

Some Tests of an Unsupervised Model of Language Acquisition

Bo Pedersen and Shimon Edelman

Department of Psychology
Cornell University
Ithaca, NY 14853, USA
{bp64,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 outline an unsupervised language acquisition algorithm and offer some psycholinguistic support for a model based on it. Our approach resembles the Construction Grammar in its general philosophy, and the Tree Adjoining Grammar in its computational characteristics. The model is trained on a corpus of transcribed child-directed speech (CHILDES). The model's ability to process novel inputs makes it capable of taking various standard tests of English that rely on forced-choice judgment and on magnitude estimation of linguistic acceptability. We report encouraging results from several such tests, and discuss the limitations revealed by other tests in our present method of dealing with novel stimuli.

1 The empirical problem of language acquisition

The largely unsupervised, amazingly fast and almost invariably successful learning stint that is language acquisition by children has long been the envy of computer scientists (Bod, 1998; Clark, 2001; Roberts and Atwell, 2002) and a daunting enigma for linguists (Chomsky, 1986; Elman et al., 1996). 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 and probabilistic learning, 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” (Langacker, 1987), 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 (Harris, 1954), as well as for some modern work (van Zaanen, 2000). 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 short paper is to illustrate some of the capabilities of the representations learned by our method vis a vis standard tests used by developmental psychologists, by second-language instructors, and by linguists. Thus, the main computational principles behind the ADIOS model are outlined here only briefly. The algorithmic details of our approach and accounts of its learning from CHILDES corpora appear elsewhere (Solan et al., 2003a; Solan et al., 2003b; Solan et al., 2004; Edelman et al., 2004).

2 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.

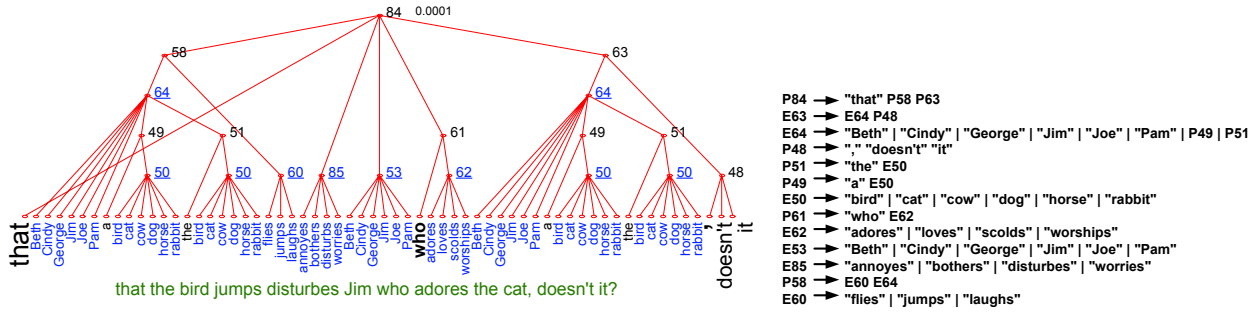


Figure 1: *Left*: a pattern (presented in a tree form), capturing a long range dependency (equivalence class labels are underscored). This and other examples here were distilled from a 400-sentence corpus generated by a 40-rule Context Free Grammar. *Right*: the same pattern recast as a set of rewriting rules that can be seen as a Context Free Grammar fragment.

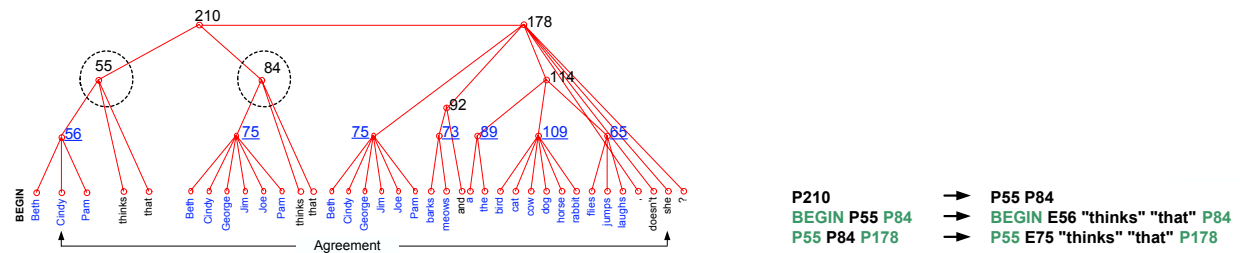


Figure 2: *Left*: because ADIOS does not rewire all the occurrences of a specific pattern, but only those that share the same context, its power is comparable to that of Context Sensitive Grammars. In this example, equivalence class #75 is not extended to subsume the subject position, because that position 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. *Right*: the context-sensitive “rules” corresponding to pattern #210.

