Manual of Information for the Lancaster Parsed Corpus

Roger Garside, Geoffrey Leech and Tamás Váradi

PART I INTRODUCTORY INFORMATION ON THE LANCASTER PARSED CORPUS

1. The Lancaster Parsed Corpus is a corpus of British English sentences excerpted from printed publications of the year 1961. The Parsed Corpus is a subset of the Lancaster-Oslo/Bergen Corpus (= LOB Corpus) (see S. Johansson, G. N. Leech and H. Goodluck, *Manual of Information to accompany the Lancaster-Oslo/Bergen Corpus of British English*, Department of English, University of Oslo, 1978). Each sentence in the Parsed Corpus has undergone a syntactic analysis in the form of a phrase marker, or labelled bracketing, following the scheme of analysis outlined in Part III below.[1]

TABLE I: CONTENTS OF LANCASTER PARSED CORPUS

Text categories		No. of sentences	No. of words
A	Press: reportage	728	8188
B	Press: editorial	781	8832
C	Press: reviews	594	7145
D	Religion	721	8803
E	Skills, trades and hobbies	563	8347
F	Popular lore	584	7526
G	Belles lettres, biography, essays	406	5987
H	Miscellaneous (government documents, etc)	441	5798
J	Learned & scientific writings	450	7537
K	General fiction	1119	11279
L	Mystery & detective fiction	1188	13010
M	Science fiction	769	7525
N	Adventure & western fiction	1344	12919
P	Romance & love story	1369	13826
R	Humour	720	8018
TOTAL:		11827	134740

The categories A-R above are those of the original LOB Corpus, which in turn were based on those of the Brown Corpus of American written English, for which the LOB Corpus was a British matching equivalent.[2]

3. The word-tagged LOB Corpus, which has provided the input to the Parsed Corpus, consists of 1,013,737 tagged running words, divided into 500 samples of approximately 2,000 words each.[3] The Parsed Corpus is a sub-corpus of it, consisting of 13.29% of the tagged LOB Corpus.

4. The Parsed Corpus contains sentences from the first 10 text samples of each Text Category of the LOB Corpus, except for Categories M and R, which contained only 6 and 9 text samples (respectively) in the original corpus. Thus, roughly speaking, the Parsed Corpus represents, in smaller quantities, the range of text samples provided in the original LOB Corpus. However, there are qualifications to be made to this statement, as indicated in the paragraphs below.

5. Since the Parsed Corpus samples are restricted to the first ten (or less) samples in each Category, they are in fact more limited in genre/domain than the original corpus. These limitations are indicated in the descriptions in Table II below, which also lists the number of text samples per category both in the Parsed Corpus, and in the LOB Corpus:

TABLE II: TYPES AND NUMBERS OF TEXT SAMPLES IN THE PARSED CORPUS

Text Categories in Parsed Corpus		Number of text samples in:	
(More precisely defined than in Table I)		Parsed Corpus	(LOB)
A	National daily papers (reportage)	10	(44)
B	National daily papers (editorial)	10	(27)
C	National daily papers (reviews)	10	(17)
D	Books on religion	10	(17)
E	Homecraft, handiman, and hobbies publications	10	(38)
F	Popular politics, psychology, sociology	10	(44)
G	Biography, memoirs	10	(77)
H	Government documents	10	(30)
J	Writing on natural sciences	10	(80)
K	Novels (general fiction)	10	(29)
L	Novels (mystery and detective)	10	(24)
M	Novels and short stories (science fiction)	6	(6)
N	Novels (adventure and western)	10	(29)
P	Novels (romance and love story)	10	(29)
R	Novels and articles from periodicals (humour)	9	(9)
	TOTAL number of text samples:	145	(500)

6. As will be evident from Tables I and II, the Parsed Corpus does not contain all the sentences of the c.2000-word text extracts in the LOB Corpus which were used as input. In fact, the sentences included in the Parsed Corpus are considerably shorter, on the average, than the average sentence length for the LOB Corpus as a whole (the average for Parsed Corpus is 11.39 words per sentence, as compared with an average of c.19 words per sentence for the entire LOB Corpus). On the whole, then, longer sentences have been excluded from the Parsed Corpus, with the result that the Parsed Corpus no longer contains LOB text extracts in their entirety.

7. The reason for the omission of longer sentences is as follows. Our original hope was that we would be able to parse the whole of the LOB Corpus automatically. The prototype probabilistic parser developed for this purpose is described in R. Garside, G. Leech and G. Sampson (eds), *The Computational Analysis of English: a Corpus-based Approach*, London: Longman, 1987, Chapter 6. In practice, this goal proved too ambitious, and we ran the parser over only the 145 text extracts as listed in Table II. Even then, the parser was unable to achieve a parse of most sentences over 20-25 words in length. These unparsed sentences were therefore omitted from the Parsed Corpus. For the remaining sentences, which had been parsed, we accepted the "winning parse" (that to which the parser had assigned the highest probability) as the input to a manual post-editing stage. During manual post-editing, many sentences which the parser had incorrectly parsed (according to the procedure mentioned above) were corrected by hand, and a number of final checks were made to ensure that the Parsed Corpus is reasonably error-free. Therefore the Lancaster Parsed Corpus can be regarded as a treebank broadly representative of the syntax of written (printed) English across a great variety of styles and text types. It may hopefully be used, for example, as a testbed for wide-coverage general-purpose grammars and parsers of English, as well as for quantitative linguistic studies of English syntax.

8. The fact that long sentences are under-represented in the corpus does, of course, limit the corpus's value as a testbed. In coverage of grammatical rules and structures, however, it will probably be more adequate than any other treebank publicly available to date.

