Polynomial Tree Substitution Grammars: an efficient framework for Data-Oriented Parsing*

Jean-Cédric Chappelier and **Martin Rajman**

EPFL
DI-LIA, IN (Écublens)
CH-1015 Lausanne, Switzerland
{Jean-Cedric.Chappelier,Martin.Rajman}@epfl.ch

Abstract

Finding the most probable parse tree in the framework of Data-Oriented Parsing (DOP), a Stochastic Tree Substitution Parsing scheme developed by R. Bod (Bod 92), has proven to be NP-hard in the most general case (Sima'an 96a). However, introducing some *a priori* restrictions on the choice of the elementary trees (i.e. grammar rules) leads to interesting DOP instances with polynomial time-complexity. The purpose of this paper is to present such an instance, based on the minimal-maximal selection principle, and to evaluate its performances on two different corpora.

1 Motivations

First introduced by R. Scha (Scha 90) and mainly developed by R. Bod (Bod 92; Bod 95; Bod 98), the Data-Oriented Parsing (DOP) approach to statistical parsing has been investigated by several researchers over the past years. However, in the most general formulation of DOP, finding the most probable parse tree (MPP) has proven to be an NP-hard problem (Sima'an 96a). Therefore various approximated MPP search have been developed (Bod 92; Goodman 96; Chappelier & Rajman 00). However, another alternative consists in restricting the set of elementary trees used in the DOP grammar in such a way that finding the MPP is no longer NP-hard. The purpose of this contribution is to present and evaluate such an approach.

The paper first provides a short introduction to Data-Oriented Parsing and introduces some notations. It then presents an example of an elementary tree selection principle that leads to a polynomial-time restriction of DOP. Finally the performance of that model is analyzed on the basis of several experiments on two different treebanks.

2 Data-Oriented Parsing

2.1 DOP Model

DOP is a Stochastic Tree Substitution Grammar (STSG) parsing scheme. A STSG is a grammar in which the productions (or "rules") consist of elementary trees which are combined with the substitution operator to give derivations of complete parse trees. In the most general (unrestricted) version of DOP, the set of elementary trees is made of *all* the subtrees extracted from the treebank used as the training corpus for the system.

Each elementary tree t is assigned a probability $p(t)$ proportional to its number of occurrences in the treebank. The probability $p(d)$ of a derivation d is then defined as $p(d) = \prod_{t\in d} p(t)$, where $t \in d$ denotes the fact that the production t occurs in derivation d.

Finally, the DOP-probability of a parse tree T is defined as the sum of the probabilities of all its derivations:

$$P_{\text{DOP}}(T) = \sum_{d\Rightarrow T} p(d) = \sum_{d\Rightarrow T} \prod_{t\in d} p(t)$$

where the subscript "$d \Rightarrow T$" means "for all derivations d leading to the parse tree T".[1]

2.2 Parsing and Most Probable Parse

Parsing usually consists of two distinct phases:

analysis, during which a compact representation of all possible derivations of the input string is built;

extraction, during which specific results are derived from the compact representation; e.g. displaying all the parse trees, extracting the MPP, etc...

In the case of STSGs, analysis can be achieved in cubic time (with respect to the input size), as it is the case for Context-Free Grammars (CFG). However, in the most general case, extracting the MPP from the resulting compact representation is an NP-hard problem (Sima'an 96a). This means that finding the MPP for an input sentence cannot (in general) be achieved in polynomial time. Notice however that finding the most probable derivation (MPD) can still be achieved in polynomial time with the standard algorithms used for Stochastic CFGs.

*Published in *Proc. of Recent Advances in Natural Language Processing (RANLP 2001)*.

[1] A given parse tree can indeed have several different derivations (even with the left-most non-terminal first rewriting convention).

