *dearth*, *wealth*, *length*, *youth*, ...
-ity	36448	477	*verticality*, *tractability*, *severity*, *seniority*, *inanity*, *electricity*, *-parity*, *scarcity*, *reality*,...

Table 1
Three Common English Suffixes

How do learners determine that *-ness* generalizes while *-th* does not. How do they determined the contexts in which *-ity* does? Of course, the three suffixes are phonologically, semantically and morphysyntactically different as well. For instance, *-ness* by and large simply concatenates with its base and nearly always signifies an abstract noun (*goodness*) while many forms in *-th* exhibit phonological irregularity and semantic non-transparency (*sloth* from *slow* and *-th*). But how much irregularity is too much for a form like *-th* to be rejected by a learner as a candidate unit ? How many semantically transparent forms are needed to accept a unit like *-ness*? At the beginning of learning, when the learner has no units at all with which to begin carving up input utterances, these problems are greatly multiplied.

In this paper, we present a theory of the principles that learners use to distinguish productive units from spurious generalizations in their language. The theory—presented in greater detail in O'Donnell (2011, 2015), O'Donnell, Snedeker, Tenenbaum, and Goodman (2011), and O'Donnell, Goodman, and Tenenbaum (2009)—is based on the idea that language learners

acquire an inventory of stored items which optimizes a *tradeoff* between a pressure to explain the input with the fewest stored items possible—a bias which pushes towards productive lexical units—and a countervailing pressure to treat each linguistic expression as simply as possible—a pressure which pushes toward the storage of a large variety specific forms. The idea behind this tradeoff-based approach is old, lying at the heart of many theories of optimal inductive inference including Bayesian approaches, minimum description length approaches, and many others.

In this paper, we illustrate these points with a detailed case study of several problems drawn from English inflectional and derivational morphology. We show that the theory is able to (i) unify several diverse classes of phenomena (productivity, processing time, competition effects, ordering) (ii) derive aspects of existing theories from more basic principles (importance of low-frequency forms, specificity preference) (iii) explain phenomena difficult for earlier accounts (productivity and ordering, paradoxical suffix combinations) (iv) generate quantitative predictions about known phenomena that can be tested with corpus data and psychological experiments (processing) and (v) makes novel predictions about language learning and change (irregularization).

2. Theoretical Setting: Lexical-Drivenness

Learning the inventory of lexical items for a language has always been an obviously important part of language learning, however in recent decades it has moved to the center stage of learning in syntax and morphology. In this time, linguistic theories have increasingly adopted a *lexically-driven* view of grammar.[3] Under this view, the linguistic structures underlying production and comprehenson are characterized by a small number of structure-building operations whose behavior is controlled by information stored on items in the lexicon, usually in the form of features. For instance, a suffix like *-ness* might be stored in the lexicon together with information specifying that it selects an adjective which is concatenated on its left and that the result is a well-formed noun.

[3] Some would say readopted, since this view was at least implicitly present in some of the earliest work in generative grammar such as Bar-Hillel (1953), Tesniére (1959), and some aspects of Chomsky (1975, 1955).

At present, every major theory of syntax and semantics and most theories of morphology are lexically-driven in this sense (Bresnan, 2001; Chomsky, 1995a, 1995b; Culicover & Jackendoff, 2005; Gamut, 1991; Gazdar, Klein, Pullum, & Sag, 1985; Heim & Kratzer, 2000; Huddleston & Pullum, 2002; Jackendoff, 2002; D. E. Johnson & Postal, 1980; McConnell-Ginet & Chierchia, 2000; Mel'čuk, 1988; Moortgat, 1997; Sag, 2012; Sag, Wasow, & Bender, 2003; Stabler, 1997; Steedman, 2000). Typically, such systems contain a core operation implementing a basic form of selection or complementation (e.g., merger in minimalist frameworks, some form of function application in categorial frameworks, selection-driven unification in others) well as some number of extensions of this core system with more powerful operations (e.g., movement, slash category percolation, adjunction, type raising, etc.).[4]

Lexically-driven theories are important because they divide labor between the lexicon and structure-building operations in such a way that most information about language-specific constraints on linguistic structures resides in the combinatorial requirements of specific lexical items. This in turns suggests an intriguing possibility: Perhaps *all* cross-linguistic variation might be reducible to differences in invntories of lexical items and, thus, language learning might be reduceable to the acquisition of the lexicon (e.g., Baker, 2008; Chomsky, 1993; Steedman, 2000).[5][6]

