BRITISH NATIONAL CORPUS
BASIC GRAMMATICAL TAGSET

From: Geoff Leech
To: Task Group D Members

I am sending herewith a draft BGTS (*basic grammatical tagset*) to be used in the basic tagging of the whole of the BNC.

The tagset (55 tags) is inevitably a compromise between LINGUISTIC DEFENSIBILITY and COMPUTATIONAL TRACTABILITY.

> **LINGUISTIC DEFENSIBILITY** means that there should be a good linguistic motivation for the tag in question.
>
> **COMPUTATIONAL TRACTABILITY:** given the limitations of grammatical tagging technology as at present, and given the need to tag 100 million words in two-and-a-quarter years, we cannot afford to make too many subtle linguistic distinctions. This in particular means that if a tag distinction leads to a high rate of error (requiring a high rate of manual intervention), we just cannot afford to make it.

However, there is one way around the error problem which we would wish to use in moderation. This is the use of so-called PORTMANTEAU tags - tags which are "two grammatical labels wrapped up into one". An example of a portmanteau tag is "VERBD/VERBN" (or perhaps "VERBDN" for short, which would appear in the final tagged version of the corpus, and would mean "The tagger does not have sufficient evidence to decide whether this word is a past tense form or a past participle". However regrettable it may seem that portmanteau tags would have to occur in the output of grammatical tagging, I can foresee no way of avoiding it, without increasing the amount of manual postediting correction beyond what our human resources could manage. There are, in my current thinking, four portmanteau tags of which I would want to make use:

> **VERBD/VERBN** (past tense or past participle)
> **PREP/PART** (preposition or adverbial particle)
> **NOUSG/VERBG** (sing. noun or -ing form of verb)
> **NOUSG/PNOUN** (sing. common noun or proper noun)

At the same time, I would emphasise that our aim would be to reduce portmanteau tagging to a minimum: in general the tagger would feel confident enough (in terms of probability) to choose between one of these tags and the other.

LABELS: In the following list, the labels could obviously be changed. The main motivation for the choice of labels is mnemonic. From the point of view of computational extraction of data from the corpus, there is some advantage in using a less mnemonic system, such as NN1 for singular common noun, where each character represents a linguistic feature (N= noun, N= common, 1= singular). However, the advantages of mnemonicity probably outweigh this. What I have presented is something of a hybrid of these two types of labelling: e.g. VBEZ for "is" retains the linguistic significanc of V (= verb) and -Z (= -s form). An optimal mnemonic label here would be "IS" itself.

RELATION TO TEI: The TEI grammatical tagging guidelines are far too general and sophisticated for our purpose. But in most cases, the 55 tags below could be mapped on to

a TEI-conformant markup. In some cases, (e.g. DET) our label does not correspond to a TEI-recognized category.

BRITISH NATIONAL CORPUS:
PROPOSAL FOR BASIC GRAMMATICAL TAGSET (*Geoffrey Leech 26.8.91*)

ADJ	**adjective (unmarked) (e.g. GOOD, OLD)**
ADJC	**comparative adjective (e.g. BETTER, OLDER)**
ADJS	**superlative adjective (e.g. BEST, OLDEST)**
ADV	**adverb (unmarked) (e.g. OFTEN, WELL)**
ADVC	**comparative adverb (e.g. OFTENER, LONGER)**
ADVQ	**wh-adverb (e.g. WHEN, HOW, WHY)**
ADVS	**superlative adverb (e.g. FURTHEST, LONGEST)**
ALPH	**alphabetical symbol (e.g. A, B, c, d)**
ART	**article (e.g. THE, AN)**
CONJ	**subordinating conjunction (e.g. ALTHOUGH, WHEN)**
COORD	**coordinator (e.g. AND, OR)**
CTHAT	**the conjunction THAT**
DET	**determiner (e.g. THESE, SOME)**
DETQ	**wh-determiner (e.g. WHOSE, WHICH)**
EXIS	**existential THERE**
GEN	**the genitive morpheme 'S or '**
ISOL	**interjection or other isolate (e.g. OH, YES, MHM)**
NEG	**the negative NOT or N'T**
NOUN	**noun (neutral for number) (e.g. AIRCRAFT, DATA)**
NOUPL	**plural noun (e.g. PENCILS, GEESE)**
NOUSG	**singular noun (e.g. PENCIL, GOOSE)**
NUM	**cardinal numeral (e.g. 3, FIFTY-FIVE, 6609) (excluding ONE)**
OF	**the preposition OF**
ONE	**the word ONE (including numeral and non-numeral uses)**
ORD	**ordinal (e.g. SIXTH, 77TH, LAST)**
PART	**adverb particle (e.g. UP, OFF, OUT)**
PERS	**personal pronoun (e.g. YOU, THEM)**
PNOUN	**proper noun (e.g. LONDON, MICHAEL, MARS)**
POSS	**possessive form (e.g. YOUR, THEIRS)**
PREP	**preposition (except for OF) (e.g. FOR, ABOVE, TO)**
PROI	**indefinite pronoun (e.g. NONE, EVERYTHING)**
PROQ	**wh-pronoun (e.g. WHO, WHOEVER)**
REFL	**reflexive pronoun (e.g. ITSELF, OURSELVES)**

TOINF	**infinitive marker (e.g. TO, IN ORDER TO)**
UNCL	**"unclassified" items which are not words of the English lexicon or do not belong to any recognized category. E.g.: formulae, such as XX61, MARKN; foreign words; BOTH when correlative with AND; etc.**
VBEB	**the base forms of the verb "BE", i.e. BE, AM, ARE**
VBED	**past form of the verb "BE", i.e. WAS, WERE**
VBEG	**-ing form of the verb "BE", i.e. BEING**
VBEN	**past participle of the verb "BE", i.e. BEEN**
VBEZ	**-s form of the verb "BE", i.e. IS, 'S**
VDOB	**base form of the verb "DO", i.e. DO**
VDOD	**past form of the verb "DO", i.e. DID**
VDOG	**-ing form of the verb "DO", i.e. DOING**
VDON	**past participle of the verb "DO", i.e. DONE**
VDOZ	**-s form of the verb "DO", i.e. DOES**
VERBB	**base form of lexical verb (e.g. TAKE, LIVE)**
VERBD	**past tense form of lexical verb (e.g. TOOK, LIVED)**
VERBG	**-ing form of lexical verb (e.g. TAKING, LIVING)**
VERBN	**past participle form of lexical verb (e.g. TAKEN, LIVED)**
VERBZ	**-s form of lexical verb (e.g. TAKES, LIVES)**
VHAVB	**base form of the verb "HAVE", i.e. HAVE**
VHAVD	**past tense form of the verb "HAVE", i.e. HAD, 'D**
VHAVG	**-ing form of the verb "HAVE", i.e. HAVING**
VHAVN	**past participle of the verb "HAVE", i.e. HAD**
VHAVZ	**-s form of the verb "HAVE", i.e. HAS, 'S**
VMOD	**modal auxiliary verb (e.g. CAN, COULD, WILL, 'LL)**

[Number of grammatical tags = 55]

Geoffrey Leech
July 31, 2009