Chapter 5

Words

5.1 Parts of Speech

Parts of speech are categories of words that are, in principle, substitutable for one another in a sentence. They're often thought of as the first level of syntactic structure. Table 5.1 lists the parts of speech used in the very widely-used Penn Treebank. Note that they include traditional parts of speech as well as things like singular and plural.

Many words have a unique POS tag, but some words are ambiguous: for example, *short* can be an adjective (*short vowel*), a noun (direct a *short*), an adverb (to throw a ball *short*) or a verb (to *short* an appliance). Figuring out which POS is the correct one depends on the context, including the POS tags of the neighboring words.

POS tagging is an example of a *sequence labeling* problem, which we'll see more of later in the course. The traditional statistical model for POS tagging was a hidden Markov model, which we saw in the chapter on weighted finite automata. The essential idea is to create a weighted finite automaton whose states are parts of speech, and given a string, to find the sequence of states that accepts the string with the highest probablity.

HMMs were replaced a long time ago by *conditional random fields* (Lafferty, McCallum, and Pereira, 2001). A CRF is a weighted finite automaton whose weights are not required to be probabilities; they're just nonnegative numbers. The weights are learned to maximize the weight of observed tag sequences and minimize the weight of other tag sequences.

At present, the state of the art model for POS tagging (when POS tagging is a separate step, which it usually isn't) is an RNN (specifically, a bidirectional LSTM) with a CRF stacked on top. We will talk about RNN+CRFs in much more detail in a later part of the course!

5.2 Morphology

Morphology is the study of how words are formed out of more basic parts, called *morphemes*, which are defined to be the smallest meaningful part of a word. For example, the word *embiggens* is formed out of several parts:

CC	Coordinating conjunction
CD	Cardinal number
DT	Determiner
EX	Existential there
FW	Foreign word
IN	Preposition or subordinating conjunction
JJ	Adjective
JJR	Adjective, comparative
JJS	Adjective, superlative
LS	List item marker
MD	Modal
NN	Noun, singular or mass
NNS	Noun, plural
NNP	Proper noun, singular
NNPS	Proper noun, plural
PDT	Predeterminer
POS	Possessive ending
PRP	Personal pronoun
PRP$	Possessive pronoun
RB	Adverb
RBR	Adverb, comparative
RBS	Adverb, superlative
RP	Particle
SYM	Symbol
TO	to
UH	Interjection
VB	Verb, base form
VBD	Verb, past tense
VBG	Verb, gerund or present participle
VBN	Verb, past participle
VBP	Verb, non-3rd person singular present
VBZ	Verb, 3rd person singular present
WDT	Wh-determiner
WP	Wh-pronoun
WP$	Possessive wh-pronoun
WRB	Wh-adverb

Table 5.1: Parts of speech in the Penn Treebank.

en- *big* *-en* *-s*

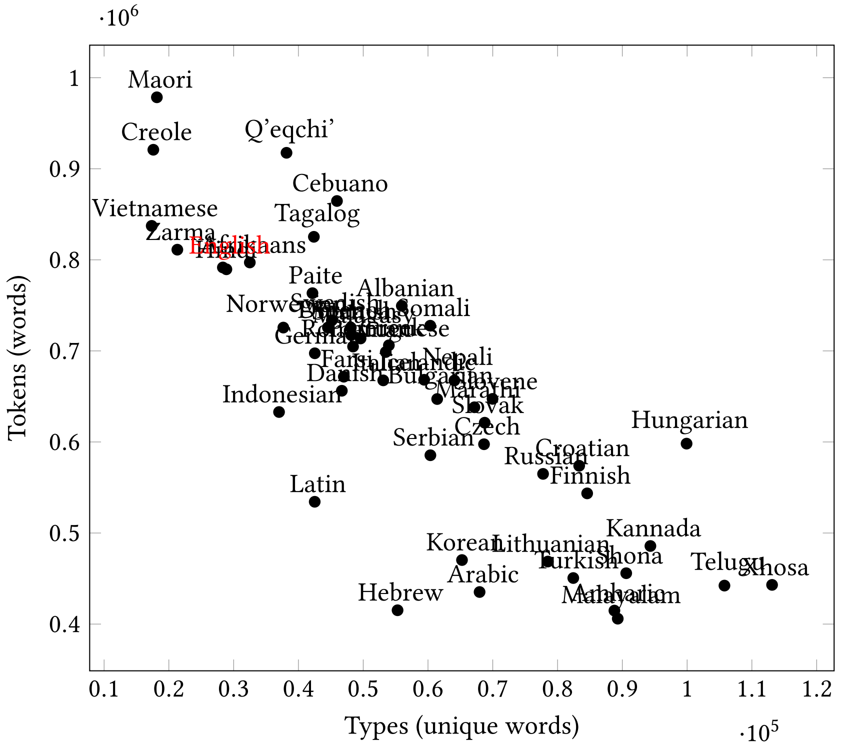
each of which contributes a little bit of meaning to the word. By contrast, the sound *b* in *big* doesn't really have any meaning on its own.