If this view is true, the problem of language learning is simply to identify the lexical units (with their features) in their language from linguistic input. This perspective suggests a simple framework in which to study language learning. Given a grammatical architecture that specifies a set of structure building operations, and a format for lexical items, the learner searches for a

[4]As defined, lexical-drivenness can emcompass a wide variety of theories which otherwise vary along major dimensions such as such as the distinction between representational/contrainst-based and derivational specification of sets of possible structures, lexicalist and single-system architectures, configurational versus relational definition of dependencies, transformational versus correspondence-based versus hypothetical-reasoning-based accounts of multiple/nonlocal dependencies, the (non)existence of abstract lexical items without a phonological component, and the requirement that morphological structure be concatenative. The framework is even broad enough to capture the majority of so-called neural approaches to linguistic structure. Thus, the problem we discuss in the next arise in many contexts.

[5]Recent progress in unsupervised grammar induction using lexically-driven formalisms lends plausibility to this idea (Bisk & Hockenmaier, 2013; Cohn, Blunsom, & Goldwater, 2010; Naseem, Chen, & Johnson, 2010; **?**; **?**).

[6]Whether it turns out that all typological variability and learning is explained by the properties of individual lexical item, or just most, this shift in emphasis focuses learning theories squarely on the problem of acquiring the set of lexical units and their combinatorial properties.

lexicon that is likely to have generated the observed data. As outline in the introduction, there will be, in general, many possible lexica consistent with any input. To decide between these, the learner must also come equipped with some way of deciding between them—an *evaluation metric* in classical linguistic terminology (Chomsky, 1975, 1955).

In this paper, we will examine an evaluation metric which is based on two simplicity preferences. First, it favors small lexica with highly reusable linguistic units. Second, it favor simple observable forms with few pieces or, equivalently, simple derivations. These two biases naturally oppose one another leading to a tradeoff. Compact lexica have small words and morphemes which lead to a larger number of morphemes per observed form. Simple observed forms have small numbers of morphemes arranged in their derivations, leading to larger inventories of stored forms. This tradeoff-based evaluation metric is not a new idea. It is a particular instantiation of the Bayesian tradeoff between a prior on grammatical simplicity (here instantiated as size of the lexicon) and a likelihood measuring fit to the data (here measured as probability assigned the corpus of observed forms). Equivalently, it is a variant of the two-part code studied in the minimum-description length framework. The learner simultaneously tries to minimize a description of the lexicon (thus optimizing for more compact lexica) and a descriptin of the data in terms of the lexicon (thus optimizing for more compact derivations of observed forms; Grünwald, 2007; Rissanen, 1978; Vitányi & Li, 2008). These ideas, in turn, go back to the foundations of the idea of an evaluation metric in linguistics as well as the foundations of theories of inductive inference by Solomonoff and others (see J. A. Goldsmith, 2011) and have been applied many times in the years since (Berwick, 1982, 1985; Brent & Cartwright, 1996; De Marcken, 1994, 1996a, 1996b,?, 1996c; Eisner, 2002; Ellison, 2004; J. Goldsmith, 2006; Goldwater, 2006; Hsu & Chater, 2010; Hsu, Chater, & Vitányi, 2011; M. Johnson, Griffiths, & Goldwater, 2007; Olivier, 1968; Stolcke & Omohundro, 1994; Villavicencio, 2002; Wolff, 1977, 1980, 1982; **?**; **?**, amongst many others).

Despite this long pedigree, the theory behind this evaluation metric and its application to linguistic problems is not well-known within linguistics. In this paper, we consider a simple

grammatical system which consists of just a single structure-building operation corresponding to feature-driven selection (the simplest case of external MERGE, function application in categorial framework, or unification on flat attribute value matrices in LFG/HPSG) and study the predictions generated by the evaluation metric for this theory on a number of problems in English morphology. Before moving on to present the details of the theory, we first discuss the learning problems raised by uncertainty about the lexicon in more detail.

References

Baker, M. C. (2008). The macroparameter in a microparametric world [Theoretical Discussion, Linguistic Analysis]. In T. Biberauer (Ed.), *The limits of syntactic variation* (p. 351). John Benjamins Publishing Company.

