Unsupervised language acquisition: syntax from plain corpus

David Horn,[1] Zach Solan,[1] Eytan Ruppin,[2] Shimon Edelman[3]
[1]School of Physics and Astronomy, Tel Aviv University, Tel Aviv 69978, Israel
[2]School of Computer Science, Tel Aviv University, Tel Aviv 69978, Israel
[3]Department of Psychology, Cornell University Ithaca, NY 14853, USA

October 14, 2004

Presented at the Newcastle Workshop on Human Language, Feb. 2004

We describe results of a novel algorithm for grammar induction from a large corpus. The ADIOS **(Automatic DIstillation of Structure) algorithm searches for significant patterns, chosen according to context dependent statistical criteria, and builds a hierarchy of such patterns according to a set of rules leading to structured generalization. The corpus is thus generalized into a context free grammar (CFG), composed of patterns, equivalence classes and words of the initial lexicon. We have evaluated our method both on corpora generated by CFG and on natural language ones. The performance of** ADIOS **is judged by searching for both good recall (acceptance of correct novel sentences) and good precision (production of correct novel sentences). The results are very encouraging.**

Introduction

Fifty years ago (Harris, 1954) suggested that partially aligned sentential contexts reveal clusters of words that correspond to various syntactic categories. With the huge advancement of computational power we are now in an era that can put this idea to a real test. In particular we can confront the interesting challenge of establishing, via statistical methods and simple algorithms, not only whether semantic grouping can be established, but also whether on the basis of pure text one can reach an understanding of the syntax with which the sentences in this text can be generated.

Based on the Harris approach, (van Zaanen, 2000) has introduced a grammar induction method that has had some limited success. We suggest a different method for the extraction of significant patterns that are then identified as new units to be added to the words of the lexicon. We go on introducing rules that ensure structured generalization, thus invoking a principle that is well accepted in linguistics (Chomsky, 1986).

While our general approach is that of Machine Learning, we do not follow the traditional method of testing a set of models and selecting the best suitable for representing the given data. Instead we allow the data to dictate the model. Clearly the rules of our algorithm are such that a special family of models is selected, however it is the data that point the way in a progressive hierarchical search. In the tradition of Machine Learning one considers a training-set, on which the machine is being trained, and a test-set on which its power of generalization is tested. Using liguistic corpora such tests usually go in one direction: seeing whether the machine accepts a new sentence it has not seen before, thus judging it to be grammatical. This defines for us the 'recall' quality. We insist however also on testing 'precision', i.e. whether new

sentences generated by the machine are indeed grammatically correct. This can be measured relatively easily if the original corpus is generated from a known grammar, e.g. some artificial CFG. We will demonstrate that ADIOS passes well such tests.

The ADIOS algorithm for grammar induction

Our algorithm, ADIOS (for Automatic DIstillation of Structure), starts out by loading a corpus of sentences onto a graph, whose vertices are the words appearing in the corpus (i.e. the units of the lexicon). The vertices are 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. 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, the available *search paths* (which initially coincide with the original corpus sentences), seeking sub-paths that exhibit coherent behavior over ordered sets of vertices, i.e., that can be partially aligned (Harris, 1954; van Zaanen, 2000). The significant patterns are selected according to a context-sensitive Motif Extraction (MEX) procedure, defined in terms of local flow quantities in the graph. An example is demonstrated in Figure 1 where the trial-path consists of the set of nodes $A \rightarrow B \rightarrow C \rightarrow D \rightarrow E$. Let us define now P_R, the right-moving ratio of fan-through (through going flux of strings) to fan-in (incoming flux of strings), which varies along the trial-path. Thus at B it will be

$$P_R(B) = \frac{\text{number of paths leading from } A \text{ to } B}{\text{total number of paths entering } A}. \tag{1}$$

This function increases along the trial-path, because the strings form a coherent bundle, but shows a decrease at E. Because of this decrease we declare D to be the end-point of the putative motif. Similarly we proceed from the right end of the trial-path starting with E and study a left- going ratio of fan-through over fan-in. Thus

$$P_L(D) = \frac{\text{number of paths leading from } D \text{ to } E}{\text{total number of paths entering } E} \tag{2}$$

and so on. This function will increase, going to the left, and the point (B) at which it shows a considerable decrease, $D_L(B) = P_L(A)/P_L(B) < \eta$, we declare to be the beginning point of the putative motif.

