



Linguistics Essentials

Instructor: Rada Mihalcea

Note: most of the material in this slide set was adapted from an NLP course taught by J. Hajic at Johns Hopkins University

-
- Levels of formal description
 - Linguistic categories
 - Words, phrases, sentences

The Description of Language

Language = Words and Rules

→ Dictionary (vocabulary) + Grammar

Dictionary

set of words defined in the language
open (dynamic)

Traditional

paper based

Electronic

machine readable dictionaries; can be obtained from paper-based

Grammar

set of rules which describe what is allowable in a language

Classic Grammars

meant for humans who know the language

definitions and rules are mainly supported by examples

no (or almost no) formal description tools; cannot be programmed

Explicit Grammar (CFG, Dependency Grammars, Link Grammars,...)

formal description

can be programmed & tested on data (texts)

Levels of (Formal) Description

6 basic levels (more or less explicitly present in most theories):

↑ and beyond (pragmatics/logic/...)

meaning (semantics)

(surface) syntax

morphology

phonology

phonetics/orthography

Each level has an input and output representation

output from one level is the input to the next (upper) level

sometimes levels might be skipped (merged) or split

Phonetics/Orthography

Input:

acoustic signal (phonetics) / text (orthography)

Output:

phonetic alphabet (phonetics) / text (orthography)

Deals with:

Phonetics:

consonant & vowel (& others) formation in the vocal tract

classification of consonants, vowels, ... in relation to frequencies, shape &

position of the tongue and various muscles

intonation

Orthography: normalization, punctuation, etc.

Phonology

Input:

sequence of phones/sounds (in a phonetic alphabet); or “normalized” text (sequence of (surface) letters in one language’s alphabet) [NB: phones vs. phonemes]

Output:

sequence of phonemes (~ (lexical) letters; in an abstract alphabet)

Deals with:

relation between sounds and phonemes (units which might have some function on the upper level)

e.g.: [u] ~ oo (as in book), [æ] ~ a (cat); i ~ y (flies)

Phonology

(Surface « Lexical) Correspondence

“symbol-based” (no complex structures)

Ex.: (stem-final change)

lexical: **b a b y + s** (+ denotes start of ending)

surface: **b a b i e s** (phonetic-related: **bébi0s**)

German (umlaut) (satz ~ sentence)

lexical: **s A t z + e** (A denotes “umlautable” a)

surface: **s ä t z e** (phonetic: **zæcʌ**, vs. **zac**)

Turkish (vowel harmony)

lexical: **e v + l A r** (~house)

surface: **e v l e r**

Morphology

Input:

sequence of phonemes (~ (lexical) letters)

Output:

sequence of pairs (lemma, (morphological) tag)

Deals with:

composition of phonemes into word forms and their underlying lemmas
(lexical units) + morphological categories (inflection, derivation,
compounding)

e.g. quotations ~ quote/V + -ation(der.V->N) + NNS.

Morphology: Morphemes & Order

Handles what is an *isolated form* in written text

Grouping of phonemes into morphemes

sequence **deliverables** ~ deliver, able and s (3 ***units***)

could as well be some “ID” numbers:

e.g. deliver ~ 23987, s ~ 12, able ~ 3456

Morpheme Combination

certain combinations/sequencing possible, other not:

deliver+able+s, but not able+derive+s; noun+s, but not noun+ing
typically fixed (in any given language)

Morphology: From Morphemes to Lemmas & Categories

Lemma: lexical unit, “pointer” to lexicon

might as well be a number, but typically is represented as the “base form”, or “dictionary headword”

possibly indexed when ambiguous/polysemous:

state¹ (verb), state² (state-of-the-art), state³ (government)

from one or more morphemes (“root”, “stem”, “root+derivation”, ...)

Categories:

small number of possible values (< 100, often < 5-10)

(Surface) Syntax

Input:

sequence of pairs (lemma, (morphological) tag)

Output:

sentence structure (tree) with annotated nodes (all lemmas, (morphosyntactic) tags, functions), of various forms

Deals with:

the relation between lemmas & morphological categories and the sentence structure

uses syntactic categories such as Subject, Verb, Object,...

e.g.: I/PP1 see/VB a/DT dog/NN ~

((I/sg)SB ((see/pres)V (a/det dog/sg)OBJ)VP)S

Syntax: Representation

Tree structure (“tree” in the sense of graph theory)
one tree per sentence

Two main ideas for the shape of the tree:

phrase structure (~ derivation tree, cf. parsing later)

using bracketed grouping

brackets annotated by phrase type

heads (often) explicitly marked

dependency structure (lexical relations “local”, functions)