9. In another respect, too, the corpus is unrepresentative: it cannot be regarded as a balanced sub-sample of the LOB Corpus. Hence any quantitative comparison between features observed in the parsed corpus and other features independently observed in the LOB Corpus are likely to be unreliable. The reason for this is (as can be gleaned from Tables I and II) that the Text Categories are not represented in the Parsed Corpus according to their proportion of the full LOB Corpus. For example (to take the most extreme cases), in Category G, the Parsed Corpus contains only 3.84% of the original LOB Corpus material, whereas in Category M, it contains 62.13%. Disparities such as this are due to (a) the greater average length of sentences in some categories than in others, and (b) variation across Text Categories in the number of text extracts in the LOB Corpus and in the number of text samples in the Parsed Corpus.

10. In practice, the consequences of the factors just mentioned are that the Parsed Corpus contains a higher proportion of material from imaginative (largely fictional) texts (Categories K-R), whereas the LOB Corpus as a whole has a higher proportion of expository text material (Categories A-J). The figures are given in Table III:

TABLE III

	Expository Categs. (A-J)		Imaginative Categs. (K-R)	
LOB	Words:	756,293	Words:	257,444
Corpus	%age of Corpus:	74.60	%age of Corpus:	25.39
Parsed	Words:	68,163	Words:	66,577
Corpus	%age of Corpus:	50.59	%age of Corpus:	49.41

Acknowledgements

We are grateful for help received from the following sources:

(a) The development of the Parsed Corpus was originally undertaken in 1983-6, with the support of Research Grant GR/C/47700 funded by the Science and Engineering Research Council (SERC).

(b) The automatic probabilistic parser which produced the parses (prior to post-editing) derived its frequency data from another parsed database of sentences from the LOB Corpus, known as the Lancaster-Leeds Treebank, manually parsed by Geoffrey Sampson (see R. Garside, G. Leech and G. Sampson, *The Computational Analysis of English: a Corpus-based Approach*, London: Longman, 1987, Chapter 7). The Lancaster Parsed Corpus implements a simplified version of the parsing scheme more fully instantiated in Sampson's treebank.

(c) The post-editing of the corpus was undertaken by a number of research students at Lancaster. We particularly acknowledge the major post-editing work undertaken by Heather Kempson and by Srikant Sarangi. Finally, the whole corpus was thoroughly checked and corrected by Tamás Váradi.

(d) Steve Fligelstone and Andrew Wilson gave invaluable help in the final stages of checking and producing the corpus.

(e) A number of errors were reported by Qiao Hong Liang of the University of Queensland. Corrections were made to the corpus in April 1995.

Notes to Part I

1. Some special symbols which occurred in the original untagged LOB Corpus remain in the parsed version. Their significance can be checked in Johansson et al. (1978). On the other hand, in the course of tagging, some symbols (especially some capital letters) were changed, and these changes remain in the parsed version.

2. For details of the composition of the Brown Corpus, see W.N. Francis and H. Ku_era, *Manual of Information to accompany a Standard Corpus of Present-day Edited American English, for use with Digital Computers*, Providence, RI: Brown University Linguistics Department, 1964 (revised 1971 and 1979).

3. Information on the Tagged LOB Corpus is provided by: S. Johansson (in collaboration with E. Atwell, R. Garside, and G. Leech), *The Tagged LOB Corpus Users' Manual*, Bergen: Norwegian Computing Centre for the Humanities, 1986. The Manual, in particular, provides details of the precise interpretation and implementation of word-tags in the LOB Corpus. (For a list of the word-tags, see Part II below.)

4. The notion of "correct parse" is, of course, not unproblematic, as on occasions true ambiguities arise: that is, there may be uncertainty about which parse best corresponds to the most likely interpretation of a sentence in its context. However, in general, this problem has been found to arise only occasionally. A problem more likely to trouble the user, despite safeguards taken by the compilers, is an inconsistency in parsing practices, in cases where it could be argued that two different bracketings are linguistically valid representations of the same form-meaning association. And finally it must be acknowledged that errors – by any definition – remain in the corpus in spite of careful and repeated post-editing. However regrettable this is, the reasons for it will be evident to anyone who has undertaken a comparable task.

PART II LIST OF GRAMMATICAL WORD-TAGS IN THE LANCASTER PARSED CORPUS

ABL	pre-qualifier in a noun phrase (QUITE, RATHER, SUCH)
ABN	pre-quantifier in a noun phrase (ALL, HALF)
ABX	pre-quantifier / double conjunction (BOTH)
AP	post-determiner (FEW, FEWER, FORMER, LAST, LATTER, LEAST, LESS, LITTLE, MANY, MORE, MOST, MUCH, NEXT, ONLY, OTHER, OWN, SAME, SEVERAL, VERY)
AP$	OTHER'S
APS	OTHERS
APS$	OTHERS'
AT	singular article (A, AN, EVERY)
ATI	singular or plural article (THE, NO)
BE	BE
BED	WERE
BEDZ	WAS
BEG	BEING
BEM	AM
BEN	BEEN
BER	ARE, 'RE
BEZ	IS, 'S
CC	coordinating conjunction (AND, AND/OR, BUT, NOR, ONLY, OR, YET)
CD	cardinal number (2, 3, etc; TWO, THREE, etc; HUNDRED, THOUSAND, etc; DOZEN, ZERO)
CS$	cardinal number + genitive
CD-CD	hyphenated pair of cardinal numbers (e.g. 1988-90)
CD1	ONE
CD1$	ONE'S
CD1S	ONES
CDS	cardinal number + plural (TENS, MILLIONS, DOZENS, etc)
CS	subordinating conjunction (AFTER, ALTHOUGH, etc)
DO	DO
DOD	DID
DOZ	DOES
DT	singular determiner (ANOTHER, EACH, THAT, THIS)
DT$	singular determiner + genitive (ANOTHER'S)
DTI	determiner neutral for number (ANY, ENOUGH, SOME)
DTS	plural determiner (THESE, THOSE)
DTX	determiner / double conjunction (EITHER, NEITHER)
EX	existential THERE
HV	HAVE
HVD	HAD, 'D (past tense)
HVG	HAVING
HVN	HAD (past participle)
HVZ	HAS, 'S
IN	preposition (general)