Generalizing the search path, one looks for an *equivalence class* of units that appear within the same context, which is then identified as a new pattern and added to the graph as a new vertex. The sub-paths that this pattern subsumes are merged into the new vertex, and the graph is rewired accordingly. This is exemplified in Fig. 2. The search for patterns and equivalence classes and their incorporation into the graph are repeated until the first iteration in which no new significant pattern emerges. A full description of the algorithm can be found in (Solan et al., 2004b).

The hierarchical steps involved in ADIOS are demonstrated in Figure 3. The different patterns (P, denoted by numbers) and equivalence classes (E, underlined numbers) are expanded into parse-trees (Hopcroft and Ullman, 1979) that should be read from top to bottom and from left to right (in the order indicated on the numbered arrows). In the generative mode, implied by this representation, only one of the descendants of an equivalence class is chosen. At the last step of the hierarchical construction, the only patterns left on the paths representing the different original sentences, are defined as root-patterns. When totally expanded they are expressed in terms of the original terminals of the graph (i.e. the original letters of the lexicon). Clearly any one sentence of the original corpus is thus generalized to many new sentences that the same

ADIOS representation can generate. The structure of the root-pattern and all its descendant Ps and Es can be represented in the format of a CFG, as shown in Figure 3D.

Testing the Algorithm

To test the algorithm we have applied it to an artificial CFG that we have constructed. It contained 50 rules, including one cyclic relationship. This has been used to generate corpora of various lengths. ADIOS has been trained on these corpora. To asses how well this ADIOS-student grasps the grammar of the teacher, we have tested both its recall and precision. This procedure is demonstrated in Figure 4. Recall is defined by the fraction of novel sentences generated by the teacher (our artificial CFG) that are accepted by the student as grammatically correct, and precision is defined as the fraction of novel sentences generated by the student that are accepted by the teacher. Acceptance in both cases means that the sentence in question can be exactly generated by the rules of the testing party.

The results displayed in Table 1 were obtained from a group of 10 ADIOS-students. Clearly the convergence of students to the teacher grows rapidly in accuracy with the size of the corpus. We find this result very satisfactory, although we know that an ADIOS-student can never coincide with the CFG, for the simple reason that the latter possesses a cyclic relationship among its Ps and Es whereas ADIOS, in its present manifestation, is limited to tree-like structures. In practice we may limit the cyclic rule to a finite number of iterations, in which case it can be instantiated by an ADIOS-teacher representation. In this case, chances are that a student may indeed recapture the correct syntax of the teacher. But these chances will always be small. Moreover, this question is in any case not applicable to real languages, where the true grammar is unknown.

Table 1: Success rate of ADIOS on an artificial CFG.

corpus size	recall	precision
800	$.85 \pm .06$	$.72 \pm .22$
1600	$.87 \pm .06$	$.63 \pm .09$
3200	$.84 \pm .05$	$.61 \pm .12$
6400	$.95 \pm .01$	$.86 \pm .08$

Applications

CHILDES

As a first application we demonstrate results obtained by applying ADIOS to a subset of the CHILDES collection (MacWhinney and Snow, 1985), which consists of transcribed speech produced by, or directed at, children. The corpus we selected contained 300,000 sentences (1.3 million tokens) produced by parents. Some of these results were presented in (Solan et al., 2004a). We have subjected the resulting ADIOS-students to a grammaticality judgment test, in the form of multiple choice questions used in English as Second Language (ESL) classes. The particular test (http://www.forumeducation.net/servlet/pages/vi/mat/gram/dia001.htm) has been administered to more than $10,000$ people in the Göteborg (Sweden) education system as a diagnostic tool when assessing students on upper secondary levels. Clearly the test does not fit the training, in

the sense that the system is trained on syntax directed at 3-year olds, and is tested at a level of 6-7 years of studying English as a second language. Nonetheless ADIOS has reached a success rate of 60%, which is considered intermediate as far as this test is considered. As a benchmark, we compared the performance of ADIOS in this test with that of a word bi-gram model. ADIOS outperformed the bi-gram model by answering 50% of the questions with 60% hits, compared to 20% of the questions with only 45% hits for the bi-gram model (note that chance performance in this test is 33%).