Probabilistic inference of pattern significance.

ADIOS represents a corpus of sentences as an initially highly redundant directed graph, which can be informally visualized as a tangle of strands that are partially segregated into *bundles*. 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 in (Solan et al., 2004), identifies significant bundles that balance high compression (small size of the bundle “lexicon”) against good generalization (the ability to generate new grammatical sentences by splicing together various strand fragments each of which belongs to a different bundle).

Context sensitivity of patterns. A pattern 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 (Fig-

ure 1). Because this variation is only allowed in the context specified by the 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 both to the Construction Grammar (discussed later) and to Langacker’s concept of the grammar as a collection of “patterns of all intermediate degrees of generality” ((Langacker, 1987), 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. Moreover, patterns may refer to themselves, which opens the door for recursion.

3 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 (Roberts and Atwell, 2002) as completely unsupervised — the GraSp model (Henrichsen, 2002) — 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 (Categorical 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 between approaches that are motivated computationally (Local Grammar and Variable Order Markov models, and Tree Adjoining Grammar, discussed elsewhere (Edelman et al., 2004), and those whose main motivation is linguistic and cognitive psychological (Cognitive and Construction grammars, discussed below).

Local Grammar and Markov models. In capturing the regularities inherent in multiple crisscrossing paths through a corpus, ADIOS superficially resembles finite-state Local Grammars (Gross, 1997) and Variable Order Markov (VOM) models (Guyon and Pereira, 1995). The VOM approach starts by postulating a maximum- n structure, which is then fitted to the data by maximizing the likelihood of the training corpus. 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 recursive, data-driven fashion, it is less susceptible to complexity issues. It can be used on huge graphs, and may yield very large patterns, which in a VOM model would correspond to an unmanageably high order n . *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 2), 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 (Joshi and Schabes, 1997). 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 (Joshi and Schabes, 1997). 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* (Goldberg, 2003), focus on the use of statistical cues such as conditional probabilities in pattern learning (Saffran et al., 1996; Gómez, 2002), and on the importance of exemplars and constructions in children’s language acquisition (Cameron-Faulkner et al., 2003). Converging evidence for the centrality of pattern-like structures is provided by corpus-based studies of *prefabs* — sequences, continuous or discontinuous, of words that appear to be prefabricated, that is, stored and retrieved as a whole, rather than being subject to syntactic processing (Wray, 2002). Similar ideas concerning the ubiquity in syntax of structural peculiarities hitherto marginalized as “exceptions” are now being voiced by linguists (Culicover, 1999; Croft, 2001).

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 (Langacker, 1987). 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 (Goldberg, 2003; Croft, 2001), 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” (Kay and Fillmore, 1999) — 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).

4 ADIOS: a psycholinguistic evaluation

To illustrate the applicability of our method to real data, we first describe briefly the outcome of running it on a subset of the CHILDES collection (MacWhinney and Snow, 1985), consisting of transcribed speech directed at children. The corpus we selected contained 300,000 sentences (1.3 million tokens) produced by parents. After 14 real-time days, the algorithm (version 7.3) identified 3400 patterns and 3200 equivalence classes. The outcome was encouraging: the algorithm found intuitively significant patterns and produced semantically adequate corresponding equivalence sets. The algorithm’s ability to recombine and reuse the acquired patterns is exemplified in the legend of Figure 3, which lists some of the novel sentences it generated.

The input module. The ADIOS system’s *input module* allows it to process a novel sentence by forming its distributed representation in terms of activities of existing patterns. *We stress that this module plays a crucial role in the tests described below, all of which require dealing with novel inputs.* Figure 4 shows the activation of two patterns (#141 and #120) by a phrase that contains a word in a novel context (*stay*), as well as another word never before encountered in any context (*5pm*).