Since in DOP, finding the MPP cannot be solved in a computationally efficient way, various heuristics and/or approximations have been used instead: Monte-Carlo Sampling (Bod 92), General Recall (Goodman 98), controlled sampling (Chappelier & Rajman 00; Chappelier & Rajman 01a). However, another alternative (developed in the present contribution) is to restrict the set of elementary trees so that the MPP can again be extracted in polynomial time. The underlying idea is to remove the complexity related to the summation of the probabilities over the numerous derivations produced for any parse tree by associating each of the parse trees with at least one equivalent derivation, the probability of which is the sum of all the derivations of that parse tree.

To do so, an equivalent CF parsing scheme is built, in which any DOP parse tree always corresponds to at least one derivation with a probability equal to the DOP-probability of that parse tree. If is is furthermore guaranteed that this probability is maximal among the probabilities of all the other derivations corresponding to the same parse tree, searching for the MPD will precisely yield that derivation, therefore allowing the selection of the (DOP-)MPP. In other words, the MPD search in the equivalent CF parsing scheme and the MPP search in the original STSG become the same.

Notice that this approach cannot be applied to the most general version of DOP as it is not possible to exhibit one single derivation that could hold the whole DOP-probability of the parse tree it corresponds to.[2] All the difficulty is therefore to find interesting restrictions of the general STSG framework in which it is possible to do so. STSGs for which the MPP search can be reduced to an equivalent SCFG MPD search will be called hereafter *polynomial STSGs*.

A trivial example of polynomial STSGs are the SCFGs, i.e the STSGs where elementary trees are limited to depth-1 trees (equivalent to CF rules). The goal of this contribution is to describe another less trivial example of polynomial STSGs, in which the set of elementary trees is restricted to depth-1 trees **and** "*complete trees*" (also called "fully lexicalized trees", i.e. trees the leaves of which are all terminals). Such a restriction of the elementary tree set is hereafter called the *minimal-maximal* selection principle.

For STSGs with such a restriction, any derivation of a given parse tree T either is the complete tree itself, when T is an elementary tree of the grammar, or shares with the other derivations of T the same initial depth-1 tree. (cf fig. 1).

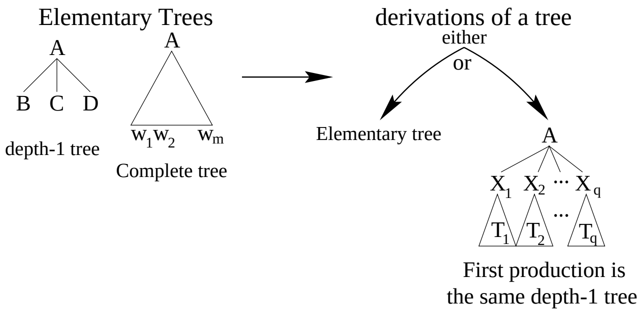


Figure 1: minimal-maximal restriction of DOP: elementary trees consist only of depth-1 trees or complete trees. In that case, all derivations of a given tree are either elementary trees themselves or start with the same depth-1 tree.

The next section shows that such STSGs are indeed polynomial, i.e. are stochastically equivalent to SCFGs.

3 A SCFG equivalent to restricted DOP

3.1 A weakly equivalent Context-Free Grammar

For any given unrestricted DOP grammar $\mathcal{G}_{\text{DOP}}$, let $\mathcal{G}_{\text{equiv}}$ be the CFG consisting of all the root-leaves rules[3] derived from all the elementary trees in $\mathcal{G}_{\text{DOP}}$.[4]

From a formal point of view (no probability for the moment) parsing with $\mathcal{G}_{\text{DOP}}$ is equivalent to parsing with $\mathcal{G}_{\text{equiv}}$.[5] This is always true, but in the special case where the set of elementary trees is restricted by the minimal-maximal selection principle, it is in addition possible to probabilize $\mathcal{G}_{\text{equiv}}$ so that finding the MPD with $\mathcal{G}_{\text{equiv}}$ is equivalent to finding the MPP with $\mathcal{G}_{\text{DOP}}$.