Comparative Syntax

Since ADIOS serves as a grammar induction tool, one may inquire whether, given this tool, one could compare different languages on the basis of their syntax. This can be done if one possesses the same text in the different languages, as is the case for the Parallel Bible (Resnik et al., 1999) corpus that is available in six different natural languages. We have applied ADIOS to the same text in the six different languages, and tested the syntax derivable from them. Here we report on one simple test which we call the spectral test: searching through all patterns that ADIOS distills from the corpus we determine their structures in terms of the three building blocks of the ADIOS representation, patterns (P) equivalence-classes (E) and terminals (T), where the latter signify the original words of the lexicon. We then ask for the probability that a given pattern is constructed out of a particular ordered combination of these elements, as shown in Figure 4.

There are two observations that are very obvious in these results. First, clearly Chinese is very different from the five European languages. Second, the histograms peak at collocation structures of two and three words, TT and TTT. Closer scrutiny of similarities between the grammars of different languages, as described by inner products of the relevant vectors of ADIOS spectra, leads to the conclusion that Spanish is closest to French and Danish is closest to Swedish. All these observations may come as no surprise, but it is satisfying to see them emerging from this system.

Summary

We have presented an algorithm that is capable to distill a CFG from a corpus. It does so in an unsupervised fashion, starting with the words of the corpus as its basic initial elements, and using the sentences as the paths on which it performs its operations.

Our algortihm is based on a statistical search for patterns within given contexts, and on a set of rules on how to generalize these patterns to include equivalence classes and to build a hierarchy of further patterns and equivalence classes. This method can be also used to construct a context sensitive grammar, however the latter calls for a much larger computational complexity (Solan et al., 2004b).

We have tested our method on artificial CFGs and on linguistic data of the type presented here. These results indicate that, for many practical purposes, ADIOS can perform the task to a large degree of accuracy. Moreover, it can achieve high levels of both precision and recall.

The method is scalable to large corpora. To the best of our knowledge, no other method achieves comparable results. We have recently succeeded in applying it to interesting problems in bioinformatics, thus showing that its power of pattern and syntax extraction may help discern not only natural languages but also the language of Nature.

