GRAMMATICAL ANALYSIS BY COMPUTER OF THE LANCASTER-OSLO/BERGEN (LOB) CORPUS OF BRITISH ENGLISH TEXTS.

Andrew David Beale
Unit for Computer Research on the English Language
Bowland College, University of Lancaster
Bailrigg, Lancaster, England LA1 4YT.

ABSTRACT

Research has been under way at the Unit for Computer Research on the English Language at the University of Lancaster, England, to develop a suite of computer programs which provide a detailed grammatical analysis of the LOB corpus, a collection of about 1 million words of British English texts available in machine readable form.

The first phrase of the project, completed in September 1983, produced a grammatically annotated version of the corpus giving a tag showing the word class of each word token. Over 93 per cent of the word tags were correctly selected by using a matrix of tag pair probabilities and this figure was upgraded by a further 3 per cent by retagging problematic strings of words prior to disambiguation and by altering the probability weightings for sequences of three tags. The remaining 3 to 4 per cent were corrected by a human post-editor.

The system was originally designed to run in batch mode over the corpus but we have recently modified procedures to run interactively for sample sentences typed in by a user at a terminal. We are currently extending the word tag set and improving the word tagging procedures to further reduce manual intervention. A similar probabilistic system is being developed for phrase and clause tagging.

THE STRUCTURE AND PURPOSE OF THE LOB CORPUS.

The LOB Corpus (Johansson, Leech and Goodluck, 1978), like its American English counterpart, the Brown Corpus (Kučera and Francis, 1964; Hauge and Hofland, 1978), is a collection of 500 samples of British English texts, each containing about 2,000 word tokens. The samples are representations of 15 different text categories: A. Press (Reportage); B. Press (Editorial); C. Press (Reviews); D. Religion; E. Skills and Hobbies; F. Popular Lore; G. Belles Lettres, Biography, Memoirs, etc.; H. Miscellaneous; J. Learned and Scientific; K. General Fiction; L. Mystery and Detective Fiction; M. Science Fiction; N. Adventure and Western Fiction, Romance and Love Story; R. Humour. There are two main sections, informative prose and imaginative prose, and all the texts contained in the corpus were printed in a single year (1961).

The structure of the LOB corpus was designed to resemble that of the Brown corpus as closely as possible so that a systematic comparison of British and American written English could be made. Both corpora contain samples of texts published in the same year (1961) so that comparisons are not distorted by diachronic factors.

The LOB corpus is used as a database for linguistic research and language description. Historically, different linguists have been concerned to a greater or lesser extent with the use of corpus citations, to some degree, at least, because of differences in the perceived view of the descriptive requirements of grammar. Jespersen (1909-49), Kruisinga and Erades (1911) gave frequent examples of citations from assembled corpora of written texts to illustrate grammatical rules. Work on text corpora is, of course, very much alive today. Storage, retrieval and processing of natural language text is a more efficient and less laborious task with modern computer hardware than it was with hand-written card files but data capture is still a significant problem (Francis, 1980). The forthcoming work, A Comprehensive Grammar of the English Language (Quirk, Greenbaum, Leech, and Svartvik, 1985) contains many citations from both LOB and Brown Corpora.

A GRAMMATICALLY ANNOTATED VERSION OF THE CORPUS

Since 1981, research has been directed towards writing programs to grammatically annotate the LOB corpus. From 1981-83, the research effort produced a version of the corpus with every word token labelled by a grammatical tag showing the word class of each word form. Subsequent research has attempted to build on the techniques used for automatic word tagging by using the output from the word tagging programs as input to phrase and clause tagging and by using probabilistic methods to provide a constituent analysis of the LOB corpus.

The programs and data files used for word tagging were developed from work done at Brown University (Greene and Rubin, 1971). Staff and research associates at Lancaster undertook the programming in PASCAL while colleagues in Oslo revised and extended the lists used by Greene and Rubin (op.cit.) for word tag assignment. Half of the corpus was post-edited at Lancaster and the other half at the Norwegian Computing Centre for the Humanities.

How word tagging works.