To do so, we first associate with each rule of $\mathcal{G}_{\text{equiv}}$ a stochastic coefficient equal to the DOP-probability P_{DOP} of the corresponding tree. Notice that this is not in general the elementary probability p usually associated with the tree.

We will show later on that this is always possible (and easy) to do while extracting the grammar from the training treebank. In the next section, we

[2]Refer to the proof of the NP-hardness of MPP extraction (Sima'an 96a).

[3]For a tree T, the corresponding root-leaves rule $R(T)$ is the CF rule, for which the left-hand side consists of the root of T and the right-hand side consists of the left-to-right ordered sequence of leaves of T.

[4]Notice that a CF rule is produced for each elementary tree. Two rules associated with two different elementary trees with the same root and leaves are distinguished by their indices. There is therefore a one-to-one mapping between the rules of $\mathcal{G}_{\text{equiv}}$ and the elementary trees of $\mathcal{G}_{\text{DOP}}$.

[5]At the end of parsing, the reconstruction of the whole tree out of the elementary trees corresponding to the CF rules in the MPD can trivially be performed.

first show that with such a probabilization, finding the MPD while parsing with $\mathcal{G}_{\text{equiv}}$ is indeed equivalent to finding the MPP while parsing with $\mathcal{G}_{\text{DOP}}$.

Notice however that this equivalence cannot directly be used as such when looking for the **two** (or more) most probable parses. Indeed, in such a case, it may occur that the second best derivation in $\mathcal{G}_{\text{equiv}}$ does not correspond to a derivation representing the second best parse tree (in $\mathcal{G}_{\text{DOP}}$) but to the second best derivations of the first best parse tree.

3.2 Stochastic equivalence

The stochastic equivalence results from the following general property of the $\mathcal{G}_{\text{equiv}}$ grammar: the probability of a derivation in $\mathcal{G}_{\text{equiv}}$ is always less than or equal to the DOP-probability of the corresponding parse tree. We refer to the appendix A for a proof of this property. Therefore, to show that finding the MPD with $\mathcal{G}_{\text{equiv}}$ is equivalent to finding the MPP with $\mathcal{G}_{\text{DOP}}$, it is sufficient to prove that for each parse tree T in $\mathcal{G}_{\text{DOP}}$ there *exists* at least one derivation in $\mathcal{G}_{\text{equiv}}$, the probability of which *is* the DOP-probability of T.

This property is proven by recursion on the depth of T:
1) The property is trivially true for all depth-1 trees, as in that case the DOP-probability is by definition equal to the elementary probability p (no other decomposition of the tree).
2) Suppose now that for every parse tree T of depth at most n, there exists in $\mathcal{G}_{\text{equiv}}$ a derivation d of T, the probability of which is the DOP-probability of T (according to $\mathcal{G}_{\text{DOP}}$).

We need to prove that this statement is also true for all parse trees (in $\mathcal{G}_{\text{DOP}}$) of depth $n+1$: Let $T \Rightarrow^* W_1^p$ now be a $(n+1)$-depth parse tree of the string W_1^p.

If T is itself a elementary tree, then its corresponding root-leaves CF rule $R(T)$ is a rule of $\mathcal{G}_{\text{equiv}}$, the probability of which is by construction the DOP-probability of T. And since $R(T)$ is a derivation of W_1^p, there exists at least one derivation representing T in $\mathcal{G}_{\text{equiv}}$ (and which has its DOP-probability).

If, on the other hand, T is not an elementary tree, then T results from a derivation $T = t_1 \circ ... \circ t_k$. Due to the nature of the elementary trees of $\mathcal{G}_{\text{DOP}}$, t_1 is necessarily a depth-1 tree, the same for all derivations of T. Let T_1, ..., T_q be the complete subtrees of T that are sons of t_1 (cf fig 1). The DOP-probability of the parse tree T is then:

$$\begin{aligned} P_{\text{DOP}}(T) &= p(t_1) \sum_{d_1 \Rightarrow T_1} \cdots \sum_{d_q \Rightarrow T_q} \prod_{i=1}^{q} p(d_i) \\ &= p(t_1) \cdot \prod_{i=1}^{q} \left(\sum_{d_i \Rightarrow T_i} p(d_i) \right) \\ &= p(t_1) \cdot \prod_{i=1}^{q} P_{\text{DOP}}(T_i) \end{aligned}$$