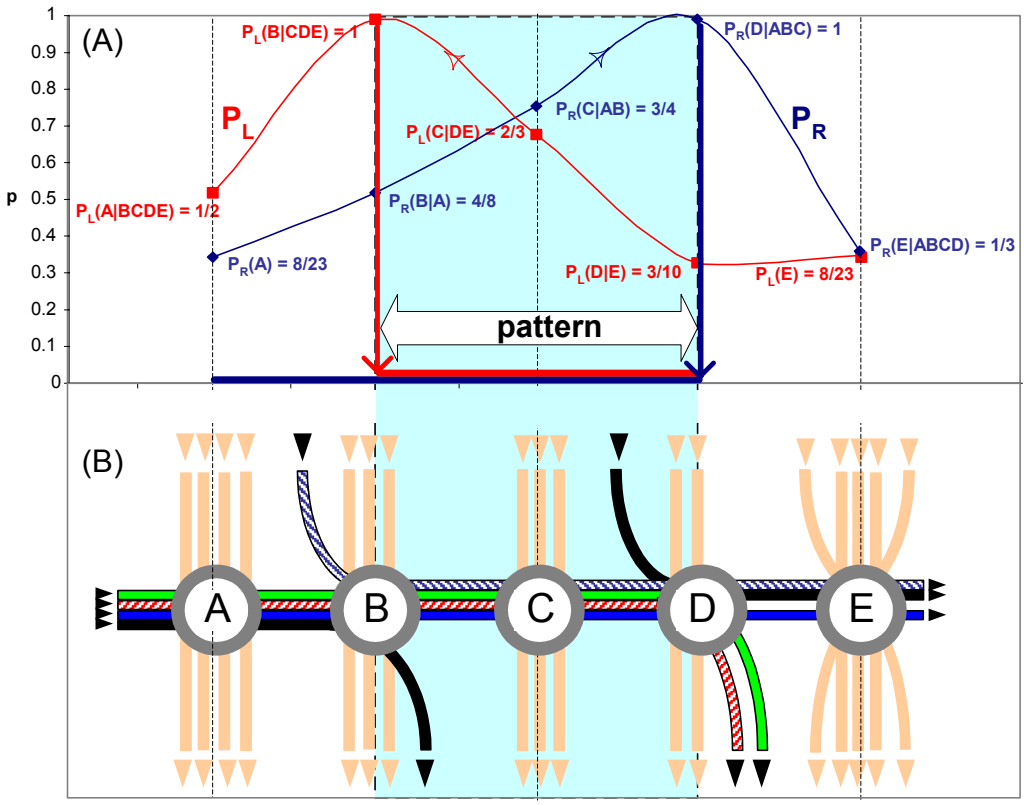

(A)
(B)
p
1
0.9
0.8
0.7
0.6
0.5
0.4
0.3
0.2
0.1
0
P_L(B|CDE) = 1
P_R(D|ABC) = 1
P_L
P_R
P_R(C|AB) = 3/4
P_L(C|DE) = 2/3
P_L(A|BCDE) = 1/2
P_R(B|A) = 4/8
P_R(E|ABCD) = 1/3
P_R(A) = 8/23
P_L(D|E) = 3/10
P_L(E) = 8/23
pattern
A
B
C
D
E

Figure 1.