Bar-Hillel, Y. (1953, January-March). A quasi-arithmetical notation for syntactic description. *Language*, *29*(1), 47–58.

Berwick, R. C. (1982). *Locality principles and the acquisition of syntactic knowledge* (Unpublished doctoral dissertation). Massachusetts Institute of Technology.

Berwick, R. C. (1985). *The acquisition of syntactic knowledge* [Monograph]. Cambridge, Massachusetts and London, England: The MIT Press.

Bisk, Y., & Hockenmaier, J. (2013, March). An HDP model for inducing combinatory categorial grammars. *Transaction of the Association for Computational Linguistics*, *1*, 63–74.

Brent, M. R., & Cartwright, T. A. (1996). Distributional regularity and phonotactic constraints are useful for segmentation. *Cognition*, *61*, 93–125.

Bresnan, J. (2001). *Lexical functional syntax* [Textbook]. Oxford: Wiley-Blackwell.

Chomsky, N. (1975, 1955). *The logical structure of linguistic theory*. New York, NY: Plenum Press.

Chomsky, N. (1993). A minimalist program for linguistic theory [Theoretical Discussion]. In K. L. Hale & S. J. Keyser (Eds.), *The view from building 20: Essays in honor of Sylvain Bromberger* (pp. 1–52). Cambridge, Massachusetts and London, England: The MIT Press.

Chomsky, N. (1995a). Bare phrase structure. In G. Webelhuth (Ed.), *Government and binding theory and the minimalist program* (pp. 383–349). Blackwell.

Chomsky, N. (1995b). *The minimalist program* [Collection]. Cambridge, Massachusetts and London, England: The MIT Press. Retrieved from http://books.google.com/books?hl=en&lr=&id=vtPQiYCNpjgC&oi=fnd&pg=PA1&dq=the+minimalist+program&ots=HX5zwVE43N&sig=cyLjWcXIl7YCKE2nNGeo-bNu2Nc

Cohn, T., Blunsom, P., & Goldwater, S. (2010). Inducing tree-substitution grammars. *Journal of Machine Learning Research*, *11*, 3053–3096.

Culicover, P., & Jackendoff, R. (2005). *Simpler syntax*. Oxford: Oxford University Press.

De Marcken, C. (1994). The acquisition of a lexicon from paired phoneme sequences and semantic representations [Computational Model]. In *Proceedings of the 2nd international colloquium on grammatical inference and applications (ICGI).* Alicante, Spain.

De Marcken, C. (1996a). Linguistic structure as composition and perturbation [Computational Model]. In *Proceedings of the 34th annual meeting on association for computational linguistics* (pp. 335–341).

De Marcken, C. (1996b). *The unsupervised acquisition of a lexicon from continuous speech* (Computational Model Nos. AI-memo-1558, CBCL-memo-129). Massachusetts Institute of Technology – Artificial Intelligence Laboratory.

De Marcken, C. (1996c). *Unsupervised language acquisition* (Dissertation). Massachusetts Institute of Technology.

Eisner, J. (2002). Discovering syntactic deep structure via Bayesian statistics. *Cognitive Science*, *26*, 255—268.

Ellison, T. M. (2004). *The machine learning of phonological structure*. Cambridge, England: Cambridge University Press.

Gamut, L. T. F. (1991). *Logic, language, and meaning volume II: Intensional logic and logical grammar*. University of Chicago Press.

Gazdar, G., Klein, E., Pullum, G. K., & Sag, I. A. (1985). *Generalized phrase structure grammar*. Harvard University Press.

Goldsmith, J. (2006). An algorithm for the unsupervised learning of morphology. *Natural Language Engineering*, *12*(4), 353–371.

Goldsmith, J. A. (2011). The evaluation metric in generative grammar [Theoretical Discussion]. In *Proceedings of the 50th anniversay celebration of the MIT department of linguistics.*

Goldwater, S. (2006). *Nonparametric bayesian models of lexical acquisition* (Unpublished

doctoral dissertation). Brown University.

Grünwald, P. D. (2007). *The minimum description length principle* [Monograph, Introduction, Review, Textbook]. Cambridge, MA: The MIT Press.

Heim, I., & Kratzer, A. (2000). *Semantics in generative grammar*. Malden, MA: Blackwell Publishing.

