Learning Syntactic Constructions from Raw Corpora

Shimon Edelman,[1] Zach Solan,[2] David Horn,[2] and Eytan Ruppin,[2]
[1]Cornell University and [2]Tel Aviv University

Construction-based approaches to syntax (Croft, 2001; Goldberg, 2003) posit a lexicon populated by units of various sizes, as envisaged by (Langacker, 1987). Constructions may be specified completely, as in the case of simple morphemes or idioms such as take it to the bank, or partially, as in the expression what's *X* doing *Y*?, where *X* and *Y* are slots that admit fillers of particular types (Kay and Fillmore, 1999). Constructions offer an intriguing alternative to traditional rule-based syntax by hinting at the extent to which the complexity of language can stem from a rich repertoire of stored, more or less entrenched (Harris, 1998) representations that address both syntactic and semantic issues, and encompass, in addition to general rules, "totally idiosyncratic forms and patterns of all intermediate degrees of generality" (Langacker, 1987, p.46). Because constructions are by their very nature language-specific, the question of acquisition in Construction Grammar is especially poignant. We address this issue by offering an unsupervised algorithm that learns constructions from raw corpora.

1 The ADIOS algorithm for grammar induction

In a standard paradigm for grammar induction, a teacher produces a sequence of strings generated by a grammar G_0, and a learner uses the resulting corpus to construct a grammar G, aiming to approximate G_0 in some sense (Adriaans and van Zaanen, 2004). Recent evidence suggests that natural language acquisition involves both statistical computation (e.g., in speech segmentation) and rule-like algebraic processes (e.g., in structured generalization) (Saffran et al., 1996; Seidenberg, 1997; Marcus et al., 1999; Peña et al., 2002; Seidenberg et al., 2002). Indeed, modern computational approaches to grammar induction integrate statistical and rule-based methods (Pereira, 2000; Geman and Johnson, 2003).

We have developed a new method that uses statistical information present in raw sequential data to identify significant segments and to distill hierarchical rule-like regularities that support structured generalization (Solan et al., 2004; Edelman et al., 2004). Our algorithm, ADIOS (for Automatic DIstillation of Structure), is data driven. Consider a corpus of sentences (more generally, sequences) over a lexicon of size N, whose units in the case of language are initially words (starting with phonemes or letters or even letters in a condensed text without spaces also works).

The algorithm starts by loading the corpus onto a directed pseudograph (a non-simple graph in which both loops and multiple edges are permitted) whose vertices are all lexicon entries, augmented by two special symbols, begin and end. Each corpus sentence defines a separate path over the graph, starting at begin and ending at end, and is indexed by the order of its appearance in the corpus. Loading is followed by an iterative search for significant *patterns*, which are added to the lexicon as new units.

The algorithm generates candidate patterns by traversing, in each iteration, a different *search path* (initially coinciding with one of the original corpus sentences), seeking sub-paths that are shared by a significant number of partially aligned (Harris, 1954; van Zaanen, 2000) paths. The significant patterns (P) are selected according to a context-sensitive probabilistic criterion defined in terms of local flow quantities in the graph. Generalizing the search path, the algorithm looks for an optional *equivalence class* (E) of units that are interchangeable in the given context, i.e., are in complementary distribution (Harris, 1954). At the end of each iteration, the most significant pattern is added to the lexicon as a new unit, the sub-paths it subsumes are merged into a new vertex, and the graph is rewired accordingly (two rewiring modes are available: a context free Mode A and a context-sensitive Mode B). The search for patterns and equivalence classes and their incorporation into the graph are repeated until no new significant patterns are found. The entire process is governed by three parameters, two of which control pattern significance, and another one sets the width L of the context window where equivalence classes are sought. Details of the algorithm appear elsewhere (Solan et al., 2004).

The final lexicon includes those of the original symbols not incorporated into larger units, and root patterns distilled by the algorithm (that is, the patterns that reside on the final graph, at the top level of the hierarchy). Due to the hierarchical process of pattern creation, each pattern is structured as a tree, whose leaves (terminals) are the original members of the lexicon and whose intermediate nodes are other patterns and equivalence classes (Figure 1). The resulting tree structure excludes cyclic recursion (loops) of patterns, although recursion may be introduced through pattern matching in a post-processing stage.