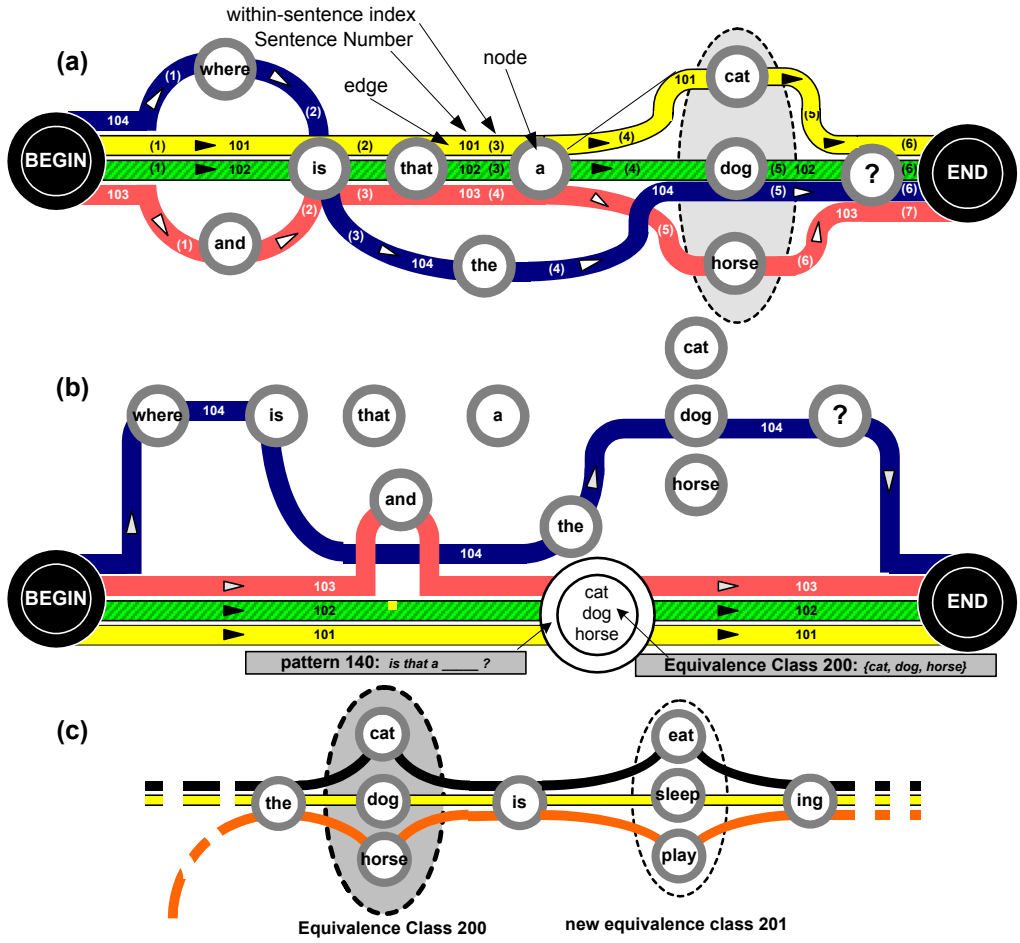
