Part of Speech Tag Set used for MT corpus

David McKelvie

April 25, 2001

1 Introduction

The part-of-speech tag set used is a modified version of the Brown Corpus tag set. The corpus was initially tagged using the Brown tag set and the tag set modified to improve the tagging performance. Some tags have been removed because they did not occur often enough in the Map Task to train effectively. A few have been added in order to model filled pauses, fragmentary words, and to make some finer distinctions among adverbs.

2 Major Word Classes

2.1 MAIN VERBS

VB	VERB BASE FORM
VBD	VERB PAST FINITE FORM eg saw, looked, went
VBG	VERB PRESENT PARTICIPLE (-ing) FORM
VBN	VERB PAST PARTICIPLE eg burnt, gone
VBZ	VERB 3rd PERSON SINGULAR FORM (+s)

Note: The verb tags used are those defined in the Brown tagset. These tags are used for main verbs (i.e. not auxiliary verbs). See also section 3.1.

2.2 NOUNS

NN	COMMON NOUN SINGULAR OR MASS
NNS	COMMON NOUN PLURAL
NP	PROPER NOUN SINGULAR

There are some differences in our noun tags from the Brown tagset. Firstly, we split genitive clitics off from nouns and treat +'s as a separate token with its own tag, thus removing the need for the Brown tags (NN$, NNS$, NP$, NPS$, and NR$). Secondly, there are no plural proper nouns or plural adverbial nouns in the Map Task, so we don't use the Brown NPS and NRS tags. Finally, we don't use the Brown tag NR for 'adverbial nouns' (e.g. east, home, monday, etc), instead we take these as either (NN noun, RB adverb or RP (see below)).

2.3 ADJECTIVES

JJ	ADJECTIVES
JJR	COMPARATIVE ADJECTIVE
JJT	SUPERLATIVE ADJECTIVE with or without -est morphology

As for nouns, we have no need for Brown's JJ$ tag (e.g. great's), and we do not make the distinction between Brown's JJS (superlative adjectives with -est morphology) and JJT (superlative adjectives without -est morphology e.g. innermost or chief) tags, mapping both to JJT.

2.4 ADVERBS

Adverbs are a real mixed bag, so we attempt to subclassify them.

QL	QUALIFIER ADVERB - premodifies adjective or adverbial particles, prepositional adverbs or prepositions
QLDT	QUALIFIER ADVERB - premodifies pronouns or determiners,
QLP	enough - when it postmodifies JJ or RB
RB	ADVERB
RBR	COMPARATIVE ADVERBS +ER
WQL	how (as qualifier as in 'how many')
WRB	how, when, whenever, where, whereabout, whereabouts, why,
NOT	not

This is basically the same as the Brown tagset for adverbs, except that we use NOT instead of *. We don't use RBT for superlative adverbs, since there are none in the Map Task lexicon. And we introduce a new tag QLDT for adverbs premodifying determiners as in 'nearly two metres'. Finally, we don't use a special tag (RN) for the word 'afar'.

3 Minor Word Classes

3.1 AUXILIARY VERBS

In comparison with the Brown tagset, we do not distinguish between positive and negative forms of auxilary verbs. So that tags BED*, etc are not used. Also some rarer forms of auxilary verbs are tagged as main verbs to avoid having rarely used tags.

TO	to (as verbal particle)
	FORMS OF THE VERB be
BE	be
BEM	am, +'m, was
BER	are, +'re, were
BEZ	is, +'s, was

BEG(being) is tagged as VBG as it is so rare in corpus.
BED(were) is tagged as BER due to rarity in corpus.
BEDZ(was) is tagged as BEM,BEZ due to rarity in corpus.
BEN(been) is tagged as VBN due to rarity in corpus.

	FORMS OF THE VERB do
DO	do, did
DOZ	does, did

Other lexical forms of DO are tagged as main verbs. DOD(did) is tagged as DO or DOZ due to rarity.

	FORMS OF THE VERB have
HV	have, +'ve, had
HVZ	has, +'s, had

HVG 'having' is tagged as VBG as it is so rare in corpus. 'had' (tagged HVD and HVN) has been retagged as HV,HVZ, or VBN due to rarity.

	MODAL VERBS
MD	MODAL +'d, +'ll, can, could, may, might, must, need, ought, shall, should, will, would PLUS negative forms of the above

3.2 DETERMINERS

The Brown tags for predeterminers (ABL, ABN, ABX) have been collapsed to one tag DPR. Similarly, the Brown tags for demonstrative determiners (DT, DT$, DTI, DTS and DTX) have been collapsed to DT. PP$ has been renamed PPG, and a new tag GEN has been defined for genitive clitics.

