

The taggers included with Gromoteur are:

- 1) the Brill parser from pattern <http://www.clips.ua.ac.be/pages/pattern-en#parser>
- 2) the Urheen parser <http://www.openpr.org.cn/index.php/NLP-Toolkit-For-Natural-Language-Processing/68-Urheen-A-Chinese/English-Lexical-Analysis-Toolkit/View-details.html>).

The taggers use the Penn Treebank tagset. The following explanation and tables are taken from: <https://www.comp.leeds.ac.uk/ccalas/tagsets/upenn.html>

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 ...
IN	preposition or conjunction, subordinating	astride among upon whether out inside pro despite on by throughout below within for towards near behind atop around if like until below next into if beside ...
JJ	adjective or numeral, ordinal	third ill-mannered pre-war regrettable oiled calamitous first separable ectoplasmic battery-powered participatory fourth still-to-be-named multilingual multi-disciplinary ...
JJR	adjective, comparative	bleaker braver breezier briefer brighter brisker broader bumper busier calmer cheaper choosier cleaner clearer closer colder commoner costlier cozier creamier crunchier cuter ...
JJS	adjective, superlative	calmest cheapest choicest classiest cleanest clearest closest commonest corniest costliest crassest creepiest crudest cutest darkest deadliest dearest deepest densest dinkiest ...
LS	list item marker	A A. B B. C C. D E F First G H I J K One SP-44001 SP-44002 SP-44005 SP-44007 Second Third Three Two * a b c d first five four one six three two

MD	modal auxiliary	can cannot could couldn't dare may might must need ought shall should shouldn't will would
NN	noun, common, singular or mass	common-carrier cabbage knuckle-duster Casino afghan shed thermostat investment slide humour falloff slick wind hyena override subhumanity machinist ...
NNP	noun, proper, singular	Motown Venneboerger Czestochwa Ranzer Conchita Trumplane Christos Oceanside Escobar Kreisler Sawyer Cougar Yvette Ervin ODI Darryl CTCA Shannon A.K.C. Meltex Liverpool ...
NNPS	noun, proper, plural	Americans Americas Amharas Amityvilles Amusements Anarcho-Syndicalists Andalusians Andes Andruses Angels Animals Anthony Antilles Antiques Apache Apaches Apocrypha ...
NNS	noun, common, plural	undergraduates scotches bric-a-brac products bodyguards facets coasts divestitures storehouses designs clubs fragrances averages subjectivists apprehensions muses factory-jobs ...
PDT	pre-determiner	all both half many quite such sure this
POS	genitive marker	's
PRP	pronoun, personal	hers herself him himself hisself it itself me myself one oneself ours ourselves ownself self she thee theirs them themselves they thou thy us
PRP\$	pronoun, possessive	her his mine my our ours their thy your
RB	adverb	occasionally unabatingly maddeningly adventurously professedly stirringly prominently technologically magisterially predominately swiftly fiscally pitilessly ...
RBR	adverb, comparative	further gloomier grander graver greater grimmer harder harsher healthier heavier higher however larger later leaner lengthier less-perfectly lesser lonelier longer louder lower more ...
RBS	adverb, superlative	best biggest bluntest earliest farthest first furthest hardest heartiest highest largest least less most nearest second tightest worst
RP	particle	aboard about across along apart around aside at away back before behind by crop down ever fast for forth from go high i.e. in into just later low more off on open out over per pie raising start teeth that through under unto up up-pp upon whole with you
SYM	symbol	% & ' " .) . * + , . < = > @ A[fj] U.S U.S.S.R * * * * * *
TO	"to" as preposition or infinitive marker	to
UH	interjection	Goodbye Goody Gosh Wow Jeepers Jee-sus Hubba Hey Kee-reist Oops amen huh howdy uh dammit whammo shucks heck anyways whodunnit honey golly man baby diddle hush sonuvabitch ...

VB	verb, base form	ask assemble assess assign assume atone attention avoid bake balkanize bank begin behold believe bend benefit bevel beware bless boil bomb boost brace break bring broil brush build ...
VBD	verb, past tense	dipped pleaded swiped regummed soaked tidied convened halted registered cushioned exacted snubbed strode aimed adopted belied figgered speculated wore appreciated contemplated ...
VBG	verb, present participle or gerund	telegraphing stirring focusing angering judging stalling lactating hankerin' alleging veering capping approaching traveling besieging encrypting interrupting erasing wincing ...
VBN	verb, past participle	multihulled dilapidated aerosolized chaired languished panelized used experimented flourished imitated reunified factored condensed sheared unsettled primed dubbed desired ...
VBP	verb, present tense, not 3rd person singular	predominate wrap resort sue twist spill cure lengthen brush terminate appear tend stray glisten obtain comprise detest tease attract emphasize mold postpone sever return wag ...
VBZ	verb, present tense, 3rd person singular	bases reconstructs marks mixes displeases seals carps weaves snatches slumps stretches authorizes smolders pictures emerges stockpiles seduces fizzes uses bolsters slaps speaks pleads ...
WDT	WH-determiner	that what whatever which whichever
WP	WH-pronoun	that what whatever whatsoever which who whom whosoever
WP\$	WH-pronoun, possessive	whose
WRB	Wh-adverb	how however whence whenever where whereby wherever wherein whereof why

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						
VBG JJ	VBG NN	VBG NN JJ						
VBN JJ	VBN TO	VBN VBD						
VBP IN	VBP PP	VBP TO	VBP VB	VBP VBD				
VB IN	VB NN							
WP IN	WP MD NP							
