CLAWS4: THE TAGGING OF THE BRITISH NATIONAL CORPUS

Geoffrey Leech, Roger Garside and Michael Bryant

UCREL, Lancaster University, UK

1 INTRODUCTION

The main purpose of this paper is to describe the CLAWS4 general-purpose grammatical tagger, used for the tagging of the 100-million-word British National Corpus, of which c.70 million words have been tagged at the time of writing (April 1994).[1] We will emphasise the goals of (a) general-purpose adaptability, (b) incorporation of linguistic knowledge to improve quality and consistency, and (c) accuracy, measured consistently and in a linguistically informed way.

The British National Corpus (BNC) consists of c.100 million words of English written texts and spoken transcriptions, sampled from a comprehensive range of text types. The BNC includes 10 million words of spoken language, c.45% of which is impromptu conversation (see Crowdy, forthcoming). It also includes an immense variety of written texts, including unpublished materials. The grammatical tagging of the corpus has therefore required the 'super-robustness' of a tagger which can adapt well to virtually all kinds of text. The tagger also has had to be versatile in dealing with different *tagsets* (sets of grammatical category labels — see 3 below) and accepting text in varied *input formats*. For the purposes of the BNC, the tagger has been required both to accept and to output text in a corpus-oriented TEI-conformant mark-up definition known as CDIF (Corpus Document Interchange Format), but within this format many variant formats (affecting, for example, segmentation into words and sentences) can be readily accepted. In addition, CLAWS allows variable *output formats*: for the current tagger, these include (a) a vertically-presented format suitable for manual editing, and (b) a more compact horizontally-presented format often more suitable for end-users. Alternative output formats are also allowed with (c) so-called 'portmanteau tags', i.e. combinations of two alternative tags, where the tagger calculates there is insufficient evidence for safe disambiguation, and (d) with simplified 'plain text' mark-up for the human reader. (See Tables 1 and 2 for examples of output formats.)

CLAWS4, the BNC tagger,[2] incorporates many features of adaptability such as the above. It also incorporates many refinements of linguistic analysis which have built up over 14 years: particularly in the construction and content of the idiom-tagging component (see 2 below). At the same time, there are still many improvements to be made: the claim that 'you can put together a tagger from scratch in a couple of months' (recently heard at a research conference) is, in our view, absurdly optimistic.

2 THE DESIGN OF THE GRAMMATICAL TAGGER (CLAWS4)

The CLAWS4 tagger is a successor of the CLAWS1 tagger described in outline in Marshall (1983), and more fully in Garside et al (1987), and has the same basic architecture. The system (if we include input and output procedures) has five major sections:

(a) segmentation of text into word and sentence units

(b) initial (non-contextual) part-of-speech assignment [using a lexicon, word-ending list, and various sets of rules for tagging unknown items]

[1] The BNC is the result of a collaboration, supported by the Science and Engineering Research Council (SERC Grant No. GR/F99847) and the UK Department of Trade and Industry, between Oxford University Press (lead partner), Longman Group Ltd., ChambersHarrap, Oxford University Computer Services, the British Library and Lancaster University. We thank Elizabeth Eyes, Nick Smith, and Andrew Wilson for their valuable help in the preparation of this paper.

[2] CLAWS4 has been written by Roger Garside, with CLAWS adjunct software written by Michael Bryant.

(c) rule-driven contextual part-of-speech assignment

(d) probabilistic tag disambiguation [Markov process]

[(c') second pass of (c)]

(e) output in intermediate form.

The intermediate form of text output is the form suitable for post-editing (see 1 above; also Table 1), which can then be converted into other formats according to particular output needs, as already noted.

The pre-processing section (a) is not trivial, since, in any large and varied corpus, there is a need to handle unusual text structures (such as those of many popular and technical magazines), less usual graphic features (e.g. non-roman alphabetic characters, mathematical symbols), and features of conversation transcriptions: e.g. false starts, incomplete words and utterances, unusual expletives, unplanned repetitions, and (sometimes multiple) overlapping speech.

