Proceedings of VEXTAL'99 – Venezia, San Servolo, V.I.U. – 22-24 November 1999

Extracting Stochastic Grammars from Treebanks

Rens Bod
School of Computer Studies
University of Leeds
Leeds LS2 9JT, UK
rens@scs.leeds.ac.uk

1. Introduction

For quite some time, language processing models dealing with non-trivial aspects of syntactic structure or semantic interpretation were always built around "competence grammars": they assumed a non-redundant, complete and consistent descrip-tion of the sentence-structures of the language. In Scha (1990, 92) and Bod (1992, 95) it has been argued that human language perception and production processes may very well work with representations of concrete past language experien-ces, and that language processing models could emulate this behavior if they would analyze new input by combining fragments of representations from an annotated corpus.

This idea was worked out in some detail in Bod (1992, 93a/b, 95), where we demonstrated how an annotated corpus can in fact be employed directly as a stochastic grammar. The first "data-oriented parsing" system described there maintains a corpus of utterances which are annotated with labeled phrase-structure trees (a "treebank"); it parses new input by combining subtrees from the corpus. The most probable analysis is estimated on the basis of the occurrence-frequencies of the subtrees in the corpus.

Subsequent work (under various labels such as "data-oriented structure processing", "corpus-based interpretation", and "treebank grammar") has shown that this approach can be instantiated in many different ways, by making different assumptions about the corpus annotations, or by employing different disambiguation strategies (cf. van den Berg et al. 1994; Bod 96, 98a/b; Bod & Kaplan 1998; Bonnema 1996; Bonnema et al. 1997; Carroll & Weir 1997; Charniak 1996, 97; Coleman & Pierrehumbert 1997; Cormons 1999; Goodman 1996, 98; Kaplan 1996; Rajman 1995a/b; Scholtes 1992, 93; Scholtes & Bloembergen 1992a/b; Sekine & Grishman 1995; Sima'an et al. 1994; Sima'an 1995, 96a/b, 97; Tugwell 1995; Way 1999).

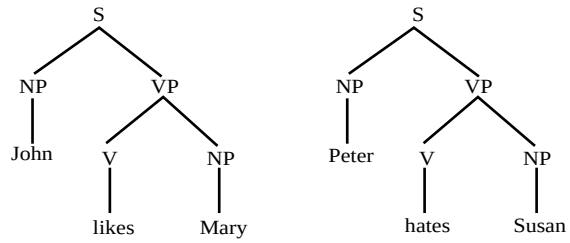
The model as originally defined imposes no constraints on the size and complexity of the corpus-subtrees that may be invoked in parsing new input. The set of subtrees that is used is thus very large and extremely redundant. Both from a theoretical and from a computational perspective we may therefore wonder whether it is possible to impose constraints on the subtrees that are used, in such a way that the performance of the model does not deteriorate or perhaps even improves. That is the main question addressed in the current paper. Moreover, by imposing different constraints on the subtree set, we can simulate several other stochastic grammars, ranging from stochastic context-free grammars to stochastic lexicalized grammars.

This paper is organized as follows. We first summarize the general data-oriented parsing (DOP) model. Then we report on a series of experiments carried out with this model to investigate several strategies for restricting the set of corpus subtrees. Finally, we go into some of the consequences of our results.

2. Summary of Data-Oriented Parsing

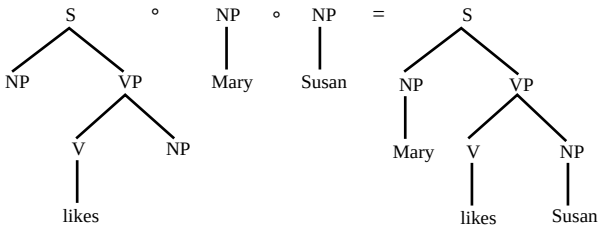
To date, the data-oriented parsing model has mainly been applied to corpora of trees labeled with syntactic annotations ("treebanks"). Let us illustrate this with a very simple example. Assume a corpus consisting of only two trees:

(1)