INF	FOR as preposition
INO	OF as preposition
INW	WITH as preposition
JJ	adjective (general)
JJB	attributive adjective
JNP	adjective with word-initial cap; e.g. WELSH, KEYNESIAN
JJR	comparative adjective
JJT	superlative adjective
MD	modal auxiliary
NC	cited word as singular noun (e.g. "LED is a verb")
NN	singular common noun
NNP	singular common noun; word-initial cap; e.g. LONDONER
NNPS	plural common noun; word-initial cap; e.g. LONDONERS
NNPS$	plur. common noun; word-init. cap; genitive LONDONERS'
NNP$	sing. common noun; word-init.cap; gen; e.g. LONDONER'S
NNS	plural common noun
NNS$	plural common noun + genitive
NNU	singular unit of measurement (e.g. IN. KG.)
NNUS	plural unit of measurement (e.g. INS. KGS.)
NNUS$	plural unit of measurement + genitive
NP	singular proper noun
NPS	plural proper noun
NPS$	plural proper noun + genitive
NP$	singular proper noun + genitive
NPL	singular locative noun; word-initial cap.; e.g. ISLAND
NPLS	plural locative noun; word-initial cap.; e.g. ISLANDS
NPLS$	plural loc. noun; word-init. cap; + gen.; e.g. ISLANDS'
NPL$	sing.loc. noun; word-init. cap; + gen.; e.g. ISLAND'S
NPT	singular titular noun; word-initial cap.; e.g. DR.
NPTS	plural titular noun; word-initial cap.; e.g. MESSRS.
NPTS$	plur. tit. noun; word-init. cap.; + gen.; e.g. QUEENS'
NR	singular adverbial noun (JANUARY, FEBRUARY, etc; SUNDAY, MONDAY, etc; EAST, WEST, etc; TODAY, TOMORROW, TONIGHT; DOWNTOWN, HOME)
NR$	singular adverbial noun + genitive
NRS	plural adverbial noun
OD	ordinal number (1ST, 2ND, etc; FIRST, SECOND, etc)
PN	nominal pronoun (ANYBODY, ANYONE, ANYTHING; EVERYBODY, EVERYONE, EVERYTHING; NOBODY, NONE, NOTHING, NO ONE; SOMEBODY, SOMEONE, SOMETHING; SO)
PN$	nominal pronoun + genitive
PP$	possessive determiner (MY, YOUR, etc)
PPS$	possessive pronoun (MINE, YOURS, etc)
PP1A	personal pronoun, 1st pers sing nom (I)
PP1AS	personal pronoun, 1st pers plur nom (WE)
PP1O	personal pronoun, 1st pers sing acc (ME)
PP1OS	personal pronoun, 1st pers plur acc (US, 'S)

PP2	personal pronoun, 2nd pers (YOU, THOU, THEE, YE)
PP3	personal pronoun, 3rd pers sing nom+acc (IT)
PP3A	personal pronoun, 3rd pers sing nom (HE, SHE)
PP3AS	personal pronoun, 3rd pers plur nom (THEY)
PP3O	personal pronoun, 3rd pers plur acc (HIM, HER)
PP3OS	personal pronoun, 3rd pers plur acc (THEM, 'EM)
PPL	singular reflexive pronoun
PPLS	plural reflexive pronoun
QL	qualifier (AS, AWFULLY, LESS, MORE, SO, TOO, VERY, etc)
QLP	post-qualifier (ENOUGH, INDEED)
RB	adverb (general)
RB$	adverb + genitive (ELSE'S)
RBR	comparative adverb
RBT	superlative adverb
RI	adverb (homograph of preposition: BELOW, NEAR, etc)
RN	nominal adverb (HERE, NOW, THERE, THEN, etc)
RP	adverbial particle (BACK, DOWN, OFF, etc)
TO	infinitival TO
UH	interjection
VB	base form of lexical verb (uninflected present tense, imperative, infinitive, subjunctive)
VBD	past tense of lexical verb
VBG	present participle or gerund of lexical verb
VBN	past participle of lexical verb
VBZ	3rd person singular of verb
WDT	WH-determiner (WHAT, WHATEVER, WHATSOEVER, WHICH, WHICHEVER, WHICHSOEVER)
WP	WH-pronoun, nom+acc (WHO, WHOEVER, THAT)
WP$	WH-pronoun, genitive (WHOSE)
WPA	WH-pronoun, nom (WHOSOEVER)
WPO	WH-pronoun, acc (WHOM, WHOMSOEVER)
WRB	WH-adverb (HOW, WHEN, WHERE, etc)
XNOT	NOT, N'T
ZZ	letter of the alphabet (E, X, etc).
!	exclamation mark (!)
&FO	formula
&FW	foreign word
(	left bracket ([
)	right bracket)]
*'	begin quote: *' *"
**'	end quote: **' **"
*-	dash
,	comma (,)
.	full stop (.)
...	ellipsis (...)
:	colon (:)
;	semicolon (;)
?	question mark (?)