The final graph includes as many paths as all the original sentences, but it can also generate many new ones. To generate a sentence from a chosen path in the graph, all its root patterns are traversed. Each recursively encountered pattern is treated as a derivation or parse tree (Hopcroft and Ullman, 1979): it is read from top (root) to bottom (terminals) and from left to right, while accruing the terminals (words from the original lexicon) and selecting one member from each encountered equivalence class (Figure 1C). Because the equivalence relations only hold in the contexts specified by their parent patterns, the ADIOS representation is inherently safer than grammars that posit globally valid categories (such as "parts of speech" in a natural language). At the same time, because each rewiring of the graph brings closer far-apart units that used to straddle the newly abstracted pat-

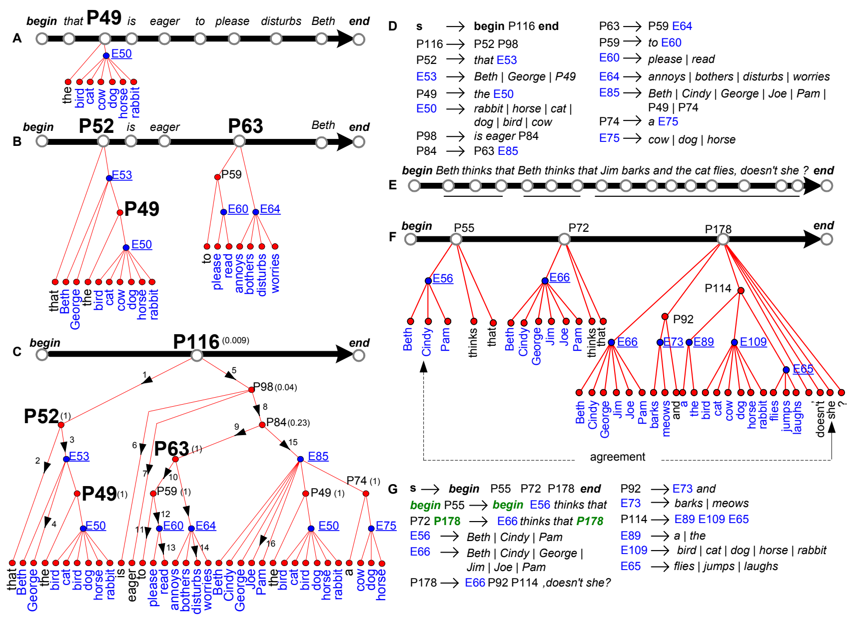

Figure 1: Progressive abstraction of patterns constructs a forest of trees rooted in vertices of the graph (training data generated by a context-free grammar, TA1, with 50 terminals and 28 rules). (A), Pattern P49, consisting of the terminal the and equivalence class E50 = {bird, cat, cow, dog, horse, rabbit}, is distilled. (B), Further application of the algorithm yields equivalence classes (underlined) such as E64, which contain some verbs. (C), Pattern P116 can generate 896 novel sentences, eight of which appear in the training corpus (the *generalization factor*, 8/896, appears in parentheses). A novel sentence, such as that George is eager to read disturbs Joe, can be read off the leaf level of the tree (numbered arrows indicate traversal order during generation). Pattern P116 is a root pattern, that is, a unit situated on a final path. (D), The set of context-free productions (rewriting rules) that is equivalent to the tree of pattern P116. (E), The initial path through a sentence to which ADIOS was applied in the context-sensitive mode B. (F), The same path after three root patterns (P55, P72 and P178) have been distilled. Note how the two similar but not identical root patterns, P55 and P72, capture the difference between the equivalence classes E56 and E66 (indeed, Beth, for example, is equivalent to Jim in the context of P72, but not of P55). In this manner, ADIOS enforces long-range agreement between E56 and the phrase doesn't she (embedded in P178), and avoids over-generalization. (G), The two context-sensitive rules in this example are [begin P55 ⇒ begin E56 thinks that] and [P72 P178 ⇒ E66 thinks that P178].

tern, the resulting representation can capture long-range structural dependencies among units.