Acceptability of correct and perturbed novel sentences. To test the quality of the representations

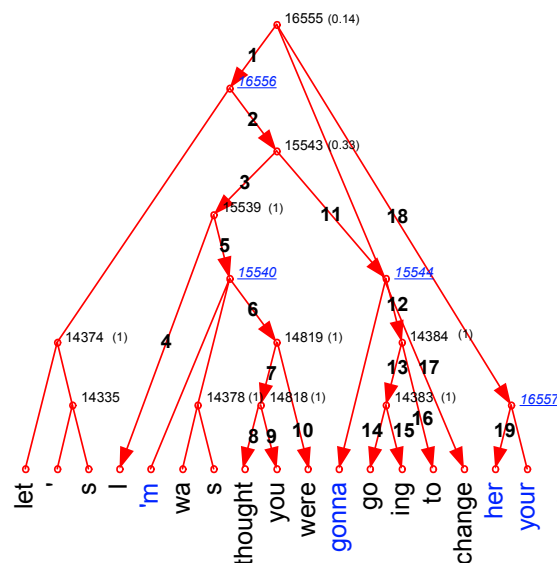


Figure 3: a typical pattern extracted from the CHILDES collection (MacWhinney and Snow, 1985). Hundreds of such patterns and equivalence classes (underscored) together constitute a concise representation of the raw data. Some of the phrases that can be described/generated by these patterns are: let’s change her...; I thought you were gonna change her...; I was going to change your...; none of these appear in the training data, illustrating the ability of ADIOS to generalize. The generation process operates as a depth-first search of the tree corresponding to a pattern. For details see (Solan et al., 2003a; Solan et al., 2004).

(patterns and their associated equivalence classes) acquired by ADIOS, we have examined their ability to support various kinds of grammaticality judgments. The first experiment we report sought to make a distinction between a set of (presumably grammatical) CHILDES sentences not seen by the algorithm during training, and the same sentences in which the word order has been perturbed. We first trained the model on 10,000 sentences from CHILDES, then compared its performance on (1) 1000 previously unseen sentences and (2) the same sentences in each of which a single random word order switch has been carried out. The results, shown in Figure 5, indicate a substantial sensitivity of the ADIOS input module to simple deviations from grammaticality in novel data, even after a very brief training.

Learnability of nonadjacent dependencies

Within the ADIOS framework, the “nonadjacent dependencies” that characterize the artificial languages used by (Gómez, 2002) translate, simply, into patterns with embedded equivalence classes.

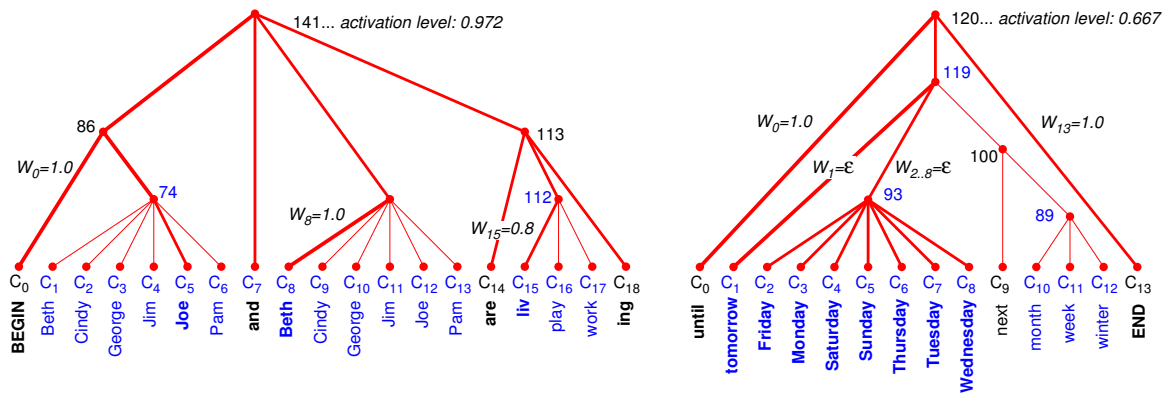


Figure 4: The two most active patterns responding to the partially novel input **Joe and Beth** are staying until 5pm. Leaf activation, which is proportional to the mutual information between input words and various members of the equivalence classes, is propagated upward by taking the average at each junction (Solan et al., 2003a).

