BRITISH NATIONAL CORPUS
TGDW08
Revised Proposal for Basic Grammatical Tagset

Geoffrey Leech, 1 April 1992

The following is the final list of tags arrived at after feedback from SALT Club members, as well as members of the BNC collaboration and the Advisory Council and Terry Langendoen (TEI).

For UCREL's purposes, this tagset will be referred to as the "C5 tagset". Each tag consists of three alphanumeric characters; however, the numeric character "0" can be thought of as an "empty character", used simply to fill out the third character where none would otherwise be present. No tag ends with an upper-case "O".

WORD-TAGS

AJ0	adjective (unmarked) (e.g. GOOD, OLD)
AJC	comparative adjective (e.g. BETTER, OLDER)
AJS	superlative adjective (e.g. BEST, OLDEST)
AT0	article (e.g. THE, A, AN)
AV0	adverb (unmarked) (e.g. OFTEN, WELL, LONGER, FURTHEST)
AVP	adverb particle (e.g. UP, OFF, OUT)
AVQ	wh-adverb (e.g. WHEN, HOW, WHY)
CJC	coordinating conjunction (e.g. AND, OR)
CJS	subordinating conjunction (e.g. ALTHOUGH, WHEN)
CJT	the conjunction THAT
CRD	cardinal numeral (e.g. 3, FIFTY-FIVE, 6609)
DPS	possessive determiner form (e.g. YOUR, THEIR)
DT0	general determiner (e.g. THESE, SOME)
DTQ	wh-determiner (e.g. WHOSE, WHICH)
EX0	existential THERE
ITJ	interjection or other isolate (e.g. OH, YES, MHM)
NN0	noun (neutral for number) (e.g. AIRCRAFT, DATA)
NN1	singular noun (e.g. PENCIL, GOOSE)
NN2	plural noun (e.g. PENCILS, GEESE)
NP0	proper noun (e.g. LONDON, MICHAEL, MARS)
ORD	ordinal (e.g. SIXTH, 77TH, LAST)
PNI	indefinite pronoun (e.g. NONE, EVERYTHING)
PNP	personal pronoun (e.g. YOU, THEM, OURS)
PNQ	wh-pronoun (e.g. WHO, WHOEVER)

PNX	reflexive pronoun (e.g. ITSELF, OURSELVES)
POS	the possessive (genitive) morpheme 'S or '
PRF	the preposition OF
PRP	preposition (except for OF) (e.g. FOR, ABOVE, TO)
TO0	infinitive marker (i.e. TO)
UNC	"unclassified" items which are not words of the English lexicon or do not belong to any recognized category. E.g.: formulae, such as "XX61"; foreign words; BOTH when correlative with AND; etc.
VBB	the base forms of the verb "BE", except infinitive, i.e. AM, ARE
VBD	past form of the verb "BE", i.e. WAS, WERE
VBG	-ing form of the verb "BE", i.e. BEING
VBI	infinitive of the verb "BE"
VBN	past participle of the verb "BE", i.e. BEEN
VBZ	-s form of the verb "BE", i.e. IS, 'S
VDB	base form of the verb "DO", except the infinitive
VDD	past form of the verb "DO", i.e. DID
VDG	-ing form of the verb "DO", i.e. DOING
VDI	infinitive of the verb "DO"
VDN	past participle of the verb "DO", i.e. DONE
VDZ	-s form of the verb "DO", i.e. DOES
VHB	base form of the verb "HAVE", except the infinitive
VHD	past tense form of the verb "HAVE", i.e. HAD, 'D
VHG	-ing form of the verb "HAVE", i.e. HAVING
VHI	infinitive of the verb "HAVE"
VHN	past participle of the verb "HAVE", i.e. HAD
VHZ	-s form of the verb "HAVE", i.e. HAS, 'S
VM0	modal auxiliary verb (e.g. CAN, COULD, WILL, 'LL)
VVB	base form of lexical verb, except the infinitive (e.g. TAKE, LIVE)
VVD	past tense form of lexical verb (e.g. TOOK, LIVED)
VVG	-ing form of lexical verb (e.g. TAKING, LIVING)
VVI	infinitive of lexical verb
VVN	past participle form of lexical verb (e.g. TAKEN, LIVED)
VVZ	-s form of lexical verb (e.g. TAKES, LIVES)
XX0	the negative NOT or N'T
ZZ0	alphabetical symbol (e.g. A, B, c, d)

[Number of grammatical word-tags = 57]

The following is the list of “PORTMANTEAU TAGS” to be declared. We are unlikely to want to use all members of this list, but perhaps it is better to declare a set which is too large, rather than one that is too small.

I rather like the idea of making significant the order of basic tags in portmanteau tags, so that (for example) VVD-VVN is understood to be a prediction that VVD is the more likely tag, where VVN-VVD is a prediction that VVN is the more likely tag. We may not end up making use of this distinction, but again, to avoid incompleteness, I am including in the declaration both orderings for each portmanteau combination.

PORTMANTEAU TAGS

AJ0-AV0	**AV0-AJ0**	adjective or adverb
AJ0-NN1	**NN1-AJ0**	adjective or singular common noun
AJ0-VVD	**VVD-AJ0**	adjective or past tense verb
AJ0-VVG	**VVG-AJ0**	adjective or -ing form of the verb
AJ0-VVN	**VVN-AJ0**	adjective or past participle
AVP-PRP	**PRP-AVP**	adverb particle or preposition
CJS-PRP	**PRP-CJS**	subordinating conjunction or preposition
CRD-PNI	**PNI-CRD**	the word “one” when ambiguous
NN1-NP0	**NP0-NN1**	singular common noun or proper noun
NN1-VVG	**VVG-NN1**	singular common noun or -ing form of the verb
VVD-VVN	**VVN-VVD**	past tense verb or past participle

[Total number of permissible “portmanteau tags” = 22]

The CLAWS tagset includes tags for labelling punctuation marks as grammatically significant. For the purposes of CLAWS, a punctuation mark is treated as equivalent to a word. The following 4 tags are proposed for the BNC:

PUNCTUATION TAGS

PUL	left bracket (i.e. (or [ )
PUN	punctuation mark - normal (i.e. . ! , : ; - ? . . .)
PUQ	quotation mark (i.e. ‘’ “ ”)
PUR	right bracket (i.e.) or])

The tag labels will be declared as SGML entities and the references appended to the relevant “words” in BNC texts. Where the grammatical tag applies to a cluster of words (e.g. “by means of”) the reference will be appended to the last word in the cluster.