PART III INFORMATION ON CONSTITUENT TAGS AND SYNTACTIC NOTATION

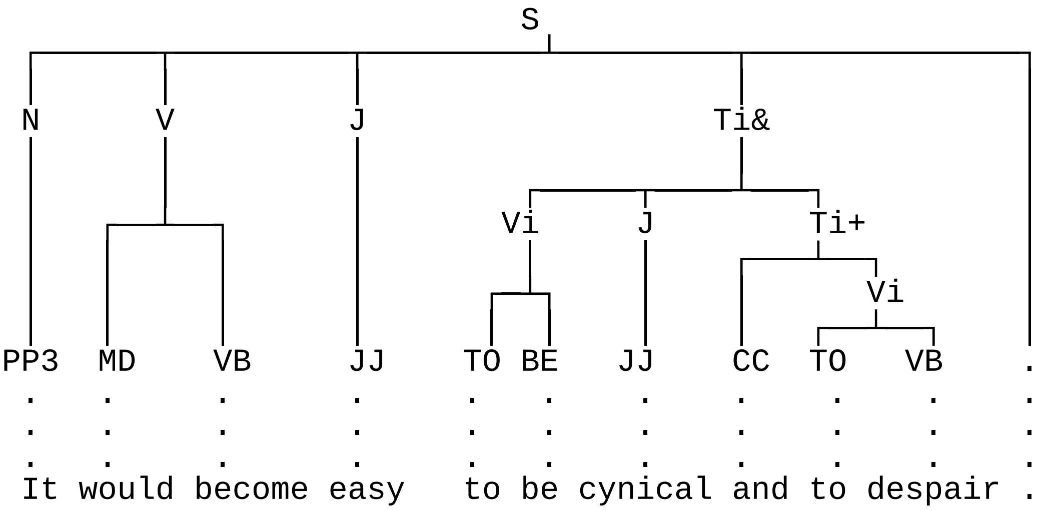
1. The General Format of the Files

The text is contained in plain ASCII text files. Each sentence is in a separate paragraph and is followed by an empty line. At the head of each paragraph is a line that serves to identify the sentence that follows. It contains an alphanumeric code standing for the genre and text sample in the LOB Corpus that the sentence belongs to. Next to this, is the ordinal number of the sentence in the present corpus. Numbering is restarted at the beginning of each genre. To ensure readability of text on 80 column wide computer screens, text lines have been formatted so that they do not reach over 72 characters. Below is a sample sentence:

```
B06 666
[S[N it_PP3 N][V would_MD become_VB V][J easy_JJ J][Ti&[Vi
to_TO be_BE Vi][J cynical_JJ J][Ti+ and_CC [Vi to_TO
despair_VB Vi]T+]Ti&] ._. S]
```

Each word in the text is followed by an underscore character and a sequence of symbols (normally capital letters) which represents a wordtag, i.e. a label giving the grammatical class of a word. (See Part II for a full list.) For example, in the sample text above PP3 means `third person personal pronoun' and MD means `modal auxiliary'. For the present purposes punctuation marks are treated as words and have their own tags.

The syntactic structure of the sentence is laid out in the form of labelled bracketing. The type of grammar used in the parsing was a Phrase Structure Grammar, in which the structure of a sentence can be shown as a tree, with `S' (= `Sentence') as the root of the tree and the wordtags `PP3', `MD', etc. as the leaves of the tree. Here, for instance, is the sentence above represented as a tree:



Alternatively, the same phrase structure (omitting the wordtags at the bottom) can be represented as a labelled bracketing as follows:

```
[S [N] [V] [J] [Ti& [Vi] [J] [Ti+ [Vi] Ti+] Ti&] S]
```

2. Constituents and Constituent Tags

The interpretation of the labelled bracketing on the output is as follows:

`[' represents the opening of a constituent, and `]' represents the closing or completion of a constituent. The symbols alongside these brackets are labels for constituents. For example, `[V' means `a verb phrase opens here' and `V]' means `a verb phrase closes here'. The labels are called constituent tags, as distinct from the wordtags, which label words and are attached to them directly. Constituent tags label higher constituents such as sentences, clauses and phrases, and are attached to the corresponding brackets.

For example, the following are the meanings of the constituent tags in our sample sentence:

S `independent sentence' (including a direct quotation or an interpolation within an including sentence)
N `noun phrase'
V `verb phrase' (in the narrower sense which excludes objects, complements, etc. following the main verb)
J `adjective phrase'
Ti& `compound infinitive clause' (i.e an infinitive clause containing two or more coordinated infinitives)
Ti+ `coordinated infinitive clause' (i.e. an infinitive clause which is the second or subsequent conjoined part of a compound infinitive clause)
Vi`infinitive verb phrase'

You will notice that these constituent tags sometimes consist of just one capital letter, and at other times there are additional symbols, such as a lower-case letter, an `&' or a `+'.

The capital letters indicate the major class of the constituent that the tag labels. Where they occur, additional lower-case letters indicate a subclassification. A `&' represents a compound constituent (e.g. `N&' indicates a compound noun phrase, in which two or more noun-phrase-like structures are joined by a coordinating conjunction). A `+' represents a coordinated constituent, e.g. `N+' indicates a coordinated noun phrase.

The grammar on which this parsing scheme is based is somewhat simplified. It is simplified (i) because very few subclassifications are included (`Vi' being an exception), and (ii) because the constituent tags indicate *formal* categories only. They do not represent *functional* concepts such as `subject', `object', `complement' and `adverbial'.

