Lec9: Lexical Categories

HUL 242

15/2

Recap

- Began with the study of sounds and then sound systems (phonology)
- Followed by the study of words and word formation (morphology)
- Organization and storage of words in the lexicon
- Move on to units larger than words ...

Arithmetic operations

(1 +2) x 3

More complex expressions are computed in a similar fashion ..

$$((23 \times 5) + 2) / (3 + (9 \times 4))$$
$$(115 + 2) / (3 + 36)$$
$$117 / 39$$
$$3$$

Arithmetic operations

(1 +2) x 3
More complex expressions are computed in a similar fashion ..

$$((23 \times 5) + 2) / (3 + (9 \times 4))$$
$$(115 + 2) / (3 + 36)$$
$$117 / 39$$
$$3$$

- We combine the interpretation of simpler parts to understand the complex expression
- Simpler parts are themselves meaningful units..

Syntax

- We put together words in a similar way: bigger units from smaller ones

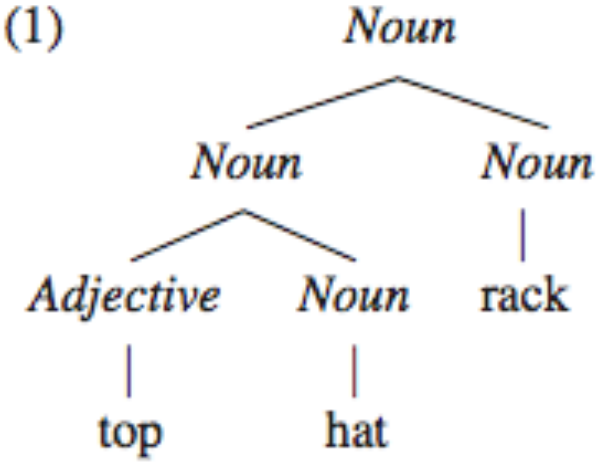
Language

- Principle of compositionality: the meaning of a complex expression depends upon constituent parts– and the rules used to combine them

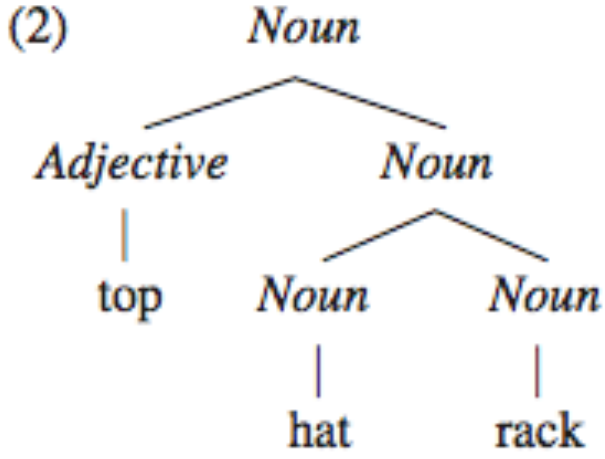
- We have seen this before with morphology and multi-word expressions ...

Structures of compounds

(1)



(2)



Bracketing ambiguity

- Unlike numbers, the structure of sequences can be ambiguous
- ((top hat) rack) or (top (hat rack))
- Often some uncertainty exists–which needs context to disambiguate
- As these elements increase in size, number of possible structures ↑
- 'state highway department public relations director'

Ways to reduce ambiguity

Interpret compounds?

((top hat) rack) or (top (hat rack)) $\rightarrow$ dominant element ?
Other ways to reduce ambiguity?

Ways to reduce ambiguity

Interpret compounds?

((top hat) rack) or (top (hat rack)) $\rightarrow$ dominant element ?
Other ways to reduce ambiguity?

- Place dominant element on the right
- Use prepositions *for* or *of*
- Prosodic information

Early language acquisition

- Evidence comes from children acquiring an understanding of structures
- Children go from one word utterances (1 year) to two words (2 yrs)
- Few prepositions or articles in this speech *mama come*, *Where horse? more cookie*
- But by 3 yrs *Why not he smiling?* or *You lookit his book*

Broca's aphasia

- A lesion/damage to an area of pre-frontal cortex that results in expressive difficulty
- `https://auditoryneuroscience.com/brocas_aphasia`
- A person's speech becomes 'telegraphic': lack of articles, prepositions auxiliaries
- Comprehension is usually intact
- Contrast with Wernicke's Aphasia: syntax is relatively less affected
- Usually taken as evidence for modular organization of language

Preliminaries

- Before we delve into syntax, we need to understand the following:

Key concepts

- Parts of Speech, Notion of constituents
- Idea of a syntactic head
- Idea of a syntactic tree

Parts of speech

- **Categories** like noun, adjective, verb, adverb etc

Definitions use semantic criteria

A noun is a name of a person or object
A verb is an action or a state ..

Parts of speech

'Twas brillig, and the slithy toves
Did gyre and gimble in the wabe:
All mimsy were the borogoves,
And the mome raths outgrabe.

Parts of speech

'Twas/VRB brillig/VBG, and/CC the/DET slithy/ADJ toves/NN
Did/AUX gyre/VRB and/CC gimble/VRB in/PRP the/DET wabe/NN:
All/QT mimsy/ADJ were/VRB the/DET borogoves/NN,
And/CC the/DET mome/ADJ raths/NN outgrabe/VRB.

