

Part-of-speech tagging

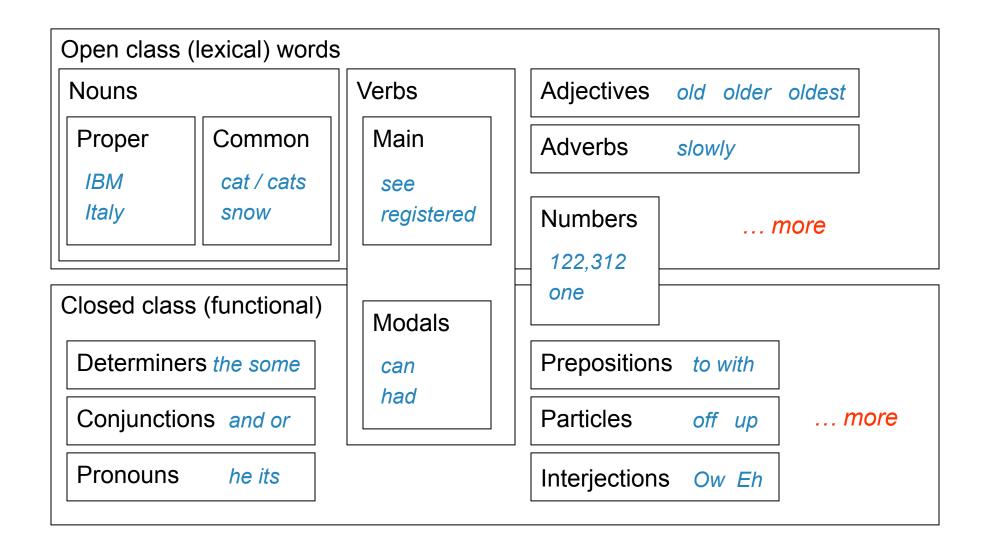
A simple but useful form of linguistic analysis

Christopher Manning



Parts of Speech

- Perhaps starting with Aristotle in the West (384–322 BCE), there was the idea of having parts of speech
 - a.k.a lexical categories, word classes, "tags", POS
- It comes from Dionysius Thrax of Alexandria (c. 100 BCE) the idea that is still with us that there are 8 parts of speech
 - But actually his 8 aren't exactly the ones we are taught today
 - Thrax: noun, verb, article, adverb, preposition, conjunction, participle, pronoun
 - School grammar: noun, verb, adjective, adverb, preposition, conjunction, pronoun, interjection





Open vs. Closed classes

- Open vs. Closed classes
 - Closed:
 - determiners: a, an, the
 - pronouns: she, he, I
 - prepositions: on, under, over, near, by, ...
 - Why "closed"?
 - Open:
 - Nouns, Verbs, Adjectives, Adverbs.



POS Tagging

- Words often have more than one POS: back
 - The <u>back</u> door = JJ
 - On my *back* = NN
 - Win the voters <u>back</u> = RB
 - Promised to <u>back</u> the bill = VB
- The POS tagging problem is to determine the POS tag for a particular instance of a word.



POS Tagging

Input: Plays well with others

Ambiguity: NNS/VBZ UH/JJ/NN/RB IN NNS

Penn Treebank POS tags

- Output: Plays/VBZ well/RB with/IN others/NNS
- Uses:
 - Text-to-speech (how do we pronounce "lead"?)
 - Can write regexps like (Det) Adj* N+ over the output for phrases, etc.
 - As input to or to speed up a full parser
 - If you know the tag, you can back off to it in other tasks



POS tagging performance

- How many tags are correct? (Tag accuracy)
 - About 97% currently
 - But baseline is already 90%
 - Baseline is performance of stupidest possible method
 - Tag every word with its most frequent tag
 - Tag unknown words as nouns
 - Partly easy because
 - Many words are unambiguous
 - You get points for them (the, a, etc.) and for punctuation marks!



Deciding on the correct part of speech can be difficult even for people

- Mrs/NNP Shaefer/NNP never/RB got/VBD around/RP to/TO joining/VBG
- All/DT we/PRP gotta/VBN do/VB is/VBZ go/VB around/IN the/DT corner/NN
- Chateau/NNP Petrus/NNP costs/VBZ around/RB 250/CD



How difficult is POS tagging?

- About 11% of the word types in the Brown corpus are ambiguous with regard to part of speech
- But they tend to be very common words. E.g., that
 - I know that he is honest = IN
 - Yes, that play was nice = DT
 - You can't go that far = RB
- 40% of the word tokens are ambiguous



Part-of-speech tagging

A simple but useful form of linguistic analysis

Christopher Manning