3. Coordination

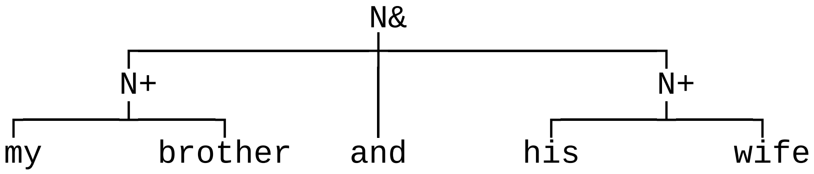
It is important to note that this parsing scheme treats coordination in a rather unorthodox way (which is, nevertheless, convenient for automatic parsing). Suppose we take a typical compound noun phrase such as:

`my brother and his wife'

Most phrase structure grammars would represent this construction roughly in the following way:

[N& [N+ my brother N+] and [N+ his wife N+] N&]

The construction is symmetrical and may be shown in the form of a tree diagram as follows:

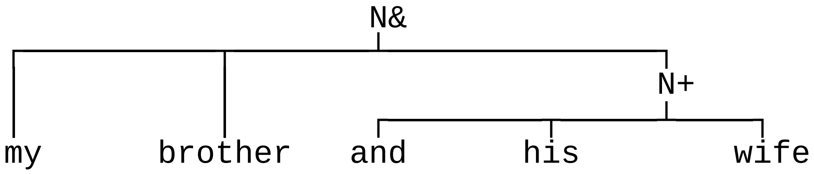


However, the present parsing scheme treats the second coordinated noun phrase as subordinate in relation to the first, as follows:

[N& my brother [N+ and his wife N+] N&]

Notice that the `and' is treated as part of the coordinated noun phrase that it introduces,

As a tree-diagram, this looks like this:



If the compound noun phrase contains three (or more) noun-phrase-like structures, the second and subsequent ones are treated as subordinated, as in:

`my brother and my sister-in-law and their children'

This is analyzed:
[N& my brother [N+ and my sister-in-law N+][N+ and their children N+] N&]

It is common, however, for the first `and' to be omitted in such structures, and in that case, the constituent tag `N-' is used for a coordinated noun phrase *not* introduced by `and':
[N& my brother [N- my sister-in-law N-][N+ and their children N+] N&]

The same convention is used in constructions with parataxis, i.e. constructions where the `and' is omitted altogether:
[J& very tasty, [J- very sweet J-] J&]

One more oddity about coordination is that if a sentence *begins* with a coordinating conjunction, the whole sentence is tagged `S' not `S+'. E.g.:
[S but in his letter 12 days later he retracted. S]

4. Details of Constituent tags

Sentence tags

Sq and **Si**

`Sq' means `a piece of direct quotation' – normally an independent piece of language which occurs in fictional dialogue enclosed in quote marks. `Si' means `an interpolated sentence' – i.e. a grammatically independent piece of language which is inserted (normally enclosed in brackets) in another sentence, but is not grammatically part of it. Note the following conventions used in handling direct quotations:

Pattern A:
"Nothing will change my mind", said Pat.
Pattern B:
Pat said, "Nothing will change my mind".
In these cases, the direct speech is analyzed as [Sq]:
Pattern A: [S "[Sq]" , [V] [N] S]
Pattern B: [S [N] [V] , "[Sq]" S]

Pattern C:
"Nothing," said Pat, "will change my mind".
In this case, Sq isn't used. Instead, the reporting clause is treated as an Si:
Pattern C: [S "..." , [Si] , "..." S]

Here is a further example of the use of Si:
That year ([Si how well I remember it! Si]) saw the beginning of my acting career.

S&, **S+** and **S-**

`S&' represents a compound sentence, `S+' represents the second or subsequent conjoin of a compound sentence, if it begins with a coordinating conjunction, and `S-' represents such a conjoin when it does not begin with a coordinating conjunction. (See the discussion on coordination in 3. above).

Finite Clause Tags

F

A finite subordinate clause – i.e. a clause which contains a finite verb, and which is grammatically included in a sentence, is symbolized `F'. Typically, the `F' is followed by another symbol as detailed below.

Fa

`Fa' is a finite adverbial clause (e.g. a finite subordinate clause of time, of condition, of reason etc.)

E.g.: `[Fa Now that I have found out Fa] it may be easier for me to say it.'

Fc
\`Fc' is a comparative clause, normally beginning with \`than' or \`as'.

E.g.: \`He is cleverer [Fc than I thought Fc].'

Fn
\`Fn' is a finite nominal clause – i.e. a finite subordinate clause which functions in the position of a noun phrase. Examples of \`Fn' are *that*-clauses and *wh*-clauses (including indirect statements and indirect questions, also including \`zero *that*-clauses', where the *that* is omitted at the beginnning of the clause),

E.g.: \`I know [Fn that you saw them Fn].'

Fr
\`Fr' is a relative clause – whether restrictive or non-restrictive,

E.g.: \`the house [Fr in which I was born Fr]'

N.B. a \`fused' or \`nominal relative clause' as in \`I will do [what you want]' is treated as \`Fn'.

F&, F+, F-, etc.
These tags, which will also occur in combination with the letters \`a', \`n', \`r' etc., are used for coordinated finite subordinate clauses in accordance with the rules detailed in 3.

Nonfinite and verbless clause tags

T
Nonfinite clauses are indicated by \`T'. However, \`T' does not normally occur alone. It is combined with the subscripts below.

Ti
\`Ti' stands for a *to*-infinitive clause (e.g. an infinitive construction in which *to*+infinitive may or may not be followed by an object, a complement and/or adverbials).

E.g.: \`It was a pity [Ti to leave them behind Ti].'

Tg
\`Tg' stands for an *-ing* clause (i.e. a participial or gerundival construction in which the *-ing* form of the verb may or may not be followed by an object, a complement, and/or adverbials).

