Parts-of-speech and the Lexicon in Natural Language

Informatics 2A: Lecture 15

John Longley

School of Informatics
University of Edinburgh

22 October 2013

Reading: Jurafsky & Martin, Chapter 5.

Part-of-speech (POS) tags: what are they?

- For each word in our input text, the morphology parsing techniques from Lecture 14 will deliver a sequence of grammatical tags, e.g. N+PL, V+Pres+3SG.
  Here, N and V are tags for *word classes* or *parts-of-speech (POS)* (nouns and verbs).
- In English (a morphologically 'simple' language), only a small number of such sequences are possible. So it's reasonable to collapse each sequence into a single tag: (e.g. NPL, VP3S). We'll sometimes call these 'POS tags' as well.
- We've been using POS tags already. But we haven't yet said much about ...
  - what these tags 'mean',
  - what others there are (and who decides),
  - what are they used for.

Distributional equivalence

Recall that for prog langs, a parser typically works entirely with tags produced by the lexer (e.g. IDENT, NUM). It won'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*.)

Parts of speech in NL grammar

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.
- In modern (English) NLP, larger (and more fine-grained) tagsets are preferred. E.g.

Penn Treebank	45 tags
Brown corpus	87 tags
C7 tagset	146 tags

Trade-off between complexity and precision ... and whatever tagset we use, there'll be some words that are hard to classify.

Criteria for classifying words

When should words be put into the same class?

Three different criteria might be considered ...

- Distributional criteria: Where can the words occur?
- Morphological criteria: What form does the word have? (E.g. -tion, -ize). What affixes can it take? (E.g. -s, -ing, -est).
- 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.

We'll look at various parts-of-speech in terms of these criteria.

Open and closed classes in natural language

There's a broad distinction between open and closed word classes:

- Open classes are typically large, have fluid membership, and are often stable under translation.
- 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.

Nouns

Notionally, nouns generally refer to living things (*mouse*), places (*Scotland*), 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.

```
>>> from nltk.book import *
>>> text2.concordance('happiness')
```

hat sanguine expectation of **happiness** *which is happiness itself*
to inform her confidante , of her **happiness** *whenever she received a letter*
early in life to despair of such a **happiness** *. Why should you be less fortunate*
and it would give me such **happiness** *, yes , almost the greatest*

Verbs

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.

```
>>> from nltk.book import *
>>> text2.concordance(marry') # Where 'marry' appears in S&S
>>> text2.similar(marry') # What else appears in such contexts?
```

Adjectives

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*.

```
>>> from nltk.book import *
>>> text2.concordance('sensible') # Where 'sensible' appears in S&S
>>> text2.similar('sensible') # What else appears in such contexts?
```

Adverbs

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.

```
>>> from nltk.book import *
>>> text2.concordance(highly') # Where 'highly' appears in S&S
>>> text2.similar(highly') # What else appears in such contexts?
```

Importance of formal and distributional criteria

Often in reading, we come across unknown words. (Especially in computing literature!)

bootloader, distros, whitelist, diskdrak, borked
(`http://www.linux.com/feature/150441`)
revved, femtosecond, dogfooding
(`http://hardware.slashdot.org/`)

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.

I really wish mandriva would redesign the diskdrak UI. The orphan bit is borked.

Clicker Question

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

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

1. adverb
2. noun
3. verb
4. adjective

Other Word Classes

Other word classes vary from language to language. English has

- determiners: *the*, *any*, *a*, ...
- prepositions: *in*, *of*, *with*, *without*, ...
- conjunctions: *and*, *because*, *after*, ...
- auxiliaries: *have*, *do*, *be*
- modals: *will*, *may*, *can*, *need*, *ought*
- pronouns: *I*, *she*, *they*, *which*, *where*, *myself*, *themselves*

English doesn't have clitics (like French *l'*) or particles (like Japanese *ga*). Russian lacks standalone reflexive pronouns.

N.B. Functions performed by words in one language may be performed by morphology in another one (e.g. reflexivity in Russian).

The tagging problem

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).

