Bridging computational, formal and psycholinguistic approaches to language

Shimon Edelman
Department of Psychology
Cornell University
Ithaca, NY 14853, USA
se37@cornell.edu

Zach Solan, David Horn, Eytan Ruppin
Faculty of Exact Sciences
Tel Aviv University
Tel Aviv, Israel 69978
{*zsolan,horn,ruppin*}*@post.tau.ac.il*

Abstract

We compare our model of unsupervised learning of linguistic structures, ADIOS [1, 2, 3], to some recent work in computational linguistics and in grammar theory. Our approach resembles the Construction Grammar in its general philosophy (e.g., in its reliance on structural generalizations rather than on syntax projected by the lexicon, as in the current generative theories), and the Tree Adjoining Grammar in its computational characteristics (e.g., in its apparent affinity with Mildly Context Sensitive Languages). The representations learned by our algorithm are truly emergent from the (unannotated) corpus data, whereas those found in published works on cognitive and construction grammars and on TAGs are hand-tailored. Thus, our results complement and extend both the computational and the more linguistically oriented research into language acquisition. We conclude by suggesting how empirical and formal study of language can be best integrated.

The empirical problem of language acquisition

The acquisition of language by children — a largely unsupervised, amazingly fast and almost invariably successful learning stint — has long been the envy of natural language engineers [4, 5, 6] and a daunting enigma for cognitive scientists [7, 8]. Computational models of language acquisition or "grammar induction" are usually divided into two categories, depending on whether they subscribe to the classical generative theory of syntax, or invoke "general-purpose" statistical learning mechanisms. We believe that polarization between classical and statistical approaches to syntax hampers the integration of the stronger aspects of each method into a common powerful framework. On the one hand, the statistical approach is geared to take advantage of the considerable progress made to date in the areas of distributed representation, probabilistic learning, and "connectionist" modeling, yet generic connectionist architectures are ill-suited to the abstraction and processing of symbolic information. On the other hand, classical rule-based systems excel in just those tasks, yet are brittle and difficult to train.

We are developing an approach to the acquisition of distributional information from raw input (e.g., transcribed speech corpora) that also supports the distillation of structural regularities comparable to those captured by Context Sensitive Grammars out of the accrued statistical knowledge. In thinking about such regularities, we adopt Langacker's notion of grammar as "simply an inventory of linguistic units" ([9], p.63). To detect potentially useful units, we identify and process partially redundant sentences that share the same word sequences. We note that the detection of paradigmatic variation within a slot in a set of otherwise identical aligned sequences (syntagms) is the basis for the classical distributional theory of language [10], as well as for some modern works [11]. Likewise, the *pattern* — the syntagm and the *equivalence class* of complementary-distribution symbols that may appear in its open slot — is the main representational building block of our system, ADIOS (for Automatic DIstillation Of Structure).

Our goal in the present paper is to help bridge statistical and formal approaches to language [12] by placing our work on the unsupervised learning of structure in the context of current research in grammar acquisition in computational linguistics, and at the same time to link it to certain formal theories of grammar. Consequently, the following sections outline the main computational principles behind the ADIOS model, and compare these to select approaches from computational and formal linguistics. The algorithmic details of our approach and accounts of its learning from CHILDES corpora and performance in various tests appear elsewhere [1, 2, 3]. In this paper, we chose to exert a tight control over the target language by using a context-free grammar (Figure 1) to generate the learning and testing corpora.

```
S:    P100 | P101 | P102 ;
P100: P1  P2  P3  P4;
P101: P18 P6  P7;
P102: P8  P2  P9  P6  P10  P2;
P2:   P11 P12 | P13;
P22:  P11 P12;
P11:  the | a;
P12:  cat | dog | cow | bird | rabbit | horse;
P13:  P14 P32 | P14;
P14:  Joe | Beth | Jim | Cindy | Pam | George;
P15:  P14 and P14 P36 | P14 P14 and P14 P36;
P16:  Beth | Pam | Cindy;
P3:   P18 and P19;
P32:  who P17 P22 | who P17 P14;
P4:   , don't they ?;
P35: believes | thinks;
P36: believe | think;
P19: P2  P20;
P18: meows | barks;
P20: laughs | jumps | flies;
P9:  is easy | is tough | is eager;
P7:  is easy | is though;
P6:  to please | to read;
P8:  that;
P10: annoys | worries | disturbs | bothers;
P17: scolds | loves | adores | worships;
P18: P14 P35 that P18 | P14 P35 that;
P5:  P16 P35 that P18;
P1:  P15 that P18;
```