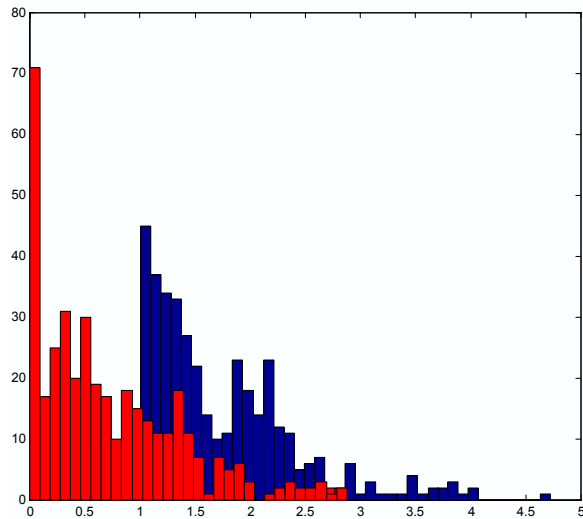


Figure 5: Grammaticality of perturbed sentences (CHILDES data). The figure shows a histogram of the input module output values for two kinds of stimuli: novel grammatical sentences (dark/blue), and sentences obtained from these by a single word-order permutation (light/red).

Gómez showed that the ability of subjects to learn a language L1 of the form $\{aXd, bXe, cXf\}$ ¹, as measured by their ability to distinguish it implicitly from $L2=\{aXe, bXf, cXd\}$, depends on the amount of variation introduced at X . We replicated this experiment by training ADIOS on 432 strings from L1, with $|X| = 2, 6, 12, 24$. The stimuli were the same strings as in the original experiment, with the individual letters serving as the basic symbols. A subsequent test resulted in

¹Symbols $a-f$ here stand for nonce words such as pel, vot, or dak, whereas X denotes a slot in which a subset of 24 other nonce words may appear.

a perfect acceptance of L1 and a perfect rejection of L2. Training with the original words (rather than letters) as the basic symbols resulted in L2 rejection rates of 0%, 55%, 100%, and 100%, respectively, for $|X| = 2, 6, 12, 24$. Thus, the ADIOS performance both mirrors that of the human subjects and suggests a potentially interesting new effect (of the granularity of the input stimuli) that may be explored in further psycholinguistic studies.

A developmental test. The CASL test (Comprehensive Assessment of Spoken Language) is widely used in the USA to assess language comprehension in children (Carrow-Woolfolk, 1999). One of its many components is a grammaticality judgment test, which consists of 57 sentences and is administered as follows: a sentence is read to the child, who then has to decide whether or not it is correct. If not, the child has to suggest a correct version of the sentence. For every incorrect sentence, the test lists 2-3 acceptable correct ones. The present version of the ADIOS algorithm can compare sentences but cannot score single sentences. We therefore ignored 11 out of the 57 sentences, which were correct to begin with. The remaining 46 incorrect sentences and their corrected versions were scored by ADIOS (which for this test had been trained on a 300,000-sentence corpus from the CHILDES database); the highest scoring sentence in each trial was interpreted as the model's choice. The model labeled 17 of the test sentences correctly, yielding a score of 108 (100 = norm) for the age interval 7-0 through 7-2. This score is the norm for the age interval 8-3 through 8-5.²

²ADIOS was undecided about the majority of the other sentences on which it did not score correctly.

benchmark	#items	ADIOS		bi-gram	
		%correct	%answered	%correct	%answered
Linebarger et al., 1983	26	65	65	42	92
Lawrence et al., 2000	70	59	73	38	63
Allen & Seidenberg, 1999	10	83	60	40	50
Martin & Miller, 2002	10	75	80	67	60
Göteborg/ESL	100	58	57	45	20
	216	61	64	42	34

Figure 6: The results of several grammaticality tests (the Göteborg ESL test is described in the text).