Types of Lexical Ambiguity

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

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

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

1. Power of understanding
2. Obtaining or dispersing secret information; also the persons engaged in obtaining or dispersing secret information

Word Frequency – Properties of Words in Use

Take any corpus of English like the Brown Corpus or Tom Sawyer and sort its words by how often they occur.

word	Freq. (f)	Rank (r)	$f \cdot r$
the	3332	1	3332
and	2972	2	5944
a	1775	3	5235
he	877	10	8770
but	410	20	8400
be	294	30	8820
there	222	40	8880
one	172	50	8600
about	158	60	9480
more	138	70	9660
never	124	80	9920
Oh	116	90	10440

Word Frequency – Properties of Words in Use

Take any corpus of English like the Brown Corpus or Tom Sawyer and sort its words by how often they occur.

word	Freq. (f)	Rank (r)	$f \cdot r$
two	104	100	10400
turned	51	200	10200
you'll	30	300	9000
name	21	400	8400
comes	16	500	8000
group	13	600	7800
lead	11	700	7700
friends	10	800	8000
begin	9	900	8100
family	8	1000	8000
brushed	4	2000	8000
sins	2	3000	6000

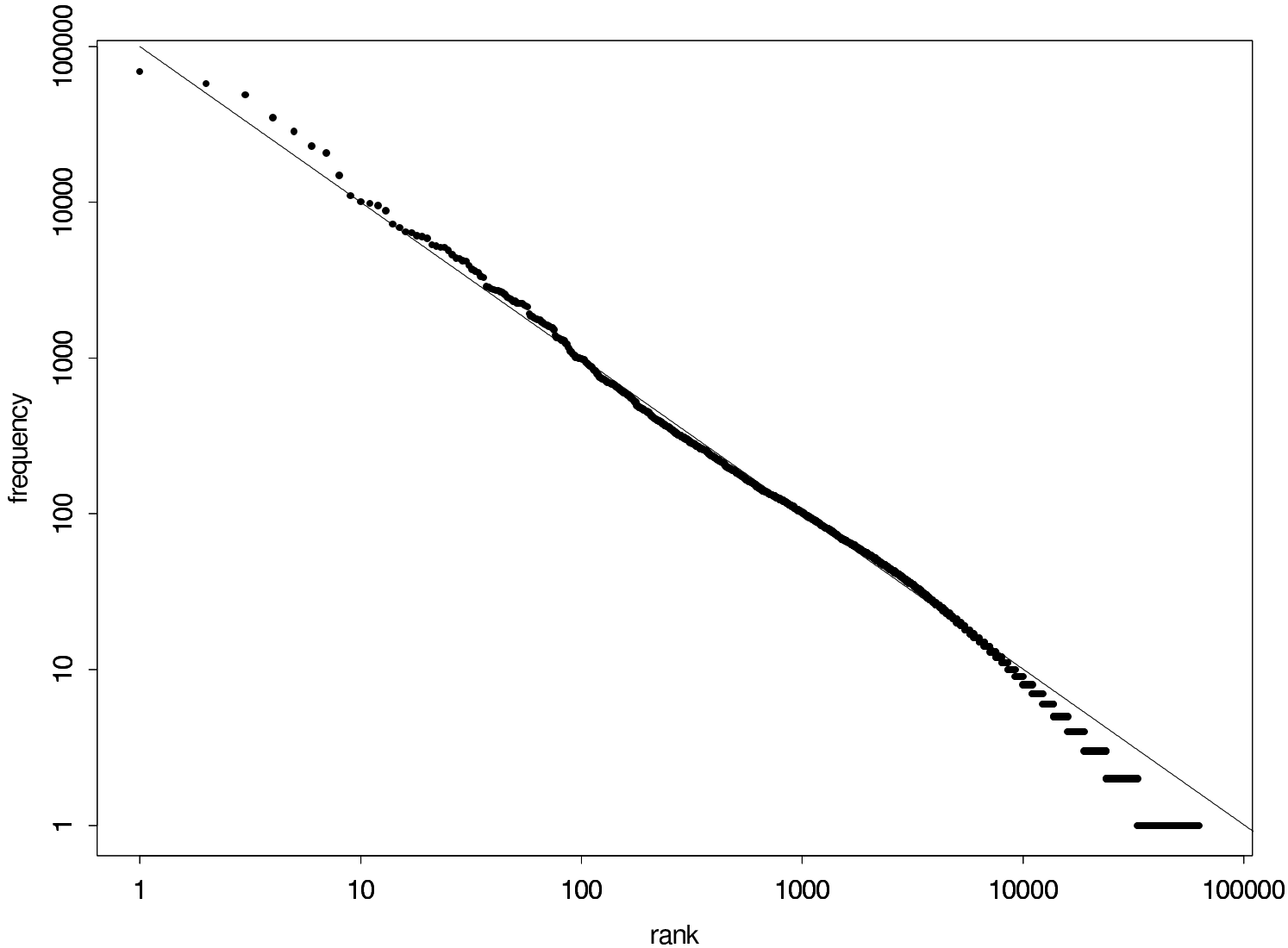
Zipf's law

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



Zipf's law

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.