Figure 1: the context free grammar used to generate the corpora for the acquisition tests described here.

The principles behind the ADIOS algorithm

The representational power of ADIOS and its capacity for unsupervised learning rest on three principles: (1) probabilistic inference of pattern significance, (2) context-sensitive generalization, and (3) recursive construction of complex patterns. Each of these is described briefly below.

Probabilistic inference of pattern significance. ADIOS represents a corpus of sentences as an initially highly redundant directed graph, in which the vertices are the lexicon entries and the paths correspond, prior to running the algorithm, to corpus sentences. The graph can be informally visualized as a tangle of strands that are partially segregated into *bundles*.

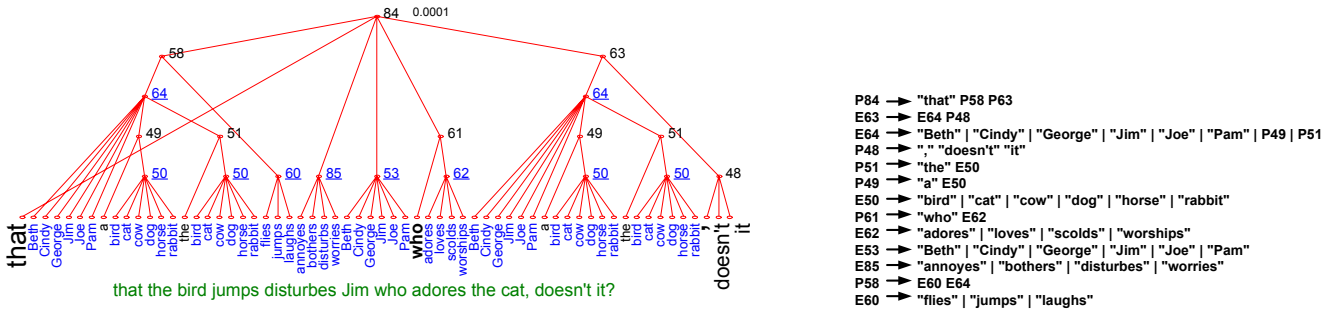

Figure 2: *Left:* a pattern (presented in a tree form), capturing a long range dependency (equivalence class labels are underscored). This and other examples here were distilled from a 400-sentence corpus generated by the grammar of Figure 1. *Right:* the same pattern recast as a set of rewriting rules that can be seen as a Context Free Grammar fragment.

Each of these consists of some strands clumped together; a bundle is formed when two or more strands join together and run in parallel, and is dissolved when more strands leave the bundle than stay in. In a given corpus, there will be many bundles, with each strand (sentence) possibly participating in several. Our algorithm, described in detail elsewhere [3],[1] identifies significant bundles iteratively, using a context-sensitive probabilistic criterion defined in terms of local flow quantities in the graph. The outcome is a set of patterns, each of which is an abstraction of a bundle of sentences that are identical up to variation in one place, where one of several symbols (the members of the equivalence class associated with the pattern) may appear (Figure 2). This representation balances high compression (small size of the pattern lexicon) against good generalization (the ability to generate new grammatical sentences from the acquired patterns).

Context sensitivity of patterns. Because an equivalence class is only defined in the context specified by its parent pattern, the generalization afforded by a set of patterns is inherently safer than in approaches that posit globally valid categories ("parts of speech") and rules ("grammar"). The reliance of ADIOS on many context-sensitive patterns rather than on traditional rules can be compared to the Construction Grammar idea (discussed later), and is in line with Langacker's conception of grammar as a collection of "patterns of all intermediate degrees of generality" ([9], p.46).

