A probabilistic model of the structural preferences of open-class words is important for accurate parsing, irrespective of the particular parsing paradigm. However, word-specific properties are not represented adequately in statistical grammars trained solely on annotated corpora, due to the Zipfian distribution of words in a corpus. The problem becomes more severe for models containing complex, fine-grained lexical categories. This dissertation presents procedures to estimate complex lexical parameters of a smoothed Penn Treebank PCFG from unlabeled data. The PCFG contains important linguistic representations for argument-adjunct distinctions and long-distance dependencies. The lexical parameters of the PCFG encode fine-grained information about structures selected by a lexical item, such as its subcategorization frames.

Values of lexical parameters of words are re-estimated from a large source of unlabeled data using the Inside-outside algorithm for PCFG induction. Re-estimation from unlabeled data is constrained by interpolating re-estimated parameters with treebank parameters; we use the intuition that certain parameters of treebank models are more accurate than others and can guide unsupervised estimation, thus avoiding heuristic constraints. Models obtained in this way are shown to be superior to models obtained with standard Inside-outside estimation. We get substantial improvements in identification of complex subcategorized structures for unseen and low-frequency verbs in the treebank, as measured by parsing-based evaluations.
This dissertation interweaves several issues related to unsupervised estimation of PCFGs: a treebank PCFG enables evaluations of re-estimated models against a high-quality gold standard (the Penn Treebank), unlike models obtained previously by Inside-outside from completely unlabeled data. Maximum-likelihood estimation (via Inside-outside) allows examination of questions regarding the relative efficacy of supervised estimation from a treebank versus unsupervised estimation on a much larger corpus. Subcategorization frames and other features in the PCFG are based solely on Penn Treebank annotation; the fine-grained, wide-coverage lexical resource obtained here is therefore aligned with Penn Treebank structures and interpretable according to Penn Treebank annotation guidelines. The framework for creating a linguistically-sophisticated PCFG can be extended to other languages having a treebank in the Penn Treebank style.