NUPOS:
A part of speech tag set for written English
from Chaucer to the present

By Martin Mueller
November 2009

1 Introduction and Summary

The following is a description of NUPOS, a part-of-speech (POS) tag set designed to accommodate the major morphosyntactic features of written English from Chaucer to the present day. The description is written for an audience not familiar with POS tagging. NUPOS is part of an enterprise to make the results of such tagging useful to humanities scholars who are not professional linguists and have not considered its utility for a wide variety of applications beyond linguistics proper.

While the NUPOS tag set can be used with any tagger that can be trained, so far it has been used only with Morphadorner (http://wordhoard.northwestern.edu) , an NLP suite developed by Phil Burns and used extensively in the MONK project. Some 2,000 texts from the 1500's to the late 1800's have been tagged with it.

2 What is POS tagging?

A part-of-speech tag set is a classification system that allows you to assign some grammatical description to each word occurrence in a text. This assignment can be done by hand or automatically. Typically you "train" an automatic tagger by giving it the results of a hand-tagged corpus. The tagger then applies to unknown text corpora what it "learned" from the training set. The "knowledge" of the automatic tagger may consist of a set of rules or of a statistical analysis of the results. Either way, a good tagger will provide accurate descriptions for 97 out of a 100 words.

Why do you want to apply POS tagging to a text in the first place? Readers might well ask this question when the sees the tagging output of the opening of Emma, which might look like this:

Emma_name Woodhouse_name, handsome_adj, clever_adj, and_conj rich_adj

This tells you nothing you did not know before. But humans are very subtle decoders who bring an extraordinary amount of largely tacit knowledge to the task of making sense of the characters on the page. The computer, however, lacks this knowledge. If you want to take full advantage of the query potential of a machine readable text you must make explicit in it at least some of the rudiments of readerly knowledge. If you do so, you can quickly and accurately perform many operations that will be difficult or practicable for human readers to do. You cannot only extract a list of adjec-

tives (or other parts of speech), you can also identify syntactic fragments, such as the sequence of three adjectives. A variety of stylistic or thematic opportunities for inquiry open up with a POS-tagged text, especially if the tagging is carried out consistently across large text archives. Analyses of this kind are based on the guiding assumption that there often is an illuminating path from low-level linguistic phenomena to larger-scale thematic or structural conclusions.

3 The concept of the LemPos

If you want to use computers for the analysis of texts that differ in time, genre, regional or social stratification you want to be in a position where the surface form of any word occurrence can be mapped to a more abstract representation that allows algorithms to identify features one surface form shares with others. For many purposes, a satisfactory mapping will consist of the combination of a part of speech tag with the lemma or the look-up form of the word in a dictionary. I call that combination a LemPos. Here are some examples:

Surface form or spelling	Lemma + POS tag or LemPos
vniuersities	university_ng1
vniuersities	university_n2
university's	university_ng1
universities	university_n2

Human readers tacitly process the ways in which these spellings stand for the same or different forms. The machine is not that bright, but once it has been presented with the 'explicitated' LemPos it can perform many operations that humans could never do with comparable speed or accuracy.

It is clear from this very simple example that the mapping of a spelling to a LemPos depends on three distinct operations:

1. the recognition of orthographic variance
2. the identification of morphosyntactic features
3. the identification of the lemma

When the NUPOS tag set is used with MorphAdorner, the text for human readers or sequence of words on the printed is supplemented with a ma-

chine-readable representation that explicitly articulates some data while ignoring others

4 About tag sets

POS tags carry some combination of morphological and syntactic pieces of information, whence they are also called morphosyntactic tags. In highly inflected languages, such as Greek, Latin, or Old English, the inspection of a word out of context will reveal much about its grammatical properties. English has shed most of its inflectional features over the centuries, and the individual word will contain ambiguities that only context can resolve. Thus the –ed form of a verb may be the past tense or the past participle. For some common verbs (put, shut, cut), the distinction between past and present is morphologically unmarked. In many cases even the distinction between verb and noun ('love') is not morphologically marked.

In English, therefore, POS tagging is a business that works with very limited morphological information (mainly the suffixes –s, -ed, -ing, -er, -est, -ly) and uses the context of preceding or following words to make sense of things. A little reflection on these facts opens one's eyes to characteristic errors of English taggers, such as the confusion of participial and past tense forms.

The most widely most used tag set for modern English is the Penn Treebank tag set. This set consists of about three dozen tags (though some of them can be combined). It offers a very crude classification system, but for many purposes it is good enough. When you are in the world of machines making decisions, crude distinctions consistently applied are more useful than error-ridden subtle distinctions.

Like other modern tag sets, the Penn Treebank set lacks important feature for the accurate tagging of written English before the twentieth century. It recognizes the third person singular of a verb (VBZ), but it does not recognize the second person singular ('thou art'). You can see the reason: the second person singular is no longer a living form. But it remains a living archaism, and it was a living form of poetic and religious usage well into the twentieth century.

Modern English taggers have a very odd way of dealing with the possessive case or genitive. In English orthography since the eighteenth century, the apostrophe has been used to distinguish between the –s suffix as a plural marker and as a possessive marker. Before the middle of the seventeenth century, this orthographical distinction is rarely or never found, and a sequence like "the kings command" is ambiguous.

The Penn Treebank set, like most other tag sets, treats the apostrophized 's' as a separate word. When the automatic tagger applies its rules, a word like "king's" is 'tokenized' as two words. The convenience of this procedure for modern English is obvious, especially since the apostrophized 's' can also stand for 'is' or 'has' in contracted forms, where it has a linguistically sounder claim to be treated as a separate word. But if you want a tag set capable of processing written English across many centuries, it is clearly preferable to find a solution that treats the 's' of the possessive case in the same way in which it treats other inflectional suffixes, such as the plural 's' or the 'ed' and 'ing' of verb forms.