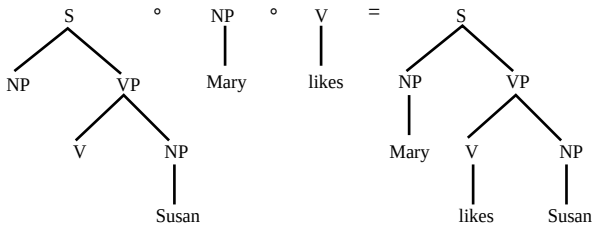
New sentences may be derived by combining subtrees from this corpus, by means of a node-substitution operation indicated as ∘. Node-substitution identifies the leftmost nonterminal frontier node of one tree with the root node of a second tree (i.e., the second tree is *substituted* on the leftmost nonterminal frontier node of the first tree). A new sentence such as *Mary likes Susan* can thus be derived as in (2):

(2)



Other derivations may yield the same parse tree; for instance:

(3)



DOP computes the probability of substituting a subtree t on a specific node as the probability of selecting t among all subtrees in the corpus that could be substituted on that node. This probability equals the number of occurrences of t, divided by the total number of occurrences of subtrees t' with the same root label as t. Let $rl(t)$ return the root label of t, then: $P(t) = \#(t) \; / \; \Sigma_{t':rl(t')=rl(t)} \#(t')$. The probability of a derivation is computed as the product of the probabilities of the subtrees it consists of. The probability of a parse tree is the sum of the probabilities of all derivations that produce it. Usually, several different parse trees can be derived for a single sentence, and in that case their probabilities provide a preference ordering.

Bod (1992) showed that DOP can use conventional context-free parsing techniques, by conver-ting subtrees into production rules. That is, every subtree t is converted into a context-free rewrite rule: $root(t) \rightarrow yield(t)$, and every such rule is indexed to maintain the link to its original subtree. The rules obtained in this way are used to construct a chart-like parse forest for the input sentence. Although the exact computation of the most probable parse tree from a parse forest is NP-hard (Sima'an 1996b), there exist randomized algorithms to estimate parse probabilities with an error that can be made arbitrarily small (cf. Bod 1995, Chappelier & Rajman 1998). Various instantiations of the DOP model have been tested on several benchmark corpora, obtaining state-of-the-art parsing and disambiguation performance -- see Bod (1998) for an overview.

3. Simulating stochastic grammars by constraining the subtree set

In this section, we go into a number of experiments that were carried out with two different treebanks: the ATIS treebank of 750 trees, and the OVIS treebank of 10,000 trees. Our main goal is to test whether all, arbitrarily large and complex subtrees are in fact relevant for predicting the correct analyses of new input utterances, or whether we can restrict the subtrees on linguistic or statistical grounds without diminishing the accuracy of predicting the appropriate analysis. By constraining the set of corpus-subtrees we can simulate various stochastic grammars, such as stochastic context-free grammars and stochastic lexicalized grammars, thus allowing for an interesting performance comparison.

3.1 The test environment

A first series of experiments was carried out with the 750 trees from the Air Travel Information System (ATIS) corpus, which were originally annotated in the Penn Treebank (Marcus et al. 1993). All analyses were checked on mistagged words, which were corrected by hand.

We used the blind testing method, dividing the 750 ATIS trees into a 90% training set of 675 trees and a 10% test set of 75 trees. The division was random except for one constraint: that all words in the test set actually occurred in the training set. The 675 training set trees were converted into their subtrees and were enriched with their corpus probabilities. The 75 sentences from the test set served as input sentences that were parsed and disambiguated by means of the subtrees of the training set trees, using the method described in section 2. We used the notion of *parse accuracy* as our accuracy metric, defined as the percentage of input sentences for which the selected parse is identical to the corresponding test set parse.

The parse accuracy obtained with the unrestricted subtree collection was 85%. We will use this number as the base line for studying the impact of the various subtree restrictions.[1]

3.2 The impact of overlapping subtrees

The DOP model as defined above chooses as the most appropriate analysis of a sentence the most probable parse of that sentence, rather than the parse generated by the most probable derivation. The main difference between the most probable parse and the most probable derivation is that by summing up over probabilities of several derivations, the most probable parse takes into account *overlapping* subtrees, while the most probable derivation does not.

Since the most probable derivation can be computed more efficiently than the most probable parse (by Viterbi optimization -- see Bod 1995, 98b), it is worth checking whether there is a difference between the predictions of these two methods.

We thus calculated the accuracies based on the analyses generated by the most probable derivations of the test sentences. The parse accuracy obtained in this way was 69%, which is lower than the 85% parse accuracy obtained by the most probable parse. We conclude that overlapping subtrees play an important role in predicting the appropriate analysis of a sentence, and should not be ignored.