ESL test (forced choice). We next used a standard test developed for English as Second Language (ESL) classes, which has been administered in Göteborg (Sweden) to more than 10,000 upper secondary levels students (that is, children who typically had 9 years of school, but only 6-7 years of English). The test consists of 100 three-choice questions, such as *She asked me __ at once* (choices: come, to come, coming) and *The tickets have been paid for, so you __ not worry* (choices: may, dare, need); the average score for the population mentioned is 65%. As before, the choice given the highest score by the algorithm won; if two choices received the same top score, the answer was “don’t know”. The algorithm’s performance in this and several other tests is summarized in Figure 6 (these tests have been conducted with an earlier version of the algorithm (Solan et al., 2003a)). In the ESL test, ADIOS scored at just under 60%; compare this to the 45% precision (with 20% recall) achieved by a straightforward bi-gram benchmark.³

ESL test (magnitude estimation). In this experiment, six subjects were asked to provide magnitude estimates of linguistic acceptability (Gurman-Bard et al., 1996) for all the 3×100 sentences in the Göteborg ESL test. The test was paper based and included the instructions from (Keller, 2000). No measures were taken to randomize the order of the sentences or otherwise control the experiment. The same 300 sentences were processed by ADIOS, whose responses were normalized by dividing the output by the sum of each triplet’s score. The results indicate a significant correlation ($R^2 = 6.3\%$, $p < 0.001$) between the scores produced by the subjects and by ADIOS. In some cases the scores of

³Chance performance in this test is 33%. We note that the corpus used here was too small to train an n -gram model for $n > 2$; thus, our algorithm effectively overcomes the problem of sparse data by putting the available data to a better use.

ADIOS are equal, which usually indicates that there are too many unfamiliar words. Omitting these sentences yields a significant $R^2 = 9.7\%$, $p < 0.001$; removing sentences for which the choices score almost equally (within 10%) results in $R^2 = 12.7\%$, $p < 0.001$.⁴

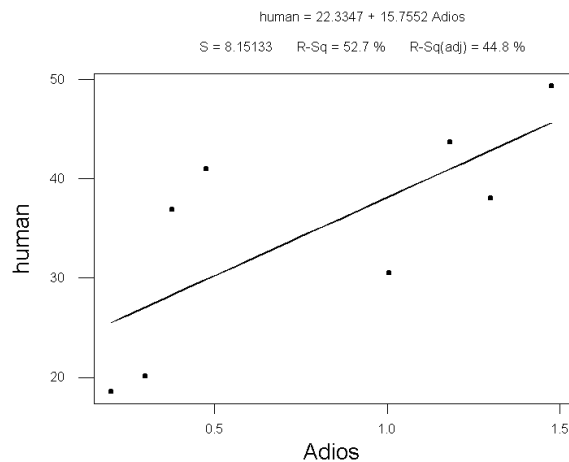


Figure 7: Magnitude estimation study from Keller, plotted against the ADIOS score on the same sentences ($R^2 = 0.53$, $p < 0.05$). The sentences (ranked by increasing score) are:

- How many men did you destroy the picture of?
- How many men did you destroy a picture of?
- How many men did you take the picture of?
- How many men did you take a picture of?
- Which man did you destroy the picture of?
- Which man did you destroy a picture of?
- Which man did you take the picture of?
- Which man did you take a picture of?

Modeling Keller’s data. A manuscript by Frank Keller lists magnitude estimation data for eight sentences.⁵ We compared these to the scores produced by ADIOS, and obtained a significant correlation (Figure 7). The input module seems capable of dealing with the substitution of *a* with *the* or of *take* with *destroy*, and it does reasonably well on the substitution of *How many men* with *Which man*. We conjecture that this performance can be improved by a more sophisticated normalization of the score produced by the input module, which should do a better job quantifying the *cover* (Edelman, 2004) of a novel sentence by the stored patterns. The limitations of the present version of the model became apparent when we tested it on the

⁴Four of the subjects only filled out the test partially (the numbers of responses were 300, 300, 186, 159, 96, 60), but the correlation was highly significant despite the missing data.

⁵<http://elib.uni-stuttgart.de/opus/volltexte/1999/81/pdf/81.pdf>

52 sentences from Keller’s dissertation, using his magnitude estimation method (Keller, 2000).⁶ For these sentences, no correlation was found between the human and the model scores. One of the more challenging aspects of this set is the central role of pronoun binding in many of the sentences, e.g., *The woman/Each woman saw Peter’s photograph of her/herself/him/himself*. Moreover, this test set contains examples of context effects, where information in an earlier sentence can help resolve a later ambiguity. Thus, many of the grammatical contrasts that appear in Keller’s test sentences are too subtle for the present version of the ADIOS input module to handle.

