Computational Linguistics

Vocabulary and other essentials from theoretical linguistics

Clayton Greenberg and Stefan Thater

Department of Computational Linguistics, Saarland University

10 May 2016

How to read these slides

Green: important terms

Blue: definitions

Blue $\approx$: approximate definitions

Purple: examples

Red: part-of-speech tags

Words are hard . . .

- Inflection, derivation, compounding
- Inflection does not significantly change meaning
- You must compound/derive before you inflect.
- Stem: writ, lemm, Lemma: write, lemma
- Lemmata may have multiple meanings!

- Lexeme $\approx$ one meaning, abstract form. LEMMA, LEMMA
- Word form $=$ any way to write a lexeme. lemma, lemmata
- Token $=$ a realization of a word form in a corpus. “ These may be considered lemmata about lemmata . ”

Binary classifications

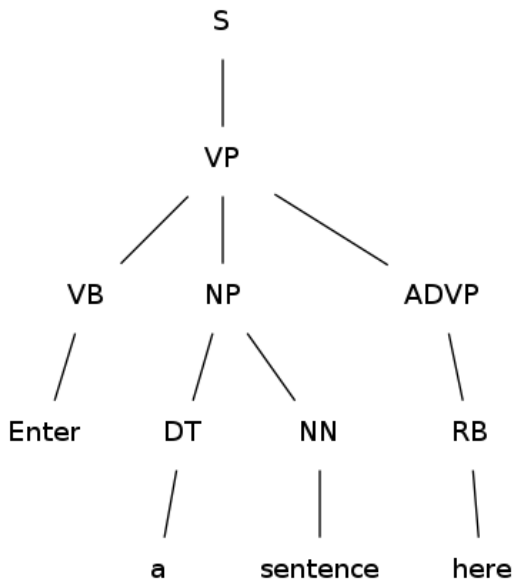
- To handle new or unseen words: find words that are closest.
- The closest words should have the same tag.

- Open class words do allow new words. (e.g. nouns and verbs)
  also known as lexical category words or content words
- Closed class words don't allow new words. (e.g. prepositions)
  also known as functional category words or stop words

- Other suggestions?
- Words can belong to different groups in different contexts!

How to group words

- Words can grouped (clustered) based on:
  Syntax ≈ grammar (tree structure) or Semantics ≈ meaning (*nym)



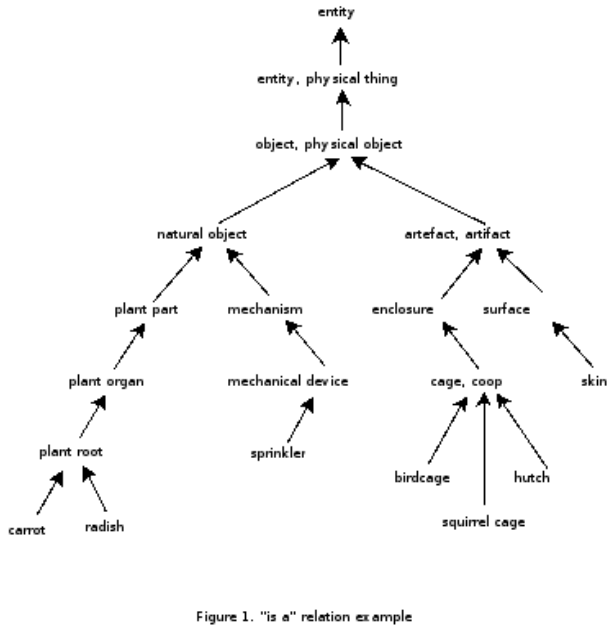


Figure 1. "is a" relation example

- A syntactic category ≈ a semantic *type*

Parts of speech

- Word class: syntactic category, grammatical category, part of speech (POS). Examples: everything in red

- Four basic groups:

Preposition	(IN)	$\{-N, -V\}$	location or θ-role	in, of, with, by
Verb	(VB)	$\{-N, +V\}$	action or state	eat, pray, love
Noun	(NN)	$\{+N, -V\}$	entity	man, Saarland, freedom
Adjective	(JJ)	$\{+N, +V\}$	property	green, furious, coherent

- $-N$ words assign Case (accusative, ergative, instrumental, . . .)
- $+V$ words assign thematic roles (agent, patient, instrument, . . .)