3.3 The impact of subtree size

Next, we tested the impact of the size of the subtrees on the parse accuracy. It may be evident that large subtrees can capture more lexico-syntactic dependencies than small ones. We are now interested in how much these dependencies actually lead to better predictions of

[1] This 85% base line accuracy lies within the range of the standard deviation obtained with 10 different training/test set splits (for which the average parse accuracy was 84.2% with a standard deviation of 2.9%).

the appropriate parse. Therefore we performed experiments with versions of DOP where the subtree collection is restricted to subtrees with a certain maximum depth (where the depth of a tree is defined as the length of the longest path from the root to a leaf). The following table shows the results of these experiments, where the parse accuracy for each maximal depth is given for both the most probable parse and for the parse generated by the most probable derivation (the accuracies are rounded off to the nearest integer).

depth of corpus-subtrees	parse accuracy	
	most probable parse	most probable derivation
1	47%	47%
≤2	68%	56%
≤3	79%	64%
≤4	83%	67%
≤5	84%	67%
≤6	84%	69%
unbounded	85%	69%

Table 1. Accuracy increases if larger corpus subtree are used

The table shows an increase in parse accuracy, for both the most probable parse and the most probable derivation, when enlarging the maximum depth of the subtrees. The table confirms that the most probable parse yields better accuracy than the most probable derivation, except for depth 1 where DOP is equivalent to a stochastic context-free grammar (and where every parse is generated by exactly one derivation). The above results clearly show that simple stochastic context-free grammars perform worse than stochastic grammars where units can cover more than one level of constituent structure.

3.4 The impact of subtree lexicalization

In this section we test the impact of lexicalized subtrees on the parse accuracy. By a lexicalized subtree we mean a subtree whose yield or frontier contains one or more words. The more words a subtree contains, the more lexical dependencies are taken into account. To test the impact of this lexical context on the parse accuracy, we performed experiments with different versions of DOP where the subtree collection is restricted to subtrees whose frontiers contain a certain maximum number of words; the maximal subtree depth was kept constant at 6. These experiments are particularly interesting since we can simulate a number of lexicalized grammars in this way. Lexicalized grammars have become increasingly popular in computational linguistics (e.g. Srinivas & Joshi 1995; Collins 1996; Charniak 1997; Carroll & Weir 1997). However, all lexicalized grammars that we know of restrict the lexical context that is taken into account. It is an interesting feature of the DOP approach that we can straightforwardly test the impact of the size of the lexical context. The following table shows the results.

number of words in frontiers	parse accuracy	
	most probable parse	most probable derivation
≤1	75%	63%
≤2	80%	65%
≤3	83%	69%
≤4	83%	72%
≤6	83%	72%
≤8	87%	72%
unbounded	84%	69%

Table 2. Accuracy increases if more words are in subtree frontiers (subtree depth ≤ 6)

The table shows an initial increase in parse accuracy, for both the most probable parse and the most probable derivation, when enlarging the lexical context. Note that the parse accuracy deteriorates if the lexicalization exceeds 8 words. Thus, there seems to be an optimal lexical context size for the ATIS corpus. The table confirms that the most probable parse yields better accuracy than the most probable derivation, also for different lexicalization sizes. In the following, we will therefore evaluate our results with respect to the most probable parse only.

3.5 The impact of subtree frequency

We may expect that highly frequent subtrees contribute to a larger extent to the prediction of the appropriate parse than very infrequent subtrees. On the other hand, many infrequent subtrees are larger than the frequent ones. They thus contain more lexical/structural context, and can parse a large piece of an input sentence at once. It is therefore interesting to see what happens if we systematically remove low-frequency subtrees. We performed a set of experiments restricting the subtree collection to subtrees with a certain minimum number of occurrences, without applying any other restrictions.

The results of these experiments (table 3) indicate that low frequency subtrees contribute significantly to the prediction of the appropriate parse: the parse accuracy seriously deteriorates if low frequency subtrees are discarded.

frequency of subtrees	parse accuracy
≥1	85%
≥2	77%
≥3	45%
≥4	28%
≥5	16%
≥6	11%

Table 3. Accuracy decreases if lower bound on subtree frequency increases.