Like other English tag sets, the Penn Treebank set consists of a somewhat inconsistent mix of syntactic and morphological markers. The tags VVZ and NN2 respectively stand for the –s forms of a verb and a noun. In each case the symbol includes information about a syntactic category (verb, noun) and a morphological condition (3rd singular, plural). But the same morphological form can operate in different syntactic environment. This is particularly true of participial forms. When a form like 'loving' is used as a verb form, the code 'VVG" provides information both about its syntactic function (VV) and its morphological form (G). But when the same word is used as an adjective or as a noun (the gerund), the codes JJ and NN ignore morphological information.

5 The NUPOS tag set

5.1 The history of the NUPOS tag set

The NUPOS tag set is a hybrid product that grew out of WordHoard, a project to create a search environment for deeply tagged corpora and includes all of Early Greek epic as well as the works of Chaucer, Spenser, and Shakespeare (http://wordhoard.northwestern.edu). The Greek texts were morphologically tagged with the help of the Morpheus tagger of the Perseus project. The Chaucer text was based on Larry Benson's Glossarial Database to the Riverside Chaucer and uses the tag set designed by Benson for that project. The Shakespeare text was tagged with the CLAWS tag set developed at Lancaster University and used for the tagging of the British National Corpus.

My original plan was to use different tag sets for Chaucer and Shakespeare. But on closer inspection I discovered that you could with hardly any

loss merge the Benson and CLAWS tags in a common set. It also turned out that that Chaucer has only two verb forms that are not found in Shakespeare: the fairly rare second person plural imperative and the quite common –n form to mark the infinitive or first and third plural present of verbs.

In other words, you need only four tags to extend a modern tag set so that it can capture the major morphosyntactic phenomena in English from Chaucer on:

1. The second person singular present
2. The second person singular past
3. The first and third plural present
4. The second plural imperative

In merging the tag sets I took from Benson a "used-as" category that is important to his scheme and compensates for a weakness in the CLAWS and Penn Treebank sets. A word will typically belong to one word class and is used in all or most cases as an instance of that class. A noun is a noun, a verb is a verb, etc. But in a phrase like "no ifs or buts" the conjunctions 'if' and 'but' are used as nouns. In the catachrestic spirit of such a phrase you can use any word class as any other word class, and much word play depends on it.

There are more systemic uses of this phenomenon. In a phrase like 'My loving lord' the present participle of the verb 'love' is used as an adjective. In 'the running of the deer' a present participle is used as a noun. Benson's tagging scheme explicitly recognizes these phenomena by creating code points like 'present participle used as adjective'. This seems to me preferable to the practice of dropping the morphological information and using JJ or NN tags, as CLAWS and the Penn Treebank set do. The utility of keeping the information is particularly apparent if you are also lemmatizing a text and want to record adjectival uses of 'loving' or 'loved' as instances of the verb 'love'.

The difficulties of classifying participial forms are worth some comment. English and its cognate languages distinguish sharply between nouns and verbs. They share number, but nouns lack voice and tense while verbs lack case and gender. But participles cross that divide. There are uses where a verbal, nominal, or attributive function clearly dominates, but there are many uses where it does not. The training data for participial forms in NUPOS follow the rule: "If in doubt it's a verbal form."

5.2 The structure of the NUPOS tag set

NUPOS owes some features to the morphological tagging scheme used in The Chicago Homer (www.library.northwestern.edu/homer). That scheme is taken over from Perseus' Morpheus but it stores the information in a very atomic fashion in a relational database so that a given word can be retrieved as an instance of any of its grammatical properties, separately or in combination.

A Greek word can be adequately defined through the categories of tense, mood, voice, case, gender, person, number, degree. In conventional grammars, a description will typically consist of a string of properties, such as aor-ind-act-3rd-sing for the Greek word 'eperse'. The VVZ tag of English tag sets does pretty much the same thing, but the 'Z' component implicitly specifies tense (present), person (3rd), and number (singular). If you keep the morphological information in a rigorously atomic and explicit fashion, you can search at different levels at granularity. For instance, any given instance of an aorist optative passive form in Greek will have person and number, but if you keep the information in what database experts call a 'normalized' fashion, you can ignore person and number (or any other atomic component) in your search.

The NUPOS tag set is implemented in a framework that supports the normalized representation of tag sets for different languages. A given form is defined by the values it holds in the categories of tense, mood, voice, case, gender, person, number, degree, wordclass and subclass, and part of speech. The categories of voice and gender are irrelevant to English, but you need both for Greek or Latin, and you need gender for French or German.

In assigning values to categories, I have made some practical decisions that may raise the linguists' eyebrows. English has a residual subjunctive (If I were…), but no tagging scheme tries to recognize it, probably because it cannot be captured with sufficient accuracy by algorithms. My mood category quite properly includes the indicative and the infinitive. Somewhat less properly, it includes participles. In the ancient and modern European languages, participles may have voice or tense, but they lack mood and may therefore be put in a 'mood' column of a database without causing damage.

5.3 Negative forms and un-words

English has some contracted forms like 'nas' (was not), 'niltow' (ne wilt thou) or "don't" whose orthographical status clearly testifies to their perception as single lexemes. If the subjunctive and optative moods are seen as modifications of the declarative indicative, why not accept a 'negative' form as a radical modification? The OED does something like it. If you look up

'cannot' you are told that it is "the ordinary modern way of writing can not." But if you look at 'can' you are taken to its inflexions, where 'cannot' is described as the negative form of can. NUPOS adds a negative category that is used to discriminate between 'will' and "won't", 'none' and 'one', or 'ever' and 'never'.

I have done something similar and perhaps more radical with 'un-words'. Do 'unforgiving' and 'unforgiven' share a common lemma? If you decide to treat 'un-' words as negative forms, the question is easy to answer, and there are very clear rules for creating 'un' forms of English lemmata. Accordingly, I have treated the prefix 'un-' as a negative modifier of a positive lemma, and its part of speech is given a -u flag. Thus 'unnatural_j-u' corresponds to 'natural-j'.