Hierarchical structure of patterns. The ADIOS graph is rewired every time a new pattern is detected, so that a bundle of strings subsumed by it is represented by a single new edge. Following the rewiring, which is context-specific, potentially far-apart symbols that used to straddle the newly abstracted pattern become close neighbors. Patterns thus become hierarchically structured in that their elements may be either terminals (i.e., fully specified strings) or other patterns. The ability of new patterns and equivalence classes to incorporate those added previously leads to the emergence of recursively structured units that support generalization (by opening paths that do not exist in the original corpus). Moreover, patterns may refer to themselves, which opens the door for true recursion (Figure 3, right; automatic detection of recursion is not currently implemented).

[1]The relevant publications can be found online at http://kybele.psych.cornell.edu/∼edelman/archive.html.

Two experiments in grammar induction

The results outlined next focus on the power of the ADIOS algorithm, which we assessed by examining the (so-called "weak") generativity of the representations it learns.

Experiment 1. In the first of the two studies described here, we trained ADIOS on 400 sentences produced by the context free grammar shown in Figure 1. We then compared a corpus C_{target} of $3,607,240$ sentences generated by this CFG (with up to three levels of recursion) with a corpus $C_{learned}$ of $1,916,061$ sentences generated by the patterns that had been learned by ADIOS from the 400-sentence training set. In both cases the sentences were generated randomly in batches of size $1.5{\cdot}10^7$ and merged until convergence, defined as 95% overlap between new and existing data. With these data, we obtained precision of 97%, with a recall value of 53% (as customary in computational linguistics, we define recall as the proportion of C_{target} sentences appearing in $C_{learned}$, and precision as the proportion of $C_{learned}$ appearing in C_{target}). In this demonstration, no attempt was made to optimize the two parameters that control pattern acquisition.

Experiment 2. The second experiment involved two ADIOS instances: a teacher and a student. In each of the four runs, the teacher was pre-loaded with a ready-made context free grammar (using the straightforward translation of CFG rules into patterns), then used to generate a series of training corpora with up to 6400 sentences, each with up to seven levels of recursion. After training in each run i ($i = [1 \ldots 4]$), a student-generated test corpus $C^{(i)}_{learned}$ of size 10000 was used in conjunction with a test corpus $C^{(i)}_{target}$ of the same size produced by the teacher, to calculate precision and recall. This was done by running the teacher as a parser on $C^{(i)}_{learned}$ and the student – as a parser on $C^{(i)}_{target}$. The results, plotted in Figure 4, indicate a substantial capacity for unsupervised induction of context-free grammars even from very small corpora.