Sections (b) and (d) apply essentially a Hidden Markov Model (HMM) to the assignment and disambiguation of tags. But the intervening section (c) has become increasingly important as CLAWS4 has developed the need for versatility across a range of text types. This task of *rule-driven contextual part-of-speech assignment* began in 1981 as an 'idiom-tagging' program for dealing, in the main, with parts of speech extending over more than one orthographic word (e.g. complex prepositions such as *according to* and complex conjunctions such as *so that*). In the more fully developed form it now has, this section utilises several different idiom lexicons dealing, for example, with (i) general idioms such as *as much as* (which is on one analysis a single coordinator, and on another analysis, a word sequence), (ii) complex names such as *Dodge City* and *Mrs Charlotte Green* (where the capital letter alone would not be enough to show that *Dodge* and *Green* are proper nouns), (iii) foreign expressions such as *annus horribilis*.

These idiom lexicons (with over 3000 entries in all) can match on both tags and word-tokens, employing a regular expression formalism at the level both of the individual item and of the sequence of items. Recognition of unspecified words with initial capitals is also incorporated. Conceptually, each entry has two parts: (a) a regular-expression-based 'template' specifying a set of conditions on sequences of word-tag pairs, and (b) a set of tag assignments or substitutions to be performed on any sequence matching the set of conditions in (a). Examples of entries from each of the above kinds of idiom lexicon entry are:

(i) did/do/does, ([XX0/AV0/ORD])2, [VVB] VVI

(ii) Monte/Mount/Mt NP0, ([WIC])2 NP0, [WIC] NP0

(iii) ad AV021 AJ021, hoc AV022 AJ022

Explanatory note:

(a) Symbolic formalism:

Let TT *be any tag, and let* ww *be any word.*
Let n,m be arbitrary integers. Then:

ww TT represents a word and its associated tag

, separates a word from its predecessor

[TT] represents an already assigned tag

[WIC] represents an unspecified word with a Word Initial Capital

TT/TT means 'either TT or TT'; ww/ww means 'either ww or ww'

ww TT TT represents an unresolved ambiguity between TT and TT

TT* represents a tag with * marking the location of unspecified characters

([TT])*n* represents the number of words (up to *n*) which may optionally intervene at a given point in the template

TT*nm* represents the 'ditto tag' attached to an orthographic word to indicate it is part of a complex sequence (e.g. *so that* is tagged *so* CJS21 , *that* CJS22). The variable *n* indicates the number of orthographic words in the sequence, and *m* indicates that the current word is in the *m*th position in that sequence.

(b) Examples of word tags (in the C5 'basic' tagset):

AJ0	adjective (unmarked)
AV0	adverb (unmarked)
CJS	subordinating conjunction
NP0	proper noun
ORD	ordinal number
VVB	finite base form of lexical verb
VVI	infinitive of lexical verb
XX0	negative marker: *not* or *n't*

(c) Explanation of the three rules above:

Rule (i) ensures that following a finite form of *do* and (optionally) up to two adverbs, negators or ordinals, a base form of the verb is tagged as an infinitive.

Rule (ii) ensures that in complex names such as *Monte Alegre, Mount Pleasant, Mount Palomar Observatory, Mt Rushmore National Memorial*, all the words with word-initial caps are tagged as proper nouns.

Rule (iii) ensures that the Latin expression *ad hoc* is tagged as a single word, either an adjective or an adverb.

We have also now moved to a more complex, two-pass application of these idiomlist entries. It is possible, on the first pass, to specify ambiguous output of an idiom assignment (as is necessary, e.g., for *as much as*, mentioned earlier), so that this can then be input to the probabilistic disambiguation process (d). On the second pass, however, after probabilistic disambiguation, the idiom entry is deterministic in both its input and output conditions, replacing one or more tags by others. In effect, this last kind of idiom application is used to correct a tagging error arising from earlier procedures. For example, a not uncommon result from Sections (a)-(d) is that the base form of the verb (e.g. *carry*) is wrongly tagged as a finite present tense form, rather than an infinitive. This can be retrospectively corrected by replacing VVB (= finite base form) by VVI (= infinitive) in appropriate circumstances.