	PREDETERMINERS (can occur before a CENTRAL DETERMINER)
DPR	all, both, double, half, just, quarter, quite, such
	CENTRAL DETERMINERS
AT	ARTICLES: a, an, no, the
DT	SING DEMONSTRATIVE: another, each, that, this QUANTIFIERS: any, some, either PLUR DEMONSTRATIVE: these, those
PPG	POSSESSIVES: her, his, its, my, our, their, your
WDT	INTERROGATIVE: what, which
	POST DETERMINERS (can appear after a central determiner)
AP	few, further, final, last, least, less, little, many more, most, much, next, only, other, same, single very
CD	CARDINAL NUMBERS e.g. one, two, three, etc
OD	ORDINAL NUMBER e.g. first, second, third
GEN	+'s Genitive clitic

3.3 PRONOUNS

EX	there (existential subject)
PD	DEMONSTRATIVE: this, that, these, those
WPS	INTERROGATIVE SUBJ: who, what, whatever, which
WPO	INTERROGATIVE OBJ: who, what, whatever, which
PPS	PERSONAL SUBJECT 3RD SING: he, she, it
PPSS	PERSONAL SUBJECT NON-3RD SING: I, we, you, they
PPO	PERSONAL OBJECT: it, us, you, me, him, her, +'s, them
PPL	PERSONAL REFL: herself, himself, itself, myself, yourself

	ourselves, yourselves, themselves
PPG2	PERSONAL POSS: hers, his, mine, mines, ours, theirs, yours
PR	RELATIVE: who, which, that
PN	OTHER PRONOUNS: any, anything, anywhere(pro-pp?)
	everything, everywhere
	none, nothing, nowhere, neither
	some, something, somewhere
	'them both', both, all, one
	RECIPROCAL 'each other' occurs twice - treat as a single token
	enough,
	ordinal numbers

3.4 PREPOSITIONS

IN	PREPOSITIONS (which take NP as complement)
RP	either ADVERBIAL PREPOSITIONS (prepositions which take NO complement)
	or VERBAL PARTICLES (prepositions which form part of verbs)
	also included here are some uses of directions e.g. north

3.5 CONJUNCTIONS

CC	and, but, either, neither, nor, or, though, yet
CS	'cause, 'til, after, as, because, before, if, like, once, since
	so, than, that, though, unless, until, whereas, whether, while

CS means a clause initial element. CC is used for clause internal conjunctions.

3.6 INTERJECTIONS

AFF	POSITIVE: right, okay, okey-dokey, mmhmm, uh-huh, yeah, yes, aye,
	fine, correct, rightee-ho, right-o, mm-mm, uh-uh,
	NEUTRAL: now, well
	NEGATIVE: no, nope
	What Quirk et.al call reaction signals/initiators
FP	FILLED PAUSE:
	eh, ehm, er, erm, hmm, mm, uh, um
UH	INTERJECTION
	ah, aha, oh,
	bang, christ, dear, fine, god,
	gosh, ha, hell, hurrah, jeez, jesus, my,
	och, oo, oops, phew, please, say, smashing, sorry,
	splat, super, ugh, whoa, whoops, why, wow
	now also includes what used to be FW (FOREIGN WORD)
	alles, culpa, es, fini, finito, gemacht, mea,
	tu, verstanden
NOI	other NOISEs made by the speaker: &noise, &laugh, &indecipherablespeech
PAU	PAUSE: ...
FRAG	FRAGMENTED WORDS (aborted words)

The Map Task microtags FP, GG, FG have been distributed among the tags AFF, FP and UH. AFF tend to appear at the beginning or end of utterances (sometimes as utterances on their own). FP tend to signal speech disfluencies. UH occur inside utterances.

Aborted words are tagged with a special tag FRAG. An alternative is to allow them to be any tag and let the tagger try and assign the most plausible tag to them.

3.7 PUNCTUATION

We don't have tags for punctuation, as they are stripped out before part-of-speech tagging. SENT is used to tag those pauses which are used to segment the speech stream into units for tagging. Currently all pauses are used in this way. In the future, once accurate pause durations have been determined, only pauses over some duration threshhold will be used as tagging unit separators.