Prepositions

- Prepositions (IN) prototypically express spatial relationships
  The rabbit was {in, on, under, below, above, near} the hat.

- Called postpositions for OV languages ($\approx$)

- Particles (RP) are IN that pair with verbs, forming phrasal verbs
  - They carry little meaning: accuse of, blame for, charge with
  - They can separate from their verbs/objects: I threw my lunch up.

Properties of verbs

- Tense: past, present, future
- Aspect:
  - $+/-$ perfect (have-en)
  - $+/-$ progressive (be-ing)
  - $+/-$ habitual
- Mood/modality: realis, irrealis, interrogative
- Voice: active, passive (be-en), middle
- Agreement: features that match properties of the arguments
- Synthetic forms (affixes) or auxiliaries (periphrastic)
- Modals (MD): shall, should, will, would, can, could, may, might, must

Main forms of a verb (in English)

1. Infinitive (VB), present tense not third singular (VBP): lie
2. Third singular present tense (VBZ): lies
3. Gerund, present participle (VBG): lying
4. Past tense (VBD): lay
5. Past/passive participle (VBN): lain

Properties of nouns and pronouns

- Standard nouns: (NN), pronouns (PP, PRP)
- Number: singular, dual (Arabic), plural (-S)
- Gender: masculine, feminine, neuter
- Case: nominative (PPS), genitive (PP$, PP$$), dative, accusative (PPO), ablative, vocative, locative, instrumental, ergative, reflexive (PPL)
- Person: first, second, third
- $+/-$ animate: squirrel, squash
- $+/-$ human: him (PPO), its (PP$)
- $+/-$ proper: Germany (NNP), country (NN)
- $+/-$ adverbial: downtown (NR), Tudors (NNPS)

Case in English

- Normal nouns in English do not decline (inflect for case).
- Exceptions: Genitives and pronouns:

	Nominative	Accusative	Possessive	2nd Possessive	Reflexive
Tag(s)	PPS (3SG) PPSS (1SG,2SG,PL)	PPO	PP$	PP$$	PPL (PPLS for PL)
1SG	I	me	my	mine	myself
2SG	you	you	your	yours	yourself
3SG MASC	he	him	his	his	himself
3SG FEM	she	her	her	hers	herself
3SG NEUT	it	it	its	its	itself
1PL	we	us	our	ours	ourselves
2PL	you	you	your	yours	yourselves
3PL	they	them	their	theirs	themselves

Adjectives

- Adjectives (JJ) describe properties of nouns
- 2 main positions: Attributive/adnominal or Predicate
- Numbers are subclass of adjectives
  - cardinal (CD): two, three, four
  - ordinal (OD): first, second, third

- Positive = the base form of an adjective, no comparison
- Comparative (-R) = affix denoting the greater of two
- Superlative (-T, -S) = affix denoting the greatest of more than two
- Periphrastic = use a separate word/phrase instead of an affix
  more JJ, most JJ

Determiners and Quantifiers

- Determiners are JJ that identify the referent(s)
  - Demonstratives: this (DT), that (DT), these (DTS), those (DTS)
  - Articles (AT): the (definite), a/an (indefinite)
  - Pre-quantifiers (ABN): many of those, all the feels
  - Interrogative determiners (WDT): what, which

- Quantifiers (DT): all, many, some, any, a/an, most
- Existential there (EX): there once was a man from Nantucket. . .