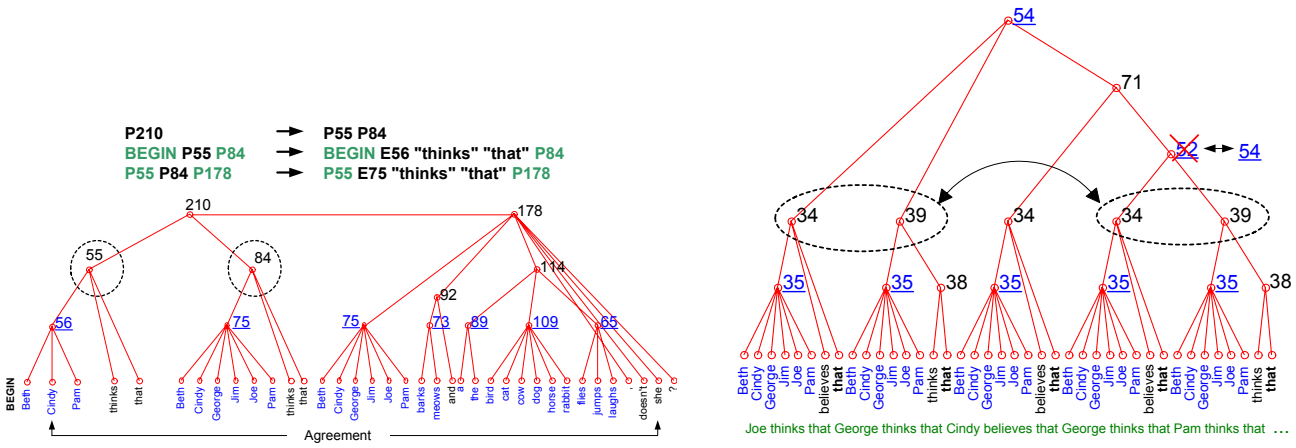

Figure 3: *Left:* because ADIOS does not rewire all the occurrences of a specific pattern, but only those that share the same context, its power is comparable to that of Context Sensitive Grammars. In this example, equivalence class #75 is not extended to subsume the subject position, because that position appears in a different context (e.g., immediately to the right of the symbol BEGIN). Thus, long-range agreement is enforced and over-generalization prevented. The context-sensitive "rules" corresponding to pattern #210 appear above it. *Right:* the ADIOS pattern representation facilitates the detection of recursive structure, exemplified here by the correspondence between equivalence classes #52 and #54.

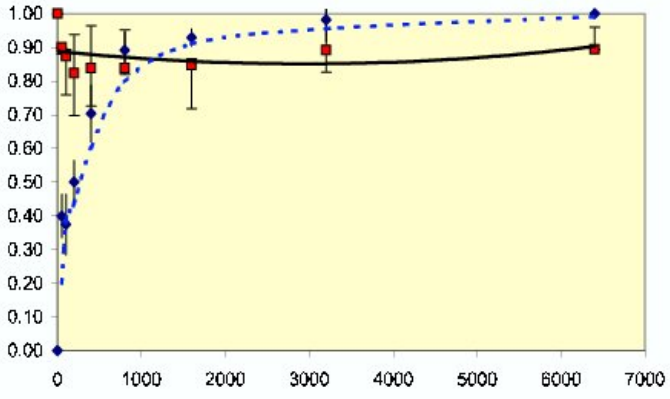

Figure 4: the results of Experiment 2; precision (squares) and recall (diamonds), plotted vs. the size of the training corpus; the error bars are std. dev. computed over four separate training/testing runs. Note that even the largest training corpus size, 6400 sentences, is a tiny proportion of the approximately $1.6 \cdot 10^8$ sentences that can be generated by the target grammar under the chosen depth constraint (7).

Related computational and linguistic formalisms and psycholinguistic findings

Unlike ADIOS, very few existing algorithms for unsupervised language acquisition use raw, unannotated corpus data (as opposed, say, to sentences converted into sequences of POS tags). The only work described in a recent review [6] as completely unsupervised — the GraSp model [13] — does attempt to induce syntax from raw transcribed speech, yet it is not completely data-driven in that it makes a prior commitment to a particular theory of syntax (Categorial Grammar, complete with a pre-specified set of allowed categories). Because of the unique nature of our chosen challenge — finding structure in language rather than imposing it — the following brief survey of grammar induction focuses on contrasts and comparisons to approaches that generally stop short of attempting to do what our algorithm does. We distinguish below between approaches that are motivated by computational considerations (Local Grammar and Variable Order Markov models, and Tree Adjoining Grammar), and those whose main motivation is linguistic and cognitive psychological (Cognitive and Construction grammars).