E.g.: \`... where he first saw light machine guns [Tg being assembled Tg].'

Tn
\`Tn' stands for a past participle clause (i.e. a construction in which the past participle form of the verb may or may not be followed by an object, a complement and/or abverbials).

E.g.: \`[Tn Disappointed by the outcome Tn], John proceeded ...'

Tb

`Tb' stands for a `bare infinitive clause' (i.e. a construction in which the `bare infinitive' — infinitive without *to* — may or may not be followed by an object, a complement and/or abverbials).

E.g.: `We saw her [Tb cross the street hurriedly Tb].'

Tf

`Tf' is used as a variant of the infinitive clause, where the subject of the infinitive is introduced by `for'.

E.g.: `That would be a lot [Tf for them to swallow Tf].'

N.B. Nonfinite clauses generally have no subject: but it is also possible for a subject to occur;

E.g.: `I never yet heard of [Tg a young lady dying of love Tg]'.

W

`W' stands for a nonfinite or verbless clause introduced by *with*.

E.g.: `... another job [W with vastly more to offer W].'
`[W With René dying so unexpectedly W], we don't know which way to turn.'
`He sauntered in [W with his hands in his pockets W].'

L

`L' stands for a verbless clause not introduced by *with* or by a subordinating conjunction.

E.g.: `[L Afraid of the consequences L], he hid the gun in a cupboard.'
`[L The Luger ready L], he walked simply back.'

NOTE: If an adverbial verbless clause or nonfinite clause is introduced by a subordinating conjunction (e.g. *if, when*), it is treated as a `Fa':

E.g.: `The liner [Fa when finished Fa] will be the largest passenger vessel built in Europe since the war.'
`[Fa If in doubt Fa], leave the decision to your superior.'

If an adverbial verbless clause or nonfinite clause is introduced by a *wh*- word *why, what, how*, it is treated as a `Fn':

E.g.: `We didn't know [Fn what to do Fn].'
`They are leaving the village. Nobody knows [Fn why Fn].'

Constituent Tags for Major Phrase Types

V

`V' means `finite Verb Phrase', in the narrow sense, in which `verb phrase' excludes objects, complements, etc. Thus `V' may include simple verb phrases such as *is, have, did* and also more complicated ones with modals, progressive aspect, perfect aspect or passive.

Vo and **Vr**
In general, no subscript is used with `V'. However, `Vo' and `Vr' are exceptions. They are used when a verb phrase is split into two parts by subject-auxiliary inversion. The first part is labelled `Vo' (o = `operator') and the second part is labelled `Vr' (r = `remainder'). E.g. in `Have you seen Mary?' `have' is `Vo' and `seen' is `Vr'.

Note that `V' includes the negative word `not' as well as adverbs. E.g. the whole of `have not seen' (or `haven't seen' or `have recently seen') is a `V'. But if the subject noun phrase occurs between the auxiliary and the main verb, this is treated as a separate noun phrase. Accordingly, `have you seen' consists of `Vo' followed by `N' followed by `Vr'.

Vi, **Vg**, **Vn**
These are labels for nonfinite verb phrases, i.e. verb phrases which are the verb phrases of nonfinite clauses `Ti', `Tg' or `Tn'.

Vi
means `*to*-infinitive verb phrase', e.g. `to eat' or `to have eaten'.
Vg
means `-*ing* participle verb phrase', e.g. `eating' or `having eaten'.
Vn
means `past participle verb phrase', e.g. `eaten'.

N
N is the label for a noun phrase, whether it is a single word (such as the pronoun *it*) or a sequence of words.

Na
In general, `N' has no subscripts. One major exception is `Na', which stands for a noun phrase marked as subject of the verb. In practice, `Na' almost always indicated one of the pronouns `I', `she', `he', `we', `they'. (N.B. `you' and `it' as subject are not marked `Na' because their status of subject is not unambiguously shown by their form.)

Nq
Another exceptional use of `N' + subscript, meaning a *wh-* noun phrase, such as `who', `which', `which car', `what time' etc.

J
`J' means an adjective phrase such as `happy', `very tall', `too happy for words', etc. If an adjective occurs as the head of a noun phrase, e.g. `the wealthy', `the unemployed', the phrase is marked `N' not `J'.

Jq
Here, as with `Nq', the `q' means `a phrase beginning with a *wh*-word', e.g. `Jq', an adjective phrase beginning with a *wh*-word, is in practice a phrase such as `How old'.

P
`P' stands for `prepositional phrase', e.g. `in London' or `on arriving at the station', `with it', `for what we are about to receive', i.e. a preposition followed by its complement or completive element. Prepositional phrases also sometimes contain adverbs like `just' in `just inside the door'.

Pq
stands for `prepositional phrase with a *wh*-word, e.g. `on whose behalf', `in which case', `for whom'.

Po
stands for a `prepositional phrase beginning with the preposition *of*'.

R
`R' is the symbol for an adverb phrase, which may be a single word such as `there' or `quickly' or may be a sequence such as `quite often', `too fast', `further than I expected', etc.

Rq
stands for an adverb phrase beginning with a *wh*-word. This would include such phrases as `how' in `How do you feel?', or `how long' in `How long have you been waiting?'

Constituent Tags for Minor Phrase Types

M
`M' stands for a `numeric phrase' when such an expression is part of a noun phrase. Examples are `five thousand' in `five thousand young people'; `another hundred' in `another hundred calories'. Numeric phrases have a numerical word at their head (e.g. `hundred'), and consist of at least two words. (N.B. if numerical expressions such as `five thousand' occur on their own as noun phrases, they are labelled `N'.)