The major difficulties to be encountered with word tagging of written English are the lack of distinctive inflectional or derivational endings and the large proportion of word forms that belong to more than one word class. Endings such as -able, -ly and -ness are graphic realizations of morphological units indicating word class, but they occur infrequently for the purposes of automatic word tag assignment; the reader will be able to establish exceptions to rules assigning word classes to words with these suffixes, because the characters do not invariably represent the same morphemes.

The solution we have adopted is to use a look up procedure to assign one or more potential tags to each input word. The appropriate word tag is then selected for words with more than one potential tag by calculating the probability of the tag's occurrence given neighbouring potential tags.

Potential word tag assignment.

In cases where more than one potential tag is assigned to the input word, the tags represent word classes of the word without taking the syntactic environment into account. A list of one to five word final characters, known as the 'suffixlist', is used for assignment of appropriate word class tags to as many word types as possible. A list of full word forms, known as the 'wordlist', is used for exceptions to the suffixlist, and, in addition, word forms that occur more than 50 times in the corpus are included in the wordlist, for speed of processing. The term 'suffixlist' is used as a convenient name, and the reader is warned that the list does not necessarily contain word final morphs; strings of between one and five word final characters are included if their occurrence as a tagged form in the Brown corpus merits it.

The 'suffixlist' used by Greene and Rubin (op.cit.) was substantially revised and extended by Johansson and Jahr (1982) using reverse alphabetical lists of approximately 50,000 word types of the Brown Corpus and 75,000 word types of both Brown and LOB corpora. Frequency lists specifying the frequency of tags for word endings consisting of 1 to 5 characters were used to establish the efficiency of each rule. Johansson and Jahr were guided by the Longman Dictionary of Contemporary English (1978) and other dictionaries and grammars including Quirk, Greenbaum, Leech and Svartvik (1972) in identifying tags for each item in the wordlist. For the version used for Lancaster-Oslo/Bergen word tagging (1983), the suffixlist was expanded to about 700 strings of word final characters, the wordlist consisted of about 7,000 entries and a total of 133 word tag types were used.

Potential tag disambiguation.

The problem of resolving lexical ambiguity for the large proportion of English words that occur in more than one word class, (BLOW, CONTACT, HIT, LEFT, RAIN, RUN, REFUSE, ROSE, WALK, WATCH ...), is solved, whenever possible by examining the local context. Word tag selection for homographs in Greene and Rubin (op. cit.) was attempted by using 'context frame rules', an ordered list of 3,300 rules designed to take into account the tags assigned to up to two words preceding or following the ambiguous homograph. The program was 77 per cent successful but several errors were due to appropriate rules being blocked when adjacent ambiguities were encountered (Marshall, 1983: 140). Moreover, about 80 per cent of rule application took just one immediately neighbouring tag into account, even though only a quarter of the context frame rules specified only one immediately neighbouring tag.

To overcome these difficulties, research associates at Lancaster have devised a transition probability matrix of tag pairs to compute the most probable tag for an ambiguous form given the immediately preceding and following tags. This method of calculating one-step transition probabilities is suitable for disambiguating strings of ambiguously tagged words because the most likely path through a string of ambiguously tagged words can be calculated.

The likelihood of a tag being selected in context is also influenced by likelihood markers which are assigned to entries with more than one tag in the lists. Only two markers, '@' and '%', are used, '@' notionally indicating that the tag is correct for the associated form less than 1 in 10 occasions, '%' notionally indicating that the tag occurs less than 1 in 100 occasions. The word tag disambiguation program uses these markers to reduce the probability of the less likely tags occurring in context; '@' results in the probability being halved, '%' results in the probability being divided by eight. Hence tags marked with '@' or '%' are only selected if the context indicates that the tag is very likely.

Error analysis.

At several stages during design and implementation of the tagging software, error analysis was used to improve various aspects of the word tagging system. Error statistics were used to amend the lists, the transition matrix entries and even the formula used for calculating transition probabilities (originally this was the frequency of potential tag A followed by potential tag B divided by the frequency of A. Subsequently, it was changed to the frequency of A followed by B divided by the product of the frequency of A and the frequency of B (Marshall, 1983: 144ff)).

