Limitations

This study conducted experiments using treebanks of 10 typologically diverse languages and showed that the optimal strategy can vary across languages. However, other factors, such as differences in annotation schemes or tokenization, could also contribute to the observed differences in the optimal strategies. Investigating the extent to which such differences actually affect optimal strategies is an important topic for future work.

Furthermore, this study used RNNG as the syntactic language model. However, various other architectures, such as PLM (Choe and Charniak, 2016) and Transformer Grammar (Sartran et al., 2022), have also been proposed. Analyzing how the inductive biases of these different architectures influence the optimal strategies is left for future research. Additionally, as mentioned in section 2, RNNG is considered to be less affected by stack size. Analyzing how the optimal strategy changes when considering models that are more strongly affected by stack size is also an interesting topic for future work.

A Datset Setting

To split the words into subwords, we applied byte pair encoding (BPE). For datasets with 13K-30K different words that appear at least twice (English, Chinese, French, German, Korean, and Hungarian), we used BPE with a vocabulary size of 5000. For the remaining datasets (Basque, Hebrew, Polish, and Swedish), which have 5K-8K words appearing at least twice, we used BPE with a vocabulary size of 1500. We used SentencePiece for subword segmentation.[13]

B Model Setting

For the hyperparameters of RNNG, we used a 2-layer LSTM (Hochreiter and Schmidhuber, 1997) for hidden state transitions, a BiLSTM as the composition model, 256-dimensional embedding vectors, 256-dimensional hidden state vectors, and a dropout rate of 0.3. For optimization, we used Adam (Kingma and Ba, 2015) with a learning rate of 0.001. Training was performed for either 80 epochs or 8000 steps, whichever was larger for each dataset. Regarding the batch size, we set it to 512 for datasets with more than 10K data points (English, Chinese, French, German, and Korean), and 128 for datasets with fewer than 10K data points (Basque, Hebrew, Hungarian, Polish, and Swedish).

C Other Results

Figure 5 shows the perplexity based on sentence probability $\tilde{p}^{\mathcal{M}}$, calculated by marginalizing the joint probability $p^{\mathcal{M}}_{\text{joint}}$ within the last beam $B_{|x|}$ to approximate $p^{\mathcal{M}}$, for each language and strategy. Figure 6 shows the perplexity calculated using the $p^{\mathcal{M}}_{\text{token}}$ for the best action sequence obtained by beam search for each language and strategy. Figure 7 shows the validation loss, i.e., the negative joint log-likelihood $-\log p^{\mathcal{M}}_{\text{joint}}$, calculated for the same data points as in Figure 2 for each language and strategy.

[13] https://github.com/google/sentencepiece

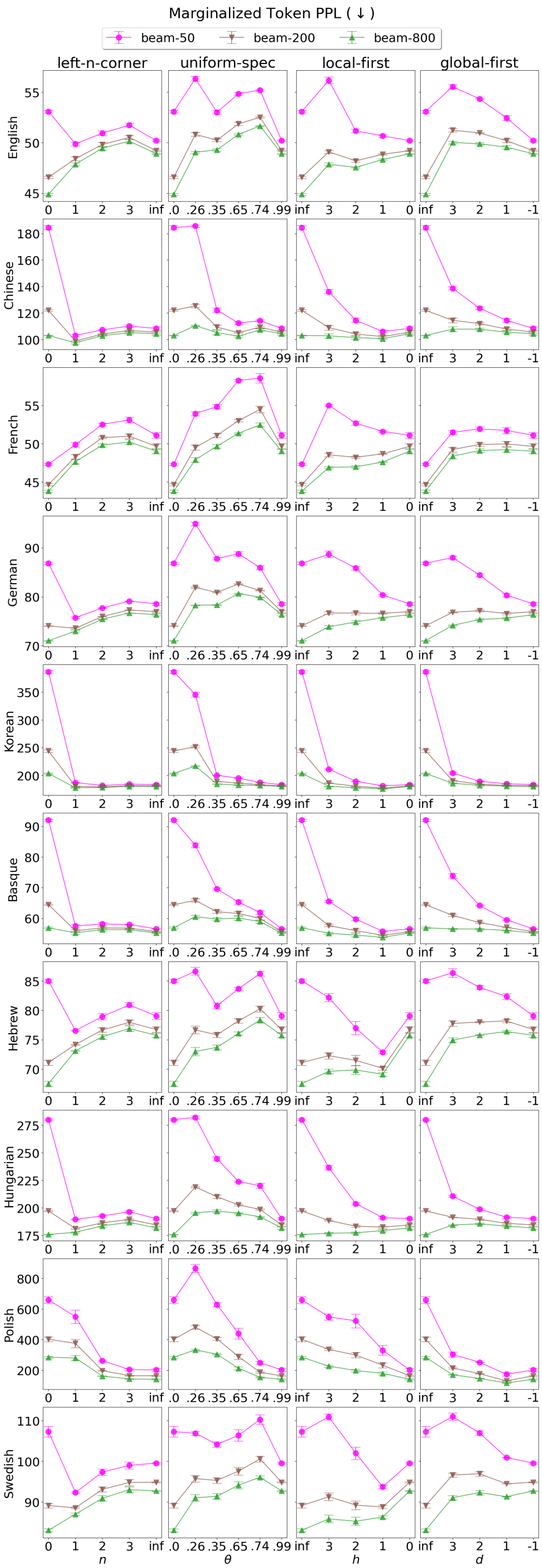

Figure 5: Perplexity based on $\tilde{p}^{\mathcal{M}}$ for all datasets. Error bars show the standard error of the mean.