Because patterns can be represented in the form of rewriting rules, which are context-free when Mode A is used (Figure 1D) and context-sensitive when Mode B is used (Figure 1G), the end product of an ADIOS run constitutes a grammar. As infinite recursion is not implemented in the current version of the algorithm, the representations learned by ADIOS are comparable in expressive power to finite Context Sensitive Grammars. This means that any grammar consisting of context sensitive rules can be loaded into an ADIOS instance (that is, translated into an ADIOS representation), provided that a limit is placed on the number of times each rule is invoked recursively. In learning, the results described in the following section show that our algorithm can acquire, from raw corpora, good operational approximations to those grammars that generate data rich with partially alignable sentences, including unconstrained natural-language data. Complex grammars in which inherent ambiguity (Hopcroft and Ullman, 1979) is exacerbated by the presence of multiple loops are dealt with effectively by acquiring more patterns.

2 Results

We tested the ADIOS algorithm both on artificial-grammar data and on natural-language corpora such as ATIS (Moore and Carroll, 2001) and the CHILDES collection (MacWhinney and Snow, 1985), and in languages as diverse as English and Chinese (Resnik et al., 1999). It is reasonable to require that the success of a learning algorithm be measured by the closeness — ideally, identity — of the learned and target grammars, G and G_0, respectively. Unfortunately, even for Context Free Grammars (CFGs), equivalence is undecidable (Hopcroft and Ullman, 1979). Moreover, for natural languages G_0 is inaccessible. We thus opt for testing our implementation for generativity[1] as follows. In the artificial-grammar experiments, which start with a target grammar, a teacher instance of the model is first pre-loaded with this grammar (using the one-to-one translation of CFG rules into ADIOS patterns), then used to generate the training corpus $C_{training}$. After training, the learner generates a test corpus $C_{learner}$ and the teacher generates a target corpus C_{target}, the latter containing only novel sentences that do not appear in $C_{training}$. The two corpora, $C_{learner}$ and C_{target}, are then used to calculate

[1] In testing a learned grammar G for *strong generativity*, the structural descriptions (parse trees) it assigns to novel strings are compared to those produced by the target grammar G_0. A *weak generativity* criterion requires merely that G accept novel G_0-grammatical strings as such, and reject the ungrammatical ones. Strong generativity of grammars acquired by unsupervised algorithms that work from raw data is in principle difficult to test, because of the unavailability of reliable "gold standard" structural descriptions for such data. At the same time, demonstrating even weak perfect generativity by an automatically acquired representation has until now proved elusive. Our results constitute significant progress on both fronts: the representations acquired by the ADIOS algorithm are (1) structured in a manner that makes certain syntactic sense, and (2) generative in that they largely encode and produce acceptable sentences.