There are always slippery cases. Since 'do' is put in the class of auxiliary verbs and the tagging does not distinguish between ordinary and auxiliary forms of the verb, the forms of 'undo' are not classified as forms of 'do', but its pos tags are given a -u flag anyhow, so that a search for -u forms will retrieve them.

If you reduce 'un-words' to their roots why not do the same thing for other prefixes, such as 'under' or 'over'? There are two reasons for this. First, un- is by far the most common prefix. Secondly, un-words have a relatively weak status as stable lemmata in their own right. The modal case of an un-word is a participial adjective or adverb (unseen, undoubtedly), while the forms of verbs beginning with 'over' or 'under' are distributed much more evenly across infinitive, present, past, and participial forms.

5.4 Comparative and superlative forms

The comparative and superlative forms of adjectives are formed with the suffixes -er and -est for short adjectives and with the periphrastic forms 'more' and 'most' for long adjectives. I have classified 'more', 'most', 'less', 'least' as comparative and superlatives determiners with -c and -s flags so that a search for pos tags with those flags will let you measure the extent of comparative and superlative markers in a text.

5.5 Word Class and POS

The word class specifies the class to which a word belongs most of the time. The assignment is made on a lexical basis without reference to a particular context. There are major word classes, and some of them have subclasses. Taggers differ in their recognition of subclasses. NUPOS is more like CLAWS than the Penn Treebank tag set in recognizing subclasses. But you can ignore the subclasses if you wish.

The Penn Treebank tag set is very Spartan when it comes to verbs and does not distinguish between the open class of common verbs and the closed class of grammatical verbs. CLAWS recognizes modal verbs and has separate tags for each of the verbs ‘be’, ‘have’ and ‘do’. NUPOS follows CLAWS in this regard, largely because digitally assisted analysis increasingly makes use of syntactic fragments created by tag sequences, and in particular by tag trigrams. If you have any interest in such analysis you will want to distinguish between auxiliaries as markers of tense or voice: 'had shot' (vhd vvn) and 'was shot' (vbds vvn) are very different constructions.

Modal verbs present some problems of classification in a diachronic corpus. In Middle English, as in modern German, modal verbs are capable of ‘full’ uses: in both languages you can say things like “I can it not,” which you cannot do in modern English, just as you know cannot use 'could' as Chaucer used it in his description of the Wife of Bath:

> Of remedies of love she knew per chaunce,
> For she koude of that art the olde daunce.

Phrases of that kind are probably not uncommon in archaizing Early Modern English. NUPOS treats all forms of ‘may’, ‘will’, ‘shall’, ‘can’ and ‘ought’ as if they were modern modals, but it does recognize modal forms that are not possible in modern English, such as a modal participles or infinitives. Quasi-modals like ‘let’ and ‘used’ are treated as common verbs.

The modal verbs ‘can’, ‘will’, ‘may’, ‘shall’ each exist in two forms, which historically are present and past forms but in practice differ in mood rather than tense. It is worth marking the difference, because a discourse rich in ‘could, would, should’ is very different from a discourse rich in ‘can, will, shall’. It is easiest, and historically accurate, to mark it as a difference in tense.

5.6 POS or part of speech proper

The part-of-speech proper of any word occurrence is the syntactic role it plays in its context regardless of any particular morphological inflection. It is usually the same as the word class of a word, but in cases like ‘my loving lord’ it is not. The POS in this narrow sense is identical with the ‘used-as’ category in Benson’s tag set for Chaucer. It provides a very coarse classification of about two dozen categories, but for many purposes it may be good enough.

It is not easy to define the conditions that make you say: this noun (or verb) is not used as a noun (or verb) in this word occurrence. In compound

nouns like 'water closet' the first noun acts as a kind of adjective; in a phrase like "the dead will rise" the adjective acts as a kind of noun. NUPOS assumes that such quasi-adjectival uses of nouns or quasi-nominal uses of adjectives are within the ordinary range of behaviour for nouns and adjectives. Therefore the POS for 'water' is noun and for 'dead' is adjective.

5.7 Ambiguous word classes

Some words cross word classes, and it is difficult for a computer program (or sometimes a human) to assign them confidently to a particular part of speech. Many of the mistakes that taggers make have to do with erroneous assignments of POS tags to such words. A particular occurrence of 'since' or 'before' may be an adverb, a preposition, or a conjunction. Many prepositions are used adverbially. The different uses of 'as' or 'like' are a nightmare to keep apart neatly.

NUPOS groups some words under the word class adverb-conjunction-preposition (ACP) and assigns its best guess to the POS tag. Thus an occurrence of 'since' may carry the tag C-ACP, which means "this is probably a conjunction but certainly an adverb, conjunction, or preposition." Such a demarcation of the boundaries of error may be useful for some purposes. The terminology makes no special claim except that the classes of these words are likely to be confused with each other but not with other classes.

In addition to the ACP word class there are three other ambiguous word classes. Conjunctive, relative, and interrogative uses of the 'wh- words' are hard to tag automatically. I have bundled these words in a CRQ class, which includes such words as 'who', 'which', 'when', 'why' 'what'.

Words like 'yesterday' or 'today' are largely adverbs, but have some nominal uses (yesterday's paper). I have classified them as AN.

The last such class is a group of words that hover systematically between adjective and noun (JN). This class includes color words, names (Albanian, Jesuit, Florentine), and an odd assortment of words that include 'evil', 'right', 'wrong', 'male', 'female', 'mercenary' etc.