Acceptability of correct and perturbed artificial sentences. In this experiment 64 random sentences was produced with a CFG. For uniformity the sentence length was kept within 15-20 words. 16 of the sentences had two adjacent words switched and another 16 had two random words switched. The 64 sentences were presented to 17 subjects, who placed each on a computer screen at a lateral position reflecting the perceived acceptability. As expected, the perturbed sentences were rated as less acceptable than the non-perturbed ones ($R^2 = 50.3\%$ with $p < 0.01$). We controlled for sentence number, for how high on the screen the sentence was placed, for the reaction time and for sentence length; only the latter had a significant contribution to the correlation. The random permutations scored significantly ($p < 0.01$) lower than the adjacent permutations. Furthermore, the variance in the scores of the randomly permuted sentences was significantly larger ($p < 0.005$), suggesting that this kind of permutation violates the sentence structure more severely, but may also sometimes create acceptable sentences by chance. Previous tests showed that ADIOS is very good at recognizing perturbed CFG-generated sentences as such, but it remains to be seen whether or not ADIOS also exhibits differential behavior on the adjacent and non-adjacent permutations.

Acceptability of ADIOS-generated sentences. ADIOS was trained on 12,700 sentences (out of a total of 12,966 sentences) in the ATIS (Air Travel Information System) corpus; the remaining 226 sentences were used for precision/recall tests. Because

⁶We remark that this methodology is not without its problems. As one of our linguistically naive subjects remarked, “The instructions were (purposefully?) vague about what I was supposed to judge — understandability, grammar, correct use of language, or getting the point through. . .”. Indeed, the scores in a magnitude experiment must be composites of several factors — at the very least, well-formedness and meaningfulness. We are presently exploring various means of acquiring and dealing with such multidimensional “acceptability” data.

ADIOS is sensitive to the presentation order of the training sentences, 30 instances were trained on randomized versions of the training set. Eight human subjects were then asked to estimate acceptability of 20 sentences from the original corpus, intermixed randomly with 20 sentences generated by the trained versions of ADIOS. The precision, calculated as the average number of sentences accepted by the subjects divided by the total number of sentences in the set (20), was 0.73 ± 0.2 for sentences from the original corpus and 0.67 ± 0.07 for the sentences generated by ADIOS. Thus, the ADIOS-generated sentences are, on the average, as acceptable to human subjects as the original ones.

5 Concluding remarks

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. Previous studies focused on the algorithm that makes such learning possible (Solan et al., 2004; Edelman et al., 2004). In the present paper, we concentrated on testing the input module that allows the acquired patterns to be used in processing novel stimuli.

The results of the tests we described here are encouraging, but there is clearly room for improvement. We believe that the most pressing issue in this regard is developing a conceptually and computationally well-founded approach to the notion of cover (that is, a distributed representation of a novel sentence in terms of the existing patterns). Intuitively, the best case, which should receive the top score, is when there is a single pattern that precisely covers the entire input, possibly in addition to other evoked patterns that are only partially active. We are currently investigating various approaches to scoring distributed representations in which several patterns are highly active. A crucial constraint that applies to such cases is that a good cover should give a proper expression to the subtleties of long-range dependencies and binding, many of which are already captured by the ADIOS learning algorithm.

Acknowledgments. Supported by the US-Israel Binational Science Foundation and by the Thanks to Scandinavia Graduate Scholarship at Cornell.