3.6 The impact of non-head words

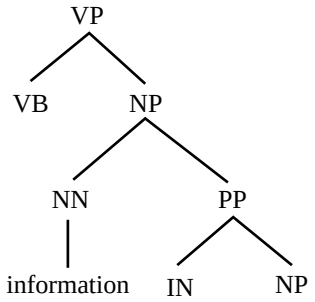
In this section we study what happens if we delete subtrees that are lexicalized only with non-head words. We start by defining the non-head words of a subtree. Given a subtree *t* with a root node category XP, by a non-head word of *t* we mean a frontier word of *t* that has a mother node category which is not of type X.[2] For the special case of subtrees rooted with a category of type S, we define a non-head word as any frontier word with a mother node category which is not a V.

In our first experiment with the non-head restriction, we deleted all subtrees whose frontier words were exclusively non-head words. We did not apply any other restrictions on the subtree collection. The parse accuracy obtained with this subtree collection was 74%. This is a deterioration with respect to the 85% parse accuracy obtained with the complete subtree collection (see table 1). Evidently, subtrees with only non-head words do contain relevant statistical dependencies that are not captured by other subtrees. This goes against common wisdom that only head-word dependencies are important (cf. Collins 1996, Charniak 1997).

We then considered the number of non-head words in the subtree frontiers. We may expect that subtrees that are lexicalized with *only* one non-head word are not important for the prediction of the appropriate parse. After eliminating these subtrees from the subtree collection, but keeping all subtrees with more than one non-head word, we obtained a parse accuracy of 80%, which is worse than the original accuracy of 85%.

Again, this shows the difficulty of finding subtree elimination criteria that do not diminish the accuracy: there is already an accuracy decrease if intuitively futile subtrees with just one non-head word are eliminated. However, the following example from the training set shows that such subtrees may not be futile at all, but can in fact describe significant relations:

(4)



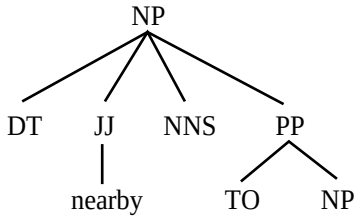
This subtree covers many ATIS substrings that are of the form VB *information* IN ... (the tag IN stands for any preposition except *to*):

find information on ...
get information from ...
give information from/on ...
have information on ...
like information on ...
list information about/on ...
need information on ...
see information from/in ...
show information from/in/on ...

In order to express the tendency that the above substrings occur with a specific structure, regardless the instantiations for VB and IN, we want to have a generalization over the verbs and prepositions but not over the noun *information*. The fragment above exactly provides this generalization together with the preference for the specific structure. Moreover, the fragment also expresses that *information* tends to occur in the object position.

An even more interesting example is the subtree in figure (5):

(5)



This subtree occurs in sentences such as *What are the nearby cities to Denver airport?* and *What are the nearby airports to Dallas?*. The key-word in these sentences is the adjective *nearby*, which determines the correct attachment of the PPs *to the airport in Atlanta* and *to Denver*. This attachment is independent of the nominal heads *cities* and *airports* of which *nearby* is a modifier. Thus it is a non-head modifier which is essential for the prediction of the appropriate parse. A similar example would be the sentence *Show the nearest airport to Dallas*.

Clearly, such sentences are problematic for models that disambiguate on the basis of constituent heads only (e.g. Collins 1996; Charniak 1997), since

[2] The words in the ATIS corpus are annotated with lexical categories that belong to a certain category *type*. For example, the lexical categories NN, NNS, NNP and NNPS are all of type noun (see Marcus et al. 1993 for an overview).

these models would neglect the crucial non-head adjectives *nearby* and *nearest*. The above examples may also be problematic for (stochastic) tree-adjoining grammar (Srinivas & Joshi 1995; Resnik 1992), where an adjectival modifier such as *nearby* is treated as a separate auxiliary tree and thus the structure in figure (5) would not be allowed as one unit.

Bod (1998b) notes that dependencies involving non-head words are much more common than one might think at first hand. They occur for instance quite frequently in the Wall Street Journal corpus in constructions such as: "*more* NNS *than*", "*same* NN *as*", etc.

3.7Validation on the OVIS treebank

The above experimental properties are derived on the basis of a relatively small treebank of only 750 trees. It is important then, to validate these properties on other, larger treebanks. We thus accomplished a series of experiments on the so-called OVIS treebank ("Openbaar Vervoer Informatie Systeem": Public Transport Information System), which consists of 10,000 trees (see Bonnema et al. 1997; Bod 1998a).