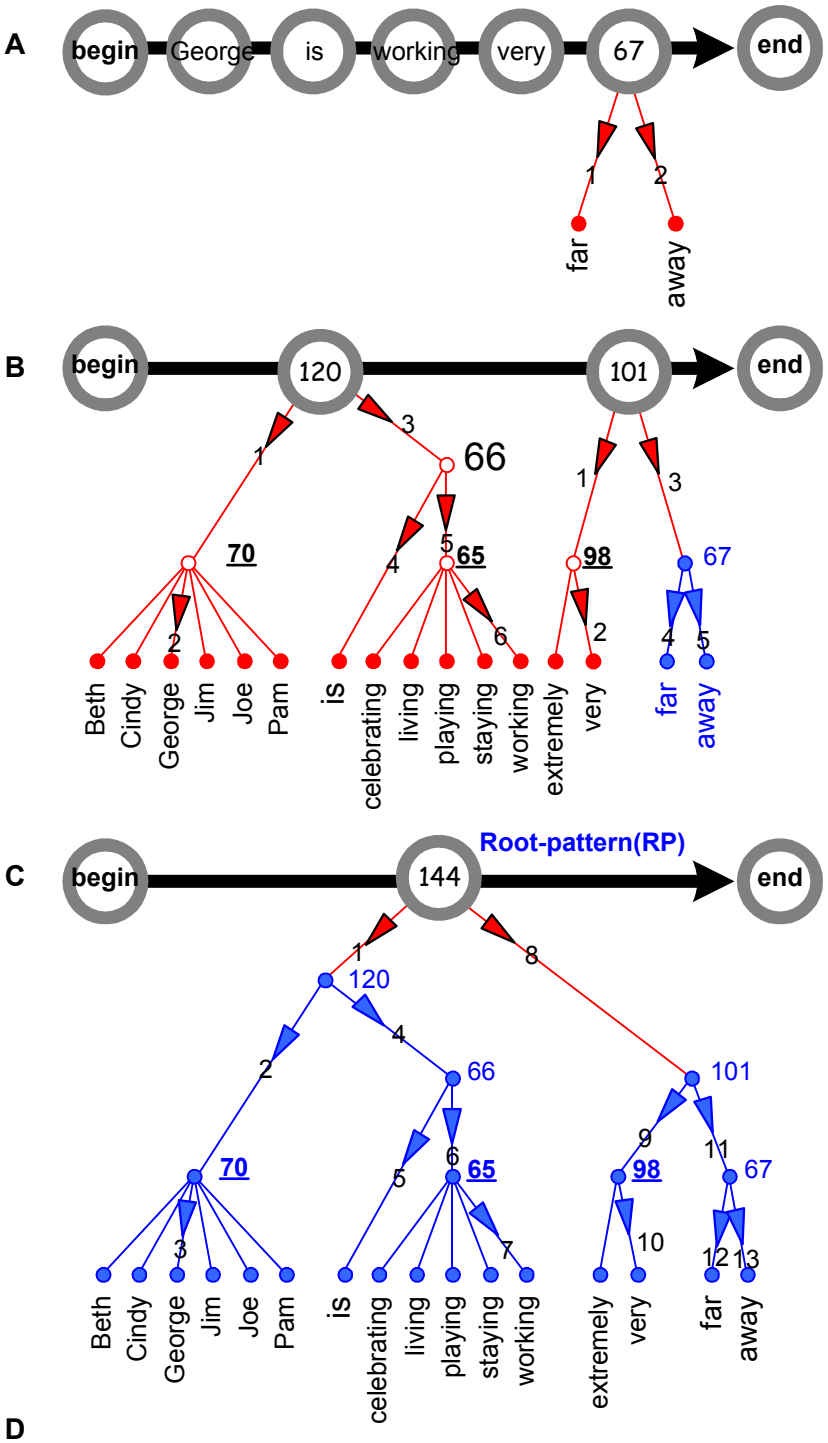

