- If $i_1 \in I$ is not a pre-terminal node, for each $\eta = i_2 i_3 \in \delta(i_1)$, let A, B, and C be non-terminal symbols of i_1, i_2, and i_3. Then,
$$q(i_1 \to \eta) = \frac{\sum_{x \in H} \sum_{y \in H} \sum_{z \in H} P_{\text{out}}(i_1[x]) \beta(A[x] \to B[y]C[z]) P_{\text{in}}(i_2[y]) P_{\text{in}}(i_3[z])}{\sum_{x \in H} P_{\text{out}}(i_1[x]) P_{\text{in}}(i_1[x])}.$$
- If $i \in I$ is a pre-terminal node above word w_k, then $q(i \to w_k) = 1$.
- If $i \in I$ is a root node, let A be the non-terminal symbol of i. Then $q_r(i) = \frac{1}{P(w)} \sum_{x \in H} \pi(A[x]) P_{\text{in}}(i[x])$.

Figure 4: Optimal parameters of approximate distribution Q.

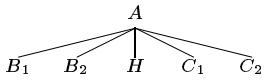

Figure 5: Original subtree.

	1	2	3	4	average $\pm$ σ
training LL	-115	-114	-115	-114	-114 $\pm$ 0.41
heldout LL	-114	-115	-115	-114	-114 $\pm$ 0.29
LR	86.7	86.3	86.3	87.0	86.6 $\pm$ 0.27
LP	86.2	85.6	85.5	86.6	86.0 $\pm$ 0.48

Table 1: Dependency on initial values.

of increase in the likelihood of the heldout data became lower than a certain threshold. Section 22 was used as test data in all parsing experiments except in the final one, in which section 23 was used. We stripped off all function tags and eliminated empty nodes in the training and heldout data, but any other pre-processing, such as comma raising or base-NP marking (Collins, 1999), was not done except for binarizations.

4.1 Dependency on initial values

To see the degree of dependency of trained models on initializations, four instances of the same model were trained with different initial values of parameters.[3] The model used in this experiment was created by CENTER-PARENT binarization and $|H|$ was set to 16. Table 1 lists training/heldout data log-likelihood per sentence (LL) for the four instances and their parsing performances on the test set (section 22). The parsing performances were obtained using the approximate distribution method in Section 3.2. Different initial values were shown to affect the results of training to some extent (Table 1).

[3]The initial value for an annotated rule probability, $\beta(A[x] \to B[y]C[z])$, was created by randomly multiplying the maximum likelihood estimation of the corresponding PCFG rule probability, $P(A \to BC)$, as follows:

$$\beta(A[x] \to B[y]C[z]) = Z_A^{-1} e^{\gamma} P(A \to BC),$$

where γ is a random number that is uniformly distributed in $[-\log 3, \log 3]$ and Z_A is a normalization constant.

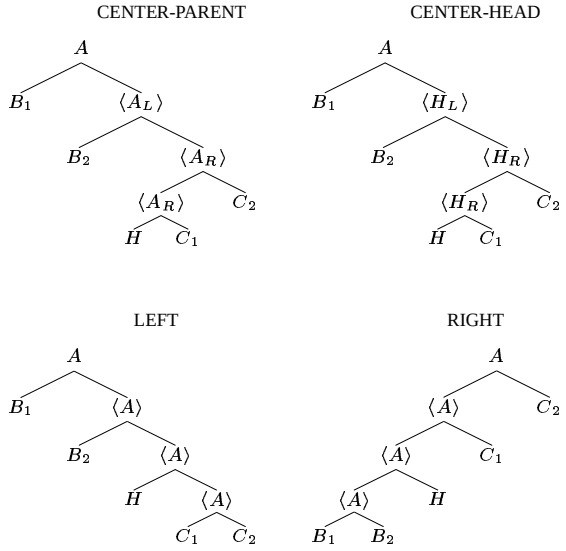

Figure 6: Four types of binarization (H: head daughter).

4.2 Model types and parsing performance

We compared four types of binarization. The original form is depicted in Figure 5 and the results are shown in Figure 6. In the first two methods, called CENTER-PARENT and CENTER-HEAD, the head-finding rules of Collins (1999) were used. We obtained an observable grammar R for each model by reading off grammar rules from the binarized training trees. For each binarization method, PCFG-LA models with different numbers of latent annotation symbols, $|H| = 1, 2, 4, 8$, and 16, were trained.