Local Grammar and Markov models. In capturing the regularities inherent in multiple criss-crossing paths through a corpus, ADIOS superficially resembles finite-state Local Grammars [14] and Variable Order Markov (VOM) models [15] that aim to produce a minimum-entropy finite-state encoding of a corpus. There are, however, crucial differences, as explained below. Our pattern significance criteria [3] involve conditional probabilities of the form $P(e_n|e_1, e_2, e_3, \ldots, e_{n-1})$, which does bring to mind an n'th-order Markov chain, with the (variable) n corresponding roughly to the length of the sentences we deal with. The VOM approach starts out by postulating a maximum-n VOM structure, which is then fitted to the data. The maximum VOM order n, which effectively determines the size of the window under consideration, is in practice much smaller than in our approach, because of computational complexity limitations of the VOM algorithms. The final parameters of the VOM are set by a maximum likelihood condition, fitting the model to the training data. The ADIOS philosophy differs from the VOM approach in several key respects. *First*, rather than fitting a model to the data, we use the data to construct a (recursively structured) graph. Thus, our algorithm naturally addresses the inference of the graph's structure, a task

that is more difficult than the estimation of parameters for a given configuration. *Second*, because ADIOS works from the bottom up in a data-driven fashion, it is not hindered by complexity issues, and can be used on huge graphs, with very large windows sizes. *Third*, ADIOS transcends the idea of VOM structure, in the following sense. Consider a set of patterns of the form $b_1[c_1]b_2[c_2]b_3$, etc. The equivalence classes $[\cdot]$ may include vertices of the graph (both words and word patterns turned into nodes), wild cards (i.e., any node), as well as ambivalent cards (any node or no node). This means that the terminal-level length of the string represented by a pattern does not have to be of a fixed length. This goes conceptually beyond the variable order Markov structure: $b_2[c_2]b_3$ do not have to appear in a Markov chain of a finite order $||b_2|| + ||c_2|| + ||b_3||$ because the size of $[c_2]$ is ill-defined, as explained above. *Fourth*, as we showed earlier (Figure 3), ADIOS incorporates both context-sensitive substitution and recursion.

Tree Adjoining Grammar. The proper place in the Chomsky hierarchy for the class of strings accepted by our model is between Context Free and Context Sensitive Languages. The pattern-based representations employed by ADIOS have counterparts for each of the two composition operations, substitution and adjoining, that characterize a Tree Adjoining Grammar, or TAG, developed by Joshi and others [16]. Specifically, both substitution and adjoining are subsumed in the relationships that hold among ADIOS patterns, such as the membership of one pattern in another. Consider a pattern $\mathcal{P}_i$ and its equivalence class $\mathcal{E}(\mathcal{P}_i)$; any other pattern $\mathcal{P}_j \in \mathcal{E}(\mathcal{P}_i)$ can be seen as substitutable in $\mathcal{P}_i$. Likewise, if $\mathcal{P}_j \in \mathcal{E}(\mathcal{P}_i)$, $\mathcal{P}_k \in \mathcal{E}(\mathcal{P}_i)$ and $\mathcal{P}_k \in \mathcal{E}(\mathcal{P}_j)$, then the pattern $\mathcal{P}_j$ can be seen as adjoinable to $\mathcal{P}_i$. Because of this correspondence between the TAG operations and the ADIOS patterns, we believe that the latter represent regularities that are best described by Mildly Context-Sensitive Language formalism [16]. Importantly, because the ADIOS patterns are learned from data, they already incorporate the constraints on substitution and adjoining that in the original TAG framework must be specified manually.

Psychological and linguistic evidence for pattern-based representations. Recent advances in understanding the psychological role of representations based on what we call patterns, or *constructions* [17], focus on the use of statistical cues such as conditional probabilities in pattern learning [18, 19], and on the importance of exemplars and constructions in children's language acquisition [20]. Converging evidence for the centrality of pattern-like structures is provided by corpus-based studies of the prevalence of "prefabricated" sequences of words [21], and of the entrenchment of such sequences in the lexicon [22]. Similar ideas concerning the ubiquity in syntax of structural peculiarities hitherto marginalized as "exceptions" are now being voiced by linguists [23, 24].

