	Accuracy		*Viable*
1. **manual** tags	*Unsupervised*	*Sky*	*Groups*
gold	50.7	*78.0*	36
mfc	**47.2**	74.5	34
mfp	40.4	76.4	160
ua	44.3	**78.4**	328
2. tagless **lexicalized** models			
full	25.8	**97.3**	49,180
partial	29.3	60.5	176
none	**30.7**	*24.5*	1
3. tags from a **flat** (Clark, 2000) clustering			
	47.8	**83.8**	197
4. prefixes of a **hierarchical** (Brown et al., 1992) clustering			
first 7 bits	46.4	73.9	96
8 bits	**48.0**	77.8	165
9 bits	46.8	**82.3**	262

Table 1: Directed accuracies for the "less is more" DMV, trained on WSJ15 (after 40 steps of EM) and evaluated also against WSJ15, using various lexical categories in place of gold part-of-speech tags. For each tag-set, we include its effective number of (non-empty) categories in WSJ15 and the oracle skylines (supervised performance).

3 Motivation and Ablative Analyses

The concepts of polysemy and synonymy are of fundamental importance in linguistics. For words that can take on multiple parts of speech, knowing the gold tag can reduce ambiguity, improving parsing by limiting the search space. Furthermore, pooling the statistics of words that play similar syntactic roles, as signaled by shared gold part-of-speech tags, can simplify the learning task, improving generalization by reducing sparsity. We begin with two sets of experiments that explore the impact that each of these factors has on grammar induction with the DMV.

3.1 Experiment #1: Human-Annotated Tags

Our first set of experiments attempts to isolate the effect that replacing gold part-of-speech tags with deterministic *one class per word* mappings has on performance, quantifying the cost of switching to a monosemous clustering (see Table 1: manual; and Table 4). Grammar induction with gold tags scores 50.7%, while the oracle skyline (an ideal, supervised instance of the DMV) could attain 78.0% accuracy.

It may be worth noting that only 6,620 (13.5%) of 49,180 unique tokens in WSJ appear with multiple part-of-speech tags. Most words, like *it*, are always tagged the same way (5,768 times `PRP`). Some words, like *gains*, usually serve as one part of speech (227 times `NNS`, as in *the gains*) but are occasionally used differently (5 times `VBZ`, as in *he gains*). Only 1,322 tokens (2.7%) appear with three or more different gold tags. However, this minority includes the most frequent word — *the* (50,959 times `DT`, 7 times `JJ`, 6 times `NNP` and once as each of `CD`, `NN` and `VBP`).[2]

token	*mfc*	*mfp*	*ua*
it	{`PRP`}	{`PRP`}	{`PRP`}
gains	{`NNS`}	{`VBZ`, `NNS`}	{`VBZ`, `NNS`}
the	{`DT`}	{`JJ`, `DT`}	{`VBP`, `NNP`, `NN`, `JJ`, `DT`, `CD`}

Table 2: Example most frequent class, most frequent pair and union all reassignments for tokens *it*, *the* and *gains*.

We experimented with three natural reassignments of part-of-speech categories (see Table 2). The first, *most frequent class* (mfc), simply maps each token to its most common gold tag in the entire WSJ (with ties resolved lexicographically). This approach discards two gold tags (types `PDT` and `RBR` are not most common for any of the tokens in WSJ15) and costs about three-and-a-half points of accuracy, in both supervised and unsupervised regimes.

Another reassignment, *union all* (ua), maps each token to the set of all of its observed gold tags, again in the entire WSJ. This inflates the number of groupings by nearly a factor of ten (effectively lexicalizing the most ambiguous words),[3] yet improves the oracle skyline by half-a-point over actual gold tags; however, learning is harder with this tag-set, losing more than six points in unsupervised training.

Our last reassignment, *most frequent pair* (mfp), allows up to two of the most common tags into a token's label set (with ties, once again, resolved lexicographically). This intermediate approach performs strictly worse than *union all*, in both regimes.

3.2 Experiment #2: Lexicalization Baselines

Our next set of experiments assesses the benefits of categorization, turning to lexicalized baselines that avoid grouping words altogether. All three models discussed below estimated the DMV *without* using the gold tags in any way (see Table 1: lexicalized).

[2] Some of these are annotation errors in the treebank (Banko and Moore, 2004, Figure 2): such (mis)taggings can severely degrade the accuracy of part-of-speech disambiguators, without additional supervision (Banko and Moore, 2004, §5, Table 1).

[3] Kupiec (1992) found that the 50,000-word vocabulary of the Brown corpus similarly reduces to ∼400 ambiguity classes.