- Many determiners/quantifiers have a *nominal* version
  - nominal pronouns (PN): one, something, anything, somebody
  - interrogative pronouns: who (WPS), whose (WP$), whom (WPO)

Other parts of speech

- AdveRBs (RB) modify things that are $+V$: often, allegedly
  - Some RB = JJ + *ly*; these specify time, manner, place
  - degree adverbs or qualifiers (QL) modify JJ or RB: very, slightly
- Coordinating Conjunctions (CC) join two equal parts: and, or, but
- Subordinating Conjunctions (CS) join a subordinate to a main
  - {that, for, NULL} are the English complementizers (CS)
  - {because, if, although, before} are CS but not complementizers
- Interjections (UH) interrupt "normal" speech/text: uh, oh, yeah

The optimal tagset size?

2: open, closed

4: IN, VB, NN, JJ

8: IN, VB, NN, JJ, PR, RB, CC, UH

45: Penn Treebank (VBD, VBG, VBN, VBP, VBZ,. . .)

226: Brown Corpus (BEZ, BER, BEDZ, BED, BE, BEN, BEM, BEG, HV, HVZ, HVD, DO, DOZ, DOD, MD,. . .)

- Corpus type?
- Intended task?
- Language dependent?

The substitution test

- Words that can replace *each other* $\longrightarrow$ same POS
- The {big, green, ugly, fat} frog with the warts is on that lily pad.

- Languages with fixed word order have constituents.
- Constituent $\approx$ string that can be replaced by one word
  Determiner, noun phrase, sentence

- The one with the warts is on that lily pad.
- The ugly one is one that lily pad.
- Kermit is on that lily pad.
- . . .

Phrases

- Prepositional Phrase (PP): a preposition with its object
- Verb Phrase (VP): the “predicate” (second half) of a sentence
- Noun Phrase (NP): an entity with all descriptors
- Adjective Phrase (AP): a phrase modifying an entity
- Relative Clause (RC): a sentence (with a gap) modifying an entity

Types of constituents

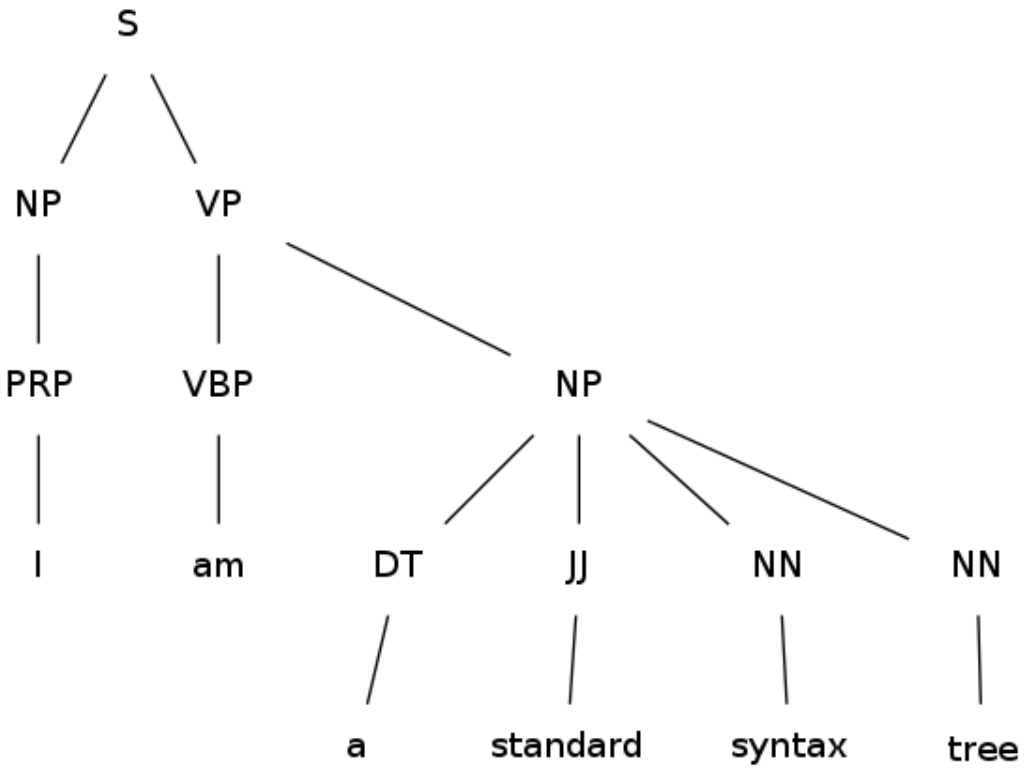
- Head $\approx$ the *important* one on the right side of the rule. **N**P $\rightarrow$ DT **N**N
- Complement $\approx$ merges with the head before the others, usually obligatory. object of a transitive verb
- Adjunct $\approx$ merges with the head after the complement(s), usually optional. JJ
- Specifier $\approx$ merges with the head last. The constituent stops projecting its features after this point. DT? Subject?