Cognitive Grammar; Construction Grammar. The main methodological tenets of ADIOS — populating the lexicon with "units" of varying complexity and degree of entrenchment, and using cognition-general mechanisms for learning and representation — fit the spirit of the foundations of Cognitive Grammar [9]. At the same time, whereas the cognitive grammarians typically face the chore of hand-crafting structures that would reflect the logic of language as they perceive it, ADIOS discovers the primitives of grammar empirically and autonomously. The same is true also for the comparison between ADIOS and the various Construction Grammars [17, 24], which are all hand-crafted. A construction grammar consists of elements that differ in their complexity and in the degree to which they are specified: an idiom such as "big deal" is a fully specified, immutable construction, whereas the expression "the X, the Y" – as in "the more, the better" [25] – is a partially specified template. The patterns learned by ADIOS likewise vary along the dimensions of complexity and specificity (e.g., not every pattern has an equivalence class).[2]

Related computational work on grammar induction

In natural language processing, a distinction is usually made between unsupervised learning methods that attempt to find good structural primitives and those that merely seek good parameter settings for predefined primitives. ADIOS, which clearly belongs to the first category, is also capable of learning from raw data, whereas most other systems start with corpora annotated by part of speech tags [26], or even rely on treebanks, or collections of hand-parsed sentences [4]. Of the many such methods, we can mention here only a few.

Global grammar optimization using tagged data. Stolcke and Omohundro (1994) learn structure (the topology of a Hidden Markov Model, or the productions of a Stochastic Context Free Grammar), by iteratively maximizing the probability of the current approximation to the target grammar, given the data. In contrast to this approach, which is global in that all the data contribute to the figure of merit at each iteration, ADIOS is local in the sense that its inferences only apply to the current bundle candidate. Another important difference is that instead of general-scope rules stated in terms of parts of speech, we seek context-specific patterns. Perhaps because of its globality and unrestricted-scope rules, Stolcke and Omohundro's method has "difficulties with large-scale natural language applications" [27]. Similar conclusions are reached by Clark, who observes that POS tags are not enough to learn syntax from ("a lot of syntax depends on the idiosyncratic properties of particular words." [5], p.36). His algorithm attempts to learn a phrase-structure grammar from tagged text, by starting with local distributional cues, then filtering spurious non-terminals using a mutual information criterion. In the final stage, his algorithm clusters the results to achieve a minimum description length (MDL) representation, by starting with maximum likelihood grammar, then greedily selecting the candidate for abstraction that would maximally reduce the description length. In its greedy approach to optimization (but not in its local search for good patterns or its ability to deal with untagged data), our approach resembles Clark's.

[2] Similarly to constructions, the ADIOS patterns carry semantic, and not just syntactic, information — an important issue that is outside the scope of the present paper.

Probabilistic treebank-based learning. Bod, whose algorithm learns by gathering information about corpus probabilities of potentially complex trees, observes that "[...] the knowledge of a speaker-hearer cannot be understood as a grammar, but as a statistical ensemble of language experiences that changes slightly every time a new utterance is perceived or produced. The regularities we observe in language may be viewed as emergent phenomena, but they cannot be summarized into a consistent non-redundant system that unequivocally defines the structures of new utterances." [4], p.145. This memory- or analogy-based language model, which is not a typical example of unsupervised learning, is mentioned here mainly because of the parallels between its data representation, Stochastic Tree-Substitution Grammar, and some of the formalisms discussed earlier.

Split and merge pattern learning. The unsupervised structure learning algorithm developed by Wolff between 1970 and 1985 stands out in that it does not need the corpus to be tagged. An excellent survey of his own and earlier attempts at unsupervised learning of language, and of much relevant behavioral data, can be found in [28]. His representations consist of SYN (syntagmatic), PAR (paradigmatic) and M (terminal) elements. Although our patterns and equivalence classes can be seen as analogous to the first two of these, Wolff's learning criterion is much simpler than that of ADIOS: in each iteration, the most frequent pair of contiguous SYN elements are joined together.[3] His system, however, had a unique provision for countering the usual propensity of unsupervised algorithms for overgeneralization: PAR elements that did not admit free substitution among all their members in some context were rebuilt in a context-specific manner. Unfortunately, Wolff's system has not been tested on unconstrained natural language.

