

Introduction to  
Natural Language Processing (600.465)

Linguistic Essentials:  
Phonology and Morphology

Dr. Jan Hajiè

CS Dept., Johns Hopkins Univ.

`hajic@cs.jhu.edu`

`www.cs.jhu.edu/~hajic`

# The Description of Language

- Grammar
  - set of rules which describe what is allowable in a language
- Classic Grammars (Quirk et al.)
  - 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, LFG, GPSG, HPSG, 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 in the v.t.
    - 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)

# 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.

# (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 & morph. 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/ind dog/sg)OBJ)VP)S

# Meaning (semantics)

- Input:
  - sentence structure (tree) with annotated nodes (lemmas, (morphosyntactic) tags, surface functions)
- Output:
  - sentence structure (tree) with annotated nodes (autosemantic lemmas, (morphosyntactic) tags, deep functions)
- Deals with:
  - relation between categories such as “Subject”, “Object” and (deep) categories such as “Agent”, “Effect”; adds other cat’s
  - e.g. ((I)SB ((was seen)V (by Tom)OBJ)VP)S ~  
(I/Sg/Pat/t (see/Perf/Pred/t) Tom/Sg/Ag/f)



## ...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/t (see/Perf/Pred/t) Tom/Sg/Ag/f) ~  
`see(Mark-Twain[SSN:...],Tom-Sawyer[SSN:...])`[Time: bef 99/9/27/14:15] [Place: 39°19'40"N76°37'10"W]

# 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: bĕbī0s*)
- Arabic: (interfixing, inside-stem doubling) (lit. ‘read’)
  - lexical: **kTb+uu+CVCCVC** (*CVCC...vowel/consonant pattern*)
  - surface: **kut tub**

# Phonology Examples

- German (umlaut) (satz ~ sentence)
  - lexical: s A t z + e (*A denotes “umlautable” a*)
  - surface: s ä t z e (*phonetic: zæcə, vs. zac*)
- Turkish (vowel harmony)
  - lexical: e v + l A r (←houses) b a ş + l A r
  - surface: e v l e r (heads→) b a ş l a r
- Czech (e-insertion & palatalization)
  - lexical: m a t E K + 0 (←mothers/gen.) m a t E K + ě
  - surface: m a t e k (mother/dat.→) m a t \_ c e

# 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:
      - state<sup>1</sup> (verb), state<sup>2</sup> (state-of-the-art), state<sup>3</sup> (government)
  - from one or more morphemes (“root”, “stem”, “root+derivation”, ...)
- Categories: non-lexical
  - small number of possible values (< 100, often < 5-10)

# Morphology Level: The Mapping

- Formally:  $A^+ \rightarrow 2^{(L, C_1, C_2, \dots, C_n)}$ 
  - A is the alphabet of phonemes ( $A^+$  denotes any non-empty sequence of phonemes)
  - L is the set of possible lemmas, uniquely identified
  - $C_i$  are morphological categories, such as:
    - grammatical number, gender, case
    - person, tense, negation, degree of comparison, voice, aspect, ...
    - tone, politeness, ...
    - part of speech (not quite morphological category, but...)
  - $2^{(L, C_1, C_2, \dots, C_n)}$  denotes the power set of  $(L, C_1, C_2, \dots, C_n)$
  - A, L and  $C_i$  are obviously language-dependent

# 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
    - valency frames
  - ...and any other! (e.g., translation)

# The Categories: Part of Speech:

## Open and Closed Categories

- Part of Speech - POS (pretty much stable set across languages)
  - not so much morphological (can be looked up in a dictionary), but:
  - 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, clitic, particle**
      - not a base for derivation (possibly only by compounding)
      - finite and (very) small number of words



# The Categories: Part of Speech, Open Categories: Verbs

- Verbs:
  - infl. categories: person, number, tense, voice, aspect, [gender, neg.], ...
  - 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: (Cz.: 5 classes + ~100 combinations)