D
`D' stands for a `determiner phrase', i.e. a phrase consisting of at least two words, in which the determiner is a head, and which is part of a noun phrase. E.g. `too many' in `too many people'; `a good few' in `a good few people'. (N.B. `too many' or `a good few' on their own, acting as a noun phrase, are labelled `N'.)

Dq
stands for a determiner phrase (as defined above) beginning with a *wh*-word. E.g. `how many' and `how much', when they are part of a noun phrase, as in `How many apples (did you buy)?'

G
`G' stands for `genitive phrase' i.e. a phrase which consists of two or more words acting as the genitive in a noun phrase. E.g.`*the earth's*' in `the earth's rotation around the sun'; `*my mother's*' in `my mother's greatest wish'; `*last Friday's*' in `last Friday's Evening Standard'; `*someone else's*' in `someone else's bedroom'; `*the Vicar of Bray's*' in `the Vicar of Bray's famous dictum'.

X
`X' is the negative word `not' when acting as an independent element of clause structure; e.g. in `He told us [Fn what not to do Fn]', *not* follows the object of the subordinate clause *what* and precedes the verb phrase *to do*. Thus, the clause *what not to do* has the three constituents `Nq', `X', and `Vi'. Generally, `not' is part of the verb phrase (see under `V' above) and therefore does not require an `X'.

E
`E' is the label used for existential `there', i.e. the unstressed `*there*' in the `there is/are` construction. E.g. `[E There E] is nothing wrong'.

U
`U' is the tag used for an exclamatory word such as `Oh' or a grammatical isolate such as `yes' or `no'.

Constituent tags with coordination

As already mentioned in 3. above, `&', `+', and `-' are suffixes used to mark constituents in coordinated constructions. They may be added to any of the constituent tags. Because of their uniform function, the above list does not contain these composite tags. Merely as an illustrative example, this is how `Tg&' and `Tg+' can be used as follows:

> `Would you mind [Tg& coming inside [Tg+ and shutting the door Tg+]Tg&]?'

Coordinating conjuncts of different types

A problem of labelling arises when the two or more elements coordinated do not belong to the same class. In such cases, we resort to the device of a `slashtag', i.e. a composite tag in which the classes of the items concerned are separated by a slash (/). E.g.:

> `[R/P& mechanically [P+ or by hand P+]R/P&]'

Notice that the slashtag is only used for the first constituent tag – the one ending in `&'. The other tags (`P+' etc.) just label themselves, whereas the first one labels the whole construction. In this composite slashtag label, the first label is the one which the first constituent would normally have (e.g. in the above example, `mechanically' would normally be an adverb phrase `R', so `R' is the first symbol of the slashtag).

No slashtag is used when the first constituent tag and the second or subsequent tags all share the same capital letter. For example, there is no need to use `N/Na&' in the parsing of `my wife and I'

> [N& my wife [Na+ and I Na+]N&]

Slashtags can get very complicated when there is a sequence of three or more coordinated elements. Here is one complex, but very rare, example:

> `Elderly people who are [J/Tg/Tn& bedridden [Tg- living alone Tg-], [Tn+ or abandoned by their families Tn+] J/Tg/Tn&]

Word level coordination

If coordination takes place on the word level only, (e.g. where two or more words inside the same constituent are linked by `and'), the coordination is indicated in the same way as described so far except for the fact that in such cases the relevant wordtags are attached to the brackets. E.g.

> `objections of [JJ& learned [JJ+ and skilful JJ+]JJ&] anatomists'

The reasoning behind this arrangement is that the set of coordinated words behave, with regard to the rest of the sentence, like a single word.

Word level coordinations of words of different classes are treated by slashtags in the same way as described in the previous section. E.g.

> `[NN/JJ/NN& health, [JJ- marital JJ-][NN+ and property NN+]NN/JJ/NN&] problems'

Note, however, that if a word is linked to a phrase of two or more words, then the word counts as a phrase for the purposes of coordination. In other words, if any of the coordinated elements is a phrase, the whole string can no longer be a word level coordination. E.g.

> `[N& John [N+ and his wife N+]N&]'

The sequence `[very young and friendly] dog', on the other hand, would be ambiguous between coordination of words and coordination of phrases; word coordination (where `very' is understood to apply to both adjectives) is more likely.

5. When are constituent tags used?

The above problem about coordination brings up the more general question of when to use constituent tags.

In general, constituent tags are used whenever, to generalize about grammatical structure, we need to recognize constituents intermediate between the `root node' `S' and the `leaf nodes' represented by grammatical wordtags such as `NN', `VB' and `IN'. This general rule needs to be supplemented by the following rules which apply to the parsing of the Lancaster Parsed Corpus.

RULE 1: No bracketing is used to identify a constituent which is recognized as such semantically, but is not formally distinguished. E.g.: compare

a) `the deep blue sea' and b) `her deep blue eyes'

A) means `the sea which is both blue and deep'. B) means `her eyes which are deep blue'. We could show the difference by treating `deep blue' as a constituent, with its own tag, in the second case:

b) `[N her [J deep blue J] eyes N]'

Of course, if `deep-blue' occurred hyphenated, which is quite likely, it would then be treated as a single word adjective (wordtagged `JJ'), rather than a phrase.

But the difference between a) and b) is difference only detectable by the criterion of meaning; it is not signalled by any special structural characteristics. Therefore, our parsing scheme does not differentiate between a) and b), and the constituent `[J deep blue J]' is not recognized.

