Automatic Mapping Among Lexico-Grammatical Annotation Models (AMALGAM)

Eric Atwell, Language research group, School of Computing, Leeds University

AMALGAM HOMEPAGE | PREVIOUS PAGE | UP A LEVEL | NEXT PAGE

The University of Pennsylvania (Penn) Treebank Tag-set

Listed alphabetically below are the standard tags used in the Penn Treebank. Each tag has examples of the tokens that were annotated with that tag. The examples are taken directly from the Penn Treebank lexicon that is supplied with Eric Brill's Transformation-Based Part-of-Speech Tagger . This is the tagger that is used as the basis for the AMALGAM e-mail tagging server.

The Penn scheme incorporates special `vertical slash' tags for occasions when the part-of-speech is ambiguous. Consider the sentence: **The duchess was entertaining last night.** (This example is taken from ``Part-of-Speech Tagging Guidelines for the Penn Treebank Project'' by Beatrice Santorini which is available from the the Penn Treebank site .) Does **entertaining** mean that she was hosting an event, in which case the word would be a present participle verb, **VBG** , or does **entertaining** act adjectively (**JJ**) implying that the Duchess was good company? Either case is possible so the Penn Treebank developers allow both tags to apply at the same time. In this case, **entertaining** is assigned the tag **JJ|VBN** . The vertical slash tags are not listed in the table below but, for completeness, those that were found to occur at least once in the Penn Treebank lexicon are listed afterwards.

The tokenisation of genitives differs in the Penn and the International Corpus of English (ICE) schemes from all of the other tagging schemes of the AMALGAM tagger. Penn and ICE do not leave the genitive marker attached to the word but strip it off. The AMALGAM tokeniser recognises this difference and removes the marker from the ends of words. There are other variations in how tokenisation is handled by the different schemes but in order to facilitate comparisons AMALGAM tokenises input text the same way, regardless of scheme, except for this one example.

If the list of examples ends with an ellipsis marker then the tag category can be assumed to be an open class.

Further reading:

Santorini, B. 1990. Part-of-speech tagging guidelines for the Penn Treebank Project. Technical report MS-CIS-90-47, Department of Computer and Information Science, University of Pennsylvania.

Tag	Description	Examples
$	dollar	$ -$ --$ A$ C$ HK$ M$ NZ$ S$ U.S.$ US$
``	opening quotation mark	` ``
''	closing quotation mark	' ''
(	opening parenthesis	([{
)	closing parenthesis	)] }
,	comma	,
--	dash	--
.	sentence terminator	. ! ?
:	colon or ellipsis	: ; ...
CC	conjunction, coordinating	& 'n and both but either et for less minus neither nor or plus so therefore times v. versus vs. whether yet
CD	numeral, cardinal	mid-1890 nine-thirty forty-two one-tenth ten million 0.5 one forty-seven 1987 twenty '79 zero two 78-degrees eighty-four IX '60s .025 fifteen 271,124 dozen quintillion DM2,000 ...
DT	determiner	all an another any both del each either every half la many much nary neither no some such that the them these this those
EX	existential there	there
FW	foreign word	gemeinschaft hund ich jeux habeas Haementeria Herr K'ang-si vous lutihaw alai je jour objets salutaris fille quibusdam pas trop Monte terram fiche oui corporis ...

The following `vertical slash' tags were found to occur at least once in the Penn lexicon:

CD|NN	CD|NNS	CD|NN|NP	CD|RB					
DT|NN	DT|RB							
IN|JJ	IN|PP	IN|RB	IN|RP					
JJR|IN	JJR|RBR							
JJ|CC	JJ|IN	JJ|JJR	JJ|NN	JJ|NP	JJ|RB	JJ|VBG	JJ|VBN	
LS|EX	LS|JJ	LS|NN	LS|NNS					
MD|VB								
NNPS|NNS	NNPS|VBZ							
NNP|CC|NP	NNP|JJ	NNP|NN	NNP|NP	NNP|NPS	NNP|POS	NNP|VB	NNP|VBN	NNP|VBZ
NNS|DT	NNS|LS	NNS|NN	NNS|NPS	NNS|VBZ				
NN|CD	NN|DT	NN|IN	NN|JJ	NN|JJ|RB	NN|NN	NN|NNS	NN|POS	NN|RB
				NN|SYM	NN|VB	NN|VBG	NN|VBP	NN|WRB
PRP|JJ	PRP|MD	PRP|VBP						
RBR|JJR	RBR|NN	RBS|JJ	RBS|JJS					
RB|CC	RB|DT	RB|IN	RB|JJ	RB|NN|JJ	RB|RBR	RB|RP	RB|VBG	RB|VBZ
RP|IN	RP|RB							
VBD|RB	VBD|VBN	VBD|VBP						

Eric Atwell, Language research group, School of Computing, Leeds University

AMALGAM HOMEPAGE | PREVIOUS PAGE | UP A LEVEL | NEXT PAGE