As T_1, ..., T_q are of depth at most n, there exists for each T_i a derivation $d_{\text{equiv}}(T_i)$ in $\mathcal{G}_{\text{equiv}}$ the probability of which (in $\mathcal{G}_{\text{equiv}}$) is equal to $P_{\text{DOP}}(T_i)$.

The derivation $(t_1, d_{\text{equiv}}(T_1), ..., d_{\text{equiv}}(T_q))$ is a derivation of T in $\mathcal{G}_{\text{equiv}}$, the probability of which is[6] $p(t_1) \cdot P_{\text{DOP}}(T_1) \cdot ... \cdot P_{\text{DOP}}(T_q)$, i.e. the DOP-probability of T.

Therefore, there exists at least one derivation of T in $\mathcal{G}_{\text{equiv}}$, the probability of which is DOP-probability of T.

This concludes the proof: finding the MPD while parsing with $\mathcal{G}_{\text{equiv}}$ is equivalent to finding the MPP while parsing with $\mathcal{G}_{\text{DOP}}$.

3.3 Effective construction of the equivalent SCFG grammar

We now have to explain how the equivalent CFG grammar $\mathcal{G}_{\text{equiv}}$ is built out of the training treebank. First, the depth-1 trees are extracted (as for the usual construction of a CFG out of a treebank). Then, for every node of every tree in the treebank the corresponding complete subtree is extracted and the corresponding root-leaves CF rule is produced (grouping together multiple occurences of the *same* elementary tree). Then the DOP-probabilities of each the complete subtrees are computed, proceeding by increasing order of depth. Indeed, if the DOP-probability of every elementary tree of depth n has been computed, it is possible to efficiently compute the DOP-probability of every elementary tree T of depth $n+1$ according to the formula

$$P_{\text{DOP}}(T) = p(T) + p(t_1) \cdot \prod_{i=1}^{q} P_{\text{DOP}}(T_i)$$

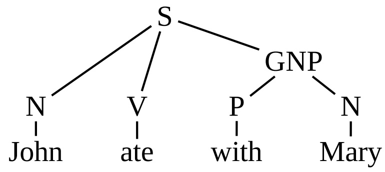
using the same notations as in the proof in the former section.

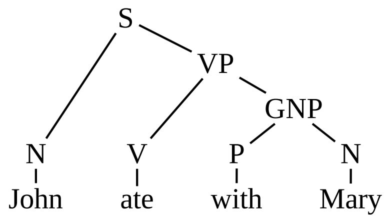
4 An example

Let us now introduce a toy example illustrating the approach presented in this paper and showing how it is different from the unrestricted DOP model.

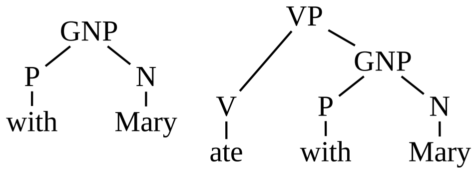
Consider the simple treebank consisting of the two following trees:

[6] by construction (SCFG derivation)

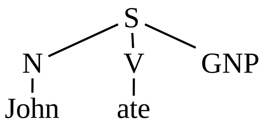
$t_1 =$ 

$t_2 =$ 

To build the STSG $\mathcal{G}_{\text{DOP}}$ according to the minimal-maximal selection principle, we select as elementary trees all the depth-1 subtrees and all the complete subtrees of t_1 and t_2. This corresponds to 8 depth-1 trees, t_1 and t_2 themselves, and the following two additionnal complete subtrees:



The grammar $\mathcal{G}_{\text{DOP}}$ therefore consists of 12 elementary trees. Notice that, for instance, a tree like



does not belong to $\mathcal{G}_{\text{DOP}}$, whereas it would have been the case in the unrestricted DOP model.

rule	P_{DOP}	p
r_1: `S -> N V GNP`	0.25	0.25
r_2: `S -> N VP`	0.25	0.25
r_3: `S -> John ate with Mary`	0.344	0.25
r_4: `S -> John ate with Mary`	0.359	0.25
r_5: `N -> John`	0.5	0.5
r_6: `N -> Mary`	0.5	0.5
r_7: `V -> ate`	1.0	1.0
r_8: `P -> with`	1.0	1.0
r_9: `GNP -> P N`	0.5	0.5
r_{10}: `GNP -> with Mary`	0.75	0.5
r_{11}: `VP -> V GNP`	0.5	0.5
r_{12}: `VP -> ate with Mary`	0.875	0.5

Table 1: The 12 rules of the equivalent CF grammar. The stochastic coefficient actually used is P_{DOP}. p corresponds to the elementary probability of the corresponding tree in the original DOP grammar.

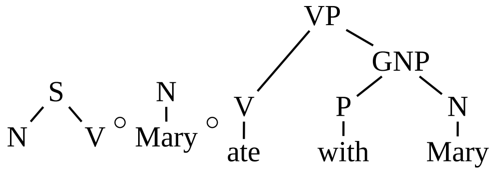
The equivalent CF grammar $\mathcal{G}_{\text{equiv}}$ consists of the 12 rules indicated in table 4. Notice that the rule

```
S -> John ate with Mary
```

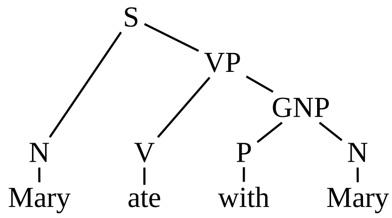
appears twice in the grammar. This has to be so since each of these two rules corresponds to a different tree in $\mathcal{G}_{\text{DOP}}$.

Notice also that $\sum_\alpha P(X \rightarrow \alpha)$ is not equal to 1 in general, i.e. $\mathcal{G}_{\text{equiv}}$ is not a *proper* SCFG (Booth & Thompson 73). However, this is not a problem in practice, since $\mathcal{G}_{\text{equiv}}$ should be regarded only as a practical tool to implement restricted DOP in polynomial time; the underlying probabilistic model remains the DOP model.

Let us now consider the sentence s="*Mary ate with Mary*". According to $\mathcal{G}_{\text{equiv}}$ the MPD for s is $d = r_2 \circ r_6 \circ r_{12}$ with the (maximum) probability $p(d) = 0.25 \cdot 0.5 \cdot 0.875 = 0.109$. This derivation corresponds to the derivation



The MPP for s (in $\mathcal{G}_{\text{DOP}}$) is therefore:



whose DOP-probability is 0.109.

5 Experiments

5.1 Corpora

To experimentally evaluate the performance of the minimal-maximal selection principle for DOP, we have tested it on two different corpora: the ATIS corpus (Hemphill *et al.* 90) and the Susanne #3 corpus (Sampson 94). As illustrated in table 2, these two corpora have quite different characteristics.

Contrary to most of the experiments performed so far (Sima'an 96b; Goodman 96), we did not turn the trees into a binary form (Chomsky Normal Form), but tried instead to keep the corpora as close to the original annotated data as possible.[7] In the same perspective, we did not restrict ourselves to parse Part-of-Speech tag sequences but worked on the original real word strings.

However, to carry out the experiments on the restricted DOP model with tractable memory conditions

[7] Only empty productions (traces) and a few obvious mistakes in the annotations (e.g. cycles) have been removed.

corpus	number of annotated sentences	number of CF rules	number of non-terminals	number of terminals	number of PoS tags	average sentence length	average nb of CF rules per sentence
ATIS	1’381	1’027	40	1’167	38	12.5	23.3
Susanne	6’728	20’302	767	17’863	130	20.4	36.0
Susanne short	4’000	8’882	469	10’284	122	12.9	23.8

Table 2: Various characteristics of the corpora used for the experiments.