One could posit for each of these word a distinct lemma as noun and adjective, just as one distinguishes between the verb and the noun 'love'. But I doubt whether 'blue' as noun or adjective is distinguished in the linguistic (un)conscious in the way in which the noun and verb 'love' are. It seems better to acknowledge that there is a class of words that systematically cross the boundaries of noun and adjective and whose properties can be described with some precision. The Oxford English Dictionary has it both ways with such words. Sometimes there are distinct entries, and sometimes you have an entry of the type "XX: adjective and noun."

My criterion for classifying an adjective as a JN word has been its potential as a singular noun. You can say 'my necessaries' but not 'my necessary'. But you can say 'my secret' or 'a deep blue'. But these are very fluid distinctions. POS tagging is a very crude exercises and always reminds me of Wallace Stevens' line from 'Connoisseurs of Chaos':

> The squirming facts exceed the squamous mind

5.8 One word or many?

Automatic tagging of words relies on the normal case that a lexical unit consists of a single word separated by a space from the next word. The normal case is statistically more frequent than right-handedness. But there are a lot of 'lefties', and they pose a lot of challenges.

The lefties come in three forms. There are lexical units that span more than one word. There are hyphenated words, and there are contractions. Of these contractions pose the problem that is hardest to ignore because it forces you to make decisions about tokenization and POS assignment that do not in that form arise with multi word units or hyphenated forms. Although phrases like "according to" or "in vain" are most easily seen as instance of a two-word preposition or adverb, you can find ways of tagging each word separately. The component parts of a hyphenated word nearly always fit comfortably into an existing POS tag, most often an adjective or noun. But contracted forms typically cross the noun/verb divide and cannot be assigned to a single POS tag.

There are two different ways of approaching this problem, each with its own difficulties. In the first approach you say that contracted forms (much more common in speech than in writing) are "really" two words and that the written record should divide what lazy speaker slurred together. Alternately you can say that the orthographic practice of marking contractions, typically by means of the apostrophe, responds to a linguistic reality in the mind of the speakers or author and that the tagger ignores that reality when it keeps apart what the author intended to keep together.

For a variety of reasons, both practical and theoretical, NUPOS takes the second route. At the simplest level, you must "tokenize" words before you can apply POS tags to them. Tokenization has a number of consequences in a digital file. It counts the number of words and will play some role in assigning to each word a unique address in a text. The closer the process of tokenization stays to the reader's naïve perception the better off you are. Readers will say that in the sentence "Don't do that" 'that' is the third word. You do not want to have to explain them that it is the fourth word. Nor do

you want to have a routine that counts it as the fourth word for some purpose and as the third word for others. Better to stick with the notion that "don't do that" is a three-word sentence of which "don't" is the first word.

Some contractions decompose easily into distinct parts, but others do not. Sometimes the apostrophe marks the division of words but sometimes it does not. In the case of "it's" the apostrophe neatly divides the parts. In "'tis" or "don't" the parts are easily identified, but the apostrophe is not the divider. In Early Modern English there are many contracted forms that are written as one word. 'Nas' for 'ne was' is one example. "Ain't" is a modern example of a contracted form that is not easily decomposed, and it has as much right to be treated as a single token as 'never' or 'none'.

Add these practical concerns to the assumption that the orthographic contraction reflects an underlying linguistic reality, and you come to the conclusion that contracted forms should be dealt with as single words as much as possible. That is the approach chosen in NUPOS.

The vast majority of contracted word occurrences—99% or more—are made up of a few very common patterns that are counted in the dozens rather than hundreds and amount to a closed class of combinations of pronouns and auxiliary/modal verbs or of auxiliary/modal verbs with the negative.

There is also an open class of verbs or nouns preceded by a contracted 'to' or 'the' (t'advance, th'earth) or a noun followed by the contracted form of 'is'. You might call these proclitic and enclitic contractions.

If you treat a contracted form as a single word you still have to account separately for its components. As said above, combinations of an auxiliary or modal verb with a negative can be expressed in a single tag as the negative form of that verb. Combinations of a pronoun with an auxiliary or modal verb have to be expressed through a compound tag that joins the tag for the pronoun to the tag for the verb. Such compound tags raises the total number of tags (compound or single) by about a third.

Compound tags make life harder for the developer who designs the data object model and the interface for the user who formulates queries that depend on the tags for their answer. "She'll" has to count for an instance of 'will' and 'she.' And the relevant form of 'will' in this case is "'ll" and not "she'll." Doing this in a consistent and user-friendly manner is not as easy as it sounds. But it is possible.

In Early Modern English, you find two-word spellings of forms that are now treated as single words. The most common cases are 'to day', 'to morrow' and reflexive pronouns like 'myself', 'themselves'. MorphAdorner can

and does tokenize these bigrams as single words so that a spelling like ‘them selues’ will appear in an XML representation of a text as

```
<w lemma="themselves" pos="pnx32">
```

5.9 The verb ‘be’

As in other languages, ‘be’ is the word with the largest and most diverse set of forms. Present tense forms include ‘art’, ‘is’, ‘are’, ‘be’, ‘be’st’ and ‘aren’. Past tense forms include ‘was’, ‘were’, ‘wast’, ‘wert’, and ‘weren’. There is only one form of the past participles, but it occurs in several orthographic variants.

In an earlier form of NUPOS, I mapped ‘is’ to ‘vbz’ and all other present forms to ‘vbb’. I mapped all the past forms to ‘vbd’. In this version, I use ‘vbr’ and ‘vbb’ to distinguish between ‘are’ and finite uses of ‘be’. I use ‘vbdr’ , ‘vbds’, ‘vbd2r’ and ‘vbd2s’ to distinguish between ‘were’, ‘was’, ‘wert’, and ‘wast’. These granular distinctions allow you to capture sutble distinctions between the forms. They also allow you to map variant spellings of the -r and -s form to standard spellings.

5.10 The ‘lempos’ and standardized spelling

With some exceptions and qualifications, the LemPos or combination of lemma and POS tag can be used to generate a standard spelling. You need an exception list of verbs and nouns that do not form their past and plural forms with -d or -s suffixes.