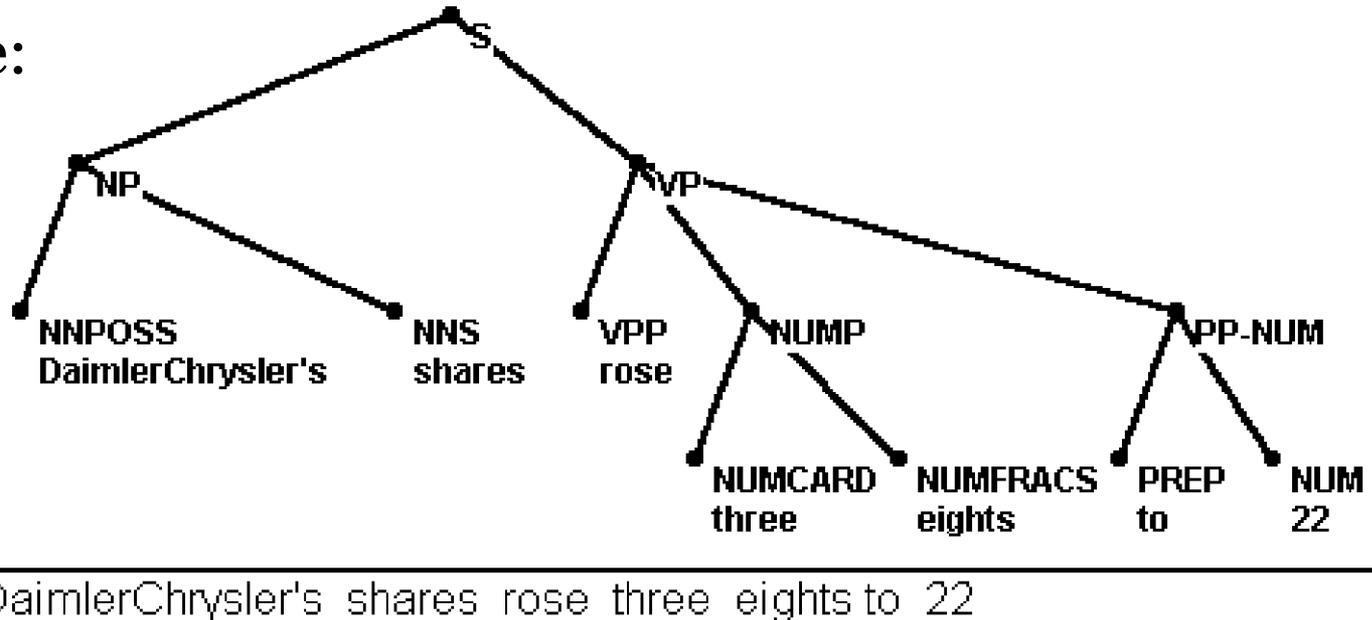
basic relation: head (governor) - dependent

links (edges) annotated by syntactic function (Sb, Obj, ...)

phrase structure: implicitly present

Syntax: Phrase Structure Tree

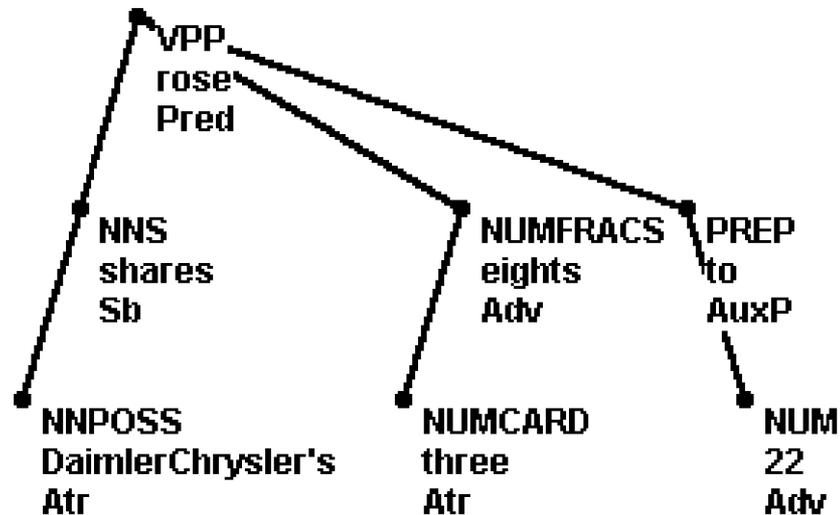
Example:



$((\text{DaimlerChrysler's shares})_{\text{NP}} (\text{rose } (\text{three eights})_{\text{NUMP}} (\text{to } 22)_{\text{PP-NUM}})_{\text{VP}})_{\text{S}}$

Syntax: Dependency Tree

Example:



DaimlerChrysler's shares rose three eights to 22

$rose_{Pred}(\text{shares}_{Sb}(\text{DaimlerChrysler's}_{Atr}), \text{eights}_{Adv}(\text{three}_{Atr}), \text{to}_{AuxP}(22_{Adv}))$

Meaning (semantics)

Input:

sentence structure (tree) with annotated nodes (lemmas, (morphosyntactic) tags, surface functions)

Output:

sentence structure (tree) with annotated nodes (semantic lemmas, (morpho-syntactic) tags, deep functions)

Deals with:

relation between categories such as “Subject”, “Object” and (deep) categories such as “Agent”, “Effect”; adds other categories

e.g. ((I)SB ((see)V (Tom)OBJ)VP)S ~

(I/Sg/Pat (see/Perf/Pred) Tom/Sg/Ag)

...and Beyond

Input:

sentence structure (tree): annotated nodes (autosemantic lemmas, (morphosyntactic) tags, deep functions)

Output:

logical form, which can be evaluated (true/false)

Deals with:

assignment of objects from the real world to the nodes of the sentence structure

e.g.: (I/Sg/Pat (see/Perf/Pred) Tom/Sg/Ag) ~

see(Mark-Twain[SSN:...],Tom-Sawyer[SSN:...])[Time:bef 99/9/27/14:15][Place:39§19'40"N76§37'10"W]

-
- Levels of formal description
 - Linguistic categories
 - Words, phrases, sentences

The Categories: Part of Speech: Open and Closed Categories

Part of Speech - POS (pretty much stable set across languages)
morphological “behavior” is typically consistent within a POS category

Open categories: (“open” to additions)

verb, noun, pronoun, adjective, numeral, adverb

subject to inflection (in general); subject to cross-category derivations

newly coined words always belong to open POS categories

potentially unlimited number of words

Closed categories:

preposition, conjunction, article, interjection, particle

not a base for derivation (possibly only by compounding)

finite and (very) small number of words

The Categories: Part of Speech, Open Categories: Nouns

Nouns: typically refer to entities

Inflection:

| | |
|--------|--|
| number | singular, plural |
| gender | feminine, masculine, neuter |
| case | nominative, genitive, accusative, dative, vocative |

semantic classification:

human/animal/(non-living) things: driver/bird/stone

concrete/abstract: computer/thought

common/proper: table/Microsoft

syntactic classification: countable/uncountable: book, water

morphological classification:

pluralia/singularia tantum: data (is), police (are)

“adverbial” nouns: afternoon, home, east (no inflection)

The Categories: Part of Speech, Open Categories: Verbs

Verbs:

Inflectional:

| | |
|----------------|--|
| subject number | singular, plural |
| subject person | first (<i>I</i> read), second (<i>you</i> read), ... |
| tense | present tense, past tense ... |
| aspect | progressive, perfect |
| modality | possibility, ... |
| voice | active, passive |

syntactic/semantic: classification:

ordinary: (to) speak, (to) write
auxiliaries: be, have, will, would, do, go (going)
modals: can, could, may, should, must, want
phasal: begin, end, start

morphological classification

conjugation type: regular/irregular, (Ge.: weak/strong/irregular)
conjugation class: (e.g. Italian: -are, -ere, -ire ...)

The Categories: Part of Speech, Open Categories: Pronouns

Pronouns:

Inflectional: number, person, gender, case
much like nouns (syntactic usage also similar)
(pro)noun ~ “stands for” a noun

classification (mostly syntactic/semantic):

personal: I, you, she, he, it, we, you, they

demonstrative: this, that

possessive: my, your, her, his, its, our, their; mine, yours, ours,...

reflexive: myself, yourself, herself, ..., oneself

interrogative: what, which, who, whom, whose, that

indefinite (“nominal”): somebody, something, one

The Categories: Part of Speech, Open Categories: Adjectives

Adjectives: describe properties of nouns

Inflectional: degree of comparison (comparative/superlative),
number, gender, case

classification:

ordinary: new, interesting, [test (equipment)]

possessive: John's, driver's

proper: Appalachian (Mountains)

often derived from verbs/nouns: teaching (assistant), trendy, stylish

morphological classification

degrees of comparison (En.: big, bigger, biggest)

usually requires agreement with the noun

The Categories: Part of Speech, Open Categories: Adverbs

Adverbs: modify a verb, and specify place, time, manner, degree

Inflectional: degree of comparison

derivation from adjectives is common:

new → newly, interesting → interestingly

non-derived adverbs:

ordinary: so, well, just, too, then, often, there

wh-adverbs (interrogative): why, when, where, how

degree adverbs/qualifiers: very, too

morphological classification (not much, really...)