References

- R. Bod. 1998. *Beyond grammar: an experience-based theory of language*. CSLI Publications, Stanford, US.
- T. Cameron-Faulkner, E. Lieven, and M. Tomasello. 2003. A construction-based analysis of child directed speech. *Cognitive Science*, 27:843–874.
- E. Carrow-Woolfolk. 1999. *Comprehensive Assessment of Spoken Language (CASL)*. AGS Publishing, Circle Pines, MN.
- N. Chomsky. 1986. *Knowledge of language: its nature, origin, and use*. Praeger, New York.
- A. Clark. 2001. *Unsupervised Language Acquisition: Theory and Practice*. Ph.D. thesis, COGS, U. of Sussex.
- W. Croft. 2001. *Radical Construction Grammar: syntactic theory in typological perspective*. Oxford U. Press, Oxford.
- P. W. Culicover. 1999. *Syntactic nuts: hard cases, syntactic theory, and language acquisition*. Oxford U. Press, Oxford.
- S. Edelman, Z. Solan, D. Horn, and E. Ruppín. 2004. Bridging computational, formal and psycholinguistic approaches to language. In *Proc. of the 26th Conference of the Cognitive Science Society*, Chicago, IL.
- S. Edelman. 2004. Bridging language with the rest of cognition: computational, algorithmic and neurobiological issues and methods. In M. Gonzalez-Marquez, M. J. Spivey, S. Coulson, and I. Mittelberg, eds., *Proc. of the Ithaca workshop on Empirical Methods in Cognitive Linguistics*. John Benjamins.
- J. L. Elman, E. A. Bates, M. H. Johnson, A. Karmiloff-Smith, D. Parisi, and K. Plunkett. 1996. *Rethinking innateness: A connectionist perspective on development*. MIT Press, Cambridge, MA.
- A. E. Goldberg. 2003. Constructions: a new theoretical approach to language. *Trends in Cognitive Sciences*, 7:219–224.
- R. L. Gómez. 2002. Variability and detection of invariant structure. *Psychological Science*, 13:431–436.
- M. Gross. 1997. The construction of local grammars. In E. Roche and Y. Schabès, eds., *Finite-State Language Processing*, pages 329–354. MIT Press, Cambridge, MA.
- E. Gurman-Bard, D. Robertson, and A. Sorace. 1996. Magnitude estimation of linguistic acceptability. *Language*, 72:32–68.
- I. Guyon and F. Pereira. 1995. Design of a linguistic postprocessor using Variable Memory Length Markov Models. In *Proc. 3rd Int'l Conf. Document Analysis and Recognition*, pages 454–457, Montreal, Canada.
- Z. S. Harris. 1954. Distributional structure. *Word*, 10:140–162.
- P. J. Henrichsen. 2002. GraSp: Grammar learning from unlabeled speech corpora. In *Proceedings of CoNLL-2002*, pages 22–28. Taipei, Taiwan.
- A. Joshi and Y. Schabès. 1997. Tree-Adjoining Grammars. In G. Rozenberg and A. Salomaa, eds., *Handbook of Formal Languages*, volume 3, pages 69 – 124. Springer, Berlin.
- P. Kay and C. J. Fillmore. 1999. Grammatical constructions and linguistic generalizations: the What's X Doing Y? construction. *Language*, 75:1–33.
- F. Keller. 2000. *Gradience in Grammar: Experimental and Computational Aspects of Degrees of Grammaticality*. Ph.D. thesis, U. of Edinburgh.
- R. W. Langacker. 1987. *Foundations of cognitive grammar*, volume I: theoretical prerequisites. Stanford U. Press, Stanford, CA.
- B. MacWhinney and C. Snow. 1985. The Child Language Exchange System. *Journal of Computational Linguistics*, 12:271–296.
- A. Roberts and E. Atwell. 2002. Unsupervised grammar inference systems for natural language. Technical Report 2002.20, School of Computing, U. of Leeds.
- J. R. Saffran, R. N. Aslin, and E. L. Newport. 1996. Statistical learning by 8-month-old infants. *Science*, 274:1926–1928.
- Z. Solan, E. Ruppín, D. Horn, and S. Edelman. 2003a. Automatic acquisition and efficient representation of syntactic structures. In S. Thrun, editor, *Advances in Neural Information Processing*, volume 15, Cambridge, MA. MIT Press.
- Z. Solan, E. Ruppín, D. Horn, and S. Edelman. 2003b. Unsupervised efficient learning and representation of language structure. In R. Alterman and D. Kirsh, eds., *Proc. 25th Conference of the Cognitive Science Society*, Hillsdale, NJ. Erlbaum.
- Z. Solan, D. Horn, E. Ruppín, and S. Edelman. 2004. Unsupervised context sensitive language acquisition from a large corpus. In L. Saul, editor, *Advances in Neural Information Processing*, volume 16, Cambridge, MA. MIT Press.
- M. van Zaanen. 2000. ABL: Alignment-Based Learning. In *COLING 2000 - Proceedings of the 18th International Conference on Computational Linguistics*, pages 961–967.
- A. Wray. 2002. *Formulaic language and the lexicon*. Cambridge U. Press, Cambridge, UK.