# NLP Notes

Carlo Ruiz

March 22, 2018

## I. English Grammar Prequisites

While word classes do have semantic tendencies–adjectives, for example, often describe properties and nouns people–parts-of-speech are traditionally defined instead based on syntactic and morphological function, grouping words that have similar neighboring words (their distributional properties) or take similar affixes (their morphological properties).

### Open classes:

Classes whose membership is continuously changing. Four major open classes occur in the languages of the world: nouns, verbs, adjectives, and adverbs.

- **Noun:** Informally, a person, place or thing. More formally, a word with the ability to occur with determiners (a goat, its bandwidth, Plato's Republic), to take possessives (IBM's annual revenue), and for most, but not all nouns, to occur in the plural form (goats, abaci).
  - **Proper Noun:** Names of specific persons or entities.
  - **Common Noun:** Are divided into two classes. *Count nouns* allow grammatical enumeration. They occur in both singular and plural and can be counted (one goat, two goats). *Mass nouns:* are used when something is conceptualized as a homogeneous group. So words like snow, salt, and communism are not counted. Mass nouns can also appear without articles where singular count nouns cannot.
  - **Gerund:** Morphologically, it is typically a verb ending in -ing. Semantically it functions as a noun, thus it can be the object of a preposition or verb, or the subject of a verb. It can often seem like an adjective since it complements the verb *be* ("The key is *understanding* the gerund"). But notice that this is different from an adjective since we can insert *the* before *understanding* and $of$ after, retaining grammatical correctness.
- **Verb:** The verb class includes most of the words referring to actions and processes, including main verbs like draw, provide, and go. Morphologically, verbs have inflections (non-third-person-sg (eat), third-person-sg (eats), progressive (eating), past participle (eaten)). We give special attention to participles. Morphologically a **participle** often ends in -ed or -ing (eating, reported). Participles are often preceded by auxiliary verbs (like adjectives). We can often differentiate between a participle and an adjective by using the very-test. Insert very before the word in question. If it does not seem grammatical, the word is a participle.
- **Adjective:** A class that includes terms for properties or qualities. Adjectives can only describe a noun.

- **Adverb:** Words that describe, and in some way, complement an adjective, verb, adverb or a full verb phrase. *Directional Adverbs* (home, here, downhill) specify the direction or location of some action. *Degree Adverbs* (extremely, very, somewhat) describe the manner of some action or process. It is often helpful to view verbs as processes and anything describing the process is an adverb. *Temporal Adverbs* (yesterday, Monday) describe the time an action, process, or event took place. Because of the heterogeneous nature of this class, some adverbs (e.g., temporal adverbs like Monday) are tagged in some tagging schemes as nouns.

### Closed classes:

Classes with relatively fixed membership such as prepositions. New prepositions are rarely coined. Closed class words are generally *function words* like of, it, and, or you, which tend to be very short, occur frequently, and often have structuring uses in grammar.

- **Preposition:** is followed by a noun phrase. Semantically they often indicate spatial or temporal relations, whether literal (on it, before then, by the house) or metaphorical (on time, with gusto, beside herself), but often indicate other relations as well, like marking the agent in (Hamlet was written by Shakespeare). A prepositional phrase never references a noun. It references either a verb or an $S$

- **Determiner:** Often mark the beginning of a noun phrase. (a, an, the, this, that)

- **Pronoun:** are forms that often act as a kind of shorthand for referring to some noun phrase or entity or event. Personal pronouns refer to persons or entities. Possessive pronouns are forms of personal pronouns that indicate either actual possession or more often just an abstract relation between the person and some object (my, your, his, her, its, one?s, our, their). Wh-pronouns (what, who, whom, whoever) are used in certain question forms, or may also act as complementizers (Frida, who married Diego. . . ).(she, who, I, others, whom)

- **Conjunction:** Joins two phrases, clauses, or sentences (and, but, or, as, if, when). Coordinating conjunctions like *and*, *or*, and *but* join two elements of equal status. Subordinating conjunctions are used when one of the elements has some embedded status. For example, *that* in ”I thought that you might like some milk” is a subordinating conjunction that links the main clause I thought with the subordinate clause you might like some milk. This clause is called subordinate because this entire clause is the ”content” of the main verb thought. Subordinating conjunctions like that which link a verb to its complementizer argument in this way are also called complementizers.