We used a random split of the OVIS treebank into a training set of 9000 trees and a test set of 1000 trees. We again studied the impact of various subtree restrictions by imposing the same constraints on the training set subtrees as above. That is, we studied the impact of overlapping subtrees, of subtree size, of subtree lexicalization, of subtree frequency, and of non-head words. The results of these experiments were in accordance with our previous results, and thus reinforced the properties that were derived for the ATIS corpus. That is, the parse accuracy decreases (1) if the maximum *size* of the subtrees decreases, (2) if the maximum *lexicalization* of the subtrees decreases, (3) if the minimal *frequency* of the subtrees increases, and (4) if constraints are imposed on the occurrence of *non-head words* in subtrees.

4 Conclusion

The experimentally derived properties presented in this paper indicate that virtually all restrictions on the subtree set diminish the parse accuracy of the test set sentences. We must keep in mind, however, that we have observed this only for one, rather impoverished, representation formalism (phrase structure trees), and for only two treebanks (ATIS, OVIS).

Our results nevertheless trigger the hypothesis that the structural units of natural language cannot be defined by a minimal set of rules, as is usually attempted in linguistic theory, but need to be defined in terms of a large set of redundant structures with virtually no restriction on form, size, lexicalization and non-head words. If this hypothesis is generally true, it has important consequences for linguistic theory. In particular, it suggests that knowledge of language is not represented as a grammar, but as a statistical ensemble of previously perceived structures that changes slightly every time a new utterance is processed. The regularities we observe in language may then be viewed as emergent phenomena, which cannot be summarized as a consistent non-redundant system that unequivocally defines the structures of new utterances.

References

M. van den Berg, R. Bod and R. Scha, 1994. "A Corpus-Based Approach to Semantic Interpretation", *Proceedings Ninth Amsterdam Colloquium*, Amsterdam, The Netherlands.

R. Bod, 1992. "Data Oriented Parsing (DOP)", *Proceedings COLING'92*, Nantes, France.

R. Bod, 1993a. "Using an Annotated Language Corpus as a Virtual Stochastic Grammar", *Proceedings AAAI'93*, Morgan Kaufmann, Menlo Park, Ca.

R. Bod, 1993b. "Monte Carlo Parsing", *Proceedings Third International Workshop on Parsing Technologies*, Tilburg/Durbuy, The Netherlands/Belgium.

R. Bod, 1995. *Enriching Linguistics with Statistics: Performance Models of Natural Language*, ILLC Dissertation Series 1995-14, University of Amsterdam.

R. Bod, 1998a. "Spoken Dialogue Interpretation with the DOP Model", *Proceedings COLING-ACL'98*, Montreal, Canada.

R. Bod, 1998b. *Beyond Grammar*, Cambridge University Press (CSLI Publications), Cambridge, UK.

R. Bod, R. Bonnema and R. Scha, 1996. "A Data-Oriented Approach to Semantic Interpretation", *Proceedings Workshop on Corpus-Oriented Semantic Analysis*, ECAI-96, Budapest, Hungary.

R. Bod and R. Kaplan, 1998. "A Probabilistic Corpus-Driven Model for Lexical-Functional Analysis", *Proceedings COLING-ACL'98*, Montreal, Canada.

R. Bonnema, R. Bod and R. Scha, 1997. "A DOP Model for Semantic Interpretation", *Proceedings ACL/EACL-97*, Madrid, Spain.

J. Carroll and D. Weir, 1997. "Encoding Frequency Information in Lexicalized Grammars", *Proceedings 5th International Workshop on Parsing Technologies*, MIT, Cambridge (Mass.).

J. Chappelier and M. Rajman, 1998. "Extraction stochastique d'arbres d'analyse pour le modèle DOP", *Proceedings TALN 1998*, Paris, France.

E. Charniak, 1996. "Tree-bank Grammars", *Proceedings AAAI'96*, Portland, Oregon.

E. Charniak, 1997. "Statistical Techniques for Natural Language Parsing", *AI Magazine*.

J. Coleman and J. Pierrehumbert, 1997. "Stochastic Phonological Grammars and Acceptability", *Proceedings Computational Phonology, Third Meeting of the ACL Special Interest Group in Computational Phonology*, Madrid, Spain.

M. Collins, 1996. "A new statistical parser based on bigram lexical dependencies", *Proceedings ACL'96*, Santa Cruz (Ca.).

B. Cormons, 1999. *Analyse et desambiguisation: Une approche purement à base de corpus (Data-Oriented Parsing) pour le formalisme des Grammaires Lexicales Fonctionnelles*, PhD thesis, Université de Rennes, France.

