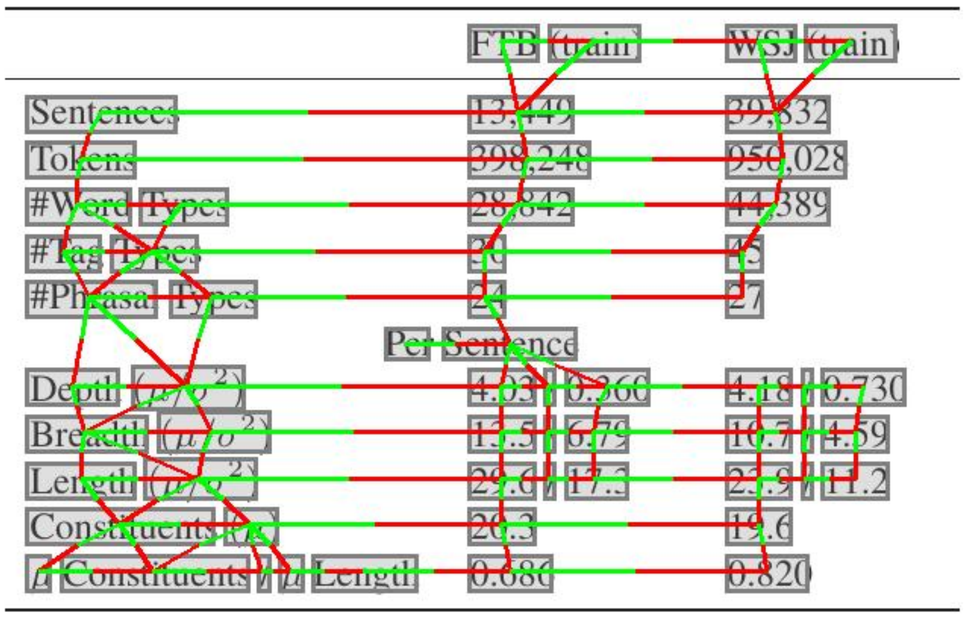

Table 2: Gross corpus statistics for the pre-processed FTB (training set) and WSJ (sec. 2-21). The FTB sentences are longer with broader syntactic trees. The FTB POS tag set has 33% fewer types than the WSJ. The FTB dev set OOV rate is 17.77% vs. 12.78% for the WSJ.

Type		#Total	#Single	%Single	%Total
MWN	*noun*	9,680	2,737	28.3	49.7
MWADV	*adverb*	3,852	449	11.7	19.8
MWP	*prep.*	3,526	342	9.70	18.1
MWC	*conj.*	814	73	8.97	4.18
MWV	*verb.*	585	243	41.5	3.01
MWD	*det.*	328	69	21.0	1.69
MWA	*adj.*	324	126	38.9	1.66
MWPRO	*pron.*	266	33	12.4	1.37
MWCL	*clitic*	59	1	1.69	0.30
MWET	*foreign*	24	18	0.75	0.12
MWI	*interj.*	4	2	0.50	0.02
		19,462	4,093	21.0%	100.0%

Table 3: Frequency distribution of the 11 MWE subcategories in the FTB (training set). MWEs account for 7.08% of the bracketings and 13.0% of the tokens in the treebank. Only 21% of the MWEs occur once ("single").

We first introduce a new instantiation of the French Treebank that, unlike previous work, does not use gold MWE pre-grouping. Consequently, our experimental results also provide a better baseline for parsing raw French text.

2 French Treebank Setup

The corpus used in our experiments is the French Treebank (Abeillé et al. (2003), version from June 2010, hereafter FTB). In French, there is a linguistic tradition of lexicography which compiles lists of MWEs occurring in the language. For example, Gross (1986) shows that dictionaries contain about 1,500 single-word adverbs but that French contains over 5,000 multiword adverbs. MWEs occur in every part-of-speech (POS) category (e.g., noun *trousse de secours* 'first-aid kit'; verb *faire main-basse* [do hand-low] 'seize'; adverb *comme dans du beurre* [as in butter] 'easily'; adjective *'à part entière'* 'wholly').

The FTB explicitly annotates MWEs (also called *compounds* in prior work). We used the subset of the corpus with functional annotations, not for those annotations but because this subset is known to be more consistently annotated. POS tags for MWEs are given not only at the MWE level, but also internally: most tokens that constitute an MWE also have a POS tag. Table 2 compares this part of the FTB to the WSJ portion of the Penn Treebank.

2.1 Preprocessing

The FTB requires significant pre-processing prior to parsing.

Tokenization We changed the default tokenization for numbers by fusing adjacent digit tokens. For example, *500 000* is tagged as an MWE composed of two words *500* and *000*. We made this *500000* and retained the MWE POS, although we did not mark the new token as an MWE. For consistency, we used one token for punctuated numbers like "17,9".

MWE Tagging We marked MWEs with a flat bracketing in which the phrasal label is the MWE-level POS tag with an "MW" prefix, and the preterminals are the internal POS tags for each terminal. The resulting POS sequences are not always unique to MWEs: they appear in abundance elsewhere in the corpus. However, some MWEs contain normally ungrammatical POS sequences (e.g., adverb *à la va vite* 'in a hurry': P D V ADV [at the goes quick]), and some words appear only as part of an MWE, such as *insu* in *à l'insu de* 'to the ignorance of'.

Labels We augmented the basic FTB label set—which contains 14 POS tags and 19 phrasal tags—in two ways. First, we added 16 finer-grained POS tags for punctuation.[1] Second, we added the 11 MWE

[1] Punctuation tag clusters—as used in the WSJ—did not improve accuracy. Enriched tag sets like that of Crabbé and Candito (2008) could also be investigated and compared to our results since Evalb is insensitive to POS tags.