Error analysis indicated that the one-step transition method for word tag disambiguation was very successful, but it was evident that further gains could be made by including a separate list of a small set of sequences of words such as according to, as well as, and so as to which were retagged prior to word tag disambiguation. Another modification was to include an algorithm for altering the values of sequences of three tags, such as constructions with an intervening adverb or simple co-ordinated constructions such that the two words on either side of a co-ordinating conjunction contained the same tag where a choice was available.

No value in the matrix was allowed to be as little as zero, by providing a minimum positive value for even extremely unlikely tag co-occurrences; this allowed at least some kind of analysis for unusual or eccentric syntax and prevented the system from grinding to a halt when confronted with a construction that it did not recognize.

Once these refinements to the suite of word tagging programs were made, the corpus was word-tagged. It was estimated that the number of manual post-editing interventions had been reduced from about 230,000 required for word tagging of the Brown corpus to about 35,000 required for the LOB corpus (Leech, Garside and Atwell, 1983: 36). The method achieves far greater consistency than could be attained by a human, were such a person able to labour through the task of attributing a tag to every word token in the corpus.

A record of decisions made at the post-editing stage was kept for the purpose of recording the criteria for judging whether tags were considered to be correct or not (Atwell, 1982b).

Improving word tagging.

Work currently being undertaken at Lancaster includes revising and extending the word tag set and improving the suite of programs and data files required to carry out automatic word tagging.

Revision of the word tag set.

The word tag set is being revised so that, whenever possible, tags are mnemonic such that the characters chosen for a tag are abbreviations of the grammatical categories they represent. This criterion for word tag improvement is solely for the benefit of human intelligibility and in some cases, because of conflicting criteria of distinctiveness and brevity, it is not always possible to devise clearly mnemonic tags. For instance, nouns and verbs can be unequivocally tagged by the first letter abbreviations 'N' and 'V', but the same cannot be said for articles, adverbs and adjectives. These categories are represented by the tags 'AT', 'RR', and 'JJ'.

It was decided, on the grounds of improving mnemonicity, to change representation of the category of number in the tag set. In the old tag set, singular forms of articles, determiners, pronouns and nouns were unmarked, and plural forms had the same tags as the singular forms but with 'S' as the end character denoting plural. As far as mnemonicity is concerned, this is confusing, especially to someone uninitiated in the refinements of LOB tagging. In the new tag set, number is now marked by having '1' for singular forms, '2' for plural forms and no number character for nouns, articles and determiners which exhibit no singular or plural morphological distinctiveness (COD, SHEEP, TROUT ...).

It is desirable, both for the purposes of human intelligibility and for mechanical processing, to make the tagged system as hierarchized as possible. In the old tag set modal verbs, and forms of the verbs BE, DO and HAVE were tagged as 'M*', 'B*', 'D*', and 'H*' (where '*' represents any of the characters used for these tags denoting subclasses of each tag class). In the new word tag set, these have been recoded 'VM*', 'VB*', 'VD*', 'VH*', to show that they are, in fact, verbs, and to facilitate verb counting in a frequency analysis of the tagged corpus; 'VV*' is the new tag for lexical verbs.

It has been taken as a design principle of the new tag set that, wherever possible, subcategories and supercategories should be retrieved by referring to the character position in the string of characters making up a tag, major word class coding being denoted by the initial character(s) of the tag and subsequent characters denoting morpho-syntactic subcategories.

Hierarchization of the new tag set is best exemplified by pronouns. 'P*' is a pronoun, as distinct from other tag initial characters, such as 'N*' for noun, 'V*' for verb and so on. 'PP*' is a personal pronoun, as distinct from 'PN*', an indefinite pronoun; 'PPI*' is a first person personal pronoun: *I*, *me*, *we*, *us*, as distinct from 'PPY*', 'PPH*' and 'PPX*' which are second, third person and reflexive pronouns; 'PPIS*' is a first person subject personal pronoun: *I* and *we*, as distinct from first person object personal pronouns, *me*, and *us*, denoted by 'PPIO*'; finally 'PPIS1:' is the first person singular subject personal pronoun, *I* (the colon is used to show that the form must have an initial capital letter).