- **Auxiliary verb :** (were, will, have, should) Cross-linguistically, auxiliaries mark certain semantic features of a main verb, including whether an action takes place in the present, past, or future (tense), whether it is completed (aspect), whether it is negated (polarity), and whether an action is necessary, possible, suggested, or desired (mood).

- **Particle:** A particle resembles a preposition or adverb and is used in combination with a verb (up, down, out, by). Particles often have extended meanings that aren't quite the same as the prepositions they resemble, as in the particle up in *she turned the volume up.*

- **Numerals:** one, two, three, first, second, third

- English also has many words of more or less unique function, including interjections (oh, hey, alas, uh, um), negatives (no, not), politeness markers (please, negative thank you), greetings (hello, goodbye), and the existential there (there are two on the table) among others. These classes may be distinguished or lumped together as interjections or adverbs depending on the purpose of the labeling.

## Reinforcing our understanding

- **Prepositions vs Adverbs vs Particles:** Prepositions are followed by noun phrases, adverbs are not. Particles often have extended meanings that aren't quite the same as the prepositions they resemble. Up in *turn it up* is a particle. Up in *he went up* is an adverb. Up in *he went up the hill* is a preposition.

- **There vs. there:** The first *there* is existential, the second is an adverb. There/EX are/VBP 70/CD children/NNS there/RB

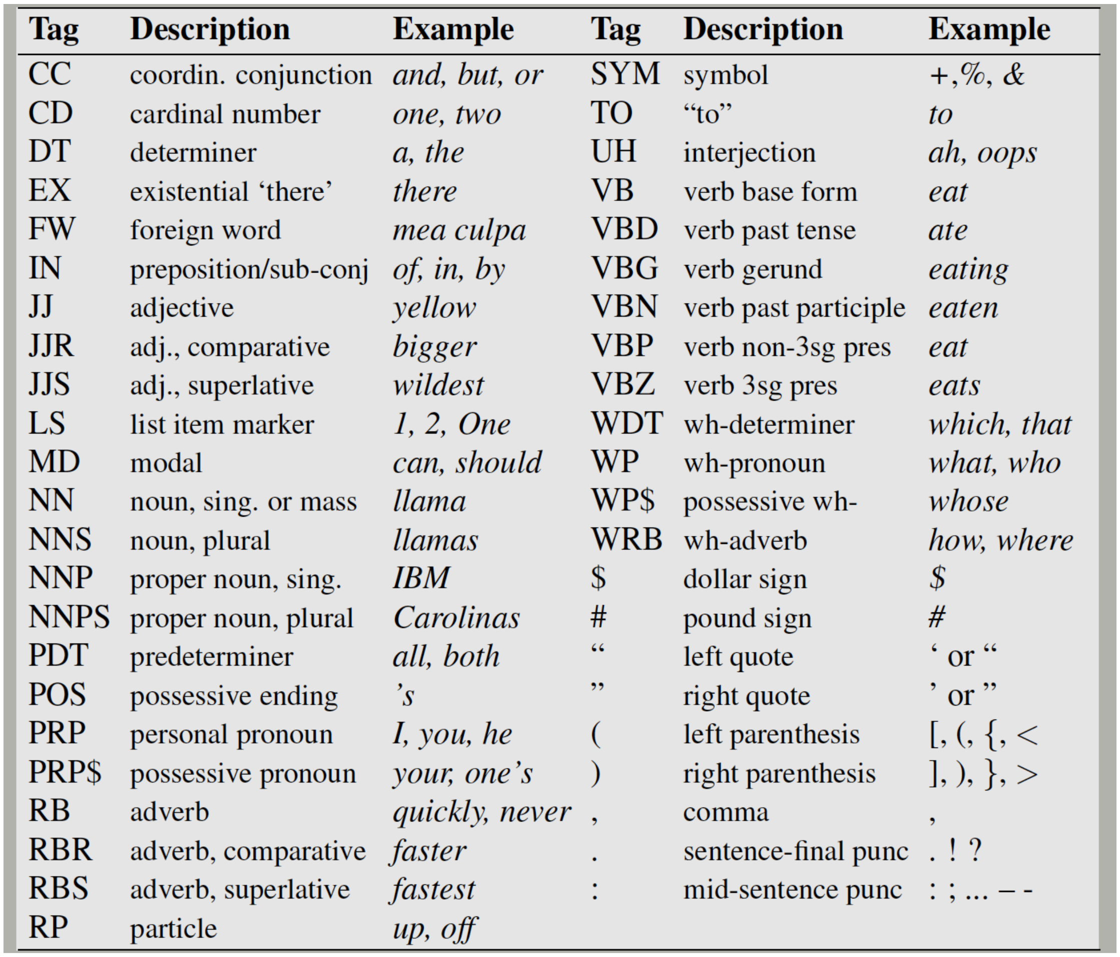
- **Participle vs. Gerund:** The gerund functions as a noun. The participle is more like a verb. A participle is often neighbored by an auxiliary verb (thus, it can look like an adjective).