degree of comparison: well, better, best

soon, sooner

The Categories: Part of Speech, Open Categories: Numerals

Numerals: used to indicate numbers

inflectional: number, gender, case, negation

open (infinite?) category: compounding (Ge.: einundzwanzig, 21)

classification:

cardinals: one, five, hundred

NB: million etc. often considered noun

ordinals: first, second, thirtieth

quantifiers: all, many, some, none

multiplicative: times, twice

multilateral: single, triple, twofold

morphological classification: as nouns/adjectives; many irregulars

The Categories: Part of Speech, Closed Categories

Closed categories: preposition, conjunction, article, interjection, clitic, particle

Morphological behavior: indeclinable

preposition: of, without, by, to;

conjunction:

 coordinating: and, but, or, however

 subordinating: that, if, because, before, after, although, as

Article (determiner): a, an, the

interjection: wow, eh, hello;

clitic: 's; may be attached to whole phrases (at the end)

particle: yes, no, not; to (+verb);

 many (otherwise) prepositions if part of phrasal verbs, e.g. (look) up

The Categories: Number and Gender

Grammatical Number: Singular, Plural

nouns, pronouns, verbs, adjectives, numerals

computer / computers; (he) goes / (they) go

In some languages (Arabic): Dual (nouns, pronouns, adjectives)

Grammatical Gender: Masculine, Feminine, Neuter

nouns, pronouns, verbs, adjectives, numerals

he/she/it;

nouns: (mostly) do not change gender for a single lexical unit

Also: animate/inanimate (gram., some genders), etc.

Mädchen (Ge.; girl, neuter); děti (Cz.; children, masc. inanim.)

The Categories: Case

Case

English: only personal pronouns/possessives, 2 forms

other languages: 4 (German), 6 (Russian), 7 (Czech,Slovak,...), 5 (Romanian)
nouns, pronouns, adjectives, numerals

most common cases (forms in singular/plural)

| | | |
|--------------|--------------------|--------------|
| nominative | I/we (work) | eu/noi (Ro) |
| genitive | (picture of) me/us | a mea/al meu |
| dative | (give to) me/us | mie |
| accusative | (see) me/us | pe mine |
| vocative | you! | tu! |
| locative | (about) me/us | (Czech) |
| instrumental | (by) me/us | (Czech) |

The Categories: Person, Tense

Person

verbs, personal pronouns

1st, 2nd, 3rd: (I) go, (you) go, (he) goes; (we) go, (you) go, (they) go

 merg, mergi, merge mergem mergeti merg (Ro)

Tense

past:

(you) went

(Ro)

ai mers

(Pol.: go)

szliście

present:

(you pl.) go

mergeti

idziecie

future (!if not “analytical”)

(you) will go

veti merge

-

concurrent (gerund)

going

mergind

idąc

Note on Tense

Examples of (traditional) tense:

infinitive: (to) write (tenseless, personless, ..., except negation (Cz.))

simple present/past: (I) write/(she) writes; (I,she) wrote

progressive present/past: (I) am writing; (I) was writing

perfect present/past: (I) have written; (I) had written

all in passive voice, too:

(the book) is being/has been/had been written etc.

all in conditional mood, too (mood: in Eng. not a morph. category)

(the book) would have been written

The Categories: Voice & Aspect

Voice

active vs. passive

(I) drive / (I am being) driven

(Ich) setzte (mich) / (Ich bin) gesetzt (Ge.: to sit down)

Aspect

imperfective vs. perfective:

покупал / купил (Ru.: I used to buy, I was buying) / I (have) bought)

imperfective continuous vs. iterative (repeating)

spal / spával (Cz.: I was sleeping / I used to sleep (every ...))