While the HMM-type process employed in Sections (b) and (d) affirms our faith in probabilistic methods, the growing importance of the contextual part-of-speech assignment in (c) and (c') demonstrates the extent to which it is important to transcend the limitations of the orthographic word, as the basic unit of grammatical tagging, and also to selectively adopt non-probabilistic solutions. The term 'idiom-tagging' originally used was never particularly appropriate for these sections, which now handle more generally the interdependence between grammatical and lexical processing which NLP systems ultimately have to cope with, and are also able to incorporate parsing information beyond the range of the one-step Markov process (based on tag bigram frequences) employed in (d).[3] Perhaps the term 'phraseological component' would be more appropriate here. The need to combine probabilistic and non-probabilistic methods in tagging has been widely noted (see, e.g., Voutilainen et al. 1992:14).

3 EXTENDING ADAPTABILITY: SPOKEN DATA AND TAGSETS

The tagging of 10 million words of spoken data (including c.4.6 million words of conversation) presents particular challenges to the versatility of the system: renderings of spoken pronunciations such as *'avin'* (for *having*) cause difficulties, as do unplanned repetitions such as *I er, mean, I mean, I mean to go*. Our solution to the latter problem has been to recognize such repetitions by a special procedure, and to disregard, in most cases, the repeated occurrences of the same word or phrase for the purposes of tagging. It has become clear that the CLAWS4 resources (lexicon, idiomlists, and tag transition matrix), developed for written English, need to be adapted if certain frequent and rather consistent errors in the tagging of spoken data are to be avoided (words such as *I, well,* and *right* are often wrongly tagged, because their distribution in conversation differs markedly from that in written texts). We have moved in this direction by allowing CLAWS4 to 'slot in' different resources according to the text type being processed, by e.g. providing a separate supplementary lexicon and idiomlist for the spoken material. Eventually, probabilistic analysis of the tagged BNC will provide the necessary information for adapting datastructures at run time to the special demands of particular types of data, but there is much work to be done before this potential benefit of having tagged a large corpus is realised.

The BNC tagging takes place within the context of a larger project, in which a major task (undertaken by OUCS at Oxford) is to encode the texts in a TEI-conformant mark-up (CDIF). Two tagsets have been employed: one, more detailed than the other, is used for tagging a 2-million-word Core Corpus (an epitome of the whole BNC), which is being post-edited for maximum accuracy. Thus tagsets, like text formats and resources, are among the features which are task-definable in CLAWS4. In general, the system has been revised to allow many adaptive decisions to be made at run time, and to render it suitable for non-specialist researchers to use.

[3] We have experimented with a two-step Markov process model (using tag trigrams), and found little benefit over the one-step model (using tag bigrams).

4 ERROR RATES AND WHAT THEY MEAN

Currently, judged in terms of major categories,[4] the system has an error-rate of approximately 1.5%, and leaves c.3.3% ambiguities unresolved (as portmanteau tags) in the output. However, it is all too easy to quote error rates, without giving enough information to enable them to be properly assessed.[5] We believe that any evaluation of the accuracy of automatic grammatical tagging should take account of a number of factors, some of which are extremely difficult to measure:

4.1 Consistency

It is necessary to measure tagging practice against some standard of what is an appropriate tag for a given word in a given context. For example, is *horrifying* in *a horrifying adventure*, or *washing* in *a washing machine* an adjective, a noun, or a verb participle? Only if this is specified independently, by an annotation scheme, can we feel confident in judging where the tagger is 'correct' or 'incorrect'. For the tagging of the LOB Corpus by the earliest version of CLAWS, the annotation scheme was published in some detail (Johansson et al 1986). We are working on a similar annotation scheme document (at present a growing in-house document) for the tagging of the BNC.

4.2 Size of Tagset

It might be supposed that tagging with a finer-grained tagset which contains more tags is more likely to produce error than tagging with a smaller and cruder tagset. In the BNC project, we have used a tagset of 58 tags (the C5 tagset) for the whole corpus, and in addition we have used a larger tagset of 138 tags (the C6 tagset)[6] for the Core Corpus of 2 million words. The evidence so far is that this makes little difference to the error rate. But size of tagset must, in the absence of more conclusive evidence, remain a factor to be considered.