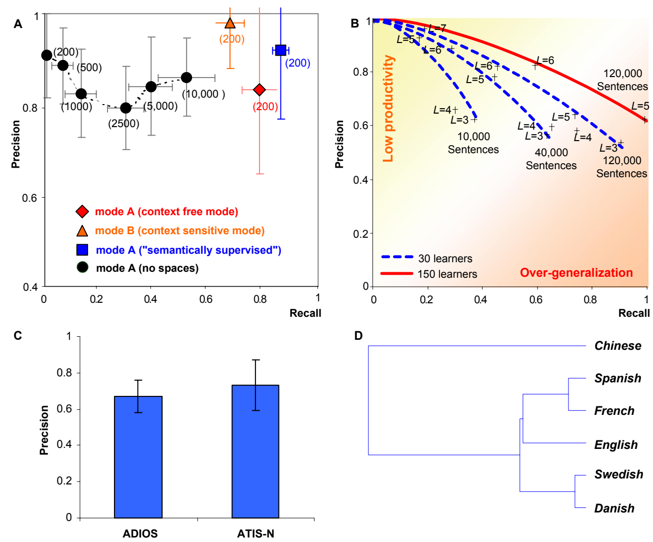

Figure 2: (A), The performance of an ADIOS model trained on extremely small corpora generated by TA1. Optimal combinations of recall and precision (single learner, 30 trials, $\eta = 0.6$, $\alpha = 0.01$, $L = 5$, maximum recursion depth for the teacher here and below set to 10) are shown for four different conditions: (i) the default learning mode A (context-free); with a 800-sentence training corpus (not shown), both precision and recall reach 90%; (ii) mode B (context-sensitive) (iii) a "semantically supervised" mode in which the equivalence classes of the target grammar are made available to the learner ahead of time; (iv) bootstrap mode, starting from a letter level and training on corpora in which all spaces between words are omitted. For comparable performance, this mode requires larger corpora (size in parentheses: 200-$10,000$ sentences). (B), Using the ATIS Context-Free Grammar (4592 rules) (Moore and Carroll, 2001) as the teacher of multiple ADIOS learners. Precision is defined by the mean over learners; for recall acceptance by one learner suffices. Several corpus sizes, context window widths L and numbers of learners are compared. (C), Output generated by an instance of ADIOS that had been trained on the natural language ATIS-N corpus was judged to be as acceptable to human subjects as sentences from ATIS-N (acceptability data, mean $\pm$ std, from eight subjects). (D), A dendrogram illustrating the relationships among six languages using pattern spectra. We define a pattern spectrum as the histogram of pattern types, with bins labeled by sequences such as (T,P) or (E,E,T), E standing for equivalence class, T for tree-terminal (original unit) and P for significant pattern. Hierarchical clustering was applied to Euclidean distances among histograms of patterns learned from a multilingual Bible (Resnik et al., 1999) (single learner per language).

precision (the proportion of $C_{learner}$ accepted by the teacher) and *recall* (the proportion of C_{target} accepted by the learner). A sentence is accepted if it precisely fits one of the paths in the ADIOS graph (that is, it can be generated by the path). In the natural language experiments, where no target grammar is available, the given corpus is split into two portions, one for training ($C_{training}$) and one for testing (C_{target}), and the same evaluation method is applied, except that precision must in this case be evaluated by an external referee (e.g., by a human subject). This evaluation method is unique in that (i) it defines precision and recall more conservatively than is standard in the literature (Klein and Manning, 2002), and (ii) it involves testing both the capability of the learner to *accept* all the grammatical sentences *and* its capability to *generate* only sentences that the teacher would deem grammatical.

2.1 Grammar induction

We have conducted a series of experiments designed to evaluate the performance of ADIOS in grammar induction (Figure 2).

Learning a simple CFG: We first replicated an experiment (Adriaans and Vervoort, 2002) that aimed to reconstruct a specific Context-Free Grammar (29 terminals and 7 rules) from a corpus of 2000 sentences. Whereas the algorithm of (Adriaans and Vervoort, 2002) generated between 3000 and 4000 rules, ADIOS (used in the default Mode A) yielded 28 patterns and 9 equivalence classes, achieving 100% precision and 99% recall. Next, we applied ADIOS to a somewhat more complex CFG (TA1 grammar, 50 terminals and 28 rules), and showed that it performs well even when only 200 sentences are used for training (see Figure 2A).

Learning a complex CFG: Because the ADIOS algorithm is greedy (the best available pattern in each iteration is immediately and irreversibly rewired), the syntax it acquires depends on the order of sentences in the training set. This is expected to affect the learning of a complex CFG, especially if it contains many loops. To assess this dependence and to mitigate it, we train multiple learners on different order-permuted versions of the corpus generated by the teacher. As Figure 2B illustrates, for the parameter values explored ($L = \{3, 4, 5, 6\}$; 30 or 150 learners; corpus size between 10,000 and 120,000 sentences), the optimal precision-recall trade-off for learning the ATIS CFG (357 terminals and 4592 rules) (Moore and Carroll, 2001) is obtained with a 150-learner cohort and L between 5 and 6. Some of the patterns learned from the ATIS corpus are shown in Figure 3.

Generativity of the learned natural language grammar: To test the ability of ADIOS to generate acceptable novel sentences after learning from a natural language corpus, we trained it on $12,700$ sentences from ATIS-N (a natural language corpus of size $13,043$ Moore and Carroll (2001)) and tested its recall level on the 343 remaining sentences. The small size of the training corpus results in a relatively low recall of 40% (under our strict definition that requires an exact match).