The Categories: Negation, Degree of Comparison

Negation:

even in English: impossible (~ not possible)

Cz: every verb, adjective, adverb, some nouns; prefix *ne-*

It: some adjectives: irregular negation (s-, non)

Degree of Comparison (non-analytical):

adjectives, adverbs:

positive (big), comparative (bigger), superlative (biggest)

Pol.: (new) nowy, nowszy, najnowszy

Combination (by prefixing):

order? both possible: (neg.: Cz./Pol.: *ne-/nie-*, sup.: nej-/naj-)

Cz.: nejnemožnější (the most impossible)

Pol.: nienajwierniejszy (the most unfaithful)

Typology of Languages

By morphological features

Analytical: using (function) words to express categories

English, also French, Italian, ..., Japanese, Chinese

I would have been going ~ (Pol.) szłabym

Inflective: using prefix/suffix/infix, combines several categories

Slavic: Czech, Russian, Polish,... (not Bulgarian); also French, German; Arabic

Latin/Slavic: Romanian

(Cz. new(acc.)) nov**ou** (Adj, Fem., Sg., Acc., Non-neg., Pos.)

Agglutinative: one category per (non-lexical) morpheme

Finnish, Turkish, Hungarian

(Fin. plural): -i-

Categories & Tags

Tagset:

list of all possible combinations of category values for a given language

typically string of letters & digits:

compact system: short idiosyncratic abbreviations:

NNS (gen. noun, plural)

positional system: each position i corresponds to C_i :

AAMP3----2A---- (gen. Adj., Masc., Pl., 3rd case (dative), comparative (2nd degree of comparison), Affirmative (no negation))

tense, person, variant, etc.: N/A (marked by “empty position”, or ‘-’)

Famous tagsets: Brown, Penn, Multext[-East], ...

The Dictionary (or Lexicon)

Repository of information about words:

Morphological:

description of morphological “behavior”: inflection patterns/classes

Syntactic:

Part of Speech

relations to other words:

subcategorization (or “surface valency frames”)

Semantic:

semantic features

frames

...and any other! (e.g., translation)

-
- Levels of formal description
 - Linguistic categories
 - Words, phrases, clauses, sentences

Words, Phrases, Clauses, Sentences

Words

smallest units on the syntax level
function/semantic

Phrases

consist of words and/or phrases; “constituents”

Clauses

have predicative meaning (single predicate)

Sentences

consist of clauses (one or more)

Words

Words

lexical units

auxiliary (function) words: have grammatical function

have meaning

idioms

fixed phrases (non-compositional) “hot dog”, “kick the bucket”

Relate to other words

dictionary: repository of information for each words about its
(idiosyncratic) relations to other words

Phrases

Phrases

sequences of words and/or phrases (i.e. of constituents)
may be discontinuous, sometimes

Types of Phrases:

Simple/Clausal (i.e. clauses, which consist of phrases, behave like phrases... recursively!)

According to head type:

Noun phrase: a new book

Adjective phrase: brand new

Adverbial phrase: so much

Prepositional phrase: in a class

Verb phrase: catch a ball

Noun Phrases

Head: noun

water

a book

new ideas

that small village

The greatest rise of interest rates since W.W.II within a single year
an operating system which, despite great efforts on the part of our
administrators, fails all too often

Adjective Phrases

Head: adjective

Simple APs very common, complex APs rare

old

very old

really very old

five times older than the oldest elephant in our ZOO

(was) sure, as far as I know, to be there first

Adverbial and Numerical Phrases

Head: adverb

three times as much

quickly

really

(... speaks) more loudly than anybody could imagine

yesterday

Numerical Phrases

(... lasted) three hours

twenty-two

Prepositional Phrases

Head: preposition

In fact, play the role of Adverbial Phrases often

in the City

at five o'clock

to a brightest future

without a glitch

to the point where neither of them could get out of it

up to five points

instead of Charles

Verb Phrases

Head: verb

(It) rains

... could ever see a large Unidentified Flying Object

..., why (we) have got so much rain

Please!

On Sunday, (he) was driven to the hospital

(It) began to snow

(...) prohibits smoking in this area

Coordination of Phrases

“Head”: conjunction, punctuation

and, or, but

cats and dogs

new or even newer

quickly and precisely

he came to the conclusion that it makes no sense to hide himself anymore and
therefore we could hear him today

(flights) from and to Dallas

eat your lunch now or at the picnic table

Clauses

Predicative function:

some activity of some subjects/objects, somewhere in time, under certain circumstances

Main clause

not part of a greater clause

Embedded clause

part of other clause, having some function (like a phrase)

A tile falling from the roof nearly killed him.

He fell asleep while listening to the news.

Function of a Clause

same as for phrase, plus some (direct speech etc.)

Sentences

Consist of a single or several main clauses

If several main clauses:

coordination, much like coordinated phrases

more coordinating conjunctions:

and, or, but, (and) therefore, ...

In written text, starts with a capital letter

Ends by period/question mark/exclamation mark

not all periods end a sentence! – example?

Sometimes even semicolon (;) might be a sentence break (...vague)