Parts of speech

'Twas/VRB brillig/VBG, and/CC the/DET slithy/ADJ toves/NN
Did/AUX gyre/VRB and/CC gimble/VRB in/PRP the/DET wabe/NN:
All/QT mimsy/ADJ were/VRB the/DET borogoves/NN,
And/CC the/DET mome/ADJ raths/NN outgrabe/VRB.

- VRB: verb; AUX: auxiliary, CC: conjunction, PRP: Preposition
- ADJ: adjective, DET: Determiner, QT: Quantifier (Pre-determiner)
- NN: noun

How did we do this?

- Use information other than semantic criteria
- What information could this include?

How did we do this?

- Use information other than semantic criteria
- What information could this include?

Distributional criteria

Morphological distribution (affixes on the word)
Syntactic distribution (words in the neighbourhood)

Morphological distribution

- Find inflectional and derivational endings
- These are endings specific to each category
- Examples ?

Morphological distribution

- Find inflectional and derivational endings
- These are endings specific to each category
- Examples ?
  - *-s* plural morpheme for nouns
  - *-ing/-ed* for verbs/gerunds
  - *-est/-er* for adjectives
  - Derivational affixes particular to these lexical categories

Syntactic distribution

- Syntactic distribution: what other words appear near the word

Syntactic distribution

- Syntactic distribution: what other words appear near the word
  - *the* or *a* before nouns
  - *did/will* auxilliaries before verbs
  - *no* before nouns, *not* before verbs
  - Adjectives between determiners and nouns *the nice girl*
  - Adverbs cannot do this e.g. **the quickly girl*

Summary of Parts of Speech

- The two main supertypes are open class and closed class (aka lexical and functional)
- The basic set of parts of speech usually includes the set below:

Open class	**Closed class**
Noun	Determiners (the, a)
Verb	Preposition (of, with)
Adjective	Conjunctions (and, or)
Adverb	Auxiliaries, Modals (do, can)
	Negation (not, no)
	Complementizers (that, if)

Many subtypes of this basic set exist-reflecting more linguistic details E.g. boy/NN, John/PN, He/PRN, himself/POSSPRO, who/WH-PRN

Ambiguity

- I like bread and butter
- She should butter the bread
- I lost a ski in the snow
- I like to ski in the snow

Ambiguity

- I like bread and butter/NN
- She should butter/VRB the bread
- I lost a ski/NN in the snow
- I like to ski/VRB in the snow

Major subtypes of POS

Nouns

three cats but not **three furniture*
Individuation and countability vs. substance or 'stuff'
Mass nouns vs. count nouns
But can I say 'Two waters and a coffee ?'
Other distinctions: proper nouns, common nouns etc

Major subtypes of POS

Nouns

three cats but not **three furniture*
Individuation and countability vs. substance or 'stuff'
Mass nouns vs. count nouns
But can I say 'Two waters and a coffee ?'
Other distinctions: proper nouns, common nouns etc

Determiner

Usually occur before nouns
Articles are a subtype of determiners (e.g. the cat, a dog)
this, that are also determiners

Major subtypes of POS

Conjunctions

Join two clauses, phrases or sentences
Examples *and, or, but* as well as *that*
Compare: *I thought that you might like some milk*
I liked the milk and the cookies

Major subtypes of POS

Conjunctions

Join two clauses, phrases or sentences
Examples *and*, *or*, *but* as well as *that*
Compare: *I thought that you might like some milk*
I liked the milk and the cookies

Auxiliaries

English auxiliaries *be*, *do*, *have*
Modals *can*, *should*,*must*

Other languages

- Often the same basic set of part of speech tags, but other challenges are there.
- Hindi: *vaha* **upar** *so raha tha*
- *vaha* **pahale** *se kamre mein baitha tha*

Other languages

- Often the same basic set of part of speech tags, but other challenges are there.
- Hindi: *vaha* **upar** *so raha tha*
- *vaha* **pahale** *se kamre mein baitha tha*

NST

Noun with spatio-temporal marking
Acts as part of post-positions: *tum ghar ke* **baahar** *baitho*

POS tagging

- Process of automatically assigning a tag, given a particular word
- Corpora labelled with part of speech tags is used a resource
- Disambiguate words having more than one part of speech

Penn tagset for English

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*			

Figure 9.1 Penn Treebank part-of-speech tags (including punctuation).

Ambiguous words in English tagged corpus

that, *back*, *down*, *put*, *set* (Brown and WSJ)

- earnings growth took a back/JJ seat
- a small building in the back/NN
- a clear majority of senators back/VBP the bill (verb-present tense)
- Dave began to back/VB toward the door (verb base)
- enable the country to buy back/RP about debt (particle)
- I was twenty-one back/RB then (adverb)

Part of speech tagging for morphologically rich languages

- Productive word formation processes: large vocabulary size, many types
- Telugu: *raamudu* 'Is it Ram ?' $\rightarrow$ Proper Noun + question (combined POS tag)
- Implies that tagsets can be extremely large (10x or 5x times that of English)
- Chinese : word segmentation problem, this needs to be applied before tagging

References

- This lecture used material from Andrew Carnie's book, Mark Liberman's Ling 001 class. (`http://www.ling.upenn.edu/courses/Fall_2015/ling001/syntax1.html`)
- Some details of POS tagging taken from Jurafsky and Martin, Ch 9, POS tagging