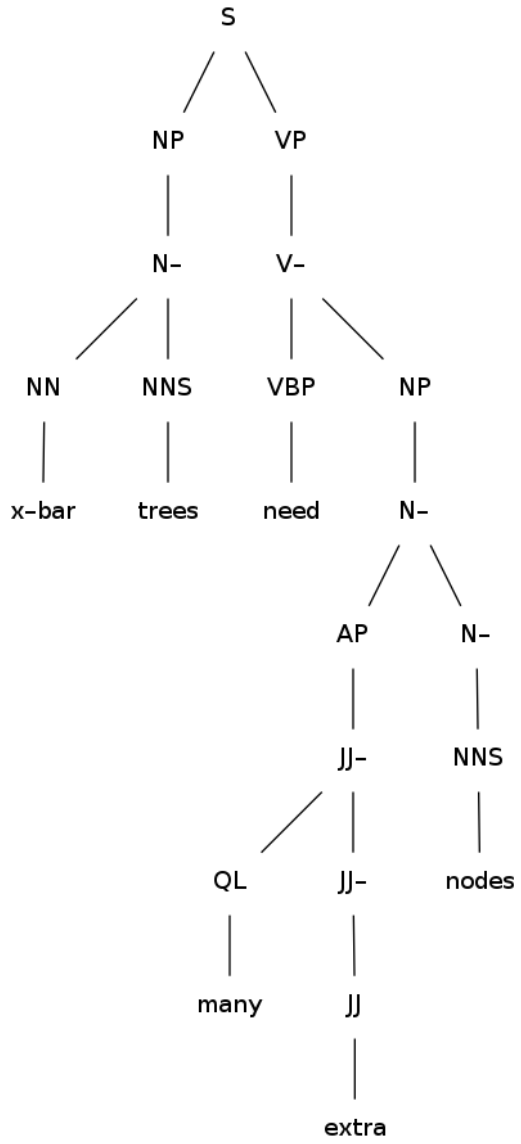
X-bar theory

- Complement (C) Rule: $X' \rightarrow X\ C$
- Adjunct (A) Rule:
  - $X' \rightarrow A\ X'$
  - $X'' \rightarrow A\ X'$
  - $X' \rightarrow X'\ A$
  - $X'' \rightarrow X'\ A$
- Specifier (Spec) Rule: $XP \rightarrow Spec\ X'$

- Why should *computational* linguists care?
  - It adds more symbols to the grammar.
  - It more precisely controls recursion.
  - It encourages unary and binary branching.

X-bar versus “standard”





Syntactic versus semantic relations

Case	Function	Thematic Role
Nominative	Subject	Agent
Genitive	Possessor	Source
Dative	Indirect object	Goal
Accusative	Direct object	Patient
Ablative	Prepositional object	Theme
		Location
		Instrument

Bridge to context-free grammars

- Chomskyan Grammar $=$ Lexicon $+$ Computational System
- Grammar in Chomsky Normal Form: A $\rightarrow$ *a*, A $\rightarrow$ B C

- Preterminals $=$ the symbols that can be rewritten as words
- Set of preterminals $=$ POS tagset

- In other formalisms, the lexicon may contain more or less information (features).

- To play with a parser, see `http://eztreesee.coli.uni-saarland.de/`