Summary, prospects and challenges

The ADIOS approach to the representation of linguistic knowledge resembles the Construction Grammar in its general philosophy (e.g., in its reliance on structural generalizations rather than on syntax projected by the lexicon), and the Tree Adjoining Grammar in its computational capacity (e.g., in its apparent ability to accept Mildly Context Sensitive Languages). The representations learned by the ADIOS algorithm are truly emergent from the (unannotated) corpus data, whereas those found in published works on cognitive and construction grammars and on TAGs are hand-tailored. Thus, our results complement and extend both the computational and the more linguistically oriented research into cognitive/construction grammar.

To further the cause of an integrated understanding of language, a crucial challenge must be met: a viable approach to the evaluation of performance of an unsupervised language learner must be developed, allowing testing both (1) neutral with respect to the linguistic dogma, and (2) cognizant of the plethora of phenomena documented by linguists over the course of the past half century (see, e.g., Figure 5).

[3] An even simpler criterion, that of mere repetition, is employed by the related approach of [29], resulting in a rule set that appears to grow linearly with the size of the corpus, rather than reaching an asymptote as in our case.

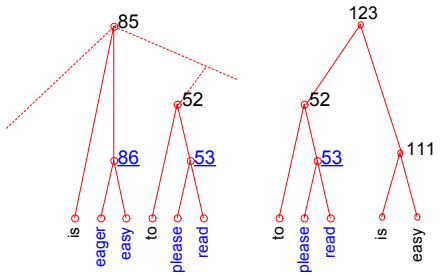

Figure 5: As a token of our intention to account, eventually, for the entire spectrum of English syntax-related phenomena described in the textbooks — agreement, anaphora, auxiliaries, *wh*-questions, passive, control, etc. [30] — we illustrate here the manner in which ADIOS treats tough movement (another phenomenon, long-range agreement, was discussed in Figure 2). When trained on sentences exemplifying "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 over-generalize to the incorrect ones (*to read is eager or *to please is eager).

Unsupervised grammar induction algorithms that work from raw data are in principle difficult to test, because any "gold standard" to which the acquired representation can be compared (such as the Penn Treebank [31]) invariably reflects its designers' preconceptions about language, which may not be valid, and which usually are controversial among linguists themselves [32]. Moreover a child "...must generalize from the sample to the language without overgeneralizing into the area of utterances which are not in the language. *What makes the problem tricky is that both kinds of generalization, by definition, have zero frequency in the child's experience.*" ([28], p.183, italics in the original). Instead of shifting the onus of explanation for this "miracle" onto some unspecified evolutionary processes (which is what the innate grammar hypothesis amounts to), we suggest that a system such as ADIOS should be tested by monitoring its acceptance of massive amounts of human-generated data, and at the same time by getting human subjects to evaluate sentences generated by the system (note that this makes psycholinguistics a crucial component in the entire undertaking).

A purely empirical approach to the evaluation problem would, however, waste the many valuable insights into the regularities of language accrued by the linguists over decades. Although some empiricists would consider this a fair price for quarantining what they perceive as a runaway theory that got out of touch with psychological and computational reality, we believe that searching for a middle way is a better idea, and that the middle way can be found, if the linguists can be persuaded to try and present their main findings in a theory-neutral manner. From recent reviews of syntax that do attempt to reach out to non-linguists (e.g., [33]), it appears that the core issues on which every designer of a language acquisition system should be focusing are dependencies (such as co-reference) and constraints (such as islands), especially as seen in a typological (cross-linguistic) perspective [24].

Acknowledgment. Supported by the US-Israel Binational Science Foundation.

References

[1] Z. Solan, E. Ruppin, D. Horn, and S. Edelman. Automatic acquisition and efficient representation of syntactic structures. In S. Thrun, ed., *Advances in Neural Information Processing (NIPS)*, vol. 15, Cambridge, MA, 2003. MIT Press.

[2] Z. Solan, E. Ruppin, D. Horn, and S. Edelman. Unsupervised efficient learning and representation of language structure. In R. Alterman and D. Kirsh, eds., *Proc. 25th Conf. of the Cognitive Science Society*, Hillsdale, NJ, 2003. Erlbaum.