SENT TAGGING UNIT SEPARATOR – NOT IN LEXICON –

4 Some problematic areas

There are a number of problems with the tag set, which require further thought. These are discussed here:

1. The word 'like'. As Jim Miller has noted, this word is used in a number of different ways and it is difficult to decide on what its POS tag should be.

2. Adjectives which subcategorise for noun phrases or prepositional phrases, such as 'like' + NP, or 'near' + NP/PP, occur rarely in the Map Task and are at present not well tagged.

3. Non-verbal uses of the word 'say', such as "does that take you up to the top right-hand corner say of the ... the ruined monastery".

5 Example of Tag Set Usage

The following is the Giver's speech from the Map Task conversation *q1ec1*, tagged using the above tag set.

Notes: Items in square brackets are pauses, with their durations in seconds. The actual durations are only rather rough estimates at the moment. Items starting with + are clitics which have been split off from the previous word. A few multi-word items have been tagged as one unit, for example *sort_of* . Tags may be followed by a slash and a number; the number shows the number of different tags this word has in total. Tags in capitals have been corrected by hand, the others were tagged using the Xerox automatic tagger. A few tags have a question mark attached; I find these problematic.

q1ec1.g.tag

```
[0.0000] okay  starting off  we   are above a  caravan park [0.9795] we   are
         aff/2 vbg/2    rp/2 ppss ber in/2  at nn      nn           ppss ber

going to   go due  south straight south and  then we   +'re going to   g--
vbg   to/3 vb ql/2 rp/4  ql/3     rp/4  cc/2 rb   ppss ber  vbg   TO/3 frag

turn straight back round and  head north past an old mill on   the right-hand
vb/2 ql/3     rp/3 rp/4  cc/2 vb/2 rp/4  in/3 at jj  nn   in/2 at  jj/2

side [3.1460] yeah south and  then straight back up   again with an old mill
nn            aff  rp/4  cc/2 rb   ql/3     rp/3 RP/2 rb    in/2 at jj  nn

on   the right and  you    +'re going to   pass on   the left-hand side of
in/2 at  nn/6  cc/2 ppss/2 ber  vbg   to/3 vb/2 in/2 at  jj/2      nn   in/2

the mill [1.5463] okay  and  then we   +'re going to   turn east [0.9900]
at  nn            aff/2 cc/2 rb   ppss ber  vbg   to/3 vb/2 rp/4

d--  not straight east slightly sort\_of northeast [1.4554] slightly slightly
frag not ql/3     rp/4 QL/2     QL/4     rp/4               RB/2     rb/2

yeah very slightly and  we   +'re going to   continue straight along erm
aff  ql/2 rb/2     cs/2 ppss ber  vbg   to/3 vb       ql/3     rp/2  fp

quite a  wee dis-- a  wee distance erm quite a  wee distance right we   +'re
dpr/3 at jj  frag  at jj  nn       fp  dpr/3 at jj  nn       aff/6 ppss ber

gonna continue along on   that course and  then we   +'re going to   turn
vbg   vb       rp/2  in/2 dt/4 nn     cc/2 rb   ppss ber  vbg   to/3 vb/2

north again [4.1689] and  immediat-- well  a  distance below that turning
rp/4  rb             cc/2 frag       aff/4 at nn       in/3  dt/4 jj/3

point there +'s   a  fenced meadow but  you    should be avoiding that by
nn    ex/4  bez/3 at jj     nn     cs/4 ppss/2 md     be vbg      pd/4 in/2

quite a  distance [1.8191] okay  so   we   +'ve turned and  we   +'re going
dpr/3 at nn               aff/2 cs/3 ppss hv   vbn/2  cs/2 ppss ber  vbg

up   north again [0.4305] continue straight up   north [0.4305] and  then
rp/2 rp/4  rb             vb       ql/3     rp/2 rp/4           cc/2 rb

we   +'re going to   turn to   the west on [0.8610] a  curvature right
ppss ber  vbg   to/3 vb/2 in/3 at  nn/4 IN/2        at nn        AFF/6

sort\_of "s"-bend [1.2765] and  immediately below that bend there is  an
jj/4     nn                cc/2 rb/2        in/3  dt/4 nn/2 ex/4  bez at