  - Skiing is my favorite sport. (gerund: subject of the verb 'love')
  - I love skiing! (gerund: object of the verb 'love')
  - I was skiing when you called. (participle: complements auxiliary verb 'was')
  - After skiing, he left. (gerund: object of the preposition 'after')
  - After skiing with her, he left. (participle: complex case. skiing is object of preposition $after$, implying skiing is a noun but since skiing is also followed by a PP and a PP cannot reference a noun, therefore skiing is a verb(participle))
  - After gracefully skiing, he left. (participle: $g$racefully is an adverb and adverbs CANNOT modify a noun, and therefore, skiing cannot be a gerund even though it is the object of the preposition.

  The examples above clearly show that syntax and parts of speech do not concern themselves with the meaning of a sentence only tangentially. They also show that some rules take precedence over others (adverb rule trumps object of preposition noun rule)

- **That vs. that:** *That* can be a conjunction or determiner (and noun?). I thought that/CC I loved that/WDT sport.

## II. Parts-of-Speech tagging



| Tag | Description | Example | Tag | Description | Example |
|---|---|---|---|---|---|
| CC | coordin. conjunction | *and, but, or* | SYM | symbol | +,%, & |
| CD | cardinal number | *one, two* | TO | “to” | *to* |
| DT | determiner | *a, the* | UH | interjection | *ah, oops* |
| EX | existential ‘there’ | *there* | VB | verb base form | *eat* |
| FW | foreign word | *mea culpa* | VBD | verb past tense | *ate* |
| IN | preposition/sub-conj | *of, in, by* | VBG | verb gerund | *eating* |
| JJ | adjective | *yellow* | VBN | verb past participle | *eaten* |
| JJR | adj., comparative | *bigger* | VBP | verb non-3sg pres | *eat* |
| JJS | adj., superlative | *wildest* | VBZ | verb 3sg pres | *eats* |
| LS | list item marker | *1, 2, One* | WDT | wh-determiner | *which, that* |
| MD | modal | *can, should* | WP | wh-pronoun | *what, who* |
| NN | noun, sing. or mass | *llama* | WP$ | possessive wh- | *whose* |
| NNS | noun, plural | *llamas* | WRB | wh-adverb | *how, where* |
| NNP | proper noun, sing. | *IBM* | $ | dollar sign | *$* |
| NNPS | proper noun, plural | *Carolinas* | # | pound sign | *#* |
| PDT | predeterminer | *all, both* | “ | left quote | ‘ or “ |
| POS | possessive ending | *’s* | ” | right quote | ’ or ” |
| PRP | personal pronoun | *I, you, he* | ( | left parenthesis | [, (, {, < |
| PRP$ | possessive pronoun | *your, one’s* | ) | right parenthesis | ], ), }, > |
| RB | adverb | *quickly, never* | , | comma | , |
| RBR | adverb, comparative | *faster* | . | sentence-final punc | . ! ? |
| RBS | adverb, superlative | *fastest* | : | mid-sentence punc | : ; ... – - |
| RP | particle | *up, off* | | | |

