AV Text Ministries

Digital-AV

Part-of-Speech (Digital-AV SDK Specification)

Revision: #i728

Adapted from https://www.ling.upenn.edu/courses/Fall_2003/ling001/penn_treebank_pos.html

Number	POS bits	Tag	Description		additional bits
1.	0x_C00	CC	Coordinating conjunction		
2.	0x_F00	CD	Cardinal number		
3.	0x_D00	DT	Determiner		
4.	0x0E00	EX	Existential *there*		
5.	0x0000	FW	Foreign word		
6.	0x0400	IN	Preposition or subordinating conjunction		
7.	0x_A00	JJ	Adjective		Unmarked: 0x0A00
8.	0x1A00	JJR	Adjective, comparative		
9.	0x2A00	JJS	Adjective, superlative		
10.	0x0700	LS	List item marker		
11.	0x_301	MD	Modal		
12.	0x401_	NN	Noun, singular or mass		
13.	0x801_	NNS	Noun, plural		
14.	0x501_	NNP	Proper noun, singular		
15.	0x901_	NNPS	Proper noun, plural		
16.	0x2D00	PDT	Predeterminer		
17.	0x0008	POS	Possessive ending		
18.	0x_03_	PRP	Personal pronoun	Nominative: 0x_07_ Oblique: 0x_0B_ Reflexive: 0x_0F_ Unmarked: 0x_03_	Neuter: 0x___1 Masculine: 0x___2 non-feminine: 0x___3 Feminine: 0x___4
19.	0x_038	PRP$	Possessive pronoun		Neuter: 0x___1 Masculine: 0x___2 non-feminine: 0x___3 Feminine: 0x___4 Genitive: 0x___8
20.	0x_B00	RB	Adverb		Unmarked: 0x0B00
21.	0x1B00	RBR	Adverb, comparative		
22.	0x2B00	RBS	Adverb, superlative		
23.	0x1D00	RP	Particle		
24.	0x_300	SYM	Symbol		
25.	0x_200	TO	*to*		
26.	0x_800	UH	Interjection		
27.	0x_101	VB	Verb, base form		
28.	0x_102	VBD	Verb, past tense		
29.	0x_104	VBG	Verb, gerund or present participle		
30.	0x_106	VBN	Verb, past participle		
31.	0x_103	VBP	Verb, non-3rd person singular present		
	0x_108		Verb, marked for singular case agreement		(EModE conjugation)
32.	0x_107	VBZ	Verb, 3rd person singular present		
33.	0xCD00	WDT	Wh-determiner		
34.	0xC030	WP	Wh-pronoun		
35.	0xC038	WP$	Possessive wh-pronoun		
36.	0xCB00	WRB	Wh-adverb		

Adapted from https://www.ling.upenn.edu/courses/Fall_2003/ling001/penn_treebank_pos.html

The Digitial-AV utilizes the part-of-speech (POS) tags as defined by the Penn Treebank. POS tagging of the bible verses themselves is performed during SDK compilation. In order to minimize the number of unknown words relative to the Penn Treebank, somewhat archaic words are modernized prior to submission to the TextBlob tagger. TextBlob itself uses the default NLTK tagger under the hood. There is not only preprocessing of text submitted for tagging, but also post-processing to handle 2nd-Person singular and associated verb conjugations. Moreover, Hitchcock's Bible Name Dictionary sets words to Proper-Nouns when they are found in that dictionary (potentially overriding the tag provided by TextBlob/NLTK).

TextBlob is housed in a Python Django web server, whereas the SDK compiler is written in C#/dot-net. The SDK compiler gets tokenized sentences from the TextBlob/NLTK library via Django as a JSON response. Source code to the entire pipeline is available on github at:
https://github.com/kwonus/avtext
and
https://github.com/kwonus/Digital-AV

The table of the previous page was cloned from the Penn Treebank website identified in the header of this document. After cloning the table, the second column was added to show how the standard Penn Treebank tags map into the bit fields of the Digital-AV.

It should be noted that while the table depicts an accurate mapping of Digital-AV bits from/to standard POS tags, some bits get set by the SDK compiler, most notably: additional bits are set for Person-Number (PN), as defined in the primary Digital-AV documentation.

Adapted from https://www.ling.upenn.edu/courses/Fall_2003/ling001/penn_treebank_pos.html

```
POS Summary (not entire range of values):
Verb:       0x_10_
Modal:      0x_30_
Noun:       0x_01_
Proper noun:0x1_1_
Pronoun:    0x_03_
WH:         0xC___
Possessive: 0x___8
ADJ:        0x_A00
ADV:        0x_B00
DET:        0x0D_0      // unmarked for number (e.g. the)
DET-SING:   0x4D_0
DET-PLURAL: 0x8D_0
Particle:   0x1D00
Prep:       0x0400
To:         0x0200
Interject.  0x0800

PN nibble (applies to verbs and nouns only):
1st Person: 0x1___
2nd Person: 0x2___
3rd Person: 0x3___
Unmarked:   0x0___
Singular:   0x4___
Plural:     0x8___
```