[From: Internat.conf. *Methodology & Techniques of Machine Translation*, British Computer Society, Cranfield, 13-15 February 1984]

The Grammatical tagging of Unrestricted English Text

R. G. Garside

University of Lancaster, U.K.

ABSTRACT

The LOB (Lancaster-Oslo/Bergen) corpus is a million-word computer-readable selection of British English texts, paralleling the Brown corpus of American English. At the University of Lancaster we are currently completing a project to tag this corpus grammatically: that is, to assign to each word in the corpus an indication of what grammatical class it belongs to (singular common noun, past tense of verb, etc.).

The assignment of tags is performed by a suite of programs which correctly assigns tags to about 97% of the words in the corpus: the remaining 3% of words are dealt with by a manual post-editing phase. We have been able to achieve this level of accuracy by making use of the already tagged Brown corpus, as a source of information about what tags occur with a given word and in what context. It should be noticed that this level of accuracy is achieved over unrestricted written English text, fiction and non-fiction, dialogue, incomplete and non-standard English, etc.

The first step is for a suggested set of tags to be assigned to each word, at this stage ignoring the context of the word. The word is first looked up in a dictionary of about 7000 words with their associated tags. If this dictionary search fails, then a list of about 700 suffices or word-endings is searched. Associated with these two main procedures are further procedures to deal with hyphenated words, numbers and formulae, and to strip the "s" off plural nouns and third person singular forms of verbs.

The next step is to attempt to select the single correct tag for each word using the context of the word. For each potential sequence of two tags A and B, there is a probability of the tag B following the tag A. We have constructed a matrix of probabilities from the tagged Brown corpus, with suitable modifications for the changes we have made to the tag system. Then a program in the tagging suite takes a sequence of words each with more than one tag and bounded at each end by a word with a single unambiguous tag.

The program computes the probability of each possible sequence of tags starting and finishing with the unambiguous ones, by multiplying together the probabilities of the individual links, and rearranges the tags associated with each word so that they are arranged (together with the computed probability of the tag) in order of decreasing likelihood. If the most likely tag has a sufficiently high probability the program deletes the remaining tags associated with the word; otherwise the choice is left to the manual post-editor. Certain modifications had to be made to this conceptually simple scheme (particularly in the areas of conjunctions and of adverbs) to improve the success rate of disambiguation.

We are currently analysing the errors produced by this system, and planning the use of similar probabilistic techniques in the syntactic analysis of the LOB corpus.