[3] Z. Solan, D. Horn, E. Ruppin, and S. Edelman. Unsupervised context sensitive language acquisition from a large corpus. In L. Saul, ed., *NIPS*, vol. 16, Cambridge, MA, 2004. MIT Press.

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

[5] A. Clark. *Unsupervised Language Acquisition: Theory and Practice*. PhD thesis, COGS, University of Sussex, 2001.

[6] A. Roberts and E. Atwell. Unsupervised grammar inference systems for natural language. TR 2002.20, School of Computing, University of Leeds, 2002.

[7] N. Chomsky. *Knowledge of language: its nature, origin, and use*. Praeger, New York, 1986.

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

[9] R. W. Langacker. *Foundations of cognitive grammar*, vol.I: theoretical prerequisites. Stanford University Press, Stanford, CA, 1987.

[10] Z. S. Harris. Distributional structure. *Word*, 10:140–162, 1954.

[11] M. van Zaanen. ABL: Alignment-Based Learning. In *COLING 2000 - Proceedings of the 18th International Conf. on Computational Linguistics*, pp. 961–967, 2000.

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

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

[14] M. Gross. The construction of local grammars. In E. Roche and Y. Schabès, eds., *Finite-State Language Processing*, pp. 329–354. MIT Press, Cambridge, MA, 1997.

[15] I. Guyon and F. Pereira. Design of a linguistic postprocessor using Variable Memory Length Markov Models. In *Proc. 3rd Int'l Conf. Document Analysis and Recogition*, pp. 454–457, Montreal, Canada, 1995.

[16] A. Joshi and Y. Schabes. Tree-Adjoining Grammars. In G. Rozenberg and A. Salomaa, eds., *Handbook of Formal Languages*, vol. 3, pp. 69–124. Springer, Berlin, 1997.

[17] A. E. Goldberg. Constructions: a new theoretical approach to language. *Trends in Cognitive Sciences*, 7:219–224, 2003.

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

[19] R. L. Gómez and L. Gerken. Infant artificial language learning and language acquisition. *Trends in Cognitive Science*, 6:178–186, 2002.

[20] T. Cameron-Faulkner, E. Lieven, and M. Tomasello. A construction-based analysis of child directed speech. *Cognitive Science*, 27:843–874, 2003.

[21] A. Wray. *Formulaic language and the lexicon*. Cambridge University Press, Cambridge, UK, 2002.

[22] C. L. Harris. Psycholinguistic studies of entrenchment. In J. Koenig, ed., *Conceptual Structures, Language and Discourse*, vol. 2. CSLI, Berkeley, CA, 1998.

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

[24] W. Croft. *Radical Construction Grammar: syntactic theory in typological perspective*. Oxford University Press, Oxford, 2001.

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

[26] D. Klein and C. D. Manning. Natural language grammar induction using a constituent-context model. In T. G. Dietterich, S. Becker, and Z. Ghahramani, eds., *NIPS 14*, Cambridge, MA, 2002. MIT Press.

[27] A. Stolcke and S. Omohundro. Inducing probabilistic grammars by Bayesian model merging. In R. C. Carrasco and J. Oncina, eds., *Grammatical Inference and Applications*, pp. 106–118. Springer, 1994.

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

[29] C. G. Nevill-Manning and I. H. Witten. Identifying hierarchical structure in sequences: A linear-time algorithm. *Journal of Artificial Intelligence Research*, 7:67–82, 1997.

[30] I. A. Sag and T. Wasow. *Syntactic theory: a formal introduction*. CSLI Publications, Stanford, CA, 1999.

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

[32] A. Clark. Unsupervised induction of Stochastic Context-Free Grammars using distributional clustering. In *Proceedings of CoNLL 2001*, Toulouse, 2001.

[33] C. Phillips. Syntax. In L. Nadel, ed., *Encyclopedia of Cognitive Science*, vol. 4, pp. 319–329. Macmillan, London, 2003.