abandoned cottage [0.9253] and  we   +'re passing above the top  of   that
jj        nn               cs/2 ppss ber  vbg     in/2  at  nn/2 in/2 pd/4

we   +'re going to   continue in   that sort\_of "s" shape a  big wide
ppss ber  vbg   to/3 vb       in/2 dt/4 jj/4     nn  nn    at jj  jj

"s" [1.5535] and  on   the sort\_of mmhmm top   erm left of   that again below
nn           cc/2 in/2 at  jj/4     aff   jjt/2 fp  nn/6 in/2 pd/4 rb    in/3
```

```
it    there +'s   a  fenced meadow but  you    +'re passing on   the top
ppo/2 ex/4  bez/3 at jj     nn     cs/4 ppss/2 ber  vbg     in/2 at  nn/2

of   that okay [0.9991] right okay  we   +'ve gone from the abandoned cottage
in/2 pd/4 aff/2         aff/6 aff/2 ppss hv   vbn  in/2 at  jj        nn

right and  we   +'re on   the sort\_of "s" shape yeah [0.9198] right and  then
AFF/6 cc/2 ppss ber  in/2 at  jj/4     nn  nn    aff           aff/6 cc/2 rb

at   the top  of   the "s" we   +'re turning north [1.1845] okay  we   +'re
in/2 at  nn/2 in/2 at  nn  ppss ber  vbg/3   rp/4           aff/2 ppss ber

going straight due  north at   the top   there there +'s   a  west
vbg   ql/3     ql/2 rp/4  in/2 at  NN/2  pn?/4 ex/4  bez/3 at jj/4

lake [1.0063] which we   +'re going to   pass on   the south erm
nn            wpo/4 ppss ber  vbg   to/3 vb/2 in/2 at  jj/4  fp

southeast [1.8785] side and  we   +'re gonna do that in   a  curve almost a
JJ/4               nn   cc/2 ppss ber  vbg   do pd/4 in/2 at nn/2  qldt/3 at

half "u" shape [1.3290] yeah [0.9769] yeah [1.7651] the southeast and
nn/6 nn  nn             aff           aff           at  nn/4      cc/2

continue up   north slightly [1.2342] but  not quite to   the tip of   that
vb       rp/2 rp/4  rb/2              cc/4 not ql/3  in/3 at  nn  in/2 dt/4

lake [0.9569] and  then we   +'re going to   turn down ove-- above a  trick
nn            cc/2 rb   ppss ber  vbg   to/3 vb/2 rp/4 frag  in/2  at jj

point and  we   +'re going to   turn immediately to   your right and
nn    cc/2 ppss ber  vbg   to/3 vb/2 ql/2        in/3 ppg  nn/6  cc/2

straight down at   an angle of   forty-five [2.8604] okay  and  gonna
ql/3     rp/4 in/2 at nn    in/2 pn/3                aff/2 cc/2 vbg

continue that wee distance down and  at   the point [0.4503] at   the end
vb       dt/4 jj  nn       rp/4 cc/2 in/2 at  nn             in/2 at  nn/2

of   that it    should be near  to   the abandoned cottage where we   went
in/2 pd/4 pps/2 md     be JJ?/4 in/3 at  jj        nn      wrb   ppss vbd/2

past miles away but  if not just carry on   and  then continue down in   that
in/3 nns   rp/2 cs/4 cs not rb/3 vb    rp/2 cc/2 rb   vb       rp/4 in/2 dt/4

forty-five degree [0.9006] and  turn round by   a  monument on   the outside
cd/3       nn              cc/2 vb/2 rp/4  in/2 at nn       in/2 at  nn/3

of   the monument [1.6423] yeah and  then a  very slight turning up   again
in/2 at  nn                aff  cc/2 rb   at ap/2 jj     nn/3    in/2 rb

north sort\_of northwest [1.1762] very slight curve sort\_of very slight
rp/4  QL/4     rp/4               ap/2 jj     nn/2  rb?/4    ap/2 jj

"s"-shaped just  a  slight curve and  then gonna proceed up   north again
jj         dpr/3 at jj     nn/2  cc/2 rb   vbg   vb      rp/2 rp/4  rb

and  on   the right-hand side there +'s   a  nuclear test site before right
```

```
cc/2 in/2 at  jj/2       nn   ex/4  bez/3 at jj      nn   nn   cs/3   QL/6

before reaching the top  of   that northbound and  then you    +'re going
cs/3   vbg      at  nn/2 in/2 dt/4 nn/2       cc/2 rb   ppss/2 ber  vbg

to   turn back west and  above that there +'s   an east lake [1.7361]
to/3 vb/2 rp/3 rp/4 cc/2 in/2  pd/4 ex/4  bez/3 at jj/4 nn

yeah [1.5388] and  that +'s   the finish
aff           cc/2 pd/4 bez/3 at  nn/2
```