Part-of-speech ambiguity arises when a token could be assigned more than one tag. Although only 15% of words in a lexicon are unambiguous (for a sufficiently large corpus), ambiguous tags are roughly 60% of a given corpus. This happens because ambiguous words occur with higher frequency (many of the most common words in english are ambiguous). However, some tokens are rarely assigned one of their possible parts of speech tag. Take *a* for example, it can be used as a noun or an article. The article tag, however, is much, much more likely. This gives rise to the baseline strategy: assign the tag with the highest frequency. This strategy can achieve accuracy rates of over 90%. With more sophisticated algorithms, we can achieve accuracy rates of over 95%.

### HMM tagging

When applying HMM to part-of-speech tagging, we often don’t use the EM algorithm to solve for parameters, rather we use maximum likelihood. This reduces parameter estimation to a counting problem. HMMs for part-of-speech tagging are trained on a fully labeled dataset?a set of sentences with each word annotated with a part-of-speech tag–setting parameters by maximum likelihood estimates on this training data. Thus the only algorithm we will need is the Viterbi algorithm for decoding, and we will also need to see how to set the parameters from training data.

### Decoding: the target equation

The goal of HMM decoding is to choose the tag sequence that is most probable given the observation sequence of $n$ words,

$$\hat{t}_1^n = \text{argmax}_{t_1^n}\ P(t_1^n \mid w_1^n)$$

$$\hat{t}_1^n = \text{argmax}_{t_1^n} \frac{P(w_1^n \mid t_1^n)P(t_1^n)}{P(w_1^n)} \qquad \text{Bayes Rule}$$

$$\hat{t}_1^n = \text{argmax}_{t_1^n} P(w_1^n \mid t_1^n)P(t_1^n) \qquad \text{drop denominator}$$

from here we make two further assumptions:

$$P(w_1^n \mid t_1^n) = \prod_{i=1}^{n} P(w_i \mid t_i) \qquad \text{prob of a word depends only on its own tag}$$

$$P(t_1^n) = \prod_{i=1}^{n} P(t_i \mid t_{i-1}) \qquad \text{bigram assumption for tags}$$

Thus we have,

$$\hat{t}_1^n = \text{argmax}_{t_1^n} \prod_{i=1}^{n} P(w_i \mid t_i)P(t_i \mid t_{i-1})$$

$P(w_i \mid t_i)$ are the emission probabilities. $P(t_i \mid t_{i-1})$ are the transition probabilities.

### Estimating the Probabilities

The maximum likelihood estimate of a transition probability is computed by counting, out of the times we see the first tag in a labeled corpus, how often the first tag is followed by the second. The emission probabilities are computed in a similar manner, counting the number of times a tag is associated with a given word, divided by the number of times the tag appears in the corpus.

$$P(t_i|t_{i-1}) = \frac{C(t_{i-1}, t)}{C(t_{i-1})}$$

$$P(w_i \mid t_i) = \frac{C(t_i, w_i)}{C(t_i)}$$

### Decoding: The Viterbi Algorithm

TODO

### Trigram HMM

TODO

### Unknown words

TODO

### Additional Tricks

TODO

## III. Formal Grammars of English

The fundamental notion underlying the idea of constituency is that of abstraction–groups of words behaving as a single units, or constituents. A significant part of developing a grammar involves discovering the inventory of constituents present in the language. In English, we have noun phrases (NP), verb phrases (VP), prepositional phrases (PP) to name a few.

