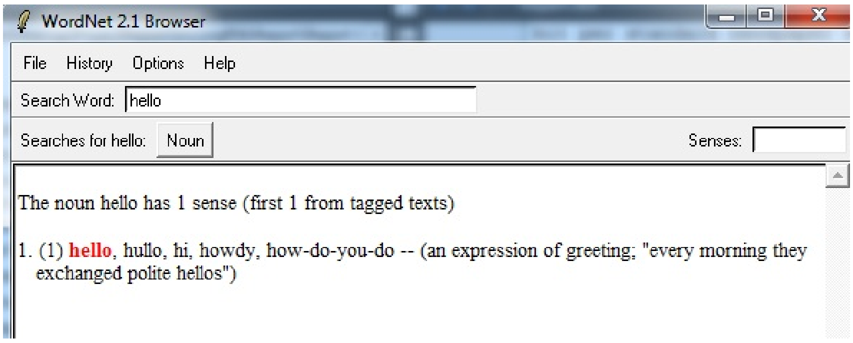
Searches for words that appear in the same synset return the same synset. It is this synset structure that makes WordNet ideal for use in this project.



Search Word: hello

The noun hello has 1 sense (first 1 from tagged texts)

1. (1) **hello**, hullo, hi, howdy, how-do-you-do -- (an expression of greeting; "every morning they exchanged polite hellos")

Figure 3.2: Screenshot of WordNet application after querying the word "hello"

Corpora

There are two sets of corpus data which are used in the algorithm. The first is from the British National Corpus (BNC) [3], and shows the frequency of a subset of words (along with their part-of-speech tags) which are found in dialogue (a subset is used as the full list of words is almost a million words long) [18]. The second set is taken from the American National Corpus (ANC) [28], and lists bigrams of words and their frequencies. The bigram data is limited to bigrams which have a frequency of more than 4, as below this the frequency is too low and also it limits the amount of data that needs to be searched (the full set is over 2 million in size, with this limit it is approximately 250 thousand).

3.1.2 Synonym Retrieval

Synonym retrieval is perhaps the most important aspect of the algorithm. If the synonyms are not retrieved properly, it would be difficult to ensure that data can be deobfuscated as there is guarantee that the synset you generate when deobfuscating is the same as when you obfuscated. A bad synset will also include words which are only very loosely synonymous with the original words, which will affect the robustness of the algorithm.

The first task in retrieving the synonyms for a word is to decide which of the four word types (noun, verb, adverb or adjective) the word is. In some cases this is simple as the word can only be of one type, but in many cases this is not the case and the word can be of two or more of the possible types. If the wrong type is chosen, then the synonym chosen may not make sense. To choose the most likely word type, the word frequency data from the BNC is used. The tags used in this list are much more detailed than in the WordNet database, so the word may appear in the list multiple times with many different tags all representing the same tag. When a word is searched in this list, the total frequencies for noun, verb adverb and adjective are all found, and the greatest frequency is used as the correct tag for the