Parts of speech in natural language
Informatics 2A: Lecture 16

Adam Lopez

School of Informatics
University of Edinburgh

25 October 2016
1. Word classes and Part of Speech tags

2. Some specific word classes

3. Lexical ambiguity and word frequency

Reading: Jurafsky & Martin, Chapter 5.
Linguists have been classifying words for a long time . . .

- **Dionysius Thrax of Alexandria** (c. 100 BC) wrote a grammatical sketch of Greek involving 8 parts-of-speech:
  - nouns
  - verbs
  - pronouns
  - prepositions
  - adverbs
  - conjunctions
  - participles
  - articles

- Thrax's list and minor variations on it dominated European language grammars and dictionaries for 2000 years.

- Do you notice any important part of speech that’s missing?
Criteria for classifying words

When should words be put into the same class?

Three different criteria might be considered . . .

- **Notional** (or semantic) criteria: What sort of concept does the word refer to? (E.g. nouns often refer to ‘people, places or things’). More problematic: less useful for us.

- **Formal** (i.e. morphological) criteria: What form does the word have? (E.g. -tion, -ize). What affixes can it take? (E.g. -s, -ing, -est).

- **Distributional** criteria: In what contexts can the words occur?

*Walk, slice, donate, and believe* don’t have much in common semantically, but morphological evidence suggests they belong to the same class: they can all combine with suffix -s (marking present tense) or -ed (marking past tense).
For programming languages, a parser typically works entirely with tags produced by the lexer (e.g. IDENT, NUM). It doesn’t care whether an identifier is $x$ or $y$, or whether a numeral is 0 or 5.

**Consequence:** $x$ and $y$ have the same distribution: $x$ can occur wherever $y$ can, and vice versa.

The idea of POS tags is much the same: group the words of a language into classes of words with the same (or similar) distributions. E.g. the words

```
crocodile    pencil    mistake
```

are very different as regards meaning, but grammatically can occur in the same contexts. So let’s classify them all as nouns.

(More specifically, as *singular, countable, common nouns*.)
We can operationalize the idea of distributional equivalence by using tests: can one word substitute for another?

*Kim saw the elephant before we did.*
*Kim saw the movie before we did.*
*Kim saw the mountain before we did.*
*Kim saw the error before we did.*

Tests can be too strict:

* Kim saw the Sam before we did
* Kim arrived the movie before we did

(n.b. In much theoretical linguistics writing, ungrammatical sentences are preceded with *.)
Notionally, nouns generally refer to living things (*mouse*), places (*Scotland*), non-living things (*harpoon*), or concepts (*marriage*).

Formally, *-ness*, *-tion*, *-ity*, and *-ance* tend to indicate nouns. (*happiness, exertion, levity, significance*).

Distributionally, we can examine the contexts where a noun appears and other words that appear in the same contexts. For example, nouns can appear with possession: “his car”, “her idea”.
Notionally, verbs refer to actions (*observe*, *think*, *give*).

Formally, words that end in *-ate* or *-ize* tend to be verbs, and ones that end in *-ing* are often the present participle of a verb (*automate*, *calibrate*, *equalize*, *modernize*; *rising*, *washing*, *grooming*).

Distributionally, we can examine the contexts where a verb appears and at other words that appear in the same contexts, which may include their arguments.

Different types of verbs have different distributional properties. For example, base form verbs can appear as infinitives: “to jump”, “to learn”.
Example of noun and verb classes

Nouns:
- Proper nouns: names such as Regina, IBM, Edinburgh
- Pronouns: he, she, it, they, we
- Common nouns
  - Count nouns: e.g. goat
  - Mass nouns: e.g. snow (? snows)

Verbs
- Base form: walk, live, leave (can combine with to)
- Past tense: walked, lived, left
- Auxiliary: can, could, may, might, must, shall, should, will, would

What about gerunds? (i.e. walking, living, leaving)
What is the part of speech for “walking”? Use linguistic tests.
What is the part of speech for “walking”? Use linguistic tests.
A classic test of verby words is whether they take direct objects.
A classic test of nouny words is appearing with a determiner.
What is the part of speech for “walking”? Use linguistic tests.
A classic test of verby words is whether they take direct objects.
A classic test of nouny words is appearing with a determiner.

Example

The not **observing** this rule is that which the world has blamed in our satirist. (Dryden)
The only mental provision she was making for the evening of life, was the **collecting** and **transcribing** all the riddles of every sort that she could meet with. (Austen)
The difficulty is in the **getting** the gold into Erewhon. (Butler)
What is the part of speech for “walking”? Use linguistic tests.

A classic test of verby words is whether they take direct objects.

A classic test of nouny words is appearing with a determiner.

Example

The not observing this rule is that which the world has blamed in our satorist. (Dryden)
The only mental provision she was making for the evening of life, was the collecting and transcribing all the riddles of every sort that she could meet with. (Austen)
The difficulty is in the getting the gold into Erewhon. (Butler)

Gerunds can also behave as adjectives (an unassuming man) and prepositions (concerning your point).
Notionally, adjectives convey properties of or opinions about things that are nouns (small, wee, sensible, excellent).