(There are some sounds in English, called *phonesthemes*, that seem to have a tiny bit of meaning. For example, many words having to do with light have a *gl* sound in them (gleam, glimmer, glitter, glow, glare, glint, gloss, etc.). These are generally not considered morphemes, and we won't have anything more to say about them here.)

This section draws heavily on the morphology chapters of Bender's textbook (Bender, 2013).

5.2.1 Why process morphology?

It's perfectly possible to write NLP tools that are ignorant of morphology, just treating each word as an atomic unit. But languages vary very widely in the sizes of their vocabularies. Below is a plot of the number of types and tokens in the Bible (excluding deuterocanonical books) in various languages. No attempt was made at tokenization (not even separating punctuation). Several languages that do not use explicit word boundaries are excluded. Vietnamese might fall into this category as well.



English (written in red) is towards the left end of the chart. But most of the languages here have vocabularies that are much larger than English. Why? Because these other languages have richer morphology that enables them to form more complex and diverse words than English can. And while it’s true that with enough data, computers can learn the meanings of *walk* and *walks* as if they were two unrelated words, this may not be true for languages that have richer morphology and/or less data (that is to say, nearly every language on the planet, except maybe Chinese).

5.2.2 Kinds of morphemes

It’s common to distinguish between *inflectional* and *derivational* morphemes. Inflectional morphemes, like *-s*, indicate features of a word (singular vs. plural, 1st, 2nd, or 3rd person, gender, case, etc.) and/or agree with features of other words. Derivational morphemes change the meaning of a word, like *un-*, and can also change the part-of-speech of a word, like *-en* changes adjectives into verbs.

5.2.3 Morphology in different languages

Analytic vs. synthetic.

Analytic: Mandarin

(5.1) wǒ shòu bu liǎo
I bear not POSSIBLE
‘I can’t bear (it).’

Synthetic: Turkish (Bender, p. 26)

(5.2) dayanamıyorum
dayan- -a- -m- -ıyor- -um
bear POSSIBLE NEGATIVE IMPERFECT 1SG
‘I can’t bear (it).’

Agglutinating vs. fusional.

Agglutinating: Turkish (same as above)

Fusional: Latin (disclaimer: my Latin’s not very good)

(5.3) non possum durare
non pos- -sum dura- -re
not can PRESENT-I bear INFINITIVE
‘I can’t bear (it).’

(5.4) non potui durare
non pot- -ui dura- -re
not can PERFECT-I bear INFINITIVE
‘I couldn’t bear (it).’

Fusional: Hebrew (Bender, p. 12)

(5.5) katav
ktb + *CaCaC*
‘he wrote’

(5.6) hixtiv
ktb + hi*CC*i*C*
‘he dictated (≈ caused to write)’

(5.7) mixtav
ktb + mi*CC*a*C*
‘a letter’

(5.8) ktav
ktb + *CC*a*C*
‘writing, alphabet’