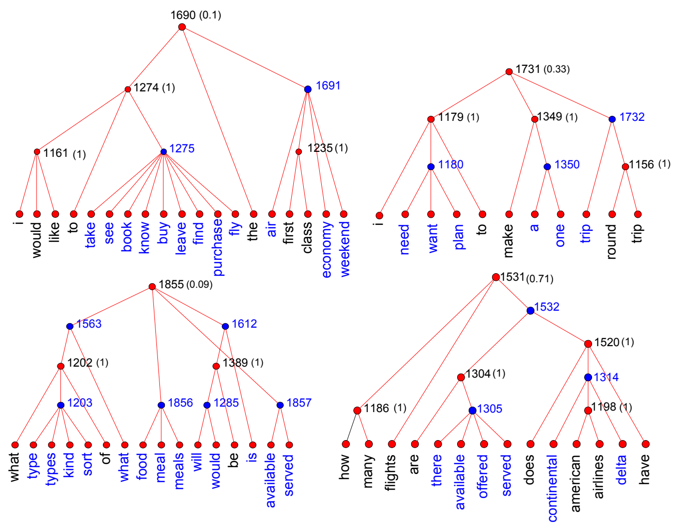

Figure 3: Four patterns extracted from the ATIS natural language corpus. Some of the sentences that can be described/generated by patterns #1690, #1731, #1855 and #1531 are: I would like to book the first class; I plan to make a round trip; what kind of food would be served ; how many flights does continental have . None of these sentences appear in the training data, illustrating the ability of ADIOS to generalize. The numbers in parentheses denote the generalization factors of the patterns and their components (e.g., pattern #1690 generates 90% new strings, while pattern #1731 generates 66% new strings).

Figure 2C compares the acceptability of ADIOS-generated sentences with original sentences from the ATIS-N corpus. Notably, the output generated by ADIOS is on the average as acceptable to human subjects as the original corpus sentences. The human-judged precision ($\approx$ 70%, as shown in the plot) is remarkable; for comparison, the ATIS-CFG grammar, hand-constructed to fit the ATIS-N corpus (with recall of 45% on same data) produces over 99% ungrammatical sentences when used in a generative fashion.

Languages other than English: Applying ADIOS to the Parallel Bible corpus (Resnik et al., 1999), we compared six different languages through a meta-analysis of their respective ADIOS grammars. The dendrogram shown in Figure 2D captures the resulting typological relationships.

2.2 Psycholinguistics

Learning "nonadjacent dependencies": Gómez (2002) showed that the ability of subjects to learn an artificial language L1 of the form $\{aXd, bXe, cXf\}$, as measured by their ability to distinguish it implicitly from L2=$\{aXe, bXf, cXd\}$, depends on the amount of variation introduced at X (symbols a through f here stand for 3- or 4-letter nonsense words, whereas X denotes a slot in which a subset of 2-24 other nonsense words may appear). Within the ADIOS framework, these non-adjacent dependencies translate into patterns with embedded equivalence classes. We replicated the Gómez study by training ADIOS on 432 strings from L1 (30 learners, $|X| = 2, 6, 12, 24$, $\eta = 0.6$, $\alpha = 0.01$). Training with the context window parameter L set to 3 resulted in performance levels (rejection rate of patterns outside of the learned language) that increased monotonically with $|X|$, in correspondence with the human behavior. Interestingly, when trained with $L = 4$, *adios* reaches perfect performance in this task.

Grammaticality judgments: A single instance of ADIOS was trained on the CHILDES (MacWhinney and Snow, 1985) corpus, using sentences spoken by parents to three year old children. It was then subjected to five grammaticality judgment tests. One of these, the Göteborg multiple-choice ESL (English as Second Language) test, consists of 100 sentences, each containing an open slot; for each sentence, the subject selects one word from a list of three choices, so as to complete the sentence grammatically. In this test, ADIOS scored at 60%, which is the average score for 9th grade ESL students. Interestingly, the average score of ADIOS on the entire collection of tests was at the same level.

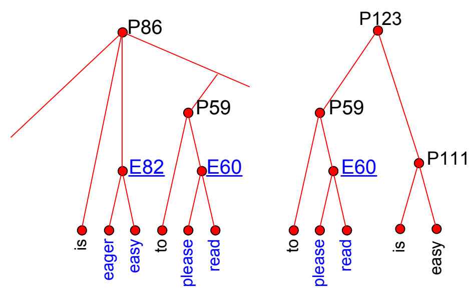

Figure 4: An illustration of the ability of ADIOS to deal with certain structure-dependent syntactic phenomena. In this example, when trained on sentences exemplifying the so-called "tough movement", ADIOS forms patterns that represent the correct phrases (...is easy to read, is easy to please, is eager to read, is eager to please, to read is easy and to please is easy), but does not overgeneralize to the incorrect ones (*to read is eager or *to please is eager).