The third criterion for revising and enlarging the word tag set is to improve and extend the linguistic categorisation. For instance, a tag for the category of predicative adjective, 'JA', has been introduced for adjectives like *ablaze*, *adrift* and *afloat*, in addition to the already existing distinction between attributive and ordinary adjectives, marked 'JB' as distinct from 'JJ'. There is an essential distributional restriction on subclasses of adjectives occurring only attributively or predicatively, and it was considered appropriate to notate this in the tag set in a consistent manner. The attributive category has been introduced for comparative adjectives, 'JBR', (UPPER, OUTER ...) and superlative adjectives, 'JBT', (UTMOST, UTTERMOST ...).

As a further example of improving the linguistic categorization without affecting the proportion of correctly tagged word forms, consider the word ONE. In the old tagging system, this word was always assigned the tag 'CD1'. This is unsatisfactory, even though ONE is always assigned the tag it is supposed to receive, because ONE is not simply a singular cardinal number. It can be a singular impersonal pronoun, *One is often surprised by the reaction of the pupils*, or a singular common noun, *He wants this one*, contrasting, for instance, with its plural form *He wants those ones*. It is therefore appropriate for ONE to be assigned 3 potential tags, 'CD1', 'PN1', and 'NN1', one of which is to be selected by the transition probability procedure.

Revision of the programs and data files.

Revision of the word tag set has necessitated extensive revision of the word- and suffixlists. The transition matrix will be adapted so that the corpus can be retagged with tags from the new word tag set. In addition, programs are being revised to reduce the need for special pre-editing and input format requirements. In this way, it will be possible for the system to tag English texts other than the LOB corpus without pre-editing.

Reducing Pre-editing.

For the 1983 version of the tagged corpus, a pre-editing stage was carried out partly by computer and partly by a human pre-editor (Atwell, 1982a). As part of this stage, the computer automatically reduced all sentence-initial capital letters and the human pre-editor recapitalized those sentence initial characters that began proper nouns. We are now endeavouring to cut out this phase so that the automatic tagging suite can process input text in its normal orthographic form as mixed case characters.

Sentence boundaries were explicitly marked, as part of the input requirements to the tagging procedures, and since the word class of a word with an initial capital letter is significantly affected by whether it occurs at the beginning of a sentence, it was considered appropriate to make both sentence boundary recognition and word class assignment of words with a word initial capital automatic. All entries in the word list now appear entirely in lower case and words which occur with different tags according to initial letter status (board, march, may, white ...) are assigned tags according to a field selection procedure: the appropriate tags are given in two fields, one for the initial upper case form (when not acting as the standard beginning-of-sentence marker) and the other for the initial lower case form. The probability of tags being selected from the alternative lists is weighted according to whether the form occurs at the beginning of the sentence or elsewhere.

Knut Hofland estimated a success rate of about 94.3 per cent without pre-editing (Leech, Garside and Atwell, 1983: 36). Hence, the success rate only drops by about 2 per cent without pre-editing. Nevertheless, the problems raised by words with tags varying according to initial capital letter status need to be solved if the system is to become completely automatic and capable of correct tagging of standard text.

Constituent Analysis.

The high success rate of word tag selection achieved by the one-step probability disambiguation procedure prompted us to attempt a similar method for the more complex tasks of phrase and clause tagging. The paper by Garside and Leech in this volume deals more fully with this aspect of the work.

Rules and symbols for providing a constituent analysis of each of the sentences in the corpus are set out in a Case-Law Manual (Sampson, 1984) and a series of associated documents give the reasoning for the choice of rules and symbols (Sampson, 1983 -). Extensive tree drawing was undertaken while the Case-Law Manual was being written, partly to establish whether high-level tags and rules for high-level tag assignment needed to be modified in the light of the enormous variety and complexity of ordinary sentences in the corpus, and partly to create a databank of manually parsed samples of the LOB corpus, for the purposes of providing a first-approximation of the statistical data required to disambiguate alternative parses.