Adverbs pose a separate problem. The standard adverbial form of an adjective uses a -ly suffix. But there is a class of spatial adjectives that use an ‘-s’ suffix (‘downwards’). There is also a zero form of adverbs (‘pretty much’, ‘real soon’). The zero and -ly forms of some adjectives may have quite different meanings, as in the case of ‘just’, ‘very’, ‘pretty’, ‘straight’, or ‘hard’. Where there is strong semantic differentiation, it makes sense to split the adverb from its original lemma. Thus adverbial ‘hard’ and ‘hardly’, ‘just’ and ‘justly’, ‘very’ and ‘verily’ are treated as different lemmata.

You could solve this problem by having different tags for the zero, -s, and -ly forms of adverbs formed from adjectives.

Yet another problem is posed by variants that hover between morphological and orthographic variance -- ‘loveth’ vs. ‘loves’ or ‘spake’ vs. ‘spoke’. Mapping ‘loveth’ to ‘loves’ or ‘spake’ to ‘spoke’ is less violent than mapping ‘wast’ to ‘wert’, but it does erase some real differences, as opposed to

mapping ‘vniuersitie’ to ‘university’, where the differences are merely and systematically orthographic.

There are problems with homonyms. Depending on the meaning of the verb, the lempos ‘lie_vvd’ maps to the spellings ‘lay’ or ‘lied’. ‘Hanged’ and ‘hung’ are participial forms with quite distinct meanings, but they are both correctly described by the lempos ‘hang_vvd’.

You can go on with the enumeration of such problems. Some of them could in principle be resolved by more granular tag sets. Others resist algorithmic treatment. But it is also true that for the vast majority of cases, a LemPos can be mapped algorithmically to a single standard spelling.

5.11 How many tags and how many errors?

A good modern tagger will tag ~97% of words correctly. This is less impressive than it sounds because you can determine the part of speech of ~90% of all word occurrences from their lexical status. So from one perspective, the POS tagger makes a difference only for the last 10%, and it makes mistakes in a third of the cases.

Mistakes come in different shapes, and some matter more than others. For instance, the infinitive and present form of the verb are morphologically indistinct. The infinitive is identified from a preceding ‘to’ or auxiliary verb. If other words intervene between the auxiliary and the verb mistakes are likely. Of 100 verb forms that are identified as VVB or VVI between 10 and 12 are likely to be classified wrongly. Perhaps wisely the Penn Treebank tag set does not even make the distinction. CLAWS and NUPOS try to make it because an infinitive always depends on another verb, and if you can exclude infinitive verbs from your count it is easier to count clauses. But for many users VVB/VVI errors are insignificant.

Another source of error is the confusion of the past participle (VVN) and the past tense (VVN). These too are morphologically indistinct except for a limited number of ‘strong’ verbs. In both NUPOS and CLAWS (at least when used with 16h century texts for which it was not designed) this error is more common than the confusion of VVB and VVI and may run as high as 15%-18%. If a form is correctly classified as a present or past participle its use may be incorrectly classified as a noun or an adjective.

Taggers using NUPOS will have trouble with identifying the possessive case of nouns where there is no apostrophe to mark it. Phrases like “the kings command” are genuinely difficult, and they involve a double error. The first mistake, classifying a possessive singular as a plural, is relatively benign. But if the tagger gets the first word wrong it may well make a mis-

take with the next word and classify a noun as a verb. That is a more consequential error: ng1-n1 is a very different syntactic construction from n2-vvb.

The coarser the classification, the lower the error rate. If you are satisfied with a broad classification of word occurrences as nouns, verbs, or adjectives, and do not worry about confusions of the VVB/VVI or VVD/VVN kind, the error rate probably drops by half.

5.12 Tagging at different levels of granularity

NUPOS is more explicit than other tagging schemes in letting users determine the granularity of the tagging. The NUPOS tag is really a "key" or unique ID that represents the classification of each morphological condition by discrete categories that users may ignore or activate. Depending on whether you classify by the strict POS tag, the combination of POS and wordclass, or the combination of all categories, you may end up with some twenty, sixty, or 250 tags.

6 Appendix

The following table shows the tag set for NUPOS. For each tag, the tag name is followed by an explanation, by an example, and by the approximate rate of occurrence per million words in 320 16h and 17th century English plays with a total word count of about six million words.

The NUPOS training data have included:

1. The complete works of Chaucer and Shakespeare
2. Spenser's *Faerie Queene*
3. North's translation of Plutarch's Lives
4. Mary Wroth's *Urania*
5. Jane Austen's *Emma*
6. Dickens' *Bleak House* and *The Old Curiosity Shop*
7. Emily Bronte's *Wuthering Heights*
8. Thackeray's *Vanity Fair*
9. Mrs. Gaskell's *Mary Barton*
10. Frances' Trollope's *Michael Armstrong*
11. George Eliot's *Adam Bede*
12. Scott's *Waverley*
13. Harriet Beecher Stowe's *Uncle Tom's Cabin*
14. Melville's *Moby Dick*

Examples are chosen for the most part from the training data.

