About \\AIHV\NLP\TREEBANK_IN_PROLOG (files stored on AI server)

These files contain data from the Penn Treebank in a form that is easy to handle in Prolog.

lexicon.pl -- Lexicon

This is derived from the lexicon of CPIDR, which is derived from the whole Penn Treebank.

Here we list each word (or punctuation mark), the most frequent tag, and all the tags that have been found with it, such as:

```
most_frequent_tag( ('run'), 'VB').
possible_tag( ('run'), 'NN').
possible_tag( ('run'), 'VB').
possible_tag( ('run'), 'VBD').
possible_tag( ('run'), 'VBN').
possible_tag( ('run'), 'VBP').
```

Note that the Treebank does contain mistakes and inconsistencies. For example, run/VBD is either a mistake or an instance of nonstandard English.

The following are derived from an approximately 47,000-word extract from the Wall Street Journal portion of the Penn Treebank, as supplied with Bird et al.'s Natural Language Toolkit:

treebank_words.pl --   The words, as a list of atoms.

treebank_tagged_words.pl -- Same, with part-of-speech tags. We use the Prolog infix operator '/' to join word to tag.

treebank_sentences.pl   -- Same, divided into sentences.

treebank_tagged_sentences.pl -- Same, with tags added.

treebank_trees.pl -- Same, with tree structure using a list notation similar to that of the original Treebank rather than the structure notation that is more usual in Prolog.

treebank_trees_simplified.pl -- Same, omitting some indications of grammatical relations, so that you have just a pure tree structure.