On the other hand, `her very blue eyes' would be analyzed as having an adjective phrase (`J') in it, because in this case the adjective phrase can be distinguished structurally by the fact that there is a qualifying adverb (`QL') before the adjective `blue':

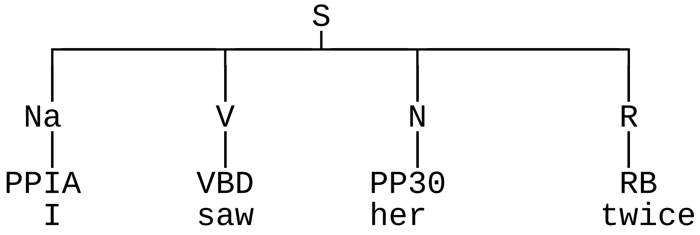
c) `[N her [J very_QL blue J] N]'

Rule 2: A constituent tag IS used where the constituent it labels is an element of clause or sentence structure.

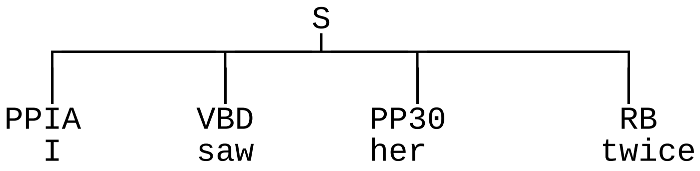
For example, in the sentence, d) `I saw her twice', each of the words is a separate element of sentence structure: `I' is the subject, `saw' is the verb element, `her' is the direct object, and `twice' is an adverbial. So each of these words is assigned a phrase tag:

d) `[S [Na I Na] [V saw V] [N her N] [R twice R] S]'

If we draw a tree diagram of this (inserting the wordtags as well), the result is that the tree contains `unary branching' beneath each of the phrase tags:



This seems wasteful and goes against the general requirement that constituent tags are inserted only when necessary. If we got rid of the phrase tags, the resulting structure would be much simpler.



But this structure would fail to recognize the potential expansion of each of these words into phrases; e.g.:

e) `[S [N My brother N] [V has seen V] [N your sister N] [N several times N] S]'

In fact, the rule that elements of clause/sentence structure are given phrase tags also extends to **E** (the tag for existential `there') and to `U' (the tag for exclamations), even though these cannot be (easily) expanded into multi-word phrases.
Exception: However, coordinating and subordinating conjunctions do not count as clause/sentence elements for these purposes. When they introduce a clause or sentence, these words (wordtagged `CC' and `CS') are direct constituents of the clause or sentence without receiving tags of their own.

Rule 3: WITHIN phrases (i.e. constituents tagged with one of the symbols `N', `V', `J', `R', `P', `M', `G', `D') phrase or clause tags are used only when the constituent that they label consists of two or more words.
In other words, except for the `unary branching' allowed in accordance with Rule 2 above, `unary branching' is not tolerated. For example, in a noun phrase (`N'), there are various constituents of two or more words which are given tags:

f) Prepositional phrases (`P')
`the city [P of New York P]'
g) Adjective phrases (`J')
`the [J most difficult J] job'
h) Genitive phrases (`G')
`[G last night's G] election'
i) Determiner phrases (`D')
`[D so much D] trouble'
j) Numeric phrases (`M')
`[M five hundred M] pounds'
k) Noun phrases (`N')
`a [N two seater N] sports car'

But if these two-words were replaced by single words, then the tags would no longer be used. Examples parallel to some of the examples above have no constituent tags:

g')`the hardest job'
h')`yesterday's election result'
i') `five pounds'
k')`a two-seater sports car'

Note that if the two words are joined by hyphenation, as in k'), this no longer counts as a phrase, but as a single word.

Note also that in k'), if there were no hyphenation, there would a *structural* rather than *semantic* reason for recognizing `two seater' as an `N': the phrase is marked structurally by the fact that the indefinite article (which is singular) is followed by the number `two' (which is plural).

EXCEPTION: There is one important exception to Rule 3: viz. that a preposition in a prepositional phrase is followed by a constituent (normally a noun phrase), which is labelled by a constituent tag, even when the constituent consists of one word. In l) and m)

l) `This present is [P for [N your sister N] P]'
m) `This present is [P for [N you N] P]'

the prepositional complements `your sister' and `you' are both treated as noun phrases, even though one of them consists of just a single word. The reason for this exception is that the nominal expression which follows a preposition has all the potentiality of a noun phrase.

6. Punctuation

Punctuation points and brackets (`.' `?' `!' `:' `;' `,' `-' `(' `)' ) are treated as words for the purposes of wordtagging. Contrary to orthodox practice, punctuation points are treated as parts of a sentence in their own right. The full stop which ends most sentences is treated as the last constituent of the highest `S'. Other punctuation points and brackets are placed as high in the constituent structure tree as they can be, subject to other parsing conventions and rules.
Quotation marks, however, are ignored in parsing except when they mark an `Sq' (direct quotation) constituting an independent sentence.

7. Idiom tags

The above paragraphs may have reinforced the idea that every wordtag is treated as significant in parsing: it is indeed true, in the main, that parsing operates on a word-by-word basis, treating the word-separating spaces of the written text as sacrosanct. However, apart from quotation marks, there is an important exception to this principle.

A sequence of words which behaves syntactically in an idiomatic way (e.g. `in order that', `as well as') is wordtagged by using the wordtag appropriate for the sequence as a whole for each of its components, together with two digits, the first of which indicates the total number of elements in the idiom, the second standing for the number of the position of that particular element within the sequence. E.g.:

`in_CS31 order_CS32 that_CS33'
`as_CC31 well_CC32 as_CC33'
`up_IN21 to _IN22'

The implication of this arrangement for parsing, of course, is that such sequence is treated as a unit for the purposes of parsing.