NUPOS Tag set

NUPOS	description	example	pos per million words
a-acp	acp word as adverb	I have not seen him since	6066.3
av	adverb	soon	35078.1
av-an	noun-adverb as adverb	go home	406.1
av-c	comparative adverb	sooner, rather	467.6
av-d	determiner/adverb as adverb	more slowly	1881.9
av-dc	comparative determiner/adverb as adverb	can less hide his love	1875.9
av-ds	superlative determiner as adverb	most often	931.7

av-dx	negative determiner as adverb	no more	854.2
av-j	adjective as adverb	quickly	8763.1
av-j-u	adjective as adverb (un)	unnaturally	90.2
av-jc	comparative adjective as adverb	he fared worse	731.7
av-jn	adj/noun as adverb	duly, right honourable	663.7
av-jn-u	un-adj/noun as adverb (un-)	unduly	0.3
av-jp	proper adjective as adverb	Christianly	0.5
av-jp-u	proper adjective as adverb (un-)	unchristianly	0.2
av-js	superlative adjective as adverb	in you it best lies	188.3
av-n	noun as adverb	had been cannibally given	0.2
av-s	superlative adverb	soonest	11.7
av-u	adverb (un-)	uneath	0.5
av-vvg	present participle as adverb	lovingly	76.9
av-vvg-u	present participle as adverb (un-)	unknowingly	1.4
av-vvn	past participle as adverb	Stands Macbeth thus amazedly	17.5
av-vvn-u	past participle as adverb (un-)	undoubtedly	6.6
av-x	negative adverb	never	1607.6
avc-jn	comparative adj/noun as adverb	deeper	8.0
avs-jn	superlative adj/noun as adverb	hee being the worthylest constant	
c-acp	acp word as conjunction	since I last saw him	8886.8
c-crq	wh-word as conjunction	when she saw	5271.7
cc	coordinating conjunction	and, or	32276.6
cc-acp	acp word as coordinating conjunction	but	6267.8
ccx	negative conjunction	nor	1234.6
crd	numeral	2, two, ii	4378.3
cs	subordinating conjunction	if	8093.1
cst	'that' as conjunction	I saw that it was hopeless	9263.7
d	determiner	that man, much money	28653.1
dc	comparative determiner	less money	946.4
dg	determiner in possessive use	the latter's	4.6
dgx	negative determiner in possessive use	neither’s	0.3
ds	superlative determiner	most money	381.5
dt	article	a man, the man	49407.5
dx	negative determiner as adverb	no money	3185.9
fw-es	Spanish word	cuerpo	21.0

fw-fr	French word	monsieur	642.4
fw-ge	German word	Herr	104.4
fw-gr	Greek word	kurios	8.6
fw-it	Italian word	cambio	42.9
fw-la	Latin word	dominus	1662.9
fw-mi	word in unspecified other language	n/a	169.0
j	adjective	beautiful	43855.4
j-av	adverb as adjective	the then king	0
j-jn	adjective-noun	the sky is blue	5647.8
j-jn-u	adjective-noun (un-)	undue	24.6
j-u	adjective (un-)	unnatural	650.2
j-vvg	present participle as adjective	loving lord	1700.5
j-vvg-u	present participle as adjective (un-)	unrelenting spirit	34.1
j-vvn	past participle as adjective	changed circumstances	2260.8
j-vvn-u	past participle as adjective (un-)	unblemished night	489.2
jc	comparative adjective	handsomer	1457.1
jc-jn	comparative adj/noun	yet she much whiter	61.9
jc-u	comparative adjective (un-)	unhappier	0.3
jc-vvg	present participles as comparative adjective	for what pleasinger then varietie, or sweeter then flatterie?	0.2
jc-vvn	past participle as comparative adjective	shall find curster than she	0.7
jp	proper adjective	Athenian philosopher	916.9
jp-u	proper adjective (un-)	unchristian	1.2
js	superlative adjective	finest clothes	1472.5
js-jn	superlative adj/noun	reddest hue	163.4
js-jn-u	superlative adj/noun (un-)	unwelcomest man	0.3
js-n	noun as superlative adjective	felonest (Spenser)	
js-u	superlative adjective (un-)	unworthiest hand	4.7
js-vvg	present participle as superlative adjective	the lyingest knave in Christendom	6.4
js-vvn	past participle as superlative adjective	deformed'st creature	4.7
js-vvn-u	past participle as superlative adjective (un-)	the unprovidest sir of all our courtesies	0.2
n-jn	adj/noun as noun	a deep blue	1239.3
n-jn-u	adj/noun as noun(un)	through myn unkonninge (Chaucer)	0
n-vdg	present participle as noun, 'do'	my doing	2
n-vhg	present participle as noun, 'have'		0
n-vvg	present participle as noun	the running of the deer	862.9

n-vvg-u	present participle as noun (un-)	the clear unfolding of my doubts	9.7
n-vvn	past participle as noun	the departed	16.8
n1	singular, noun	child	140905.8
n1-an	noun-adverb as singular noun	my home	169.5
n1-j	adjective as singular noun	an important good	0.2
n1-u	singular, noun (un-)	unthrift	64.9
n2	plural noun	children	35795.9
n2-acp	acp word as plural noun	and many such-like "As'es" of great charge	0.2
n2-an	noun-adverb as plural noun	all our yesterdays	6.9
n2-av	adverb as plural noun	and are etcecteras no things	0.3
n2-cc	coordinating conjunction used as noun	and’s	0.3
n2-crq	wh-word used as noun	why’s	0.3
n2-dx	determiner/adverb negative as plural noun	yeas and honest kerysey noes	0.5
n2-j	adjective as plural noun	give me particulars	185.1
n2-jn	adj/noun as plural noun	the subjects of his substitute	669.2
n2-sy	character used as plural noun	her C's	1.9
n2-u	plural noun (un-)	serious untruths	7.1
n2-uh	interjection used as noun	in russet yeas	0.8
n2-vdg	present participle as plural noun, 'do'	doings	9.8
n2-vhg	present participle as plural noun, 'have'	my present havings	0.3
n2-vvg	present participle as plural noun	the desperate languishings	164.1
n2-vvg-u	present participle as plural noun (un-)	undoings	0.2
n2-vvn	past participle as plural noun	there was no necessity of a Letter of Slains for Mutilation	0
ng1	singular possessive, noun	child's	3308.5
ng1-an	noun-adverb in singular possessive use	Tomorrow's vengeance	1.7
ng1-j	adjective as possessive noun	the Eternal's wrath	0.7
ng1-jn	adj/noun as possessive noun	our sovereign's fall	45.1
ng1-vvn	past participle as possessive noun	knock at the closed door of the late lamented's house	0.2
ng2	plural possessive, noun	children's	349.0
ng2-j	adjective as plural possessive noun	the poors' cries	1.2