### Formal Definition of CFG

We define a context-free grammer $G$ as follows:

***N*** a set of **non-terminal symbols** (NP, VP, JJ, etc)
**Σ** a set of **terminal symbols** (house, John, jumping)
***R*** a set of **rules** or productions, each of the form $A \to \beta$, where $A$ is a non-terminal $\beta$ is a string of symbols from the infinite set of strings $(\Sigma \cup N)*$
***S*** a designated **start symbol** and a member of $N$.

We adopt the following convention:

- Capital letters (e.g. *A,B, S)* are non-terminals.
- $S$ is the start symbol
- lower-case Greek letters like $\alpha,\ \beta,\ \gamma$ are strings drawn from $(\Sigma \cup N)^*$.
- Lower case Roman letter like *u, v, w* are strings of terminals.

$$\mathcal{L}_G = \{w \mid w \in \Sigma^* \text{and } S \overset{*}{\Rightarrow} w\}$$

### Sentence-Level Constructions

- **Declarative** S → NP VP
- **Imperative** S → VP
- **Yes-no question** S → Aux NP VP
- **wh-subject-question** S → Wh-NP VP
- **wh-non-subject-question** S → Wh-NP Aux NP VP

We note that S rules are special in the sense that an S constituent is complete. In other words, the verb contains all its arguments. Verbal arguments are defined later, but for now, we can define them to be the subject and object of a verb.

## The Noun Phrase

There are many derivations from a noun phrase. However, the central focus will be the rule *NP → Det Nominal* since it is the most complex. Determiners are words or phrases that precede a noun or noun phrase and serve to express its reference in the context. Among others, *D*et can be filled by simple lexical determiners (a, the, an, some) or by complex possessive expressions such as in "Denver's mayor's mother's canceled flight". We can model this last example with the recursive rule *Det → NP's.* Some nouns, such as mass nouns, require no determiners.

The **Nominal** can derive complex structures.
Base case: *Nominal → Noun.*
Postdeterminer: *Nominal → PostDet Nominal*
Postnominal: *Nominal → Nominal (PP | GerundVP | RelClause)*
*GerundVP → GerundV | GerundV (NP | PP | NP PP)*
(A GerundVP is just a VP that starts with a gerundive form verb. )
*GerundV →* verbs ending in -ing (participle or gerund)
*RelClause → (who | that) VP*
Predeterminer: *NP → PreDet NP* (PreDet: all, some, two, etc.)

## The Verb Phrase

*VP → Verb | Verb (S | VP | NP | NP PP | PP)*
While a verb phrase can have many possible kinds of constituents, not every verb is compatible with every verb phrase. For example, the verb want can be used either with an NP complement (I want a flight ...) or with an infinitive VP complement (I want to fly to...). By contrast, a verb like find cannot take this sort of VP complement (*I found to fly to Dallas). Traditional grammar distinguishes between transitive verbs like find and intransitive verbs like vanish (*I vanished a flight). Modern grammars can have as many as 100 subcategories. We say that a verb like find subcategorizes for an NP, and a verb like want sub categorizes for either an NP or a non-finite VP. We also call these constituents the complements of the verb (hence our use of the term sentential complement above). So we say that want can take a VP complement. These possible sets of complements are called the **subcategorization frame** for the verb.

## Coordination

The major phrase types can all be conjoined with **conjunctions** like *and, or,* and *but.* Since all the major phrase types can be conjoined, it is possible to represent this conjunction fact more generally; a number of grammar formalisms do this using metarules such as the following:

$$X \rightarrow X \text{ and } X$$