3 Discussion

The ADIOS algorithm differs from other methods of grammar induction in the data it requires and in the representations it builds, as well as in its algorithmic approach. Most existing approaches require corpora tagged with part-of-speech (POS) information (Clark, 2001); the very few exceptions (Wolff, 1988; Henrichsen, 2002; Adriaans and Vervoort, 2002) are not known to scale up. The extraction of grammatical primitives in published methods may rely on collocation frequencies (Wolff, 1988), or on global criteria such as the likelihood of the entire corpus given the grammar (Lari and Young, 1990; Stolcke and Omohundro, 1994; de Marcken, 1996; Clark, 2001; Henrichsen, 2002). In comparison, ADIOS carries out its inferences locally, in the context provided by the current search path, alleviating the credit assignment problem in learning, and making productive use of learned structures safer. Furthermore, ADIOS works with raw text or transcribed speech, and makes no prior assumptions about the structures it seeks. At the same time, the patterns and equivalence classes it learns can be translated in a straightforward manner into the form of context-sensitive rewriting rules. These representations are both expressive enough to support extensive generativity, and, in principle, restrictive enough to capture many of the structure-sensitive aspects of syntax (Phillips, 2003) documented by linguists; examples include long-range agreement (Figure 1F) and tough movement (Figure 4).

The massive, largely unsupervised, effortless and fast feat of learning that is the acquisition of language by children has long been a daunting challenge for cognitive scientists (Chomsky, 1986; Elman et al., 1996) and for natural language engineers (Bod, 1998; Clark, 2001; Roberts and Atwell, 2002). Because a completely bias-free unsupervised learning is impossible (Chomsky, 1986; Nowak et al., 2001; Baum, 2004), the real issue in language acquisition is to determine the constraints that a model of "grammar induction" should impose — and to characterize those constraints that infants acquiring language do in fact impose — on the learning procedure. In our approach, the constraints are defined algorithmically, in the form of a method for detecting, in sequential symbolic data, of units (patterns and equivalence classes) that are hierarchically structured and are supported by context-sensitive statistical evidence.

Our method should be of interest to linguists of various theoretical persuasions who construe grammars as containing — in addition to general and lexicalized (Daumé et al., 2002; Geman and Johnson, 2003) rules — "inventories" of units of varying kinds and sizes (Langacker, 1987; Daelemans, 1998) such as: idioms and semi-productive forms (Jackendoff, 1997; Erman and Warren, 2000), prefabricated expressions (Makkai, 1995; Wray, 2000), "syntactic nuts" (Culicover, 1999), frequent collocations (Bybee and Hopper, 2001), multiword expressions (Sag et al., 2002; Baldwin et al., 2003), and constructions (Kay and Fillmore, 1999; Croft, 2001; Goldberg, 2003; Tomasello, 2003). In addition, the growing collection of patterns revealed by our algorithm in various corpora should complement both syntax-related resources such as the Penn Treebank (Marcus et al., 1994) and semantics-oriented resources such as the WordNet (Miller and Fellbaum, 1991), the PhraseNet (Li et al., 2003), and the Berkeley FrameNet (Baker et al., 1998, 2003).

Acknowledgments. Thanks to the US-Israel Binational Science Foundation, the Dan David Prize Foundation and the Horowitz Center for Complexity Science for support; to A. Clark, B. Finlay, M. Johnson, R. Katzir, D. Klein, Y. Oshrat, B. Sanbank, R. Sharan and E. Stabler for comments, and to Bo Pedersen for help with psycholinguistic testing of trained ADIOS instances.

References

Adriaans, P. and van Zaanen, M. (2004). Computational grammar induction for linguists. *Grammars*, 7:57-68.

Adriaans, P. and Vervoort, M. (2002). The EMILE 4.1 grammar induction toolbox. In Adriaans, P., Fernau, H., and van Zaanen, M., editors, *Grammatical Inference: Algorithms and Applications: 6th International Colloquium: ICGI 2002*, volume 2484 of *Lecture Notes in Computer Science*, page 293, Heidelberg. Springer-Verlag.

Baker, C. F., Fillmore, C. J., and Cronin, B. (2003). The structure of the FrameNet database. *International Journal of Lexicography*, 16:281-296.

Baker, C. F., Fillmore, C. J., and Lowe, J. B. (1998). The Berkeley FrameNet project. In *Proceedings of the COLING-ACL*, Montreal, Canada.

Baldwin, T., Bannard, C., Tanaka, T., and Widdows, D. (2003). An empirical model of multiword expression decomposability. In *Proceedings of the ACL-2003 Workshop on Multiword Expressions: Analysis, Acquisition and Treatment*, pp.89-96, Sapporo, Japan.