ng2-jc	comparative adjective as possessive plural noun	hindering the greaters' growth	0.2
ng2-jn	adj/noun as plural possessive noun	mortals' chiefest enemy	32.9
njp	proper adjective as noun	a Roman	57.6
njp2	proper adjective as plural noun	The Romans	196.4
njpg1	proper adjective as possessive noun	The Roman's courage	7.6
njpg2	proper adjective as plural possessive noun	The Romans' courage	17.6
np1	singular, proper noun	Paul	16703.6
np1-n	singular noun as proper noun	at the Porpentine	43.1
np2	plural, proper noun	The Nevils are thy subjects	232.7
np2-n	plural noun as proper noun	such Brooks are welcome to me	0.3
npg1	singular possessive, proper noun	Paul's letter	1383.2
npg1-n	singular possessive noun as proper noun	and through Wall's chink	3.2
npg2	plural possessive, proper noun	will take the Nevils' part	5.1
ord	ordinal number	fourth	1862.5
p-acp	acp word as preposition	to my brother	64612.9
pc-acp	acp word as particle	to do	14699.0
pi	singular, indefinite pronoun	one, something	1261.4
pi2	plural, indefinite pronoun	from wicked ones	68.8
pi2x	plural, indefinite pronoun	To hear my nothings monstered	5.3
pig	singular possessive, indefinite pronoun	the pairings of one's nail	12.2
pigx	possessive case, indefinite pronoun	nobody's	0
pix	indefinite pronoun	none, nothing	1394.7
pn22	2nd person, personal pronoun	you	18844.4
pn31	3rd singular, personal pronoun	it	8254.1
png11	1st singular possessive, personal pronoun	a book of mine	476.1
png12	1st plural possessive, personal pronoun	this land of ours	78.8
png21	2nd singular possessive, personal pronoun	this is thine	
png22	2nd person, possessive, personal pronoun	this is yours	267.3
png31	3rd singular possessive, personal pronoun	a cousin of his	304.4

png32	3rd plural possessive, personal pronoun	this is theirs	30.3
pno11	1st singular objective, personal pronoun	me	9589.0
pno12	1st plural objective, personal pronoun	us	1904.1
pno21	2nd singular objective, personal pronoun	thee	3070.5
pno31	3rd singular objective, personal pronoun	him, her	7820.2
pno32	3rd plural objective, personal pronoun	them	2560.3
pns11	1st singular subjective, personal pronoun	I	26062.5
pns12	1st plural subjective, personal pronoun	we	4069.0
pns21	2nd singular subjective, personal pronoun	thou	4814.7
pns31	3rd singular subjective, personal pronoun	he, she	9647.8
pns32	3rd plural objective, personal pronoun	they	3104.9
po11	1st singular, possessive pronoun	my	15833.9
po12	1st plural, possessive pronoun	our	3379.5
po21	2nd singular, possessive pronoun	thy	4370.3
po22	2nd person possessive pronoun	your	9585.3
po31	3rd singular, possessive pronoun	its, her, his	10050.7
po32	3rd plural, possessive pronoun	their	2675.1
pp-f	preposition 'of'	of	18369.2
px11	1st singular reflexive pronoun	myself	762.2
px12	1st plural reflexive pronoun	ourselves	116.8
px21	2nd singular reflexive pronoun	thyself, yourself	620.3
px22	2nd plural reflexive pronoun	yourselves	89.5
px31	3rd singular reflexive pronoun	herself, himself, itself	736.3
px32	3rd plural reflexive pronoun	themselves	179.3
pxg21	2nd singular possessive, reflexive pronoun	yourself's remembrance	0.2
q-crq	interrogative use, wh-word, subject	Who? What? How?	5915.6
qg-crq	interrogative use, wh-word, possessive	Whose?	12.7

qo-crq	interrogative use, wh-word, object	Whom?	38.1
r-crq	relative use, wh-word, subject	the girl who ran	5601.9
rg-crq	relative use, wh-word, possessive	to such, whose faces are all zeal	782.0
ro-crq	relative use, wh-word, object	a wretched maid, whom ye have pursued	640.3
sy	alphabetical or other symbol	A, @	233.6
uh	interjection	oh!	6484.7
uh-av	adverb as interjection	Well!	475.8
uh-crq	wh-word as interjection	Why, there were but four	827.5
uh-dx	negative interjection	No!	889.7
uh-j	adjective as interjection	Grumio, mum!	13.4
uh-jn	adjective/noun as interjection	And welcome, Somerset	82.5
uh-n	noun as interjection	Soldiers, adieu!	315.1
uh-np	proper noun as interjection	Jesu	0.2
uh-v	verb as interjection	My gracious silence, hail	155.4
uh-x	negative interjection	No!	843.6
vb2-imp	2nd plural present imperative, 'be'	Beth pacient	
vb2r	2nd singular present of 'be'	thou art	711.7
vb2rx	2nd singular present, 'be'	thow nart yit blisful	
vb2s	2nd singular present of 'be'	thou beest	23.6
vbb	present tense, 'be'	they be	2559.0
vbbx	present tense negative, 'be'	aren't, ain't, beant	0.5
vbd2r	2nd singular past of 'be'	wert	93.6
vbd2s	2nd singular past of 'be'	wast	32.7
vbd2x	2nd singular past, 'be'	weren't	
vbdp	plural past tense, 'be'	whose yuorie shoulders weren couered all	
vbdr	past tense, 'be'	were	1903.6
vbdrx	past tense negative, 'be'	weren't, nere (Chaucer)	
vbds	past tense, 'be'	was	2588.5
vbdsx	past tense negative, 'be'	wasn't, nas (Chaucer)	
vbg	present participle, 'be'	being	650.0
vbi	infinitive, 'be'	be	6414.1
vbm	1st singular, 'be'	am	2705.1
vbmx	1st singular negative, 'be'	I nam nat lief to gabbe	0.2
vbn	past participle, 'be'	been	999.7
vbp	plural present, 'be'	Thise arn the wordes	0.2
vbr	present tense , 'be', 'are'	they are	4674.2
vbrx	present tense negative, 'be', are	they aren't	0.2
vbz	3rd singular present, 'be'	is	8820.2
vbzx	3rd singular present negative, 'be'	isn't	0
vd2	2nd singular present of 'do'	dost	431.5