(imposing an upper bound on the number of elementary trees), we had to reduce the Susanne corpus to a subset (hereafter called “*Susanne short*”): among the 6,803 original sentences, we finally kept 4,000 sentences for which we were able to build the restricted DOP grammar. The characteristics of this corpus are also given in table 2.

5.2 Methodology and Results

The results presented in this section were obtained from the ATIS and the Susanne short corpora previously described. The restricted DOP models were produced according to the minimal-maximal selection principle.

As a baseline reference, we also extracted the SCFG and measured its performance on the same corpora.

We were unfortunately unable to evaluate the unrestricted DOP model on the selected corpora as there were far too many elementary trees to be handled to build the complete grammar. We plan to evaluate and compare restricted and unrestricted DOP models on smaller corpora in a near future.

The evaluation protocol used for the production of the results was the same for all experiments: we compute an average performance on at least 10 runs consisting of

- (randomly) splitting the treebank into a 90% training set and 10% test set;
- extracting the grammars from the training set (the lexicon however was always the same and has been extracted from the whole original corpus; i.e. there were no unknown words in the tests: we did not try to evaluate the performance on Out-of-Vocabulary forms);
- evaluating performance on the test set. The performance measure used was the exact tree match.

The results obtained for the Susanne short and ATIS corpora are summarized in table 3. For each of the models, ”*% parsed*” indicates the coverage of the model (i.e. the percentage of sentences in the test set that received at least one parse), ”*% correct over parsed*” indicates its precision (i.e. the percentage of parsed sentences the most probable parse of which is the right parse) and “*% correct overall*” indicates the overall accuracy of the model (i.e. the percentage of sentences correctly parsed in the whole test set).

The general conclusion we can derive from the obtained results are:

1. the restricted DOP model outperforms in most of the cases the SCFG model.
2. As far as the coverage is concerned, there is a very strong difference in the nature of the two corpora that were used.
3. the precision on the ATIS corpus is very low, for both models.

The low performances on the ATIS corpus are due to the highly generative behaviour of the ATIS grammar[8]. The very low coverage obtained on the Susanne short corpus is related to the huge amount of happax CF rules (i.e. rules appearing only once in the corpus): 77 %.

6 Conclusion

This paper presents a new important aspect of Data-Oriented Parsing: a non trivial special case for which finding the MPP can be achieved in polynomial time (as opposed to the NP-hardness of the unrestricted DOP).

This is acheived by restricting the elementary trees to be either depth-1 trees or complete (fully lexicalized) trees and building an equivalent SCFG with which computing the most probable derivation (MPD) is equivalent to computing the most probable parse (MPP) with the restricted DOP model. This equivalence is made possible because, with the considered elementary tree set, there always exists for each parse tree at least one derivation which can concentrate the whole DOP-probability of that parse tree. This is not possible in the general (unrestricted) case where

[8] As already pointed out in the literature.

		Susanne short			ATIS		
		% parsed	% correct on parsed	% correct overall	% parsed	% correct on parsed	% correct overall
test (10%)	CF	45.5	23.0	10.5	99.6	25.4	25.3
	min-max DOP	45.5	24.4	11.1	99.6	21.0	20.9
self	CF	100	61.2	61.2	100	33.8	33.8
	min-max DOP	100	87.9	87.9	100	76.1	76.1

Table 3: Experimental results: percentage of exact match sentences both in test conditions (90% learning – 10% test random spliting of the corpus) and in self-evaluation (100% of the corpus).

the complexity precisely arises from the grouping of derivations producing the same parse tree.

The presented model appears to be an interesting compromise between the original DOP and SCFG models, from both of which it combines the advantages: it is as "simple" as a SCFG to parse (polynomial time) but still provides a richer probabilisation than SCFG, its parameters being able to capture more of the properties of the training corpus.