4.3 Discriminative Value of Tags

The difficulty of grammatical tagging is directly related to the number of words for which a given tag distinction is made. This measure may be called 'discriminative value'. For example, in the C5 tagset, one tag (VDI) is used for the infinitive of just one verb — *to do* — whereas another tag (VVI) is used for the infinitive of all lexical verbs. On the other hand, VDB is used for finite base forms of *to do* (including the present tense, imperative, and subjunctive), whereas VVB is used of finite base forms of all lexical verbs. It is clear the tags VDI and VDB have a low discriminative value, whereas VVI and VVB have a high one — since there are thousands of lexical verbs in English. It is also clear that a tagset of the lowest possible discriminative value — one which assigned a single tag to each word and a single word to each tag — would be utterly valueless.

4.4 Linguistic Quality

This is a very elusive, but crucial concept. How far are the tags in a particular tagset valuable, by criteria either of linguistic theory/description, or of usefulness in NLP? For example, the tag VDI, mentioned in c. above, appears trivial, but it can be argued that this is nevertheless a useful category for English grammar, where the verb *do* (unlike its equivalent in most other European languages) has a very special function, e.g. in forming questions and negatives. On the other hand, if we had decided to assign a special tag to the verb *become*, this would have been more questionable. Linguistic quality is,

[4] The error rate and ambiguity rate are less favourable if we take account of errors and ambiguities which occur within major categories. E.g. the portmanteau tag NP0-NN1 records confidently that a word is a noun, but not whether it is a proper or common noun. If such cases are added to the count, then the estimated error rate rises to 1.78%, and the estimated ambiguity rate to 4.60%.

[5] One reasonable attempt to evaluate competing accuracy of different taggers is that in Voutilainen et al (1992:11–13), where it is argued, on the basis of the tagging of sample written texts, that the performance of the Helsinki constraint grammar parser ENGCG is superior to that of CLAWS1 (the earliest version of CLAWS, completed in 1983), which is in turn is somewhat superior to Church's Parts tagger. While recognizing that the accuracy of the Helsinki system is impressive, we note also that the method of evaluation (in terms of 'precision' and 'recall') employed by Voutilainen et al in not easy to compare with the method employed here. Further, a strict attempt at measuring comparability would have to take fuller account of the 'consistency' and 'quality' criteria we mention, and of the need to compare across a broader range of texts, spoken and written. This issue cannot be taken further in this paper, but will hopefully be the basis of future research.

[6] The tagset figures exclude punctuation tags and portmanteau tags.

on the face of it, determined only in a judgemental manner. Arguably, in the long term, it can be determined only by the contribution a particular tag distinction makes to success in particular applications, such as speech recognition or machine-aided translation. At present, this issue of linguistic quality is the Achilles' heel of grammatical tagging evaluation, and we must note that without judgement on linguistic quality, evaluation in terms of b. and c. is insecurely anchored.

It seems reasonable, therefore, to lump criteria b.-d. together as 'quality criteria', and to say that evaluation of tagging accuracy must be undertaken in conjunction with (i) consistency [How far has the annotation scheme been consistently applied?], and (ii) quality of tagging [How good is the annotation scheme?].[7] Error rates are useful interim indications of success, but they have to be corroborated by checking, if only impressionistically, in terms of qualitative criteria. Our work, since 1980, has been based on the assumption that qualitative criteria count, and that it is worth building 'consensual' linguistic knowledge into the datastructures used by the tagger, to make sure that the tagger's decisions are fully informed by qualitative considerations.

REFERENCES

Crowdy, S. (forthcoming). The BNC Spoken Corpus. In G. Leech, G. Myers and J. Thomas (Ed.). *Spoken English on Computer*. London: Longman.

Garside, R., G. Leech and G. Sampson (Eds). (1987). *The Computational Analysis of English: A Corpus-based Approach*. London: Longman.

Johansson, S., E. Atwell, R. Garside and G. Leech. (1986). *The Tagged LOB Corpus: User's Manual*. Bergen: Norwegian Computing Centre for the Humanities.