Baum, E. B. (2004). *What is thought?* MIT Press, Cambridge, MA.

Bod, R. (1998). *Beyond grammar: an experience-based theory of language*. CSLI Publications, Stanford, US.

Bybee, J. and Hopper, P. (2001). Introduction. In Bybee, J. and Hopper, P., editors, *Frequency and the emergence of linguistic structure*, pp.1-24. John Benjamins, Amsterdam.

Chomsky, N. (1986). *Knowledge of language: its nature, origin, and use*. Praeger, New York.

Clark, A. (2001). *Unsupervised Language Acquisition: Theory and Practice*. PhD thesis, COGS, University of Sussex.

Croft, W. (2001). *Radical Construction Grammar: syntactic theory in typological perspective*. Oxford University Press, Oxford.

Culicover, P. W. (1999). *Syntactic nuts: hard cases, syntactic theory, and language acquisition*. Oxford University Press, Oxford.

Daelemans, W. (1998). Toward an exemplar-based computational model for cognitive grammar. In van der Auwera, J., Durieux, F., and Lejeune, L., editors, *English as a human language*, pp.73-82. LINCOM Europa, Munchen.

Daumé, H., Knight, K., Langkilde-Geary, I., Marcu, D., and Yamada, K. (2002). The importance of lexicalized syntax models for natural language generation tasks. In *Proceedings of the 2002 International Conference on Natural Language Generation (INLG - 2002)*, pp.9-16, Harriman, NY.

de Marcken, C. G. (1996). *Unsupervised Language Acquisition*. PhD thesis, MIT.

Edelman, S., Solan, Z., Horn, D., and Ruppin, E. (2004). Bridging computational, formal and psycholinguistic approaches to language. In *Proc. of the 26th Conference of the Cognitive Science Society*, Chicago, IL.

Elman, J. L., Bates, E. A., Johnson, M. H., Karmiloff-Smith, A., Parisi, D., and Plunkett, K. (1996). *Rethinking innateness: A connectionist perspective on development*. MIT Press, Cambridge, MA.

Erman, B. and Warren, B. (2000). The idiom principle and the open-choice principle. *Text*, 20:29-62.

Geman, S. and Johnson, M. (2003). Probability and statistics in computational linguistics, a brief review. In Johnson, M., Khudanpur, S., Ostendorf, M., and Rosenfeld, R., editors, *Mathematical foundations of speech and language processing*, volume 138 of *IMA Volumes in Mathematics and its Applications*, pages 1-26. Springer-Verlag, New York.

Goldberg, A. E. (2003). Constructions: a new theoretical approach to language. *Trends in Cognitive Sciences*, 7:219-224.

Gómez, R. L. (2002). Variability and detection of invariant structure. *Psychological Science*, 13:431-436.

Harris, C. L. (1998). Psycholinguistic studies of entrenchment. In Koenig, J., editor, *Conceptual Structures, Language and Discourse*, volume 2. CSLI, Berkeley, CA.

Harris, Z. S. (1954). Distributional structure. *Word*, 10:140-162.

Henrichsen, P. J. (2002). GraSp: Grammar learning form unlabeled speech corpora. In *Proceedings of CoNLL-2002*, pp.22-28. Taipei, Taiwan.

Hopcroft, J. E. and Ullman, J. D. (1979). *Introduction to Automata Theory, Languages, and Computation*. Addison-Wesley, Reading, MA.

Jackendoff, R. (1997). *The Architecture of the Language Faculty*. MIT Press, Cambridge, MA.

Kay, P. and Fillmore, C. J. (1999). Grammatical constructions and linguistic generalizations: the What's X Doing Y? construction. *Language*, 75:1-33.

Klein, D. and Manning, C. D. (2002). Natural language grammar induction using a constituent-context model. In Dietterich, T. G., Becker, S., and Ghahramani, Z., editors, *Advances in Neural Information Processing Systems 14*, Cambridge, MA. MIT Press.

Langacker, R. W. (1987). *Foundations of cognitive grammar*, volume I: theoretical prerequisites. Stanford University Press, Stanford, CA.

Lari, K. and Young, S. (1990). The estimation of stochastic context-free grammars using the inside-outside algorithm. *Computer Speech and Language*, 4:35-56.

