is the sum of the probabilities of its distinct derivations. Let $t_{id}$ be the $i$-th subtree in the derivation $d$ that produces tree $T$, then the probability of $T$ is given by

$$P(T) = \Sigma_d \Pi_i P(t_{id})$$

As we will explain under step 3, the most probable parse tree of a sentence is estimated by Viterbi $n$-best summing up the probabilities of derivations that generate the same tree.

It may be evident that had we only the sentence *Investors suffered heavy losses* in our corpus, there would be no difference in probability between the five parse trees in figure 1, and U-DOP would not be able to distinguish between the different trees. However, if we have a different sentence where JJ NNS (*heavy losses*) appears in a different context, e.g. in *Heavy losses were reported*, its covering subtree gets a relatively higher frequency and the parse tree where *heavy losses* occurs as a constituent gets a higher total probability than alternative parse trees. Of course, it is left to the experimental evaluation whether *non*-constituents ("distituents") such as VBD JJ will be ruled out by U-DOP (section 3).

An important feature of (U-)DOP is that it considers counts of subtrees of a wide range of sizes: everything from counts of single-level rules to entire trees. A disadvantage of the approach is that an extremely large number of subtrees (and derivations) must be taken into account. Fortunately, there exists a rather compact PCFG-reduction of DOP which can also be used for U-DOP (Goodman 2003). Here we will only give a short summary of this PCFG-reduction. (Collins and Duffy 2002 show how a tree kernel can be used for an all-subtrees representation, which we will not discuss here.)

Goodman's reduction method first assigns every node in every tree a unique number which is called its address. The notation $A@k$ denotes the node at address $k$ where $A$ is the nonterminal labeling that node. A new nonterminal is created for each node in the training data. This nonterminal is called $A_k$. Let $a_j$ represent the number of subtrees headed by the node $A@j$. Let $a$ represent the number of subtrees headed by nodes with nonterminal $A$, that is $a = \Sigma_j a_j$. Goodman then gives a small PCFG with the following property: for every subtree in the training corpus headed by $A$, the grammar will generate an isomorphic subderivation with probability $1/a$. For a node $A@j(B@k, C@l)$, the following eight PCFG rules in figure 3 are generated, where the number in parentheses following a rule is its probability.

| | | | |
|---|---|---|---|
| $A_j \rightarrow BC$ | $(1/a_j)$ | $A \rightarrow BC$ | $(1/a)$ |
| $A_j \rightarrow B_kC$ | $(b_k/a_j)$ | $A \rightarrow B_kC$ | $(b_k/a)$ |
| $A_j \rightarrow BC_l$ | $(c_l/a_j)$ | $A \rightarrow BC_l$ | $(c_l/a)$ |
| $A_j \rightarrow B_kC_l$ | $(b_kc_l/a_j)$ | $A \rightarrow B_kC_l$ | $(b_kc_l/a)$ |

Figure 3. PCFG-reduction of DOP

In this PCFG reduction, $b_k$ represents the number of subtrees headed by the node $B@k$, and $c_l$ refers to the number of subtrees headed by the node $C@l$. Goodman shows by simple induction that his construction produces PCFG derivations isomorphic to (U-)DOP derivations with equal probability (Goodman 2003: 130-133). This means that summing up over derivations of a tree in DOP yields the same probability as summing over all the isomorphic derivations in the PCFG.[1]

The PCFG-reduction for U-DOP is slightly simpler than in figure 3 since the only labels are $S$ and $X$, and the part-of-speech tags. For the tree-set of $8.23 * 10^5$ binary trees generated under step 1, Goodman's reduction method results in a total number of $14.8 * 10^6$ distinct PCFG rules. While it is still feasible to parse with a rule-set of this size, it is evident that our approach can deal with longer sentences only if we further reduce the size of our binary tree-set.

It should be kept in mind that while the probabilities of all parse trees generated by DOP sum up to 1, these probabilities do not converge to the "true" probabilities if the corpus grows to infinity (Johnson 2002). In fact, in Bod et al. (2003) we showed that the most probable parse tree as defined above has a tendency to be constructed by the *shortest derivation* (consisting of the fewest and thus largest subtrees). A large subtree is overruled only if the combined relative frequencies of smaller subtrees yields a larger score. We refer to Zollmann and Sima'an (2005) for a recently proposed estimator that *is* statistically consistent (though it is not yet known how this estimator performs on the WSJ) and to Zuidema (2006) for a theoretical comparison of existing estimators for DOP.

[1] As in Bod (2003) and Goodman (2003: 136), we additionally use a correction factor to redress DOP's bias discussed in Johnson (2002).