One open question we are currently working on is to identify other (possibly more general) selection principles providing new additional instances of Polynomial Tree Substitution Grammars.

References

(Bod 92) R. Bod. Applying Monte Carlo techniques to Data Oriented Parsing. In *Proceedings Computational Linguistics in the Netherlands*, Tilburg (The Netherlands), 1992.

(Bod 95) R. Bod. *Enriching Linguistics with Statistics: Performance Models of Natural Language*. Academische Pers, Amsterdam (The Netherlands), 1995.

(Bod 98) R. Bod. *Beyond Grammar, An Experience-Based Theory of Language*. Number 88 in CSLI Lecture Notes. CSLI Publications, Standford (CA), 1998.

(Booth & Thompson 73) T. L. Booth and R. A. Thompson. Applying probability measures to abstract languages. *IEEE Transactions on Computers*, C-22(5):442–450, may 1973.

(Chappelier & Rajman 00) J.-C. Chappelier and M. Rajman. Monte-Carlo sampling for NP-hard maximization problems in the framework of weighted parsing. In D.N. Christodoulakis, editor, *Natural Language Processing – NLP 2000*, number 1835 in Lecture Notes in Artificial Intelligence, pages 106–117. Springer, 2000.

(Chappelier & Rajman 01a) J.-C. Chappelier and M. Rajman. Parsing DOP with Monte-Carlo techniques. In R. Bod, R. Scha, and K. Sima'an, editors, *Data-Oriented Parsing*. CSLI Publications, 2001.

(Chappelier & Rajman 01b) J.-C. Chappelier and M. Rajman. Polynomial tree substitution grammars: an efficient framework for data-oriented parsing. Technical Report 01/tocome, Département Informatique, EPFL, Lausanne (Switzerland), May 2001.

(Goodman 96) J. Goodman. Efficient algorithms for parsing the dop model. In *Proc. of the Conf. on Empirical Methods in Natural Language Processing*, pages 143–152, May 1996.

(Goodman 98) J. Goodman. *Parsing Inside-Out*. Unpublished PhD thesis, Harvard University, May 1998. cmp-lg/9805007.

(Hemphill *et al.* 90) C. T. Hemphill, J; J. Godfrey, and G. R. Doddington. The ATIS spoken language systems pilot corpus. In Morgan Kaufmann, editor, *DARPA Speech and Natural Language Workshop*, June 1990.

(Sampson 94) G. Sampson. The Susanne corpus, release 3. In *School of Cognitive & Computing Sciences*, Brighton (England), 1994. University of Sussex, Falmer.

(Scha 90) R. Scha. Language theory and language technology: competence and performance. In de Kort and Leerdam, editors, *Computertoepassingen in de Neerlandistiek*. LVVN-jaarboek, Almere (The Netherlands), 1990. in Dutch.

(Sima'an 96a) K. Sima'an. Computational complexity of probabilistic disambiguation by means of tree grammars. In *Proceedings of COLING'96*, Copenhagen (Denmark), 1996. cmp-lg/9606019.

(Sima'an 96b) K. Sima'an. Efficient disambiguation by means of stochastic tree substitution grammars. In R. Mitkov and N. Nicolov, editors, *Recent Advances in NLP*, volume 136 of *Current Issues in Linguistic Theory*. Benjamins, Amsterdam (The Netherlands), 1996.

A Complementary Proof

The purpose of this appendix is to outsketch the proof of the fact that:
the probability of a derivation (in $\mathcal{G}_{\text{equiv}}$) is always less or equal to the DOP-probability of the corresponding parse tree (in $\mathcal{G}_{\text{DOP}}$).
A complete formal proof can be found in (Chappelier & Rajman 01b).