To date, about 35,000 words (1,500 sentences) have been manually parsed and keyed into an ICL VME 2900 machine. We are presently aiming for a tree bank of about 50,000 words of evenly distributed samples taken from different corpus categories representing a cross-section of about 5 per cent of the word tagged corpus.

The future.

It should be made clear to the reader that several aspects of the research are cumulative. For instance, the statistics derived from the tagged Brown corpus were used to devise the one-step probability program for word tag disambiguation. Similarly, the word tagged LOB corpus is taken as the input to automatic parsing.

At present, we are attempting to provide constituent structures for the LOB corpus. Many of these constructions are long and complex; it is notoriously difficult to summarise the rich variety of written English, as it actually occurs in newspapers and books, by using a limited set of rewrite rules. Initially, we are attempting to parse the LOB corpus using the statistics provided by the tree bank and subsequently, after error analysis and post-editing, statistics of the parsed corpus can be used for further research.

ACKNOWLEDGEMENTS

The work described by the author of this paper is currently supported by Science and Engineering Research Council Grant GR/C/47700.

REFERENCES

Abbreviation:
ICAME = International Computer Archive of Modern English.

Atwell, E.S. (1982a). LOB Corpus Tagging Project: Manual Pre-edit Handbook. Unpublished document: Unit for Computer Research on the English Language, University of Lancaster.
(1982b). LOB Corpus Tagging Project: Manual Post-edit Handbook. (A mini-grammar of LOB Corpus English, examining the types of error commonly made during automatic (computational) analysis of ordinary written English). Unpublished document: Unit for Computer Research on the English Language, University of Lancaster.

Francis, W.N. (1980). 'A tagged corpus - problems and prospects', in Studies in English linguistics for Randolph Quirk (1980) edited by S. Greenbaum, G.N. Leech and J. Svartvik, 192-209. London: Longman.

Greene, B.B. and Rubin, G.M. (1971). 'Automatic Grammatical Tagging of English', Providence, R.I.: Department of Linguistics, Brown University.

Hauge, J. and Hofland, K. (1978). Microfiche version of the Brown University Corpus of Present-Day American English. Bergen: NAVF's EDB-Senter for Humanistisk Forskning.

Jespersen, O. (1909-49). A Modern English Grammar on Historical Principles, Munksgaard.

Johansson, S. (1982) (editor). Computer Corpora in English language research. Bergen: Norwegian Computing Centre for the Humanities.

Johansson, S. and Jahr, M-C. (1982). 'Grammatical Tagging of the LOB Corpus: Predicting Word Class from Word Endings', in S. Johansson (1982), 118-134.

Johansson, S., Leech, G. and Goodluck, H. (1978). Manual of information to accompany the Lancaster-Oslo/Bergen corpus of British English, for use with digital computers. Unpublished document: Department of English, University of Oslo.

Kruisinga, E. and Erades, P.A. (1911). An English Grammar. Nordhoof.

Kučera, H. and Francis, W.N. (1964, revised 1971 and 1979). Manual of Information to accompany A Standard Corpus of Present-Day Edited American English, for use with Digital Computers. Providence, Rhode Island: Brown University Press.

Leech, G.N., Garside, R., and Atwell, E. (1983). 'Recent Developments in the use of Computer Corpora in English Language Research', Transactions of the Philological Society, 23-40.

Longman Dictionary of Contemporary English (1978). London: Longman.

Marshall, I. (1983). 'Choice of Grammatical Word-Class without Global Syntactic Analysis: Tagging Words in the LOB Corpus', Computers and the Humanities, Vol. 17, No. 3, 139-150.

Quirk, R., Greenbaum, S., Leech., G.N. and Svartvik, J. (1972). A Grammar of Contemporary English. London: Longman.
(1985). A Comprehensive Grammar of the English Language. London: Longman.

Sampson, G.R. (1984). UCREL Symbols and Rules for Manual Tree-Drawing. Unpublished document: Unit for Computer Research on the English Language, University of Lancaster.
(1983 -). Tree Notes I - XIV. Unpublished documents: Unit for Computer Research on the English Language, University of Lancaster.