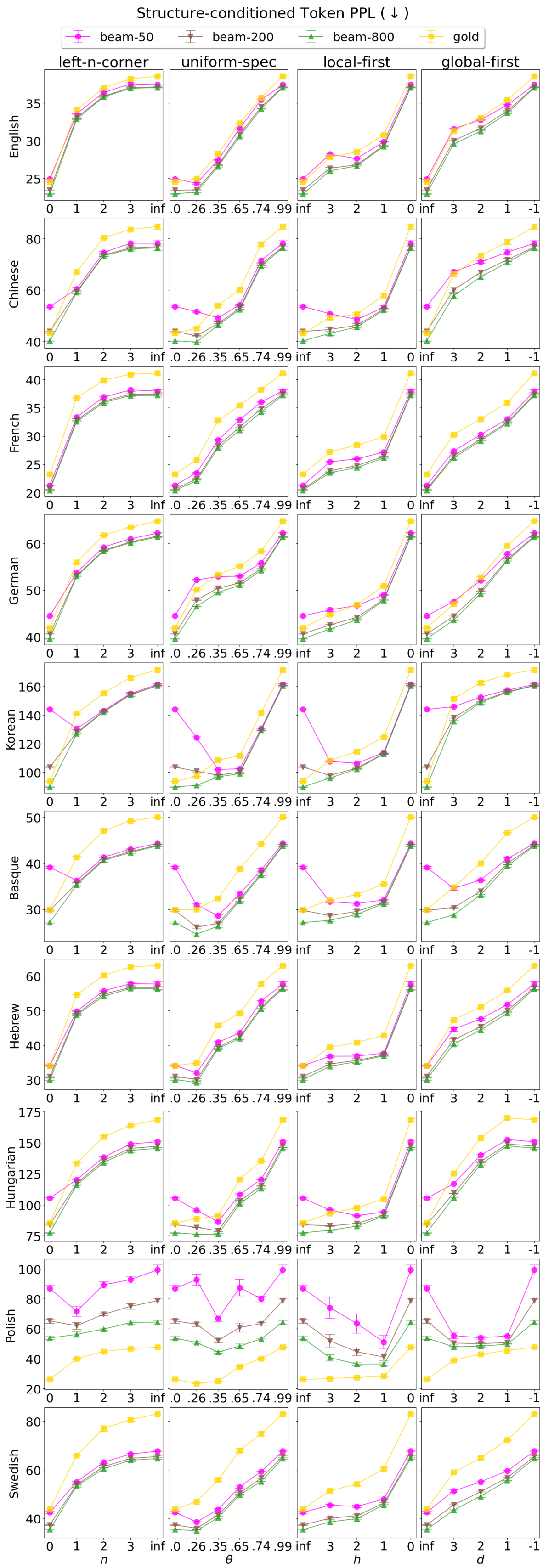

Figure 6: Perplexity based on $p^{\mathcal{M}}_{\text{token}}$ for all datasets. Error bars show the standard error of the mean.

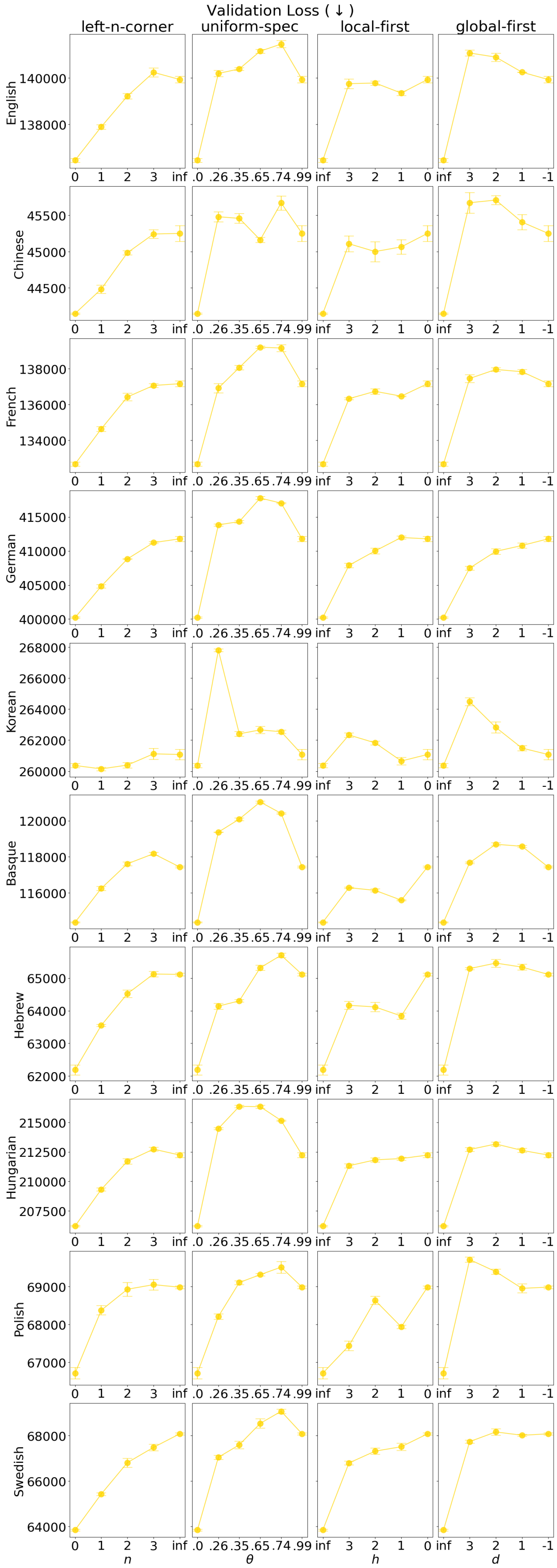

Figure 7: Validation loss, i.e., $-\log p^{\mathcal{M}}_{\text{joint}}$ for all datasets. Error bars show the standard error of the mean.