	System Description		*Accuracy*
#1 (§3.1)	"less is more"	(Spitkovsky et al., 2009)	44.0
#3 (§4.1)	"less is more" with monosemous induced tags		41.4 (-2.6)

Table 4: Directed accuracies on Section 23 of WSJ (all sentences) for two experiments with the base system.

clusterings — one flat and one hierarchical — can be better for dependency grammar induction than monosemous syntactic categories derived from gold part-of-speech tags. And we confirmed that the unsupervised tags are worse than the actual gold tags, in a simple dependency grammar induction system.

5 State-of-the-Art without Gold Tags

Until now, we have deliberately kept our experimental methods simple and nearly identical to Klein and Manning's (2004), for clarity. Next, we will explore how our main findings generalize beyond this toy setting. A preliminary test will simply quantify the effect of replacing gold part-of-speech tags with the monosemous flat clustering (as in experiment #3, §4.1) on a modern grammar inducer. And our last experiment will gauge the impact of using a polysemous (but still unsupervised) clustering instead, obtained by executing standard sequence labeling techniques to introduce context-sensitivity into the original (independent) assignment or words to categories.

These final experiments are with our latest state-of-the-art system (Spitkovsky et al., 2011) — a partially lexicalized extension of the DMV that uses constrained Viterbi EM to train on nearly all of the data available in WSJ, at WSJ45 (48,418 sentences; 986,830 non-punctuation tokens). The key contribution that differentiates this model from its predecessors is that it incorporates punctuation into grammar induction (by turning it into parsing constraints, instead of ignoring punctuation marks altogether). In training, the model makes a simplifying assumption — that sentences can be split at punctuation and that the resulting fragments of text could be parsed independently of one another (these parsed fragments are then reassembled into full sentence trees, by parsing the sequence of their own head words). Furthermore, the model continues to take punctuation marks into account in inference (using weaker, more accurate constraints, than in training). This system scores 58.4% on Section 23 of WSJ$^\infty$ (see Table 5).

5.1 Experiment #5: A Monosemous Clustering

As in experiment #3 (§4.1), we modified the base system in exactly one way: we swapped out gold part-of-speech tags and replaced them with a flat distributional similarity clustering. In contrast to simpler models, which suffer multi-point drops in accuracy from switching to unsupervised tags (e.g., 2.6%), our new system's performance degrades only slightly, by 0.2% (see Tables 4 and 5). This result improves over substantial performance degradations previously observed for unsupervised dependency parsing with induced word categories (Klein and Manning, 2004; Headden et al., 2008, *inter alia*).[7]

One risk that arises from using gold tags is that newer systems could be finding cleverer ways to exploit manual labels (i.e., developing an over-reliance on gold tags) instead of actually learning to acquire language. Part-of-speech tags are *known* to contain significant amounts of information for unlabeled dependency parsing (McDonald et al., 2011, §3.1), so we find it reassuring that our latest grammar inducer is *less* dependent on gold tags than its predecessors.

5.2 Experiment #6: A Polysemous Clustering

Results of experiments #1 and 3 (§3.1, 4.1) suggest that grammar induction stands to gain from relaxing the *one class per word* assumption. We next test this conjecture by inducing a polysemous unsupervised word clustering, then using it to induce a grammar.

Previous work (Headden et al., 2008, §4) found that simple bitag hidden Markov models, classically trained using the Baum-Welch (Baum, 1972) variant of EM (HMM-EM), perform quite well,[8] on average, across different grammar induction tasks. Such sequence models incorporate a sensitivity to context via state transition probabilities $\mathbb{P}_{\texttt{TRAN}}(t_i \mid t_{i-1})$, capturing the likelihood that a tag t_i immediately follows the tag t_{i-1}; emission probabilities $\mathbb{P}_{\texttt{EMIT}}(w_i \mid t_i)$ capture the likelihood that a word of type t_i is w_i.

[7] We also briefly comment on this result in the "punctuation" paper (Spitkovsky et al., 2011, §7), published concurrently.

[8] They are also competitive with Bayesian estimators, on larger data sets, with cross-validation (Gao and Johnson, 2008).