Marshall, I. (1983). Choice of grammatical word-class without global syntactic analysis: tagging words in the LOB Corpus. *Computers and the Humanities*, *17*, 139-50.

Voutilainen, A., J. Heikkilä and A. Anttila. (1992). *Constraint Grammar of English: A Performance-Oriented Introduction*. University of Helsinki: Department of General Linguistics.

[7] An example of a consistency issue is: How consistently is *Time* [the name of a magazine] tagged in the corpus? Is it tagged always NP0 (as a proper noun), or always NN1 (as a common noun), or sometimes NP0 and sometimes NN1? An example of a quality issue is: Is it worth distinguishing between proper nouns and common nouns, anyway?

Table 1: Tagged Spoken Data from the BNC — Vertical Format

```
------------------------------------------------------------------
0000203 070 I                                  03 PNP
0000203 080 can                                03 [VM0/100] NN1%/0
0000203 090 just                               03 [AV0/100] AJ0%/0
0000203 100 take                               98 VVI
0000204 010 note                               03 [NN1/99] VVB/1
0000204 020 of                                 03 PRF
0000204 030 any                                03 [DT0/100] AV0%/0
0000204 040 other                              03 [AJ0/99] NN1@/1
0000204 050 er                                 03 UNC
0000204 060 personal                           03 AJ0
0000204 070 pension                            03 [NN1/100] VVB@/0
0000204 071 ,                                  03 ,
0000204 080 not                                03 XX0
0000204 090 personal                           03 AJ0
0000204 100 pension                            03 [NN1/100] VVB@/0
0000204 101 ,                                  03 ,
0000204 110 any                                03 [DT0/97] AV0%/3
0000204 120 erm                                03 UNC
0000205 010 other                              03 [AJ0/98] NN1@/2
0000205 020 insurance                          03 NN1
0000205 030 you                    >           03 PNP
0000205 031 've                    <           03 VHB
0000205 040 got                                98 VVN
0000205 041 ,                                  03 ,
0000205 050 just                               03 [AV0/100] AJ0%/0
0000205 060 put                                03 [VVB/66] VVD@/22 VVN@/13
0000205 070 it                                 03 PNP
0000205 080 on                                 03 [AVP@/62] PRP/38
0000205 090 there                              03 [AV0/100] EX0/0
0000205 100 and                                97 CJC
0000205 101 ,                                  96 ,
0000205 110 and                                96 CJC
0000205 120 that                   >           97 DT0
0000205 121 's                     <           97 [VBZ/100] VHZ@/0
0000206 001 <unclear>                          01 NULL
0000206 002 </u>                               01 NULL
0000207 001 **22;2679;u                        01 NULL
0000207 002 ------------------------------------------------------
```

Table 2: Tagged Spoken Data from the BNC — Horizontal Format

```
<s c="0000201 002" n=00061>
<ptr t=P13> That&DT0;'s&VBZ; what&DTQ; <ptr t=P14> I&PNP; was&VBD;
told&VVN; to&TO0; bring&VVI; and&CJC; that&DT0;'s&VBZ; what&DTQ;
I&PNP; have&VHB; brought&VVN;.&PUN;
</u>
<u id=D0027 who=W0000>
<s c="0000203 002" n=00062>
Yeah&ITJ;,&PUN; I&PNP;'m&VBB;,&PUN; I&PNP;'ve&VHB; got&VVN;
another&DT0; form&NN1;,&PUN; I&PNP; can&VM0; just&AV0; take&VVI;
note&NN1; of&PRF; any&DT0; other&AJ0; er&UNC; personal&AJ0;
pension&NN1;,&PUN; not&XX0; personal&AJ0; pension&NN1;,&PUN; any&DT0;
erm&UNC; other&AJ0; insurance&NN1; you&PNP;'ve&VHB; got&VVN;,&PUN;
just&AV0; put&VVB; it&PNP; on&AVP-PRP; there&AV0; and&CJC;,&PUN;
and&CJC; that&DT0;'s&VBZ; <unclear>
</u>
</u>
```