vd2-imp	2nd plural present imperative, 'do'	Dooth digne fruyt of Penitence	0
vd2x	2nd singular present negative, 'do'	thee dostna know the pints of a woman	0.2
vdb	present tense, 'do'	do	3093.9
vdbx	present tense negative, 'do'	don't	2.7
vdd	past tense, 'do'	did	1416.8
vdd2	2nd singular past of 'do'	didst	155.3
vdd2x	2nd singular past negative, verb	"Why, thee thought'st Hetty war a ghost, didstna?	0
vddp	plural past tense, 'do'	on Job , whom that we diden wo	0
vddx	past tense negative, 'do'	didn't	0
vdg	present participle, 'do'	doing	52.2
vdi	infinitive, 'do'	to do	1003.2
vdn	past participle, 'do'	done	766.3
vdp	plural present, 'do'	As freendes doon whan they been met	0
vdz	3rd singular present, 'do'	does	1185.1
vdzx	3rd singular present negative, 'do'	doesn't	0
vh2	2nd singular present of 'have'	thou hast	559.8
vh2-imp	2nd plural present imperative, 'have'	O haveth of my deth pitee!	0
vh2x	2nd singular present negative, 'have'	hastna	0
vhb	present tense, 'have'	have	5394.4
vhbx	present tense negative, 'have'	haven't	4.2
vhd	past tense, 'have'	had	1821.0
vhd2	2nd singular past of 'have'	thou hadst	92.4
vhdp	plural past tense, 'have'	Of folkes that hadden grete fames	0
vhdx	past tense negative, 'have'	hadn't	0.2
vhg	present participle, 'have'	having	157.6
vhi	infinitive, 'have'	to have	2239.8
vhn	past participle, 'have'	had	155.1
vhp	plural present, 'have'	They han of us no jurisdiccioun,	0
vhz	3rd singular present, 'have'	has, hath	2753.6
vhzx	3rd singular present negative, 'have'	Ther loveth noon, that she nath why to pleyne.	0
vm2	2nd singular present of modal verb	wilt thou	921.7
vm2x	2nd singular present negative, modal verg	O deth, allas, why nyltow do me deye	0
vmb	present tense, modal verb	can, may, shall, will	17429.8

vmb1	1st singular present, modal verb	Chill not let go, zir, without vurther 'cagion	0.7
vmbx	present tense negative, modal verb	cannot; won't; I nyl nat lye	1039.8
vmd	past tense, modal verb	could, might, should, would	6475.3
vmd2	2nd singular past of modal verb	couldst, shouldst, wouldst; how gret scorn woldestow han	264.2
vmd2x	2nd singular present, modal verb	Why noldest thow han writen of Alceste	0
vmdp	plural past tense, modal verb	tho thinges ne scholden nat han ben doon.	0
vmdx	past negative, modal verb	couldn't; She nolde do that vileynye or synne	1.2
vmi	infinitive, modal verb	Criseyde shal nought konne knowen me.	0
vmn	past participle, modal verb	I had oones or twyes ycould	0
vmp	plural present tense, modal verg	and how ye schullen usen hem	0
vv2	2nd singular present of verb	thou knowest	975.6
vv2-imp	2nd present imperative, verb	For, sire and dame, trusteth me right weel,	0
vv2-u	2nd singular present of verb (un-)	thou unbendest	0.3
vv2x	2nd singular present negative, verb	“Yee!” seyde he, “thow nost what thow menest;	0
vvb	present tense, verg	they live	38328.6
vvb-u	present tense, verb (un-)	they unfold	56.6
vvbx	present tense negative, verb	What shall I don? For certes, I not how	0.2
vvd	past tense, verb	knew	10730.8
vvd-u	past tense, verb (un-)	he unlocked the horse	7.3
vvd2	2nd singular past of verb	knewest	159.5
vvd2-u	2nd singular past of verb (un-)	thy treacherous blade unrippedest the bowels	0.2
vvd2x	2nd singular past negative, verb	thou seidest that thou nystist nat	
vvdp	past plural, verb	They neuer strouen to be chiefe	
vvdx	past tense negative, verb	she caredna to gang into the stable	
vvg	present participle, verb	knowing	4715.1
vvg-u	present participle, verb (un-)	without unveiling herself	7.6
vvi	infinitive, verb	to know	44589.5
vvi-u	infinitive, verb (un-)	I must unclasp me	96.6
vvn	past participle, verb	known	20285.1

vvn-u	past participle, verb (un-)	would you be thus unclothed	147.5
vvp	plural present, verb	Those faytours little regarden their charge	1.0
vvp-u	plural present, verb(un-)	Tthey unsowen the semes of freendshipe (Chaucer)	
vvz	3rd singular preseent, verb	knows	10287.8
vvz-u	3rd singular preseent, verb	he that unbuckles this	7.8
vvzx	3rd singular present negative, verb	She caresna for Seth.	0
wd	word wrongly split or joined in text		546.4
xx	negative	not	10210.2
zf	English word wrongly used by foreign speaker		102.2
zz	unknown or unparsable token		2312.4