Formally, words that end in -al, -ble, and -ous tend to be adjectives (formal, gradual, sensible, salubrious, parlous)

Distributionally, adjectives usually appear before a noun or after a form of be.
Notionally, adverbs convey properties of or opinions about actions or events (*quickly, often, possibly, unfortunately*) or adjectives (*really*).

Formally, words that end in *-ly* tend to be adverbs.

Distributionally, adverbs can appear next to a verb, or an adjective, or at the start of a sentence.
Closed classes

- **pronouns**: I, you, he, she, it, they
- **prepositions**: on, under, over, near, by, at, from, to, with
- **determiners**: a, an, the
- **conjunctions**: and, but, or, as, if, when
- **numerals**: one, two, three, first, second, third
- **particles**: up, down, on, off, in, out, at, by
Closed classes

- **pronouns:** I, you, he, she, it, they
- **prepositions:** on, under, over, near, by, at, from, to, with
- **determiners:** a, an, the
- **conjunctions:** and, but, or, as, if, when
- **numerals:** one, two, three, first, second, third
- **particles:** up, down, on, off, in, out, at, by

“Particle” is the technical term for “we don’t know what the hell this is”

—Bender, 2013
There’s a broad distinction between open and closed word classes:

- **Open classes** are typically large, have fluid membership, and can absorb new words (e.g. *mansplain*).
- Four major open classes are widely found in languages worldwide: *nouns, verbs, adjectives, adverbs*.
  - Virtually all languages have at least the first two.
  - All Indo-European languages (e.g. English) have all four.

- **Closed classes** are typically small, have relatively fixed membership, and the repertoire of classes varies widely between languages. E.g. *prepositions* (English, German), *post-positions* (Hungarian, Urdu, Korean), *particles* (Japanese), *classifiers* (Chinese), etc.

- Closed-class words (e.g. *of, which, could*) often play a structural role in the grammar as *function words*.
In modern (English) NLP, larger (and more fine-grained) tagsets are often used. E.g.


More recently, a “universal” tagset has been developed and applied to dozens of languages: noun, verb, adjective, adverb, pronoun, determiner, adposition, numeral, conjunction, particle, punctuation, and other.

However, there is no tagset that recognizes all relevant distinctions and for which all tags are attested in every language! There is a trade-off between complexity and precision . . . and whatever tagset we use, there will be some words that are hard to classify.
Types of Lexical Ambiguity

Part of Speech (PoS) Ambiguity: e.g., *still*:

1. *adverb*: at present, as yet ("still unknown")
2. *noun*: (1) silence; (2) individual frame from a film; (3) vessel for distilling alcohol ("whisky still")
3. *adjective*: motionless, quiet ("still water")
4. *transitive verb*: to calm ("still the horses")

Sense Ambiguity: e.g., *intelligence*:

1. Power of understanding ("human intelligence")
2. Obtaining or dispersing secret information; also the persons engaged in obtaining or dispersing secret information ("military intelligence")
Buffalo buffalo Buffalo buffalo buffalo buffalo buffalo Buffalo buffalo
Buffalo buffalo Buffalo buffalo buffalo buffalo buffalo buffalo
Bison from Buffalo, which bison from Buffalo bully, themselves bully bison from Buffalo.
Buffalo buffalo Buffalo buffalo buffalo buffalo Buffalo buffalo
Bison from Buffalo, which bison from Buffalo bully, themselves bully bison from Buffalo.

(Buffalo buffalo) ((that) (Buffalo buffalo) buffalo) buffalo (Buffalo buffalo)
Often in reading, we come across unknown words.

And, as in uffish thought he stood,
The Jabberwock, with eyes of flame,
Came whiffling through the tulgey wood,
And burbled as it came!

Even if we don’t know its meaning, formal and distributional criteria help people (and machines) recognize which (open) class an unknown word belongs to.
Example of POS inference

Those *zorls* you *splarded* were *malgy*.

What is the part of speech of the word *malgy*?

1. adverb
2. noun
3. verb
4. adjective
Example of POS inference

The highly-valued share plummeted over the course of the busy week.

Can you decide on the tags of each word?
The highly-valued share plummeted over the course of the busy week.

Can you decide on the tags of each word?

The/ highly-valued/ share/ plummeted/ over/ the/ course/ of/ the/ busy/ week/ .
The highly-valued share plummeted over the course of the busy week.

Can you decide on the tags of each word?

The/DT highly-valued/JJ share/NN plummeted/VBD over/IN the/DT course/NN of/IN the/DT busy/JJ week/NN ./.
Given an input text, we want to tag each word correctly:

The DT grand JJ jury NN commented VBD on IN a DT number NN of IN other JJ topics NNS . /

There EX was VBD still JJ lemonade NN in IN the DT bottle NN . /

(Many Brown/Penn tags are quite counterintuitive!)

- In the above, number and bottle are nouns not verbs — but how does our tagger tell?
- In the second example, still could be an adjective or an adverb — which seems more likely?