J. Eisner, 1997. "Bilexical Grammars and a Cubic-Time Probabilistic Parser", *Proceedings Fifth International Workshop on Parsing Technologies*, Boston, Mass.

J. Goodman, 1996. "Efficient Algorithms for Parsing the DOP Model", *Proceedings Empirical Methods in Natural Language Processing*, Philadelphia, PA.

J. Goodman, 1998. *Parsing Inside-Out*, Ph.D. thesis, Harvard University, Mass.

R. Kaplan, 1996. "A Probabilistic Approach to Lexical-Functional Analysis", *Proceedings of the 1996 LFG Conference and Workshops*. CSLI Publications, Stanford, CA.

M. Marcus, B. Santorini and M. Marcinkiewicz, 1993. "Building a Large Annotated Corpus of English: the Penn Treebank", *Computational Linguistics* 19(2).

M. Rajman, 1995a. *Apports d'une approche a base de corpus aux techniques de traitement automatique du langage naturel*, PhD thesis, Ecole Nationale Superieure des Telecommunications, Paris.

M. Rajman, 1995b. "Approche Probabiliste de l'Analyse Syntaxique", *Traitement Automatique des Langues*, vol. 36(1-2).

P. Resnik, 1992. "Probabilistic Tree-Adjoining Grammar as a Framework for Statistical Natural Language Processing", *Proceedings COLING'92*, Nantes, France.

R. Scha, 1990. "Taaltheorie en Taaltechnologie; Competence en Performance", in Q.A.M. de Kort and G.L.J. Leerdam (eds.), *Computertoepassingen in de Neerlandistiek*, Almere: Landelijke Vereniging van Neerlandici (LVVN-jaarboek).

R. Scha, 1992. "Virtuele Grammatica's en Creatieve Algoritmen", *Gramma/TTT* 1(1).

J. Scholtes, 1992. "Resolving Linguistic Ambiguities with a Neural Data-Oriented Parsing (DOP) System", in I. Aleksander and J. Taylor (eds.), *Artificial Neural Networks 2*, Vol. 2, Elsevier Science Publishers.

J. Scholtes and S. Bloembergen, 1992a. "The Design of a Neural Data-Oriented Parsing (DOP) System", *Proceedings of the International Joint Conference on Neural Networks*, (IJCNN), Baltimore, MD.

J. Scholtes and S. Bloembergen, 1992b. "Corpus Based Parsing with a Self-Organizing Neural Net", *Proceedings of the International Joint Conference on Neural Networks*, (IJCNN), Bejing, China.

S. Sekine and R. Grishman, 1995. "A Corpus-based Probabilistic Grammar with Only Two Non-terminals", *Proceedings Fourth International Workshop on Parsing Technologies*, Prague, Czech Republic.

K. Sima'an, R. Bod, S. Krauwer and R. Scha, 1994. "Efficient Disambiguation by means of Stochastic Tree Substitution Grammars", *Proceedings International Conference on New Methods in Language Processing*, UMIST, Manchester, UK.

K. Sima'an, 1995. "An optimized algorithm for Data Oriented Parsing", *Proceedings International Conference on Recent Advances in Natural Language Processing*, Tzigov Chark, Bulgaria.

K. Sima'an, 1996a. "An optimized algorithm for Data Oriented Parsing", in R. Mitkov and N. Nicolov (eds.), *Recent Advances in Natural Language Processing 1995*, volume 136 of *Current Issues in Linguistic Theory*. John Benjamins, Amsterdam.

K. Sima'an, 1996b. "Computational Complexity of Probabilistic Disambiguation by means of Tree Grammars", *Proceedings COLING-96*, Copenhagen, Denmark. (cmp-lg/9606019)

K. Sima'an, 1997. "Explanation-Based Learning of Data-Oriented Parsing", in T. Ellison (ed.) *CoNLL97: Computational Natural Language Learning*, ACL'97, Madrid, Spain.

B. Srinivas and A. Joshi, 1995. "Some novel applications of explanation-based learning to parsing lexicalized tree-adjoining grammars", *Proceedings ACL'95*, Cambridge (Mass.).

D. Tugwell, 1995. "A State-Transition Grammar for Data-Oriented Parsing", *Proceedings EACL'95*, Dublin, Ireland.

A. Way, 1999. "A Hybrid Archtecture for Robust MT using LFG-DOP", to appear in *Journal of Experimental and Theoretical Artificial Intelligence* (Special Issue on Memory-Based Language Processing).