Li, X., Roth, D., and Tu, Y. (2003). Phrasenet: towards context sensitive lexical semantics. In Daelemans, W. and Osborne, M., editors, *Proceedings of CoNLL-2003*, pp.87-94, Edmonton, Canada.

MacWhinney, B. and Snow, C. (1985). The Child Language Exchange System. *Journal of Computational Lingustics*, 12:271-296.

Makkai, A. (1995). *Homo Loquens* as 'sender-receiver' (i.e., *tranceiver*) and the *raison d'être* of sememic, lexemic and morphemic prefabs in natural language structures and language use. In Landsberg, M. E., editor, *Syntactic iconicity and linguistic freezes*, pp.91-116. Mouton de Gruyter, Berlin.

Marcus, G. F., Vijayan, S., Rao, S. B., and Vishton, P. M. (1999). Rule learning by seven-month-old infants. *Science*, 283:77-80.

Marcus, M. P., Santorini, B., and Marcinkiewicz, M. A. (1994). Building a large annotated corpus of English: The Penn Treebank. *Computational Linguistics*, 19(2):313-330.

Miller, G. A. and Fellbaum, C. (1991). Semantic networks of English. *Cognition*, 41:197-229.

Moore, B. and Carroll, J. (2001). Parser comparison – context-free grammar (CFG) data. online at http://www.informatics.susx.ac.uk/research/nlp/carroll/cfg-resources/.

Nowak, M. A., Komarova, N. L., and Niyogi, P. (2001). Evolution of universal grammar. *Science*, 291:114-119.

Peña, M., Bonatti, L. L., Nespor, M., and Mehler, J. (2002). Signal-driven computations in speech processing. *Science*, 298:604-607.

Pereira, F. (2000). Formal grammar and information theory: Together again? *Philosophical Transactions of the Royal Society*, 358(1769):1239-1253.

Phillips, C. (2003). Syntax. In Nadel, L., editor, *Encyclopedia of Cognitive Science*, volume 4, pp.319-329. Macmillan, London.

Resnik, P., Olsen, M. B., and Diab, M. (1999). The Bible as a parallel corpus: annotating the ‘Book of 2000 Tongues’. *Computers and the Humanities*, 33:129-153. parallel Bible online at http://www.umiacs.umd.edu/∼resnik/parallel/bible.html.

Roberts, A. and Atwell, E. (2002). Unsupervised grammar inference systems for natural language. Technical Report 2002.20, School of Computing, University of Leeds.

Saffran, J. R., Aslin, R. N., and Newport, E. L. (1996). Statistical learning by 8-month-old infants. *Science*, 274:1926-1928.

Sag, I. A., Baldwin, T., Bond, F., Copestake, A., and Flickinger, D. (2002). Multiword expressions: A pain in the neck for NLP. In *Proceedings of the Third International Conference on Intelligent Text Processing and Computational Linguistics (CICLING 2002)*, pp.1-15, Mexico City, Mexico.

Seidenberg, M. S. (1997). Language acquisition and use: Learning and applying probabilistic constraints. *Science*, 275:1599-1603.

Seidenberg, M. S., MacDonald, M. C., and Saffran, J. R. (2002). Does grammar start where statistics stop? *Science*, 298:553-554.

Solan, Z., Horn, D., Ruppin, E., and Edelman, S. (2004). Unsupervised context sensitive language acquisition from a large corpus. In Saul, L., editor, *Advances in Neural Information Processing*, volume 16, Cambridge, MA. MIT Press.

Stolcke, A. and Omohundro, S. (1994). Inducing probabilistic grammars by Bayesian model merging. In Carrasco, R. C. and Oncina, J., editors, *Grammatical Inference and Applications*, pp.106-118. Springer.

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

van Zaanen, M. (2000). ABL: Alignment-Based Learning. In *COLING 2000 - Proceedings of the 18th International Conference on Computational Linguistics*, pp.961-967.

Wolff, J. G. (1988). Learning syntax and meanings through optimization and distributional analysis. In Levy, Y., Schlesinger, I. M., and Braine, M. D. S., editors, *Categories and Processes in Language Acquisition*, pp.179-215. Lawrence Erlbaum, Hillsdale, NJ.

Wray, A. (2000). Holistic utterances in protolanguage: the link from primates to humans. In Knight, C., Studdert-Kennedy, M., and Hurford, J. R., editors, *The evolutionary emergence of language*, pp.285-302. Cambridge University Press, Cambridge.