These issues lead us to consider word frequencies (among other things).
Ambiguity by part-of-speech tags:

<table>
<thead>
<tr>
<th>Language</th>
<th>Type-ambiguous</th>
<th>Token-ambiguous</th>
</tr>
</thead>
<tbody>
<tr>
<td>English</td>
<td>13.2%</td>
<td>56.2%</td>
</tr>
<tr>
<td>Greek</td>
<td>&lt;1%</td>
<td>19.14%</td>
</tr>
<tr>
<td>Japanese</td>
<td>7.6%</td>
<td>50.2%</td>
</tr>
<tr>
<td>Czech</td>
<td>&lt;1%</td>
<td>14.5%</td>
</tr>
<tr>
<td>Turkish</td>
<td>2.5%</td>
<td>35.2%</td>
</tr>
</tbody>
</table>

Taken from real data for treebanks annotated with their POS tags
Take any corpus of English like the Brown Corpus or Tom Sawyer and sort its words by how often they occur.

<table>
<thead>
<tr>
<th>word</th>
<th>Freq. ($f$)</th>
<th>Rank ($r$)</th>
<th>$f \cdot r$</th>
</tr>
</thead>
<tbody>
<tr>
<td>the</td>
<td>3332</td>
<td>1</td>
<td>3332</td>
</tr>
<tr>
<td>and</td>
<td>2972</td>
<td>2</td>
<td>5944</td>
</tr>
<tr>
<td>a</td>
<td>1775</td>
<td>3</td>
<td>5235</td>
</tr>
<tr>
<td>he</td>
<td>877</td>
<td>10</td>
<td>8770</td>
</tr>
<tr>
<td>but</td>
<td>410</td>
<td>20</td>
<td>8400</td>
</tr>
<tr>
<td>be</td>
<td>294</td>
<td>30</td>
<td>8820</td>
</tr>
<tr>
<td>there</td>
<td>222</td>
<td>40</td>
<td>8880</td>
</tr>
<tr>
<td>one</td>
<td>172</td>
<td>50</td>
<td>8600</td>
</tr>
<tr>
<td>about</td>
<td>158</td>
<td>60</td>
<td>9480</td>
</tr>
<tr>
<td>more</td>
<td>138</td>
<td>70</td>
<td>9660</td>
</tr>
<tr>
<td>never</td>
<td>124</td>
<td>80</td>
<td>9920</td>
</tr>
<tr>
<td>Oh</td>
<td>116</td>
<td>90</td>
<td>10440</td>
</tr>
</tbody>
</table>
Take any corpus of English like the **Brown Corpus** or **Tom Sawyer** and sort its words by how often they occur.

<table>
<thead>
<tr>
<th>word</th>
<th>Freq. ((f))</th>
<th>Rank ((r))</th>
<th>(f \cdot r)</th>
</tr>
</thead>
<tbody>
<tr>
<td>two</td>
<td>104</td>
<td>100</td>
<td>10400</td>
</tr>
<tr>
<td>turned</td>
<td>51</td>
<td>200</td>
<td>10200</td>
</tr>
<tr>
<td>you'll</td>
<td>30</td>
<td>300</td>
<td>9000</td>
</tr>
<tr>
<td>name</td>
<td>21</td>
<td>400</td>
<td>8400</td>
</tr>
<tr>
<td>comes</td>
<td>16</td>
<td>500</td>
<td>8000</td>
</tr>
<tr>
<td>group</td>
<td>13</td>
<td>600</td>
<td>7800</td>
</tr>
<tr>
<td>lead</td>
<td>11</td>
<td>700</td>
<td>7700</td>
</tr>
<tr>
<td>friends</td>
<td>10</td>
<td>800</td>
<td>8000</td>
</tr>
<tr>
<td>begin</td>
<td>9</td>
<td>900</td>
<td>8100</td>
</tr>
<tr>
<td>family</td>
<td>8</td>
<td>1000</td>
<td>8000</td>
</tr>
<tr>
<td>brushed</td>
<td>4</td>
<td>2000</td>
<td>8000</td>
</tr>
<tr>
<td>sins</td>
<td>2</td>
<td>3000</td>
<td>6000</td>
</tr>
</tbody>
</table>
Given some corpus of natural language utterances, the frequency of any word is inversely proportional to its rank in the frequency table (observation made by Harvard linguist George Kingsley Zipf).

Zipf’s law states that: \( f \propto \frac{1}{r} \)

There is a constant \( k \) such that: \( f \cdot r = k \).
Zipf’s law for the Brown corpus
According to Zipf’s law:

- There is a very small number of very common words.
- There is a small-medium number of middle frequency words.
- There is a very large number of words that are infrequent.

(It’s not fully understood why Zipf’s law works so well for word frequencies.)

In fact, many other kinds of data conform closely to a Zipfian distribution:

- Populations of cities.
- Sizes of earthquakes.
- Amazon sales rankings.
Why do we need POS tags?

- They are often an essential ingredient in natural language applications
- Usually appear at the “bottom” of the pipeline
- For example: most of the syntactic variability (we will learn about that later) is determined by the sequence of POS tags in a sentence. POS tags are easier to predict than the full syntax, and therefore, by predicting the POS tags, we pave the way for identification of full phrases: noun phrases, verb phrases, etc.