We first prove the following more general proposition:

Proposition 1: Given a SCFG $\mathcal{G}$, for any parse tree T and any derivation $d = t_1 \circ ... \circ t_m$ of T,

$$P_{\text{DOP}}(T) \geq \prod_{t_i} P_{\text{DOP}}(t_i)$$

To proceed with the proof of proposition 1, we first need to precisely define the terms $P_{\text{DOP}}(t_i)$ which correspond to an extension of the DOP-probability to trees that are not necessarily derivation trees.

This definition needs to extend the notion of tree derivation to the more general notion of tree decomposition (in elementary trees).[9] For any parse tree T (or any tree T resulting from a partial derivation in $\mathcal{G}$), there is a one-to-one mapping between derivations of T and decompositions of T. Therefore for such trees, we will indifferently use either the derivation or the decomposition point of view. However, there exist trees which cannot be written as a left-most derivation but still have a decomposition in elementary trees of $\mathcal{G}$.

A decomposition ξ of tree T into the elementary trees t_1, ..., t_m is denoted by $\xi = \langle t_1, ..., t_m \rangle_T$. The probability of such a decomposition is defined as $p(\xi) = \prod_{t_i} p(t_i)$.

The generalization of the DOP-probability to trees that do not result from a left-most derivation in $\mathcal{G}$ is given by:

$$P_{\text{DOP}}(T) = \sum_{\xi \in \mathcal{D}(T)} p(\xi)$$

where $\mathcal{D}(T)$ is the set of all decompositions of T over $\mathcal{G}$ (which might be empty if T cannot be decomposed into elementary trees of $\mathcal{G}$).

To proceed with the proof of proposition 1, we need one more definition: For any two decompositions ξ and ξ' of a tree T, ξ is said to be *finer* than ξ' iff every tree appearing in ξ is a subtree of a tree appearing in ξ' . This will be noted $\xi \leq \xi'$.[10]

The important property is that there is a one-to-one mapping between the set of all possible decompositions of elementary trees of a given derivation d of a tree T and the set of all the derivations of T finer than d.

Proof of Proposition 1:[11]
For any derivation $d = t_1 \circ ... \circ t_m$ of a parse tree T,

$$\begin{aligned} P_{\text{DOP}}(T) &= \sum_{d' \in \mathcal{D}(T)} p(d') \\ &= \sum_{d' \leq d} p(d') + \sum_{d' > d} p(d') \\ &\geq \sum_{d' \leq d} p(d') \end{aligned}$$

Due to the one-to-one mapping mentioned above, any d' finer than d corresponds uniquely to a tuple $(\xi_1, ..., \xi_m)$ of decompositions of the elementary trees of d. Therefore:

$$\begin{aligned} \sum_{d' \leq d} p(d') &= \sum_{\xi_1 \in \mathcal{D}(t_1)} \cdots \sum_{\xi_m \in \mathcal{D}(t_m)} \prod_{i=1}^{m} p(\xi_i) \\ &= \prod_{i=1}^{m} \sum_{\xi_i \in \mathcal{D}(t_i)} p(\xi_i) = \prod_{i=1}^{m} P_{\text{DOP}}(t_i) \end{aligned}$$

since $p(d' = t'_1 \circ ... \circ t'_q) = \prod_{i=1}^{q} p(t'_i) = \prod_{j=1}^{m} p(\xi_j)$ (simply regrouping the terms in the product).

$\square$

Finally, with Proposition 1, it is easy to prove that the probability of a derivation (in $\mathcal{G}_{\text{equiv}}$) is always less or equal to the DOP-probability of the corresponding parse tree (in $\mathcal{G}_{\text{DOP}}$): by construction indeed, the probability of a derivation in $\mathcal{G}_{\text{equiv}}$ is the product of the DOP-probabilities of the corresponding elementary trees.

[9] The formal definition of a tree decomposition is given in (Chappelier & Rajman 01b). It is basically a derivation with some extra notations to be able to handle trees which are not parse trees, i.e. trees for which the ”rewrite the left most non-terminal first” rule does not hold.

[10] This is actually an order relation on the set of decompositions of a given tree.

[11] where $d > d'$ stands for “not($d \leq d'$)”