Hsu, A. S., & Chater, N. (2010). The logical problem of language acquisition goes probabilistic: No negative evidence as a window on language acquisition [Theoretical Discussion, Computational Model]. *Cognitive Science*, *34*, 972–1016.

Hsu, A. S., Chater, N., & Vitányi, P. M. B. (2011). The probabilistic analysis of language acquisition: Theoretical, computational, and experimental analysis. *Cognition*, *120*, 380–390.

Huddleston, R., & Pullum, G. K. (2002). *The cambridge grammar of English language*. Cambridge: Cambridge University Press.

Jackendoff, R. (2002). *Foundations of language*. New York: Oxford University Press.

Johnson, D. E., & Postal, P. M. (1980). *Arc pair grammar* [Monograph]. Princeton, New Jersey: Princeton University Press.

Johnson, M., Griffiths, T. L., & Goldwater, S. (2007). Adaptor Grammars: A framework for specifying compositional nonparametric Bayesian models. In *Advances in neural information processing systems 19*. Cambridge, MA: MIT Press.

McConnell-Ginet, S., & Chierchia, G. (2000). *Meaning and grammar: An introduction to semantics*. MIT Press.

Mel'čuk, I. (1988). *Dependency syntax : Theory and practice,*. Albany, N.Y.: The SUNY Press.

Moortgat, M. (1997). Categorial type logics. In *Handbook of logic and language* (pp. 93–177). Elsevier.

Naseem, T., Chen, H., & Johnson, M. (2010). Using universal linguistic knowledge to guide grammar induction. In *Proceedings of the 2010 conference on empirical methods in natural language processing*.

O'Donnell, T. J. (2011). *Productivity and reuse in language* (Unpublished doctoral dissertation). Harvard University, Cambridge, MA.

O'Donnell, T. J. (2015). *Productivity and reuse in language: A theory of linguistic computation and storage*. Cambridge, Massachusetts: The MIT Press.

O'Donnell, T. J., Goodman, N. D., & Tenenbaum, J. B. (2009). *Fragment grammars: Exploring computation and reuse in language* (Tech. Rep. No. MIT-CSAIL-TR-2009-013). Cambridge, MA: MIT Computer Science and Artificial Intelligence Laboratory Technical Report Series.

O'Donnell, T. J., Snedeker, J., Tenenbaum, J. B., & Goodman, N. D. (2011). Productivity and reuse in language. In *Proceedings of the 33rd annual conference of the cognitive science society.* Boston, MA.

Olivier, D. C. (1968). *Stochastic grammars and language acquisition mechanisms* (Unpublished doctoral dissertation). Harvard University.

Rissanen, J. (1978). Modeling by shortest data description [Theoretical Discussion, Computational Model, Mathematical Result]. *Automaticata*, *14*(5), 465–471.

Sag, I. A. (2012). Sign-based construction grammar: An informal synopsis. In I. A. Boas Hans abd Sag (Ed.), *Sign–based construction grammar* (pp. 101–107). CSLI Publications.

Sag, I. A., Wasow, T., & Bender, E. M. (2003). *Syntactic theory: A formal introduction* (Second ed.) [Textbook]. Stanford, CA: CSLI.

Stabler, E. P. (1997). Derivational minimalism. In *Logical aspects of computational linguistics.* Berlin, Germany: Springer.

Steedman, M. (2000). *The syntactic process* [Monograph]. Cambridge, Ma: MIT Press.

Stolcke, A., & Omohundro, S. (1994). Inducing probabilistic grammars by Bayesian model merging. In *Proceedings of the international conference on grammatical inference.*

Tesniére, L. (1959). *Éléments de syntaxe structurale*. Kilncksieck.

Villavicencio, A. (2002). *The acquisition of a unification-based generalised categorial grammar* (Dissertation No. UCAM-CL-TR-533). University of Cambridge.

Vitányi, P. M. B., & Li, M. (2008). *An introduction to Kolmogorov complexity and its applications* (3rd ed.) [Reference, Texbook]. Berlin, Germany: Springer.

Wolff, J. G. (1977). The discovery of segments in natural language. *British Journal of Psychology*, *68*, 97–106.

Wolff, J. G. (1980). Language acquisition and the discovery of phrase structure. *Language and Speech*, *23*(3), 255–269.

Wolff, J. G. (1982). Language acquisition, data compression, and generalisation. *Language and Communication*, *2*(1), 57–89.