(a)
within-sentence index
Sentence Number
edge
node
where
is
that
a
cat
dog
?
and
the
horse
BEGIN
END
(b)
pattern 140: is that a ____ ?
Equivalence Class 200: {cat, dog, horse}
(c)
eat
sleep
play
ing
Equivalence Class 200
new equivalence class 201

Figure 2.



D

s ⟶ *begin* P144 *end*
P144 ⟶ P120 P101
P120 ⟶ E70 P66
E70 ⟶ *Beth | Cindy...*
P66 ⟶ *is* E65
E65 ⟶ *celebrating | living...*
P101 ⟶ E98 P67
E98 ⟶ *extremely | very*
P67 ⟶ *far away*

Figure 3.

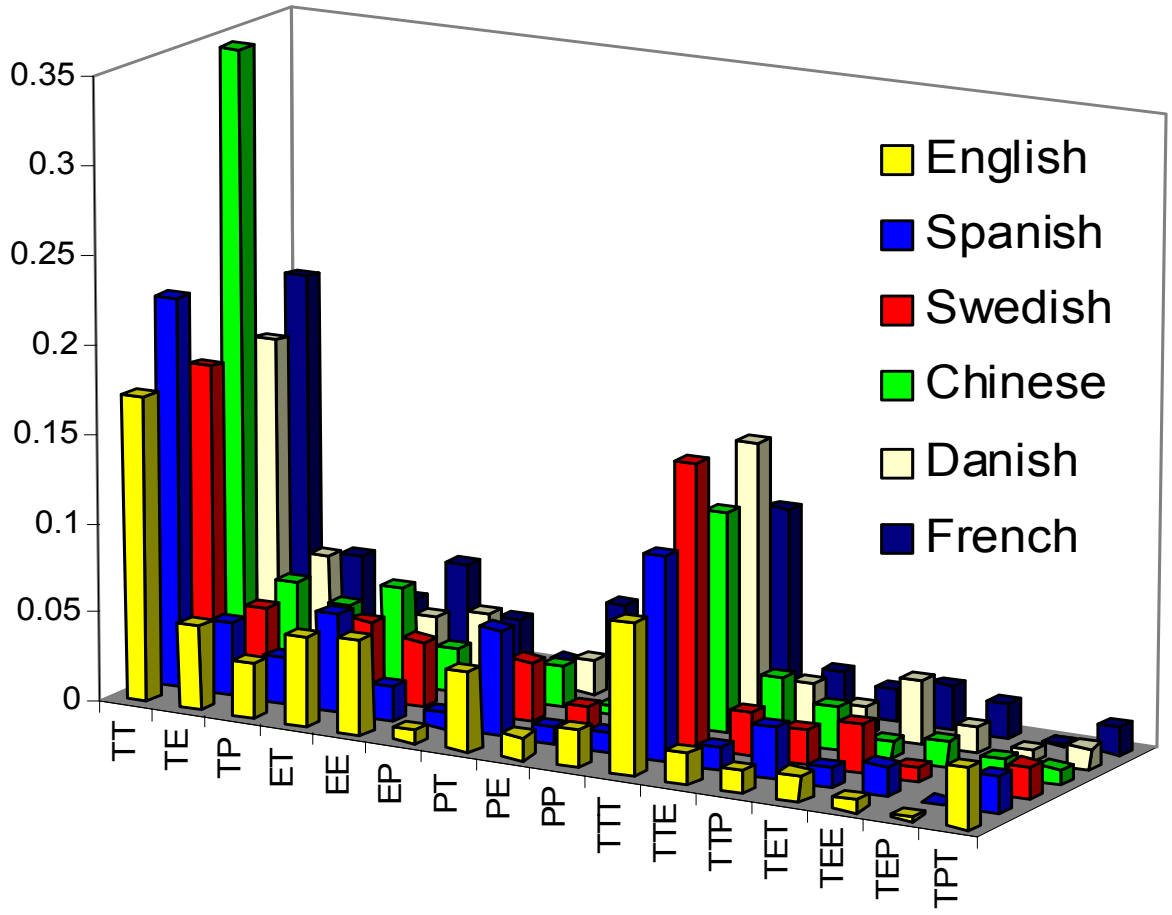

0.35
0.3
0.25
0.2
0.15
0.1
0.05
0
English
Spanish
Swedish
Chinese
Danish
French
TT
TE
TP
ET
EE
EP
PT
PE
PP
TTT
TTE
TTP
TET
TEE
TEP
TPT

Figure 4.

Figure legends

1. The definition of a significant pattern within the MEX approach. Note that the maxima of P_L and P_R define the beginning and the end of the pattern. Drops following the maxima signify divergence of paths at these points.

2. (a) A small portion of the adios-graph, for a simple corpus containing sentences #101 (is that a cat?) #102 (is that a dog?) #103 (and is that a horse?) #104 (where is the dog?). Each sentence is depicted by a solid line; edge direction is marked by arrows and is labeled by the sentence number and within-sentence index. The sentences in this example join a pattern is that a {dog, cat, horse} ?. (b). The abstracted pattern and the equivalence class associated with it are highlighted (edges that belong to sequences not subsumed by this pattern, e.g., #104, are untouched). (c) The identification of new significant patterns is done using the acquired equivalence classes (e.g., #200). In this manner, the system "bootstraps" itself, recursively distilling more and more complex patterns. This kind of abstraction also supports generalization: the original three sentences (shaded paths) form a pattern with two equivalence classes, which can then potentially generate six new sentences (e.g., the cat is play-ing and the horse is eat-ing). Taken from (Solan et al., 2004a).

3. Hierarchy of patterns following the progressive abstraction procedure of ADIOS. (A) A pattern (#67, consisting of far and away) is distilled. (B) Further abstraction yields equivalence classes (underlined in this figure) such as #70, which contains some proper names. (C) Pattern #144 can generate entire sentences, such as Joe is playing very far away, which can be read off the terminal-level of the tree (numbered arrows indicate traversal order during generation). Note that this is a root-pattern, i.e., it is not incorporated into other patterns. (D) The set of context-free productions (rewriting rules) that is equivalent to the tree of pattern #144.

4. The spectra of six different natural languages as extracted by ADIOS from online multi-lingual Bible texts (Resnik et al., 1999), consisting of 33,000 sentences. We define the pattern spectrum as the histogram of pattern types, whose bins are labeled by sequences such as (T,P) or (E,E,T), E standing for equivalence class, T for tree-terminal (original word) and P for significant pattern.

References

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

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

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

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

Resnik, P., Olsen, M. B., and Diab, M. (1999). *The Bible as a parallel corpus: annotating the 'Book of 2000 Tongues'*, volume 33. online at http://www.umiacs.umd.edu/ resnik/parallel/bible.html.

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

Solan, Z., Horn, D., Ruppin, E., and Edelman, S. (2004b). Unsupervised learning of natural languages. *preprint*.

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