# The Categories: Part of Speech, Open Categories: Nouns

- NOUNS: infl. categories: number, [gender, case, negation, ...]
  - semantic classification:
    - human/animal/(non-living) things: driver/bird/stone
    - concrete/abstract: computer/thought
    - common/proper: table/Hopkins
  - syntactic classification: countable/unc.: book, water
  - morphological classification:
    - pluralia/singularia tantum: data (is), police (are)
    - ***declension*** type (“pattern” or “class”) (Cz.: 14 basic patterns, plus deviations: ~300 patterns, + irregular inflection)
    - “adverbial” nouns: afternoon, home, east (no inflection)

# The Categories: Part of Speech, Open Categories: Pronouns

- **Pronouns**: infl. categories: number, gender, case, negation; person
  - 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
  - morphological classification: mostly idiosyncratic pattern

# The Categories: Part of Speech,

## Open Categories: Adjectives

- Adjectives:
  - infl. categories: degree of comp., [number, gender, case, negation]
  - 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
    - mostly regular declension (Cz.: 4 basic patterns, ~ 10 total)
    - degrees of comparison (En.: big, bigger, biggest)
    - but: large number of forms (agreement, cf. section on syntax)

# The Categories: Part of Speech, Open Categories: Adverbs

- Adverbs: “infl.” categories: degree of comp., [negation]
  - open cat.: regular derivation from adjectives 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 (other lang.: all 3 degrees regular)

# The Categories: Part of Speech, Open Categories: Numerals

- Numerals: infl. categories: number, gender, case, negation
  - open cat.: compounding (Ge.: einundzwanzig, 21)
  - classification:
    - cardinals: one, five, hundred
      - NB: million etc. often considered noun
    - ordinals/fractionals: first, second, thirtieth
    - quantifiers: all, many, some, none
    - multiplicative: times, twice (Cz.: dvaadvacetkrát, 22-times)
    - multilateral: single, triple, twofold
  - morphological classification: as nouns/adjectives; many irreg.

# 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: a, 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 (Czech): Dual (nouns, pronouns, adjectives)
    - (Pl.) nohami / (Dl.) nohama (Cz.; (by) legs (of sth)/(by) legs (of sb))
- Grammatical Gender: Masculine, Feminine, Neuter
  - nouns, pronouns, verbs, adjectives, numerals
    - he/she/it; читал, читала, читало (Ru.; (he/she/it) was-reading)
    - 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,...)
  - nouns, pronouns, adjectives, numerals
- most common cases (forms in singular/plural)
  - **nominative**    I/we (work)                    **tóida/tóidy** (Cz.; class)
  - **genitive**        (picture of) me/us                    **tóidy/tóid**
  - **dative**            (give to) me/us                    **tóidi/tóidám**
  - **accusative**      (see) me/us                    **tóidu/tóidy**
  - **vocative**        -/-                    **tóido/tóidy**
  - **locative**        (about) me/us                    **tóidi/tóidách**
  - **instrumental**    (by) me/us                    **tóidou/tóidami**

# The Categories: Person, Tense

- Person
  - verbs, personal pronouns
    - 1st, 2nd, 3rd: (I) go, (you) go, (he) goes; (we) go, (you) go, (they) go
    - jdu, jdeš, jde, jdeme, jdete, jdou (Cz.)
- Tense
  - past: (you) went - szliście
  - present: (you pl.) go jdete idziecie
  - future (!if not “analytical”) - pùjdete -
  - concurrent (gerund) going jda id<sup>1</sup>e
  - preceding - - szed<sup>3</sup>szy

# Note on Tense

- Grammars: more (syntactic/sematic) tenses
  - but: morphology handles isolated words → some tenses can be defined & handled only at an upper level (surface syntax)
- Examples of (traditional) tense (synthetical ***and*** analytical):
  - 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 (cf. later), 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-*
- 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.: nejnemo□nijši (the most impossible)
    - Pol.: *nie*najwierniejszy (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.) *sz³abym*
  - Inflective: using prefix/suffix/infix, combines several categ.
    - Slavic: Czech, Russian, Polish,... (not Bulgarian); also French, German; Arabic
      - (Cz. *new*(acc.)) *novou* (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
  - $T \subset C_1 \times C_2 \times \dots \times C_n$
  - 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], ...