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 vocal track.
    - intonation
  - Orthography: normalization, punctuation, etc.
Phonology/lexicon

Input:
- sequence of phones/sounds (in a phonetic alphabet); or “normalized” text (sequence of (surface) letters in one language’s alphabet) [NB nota bene (note well): 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) (or surface and lexical)
- e.g.: [u] ~ oo (as in book), [æ] ~ a (cat); i ~ y (flies)
Phonology Examples

- **German (umlaut) (satz ~ sentence)**
  - lexical: \( s \, A \, t \, z \, + \, e \) (\( A \) denotes “umlautable” \( a \))
  - surface: \( s \, ä \, t \, z \, e \) (phonetic: \( zæc\Theta \), vs. \( zac \))

- **Turkish (vowel harmony)**
  - lexical: \( e \, v \, + \, l \, A \, r \) (←houses) \( b \, a \, š \, + \, l \, A \, r \)
  - surface: \( e \, v \, 1 \, e \, r \) (heads→) \( b \, a \, š \, l \, a \, r \)

- **Czech (e-insertion & palatalization)**
  - lexical: \( m \, a \, t \, E \, K \, + \, 0 \) (←mother’s/gen.) \( m \, a \, t \, E \, K \, + \, ě \)
  - surface: \( m \, a \, t \, e \, k \) (mother/dat. →) \( m \, a \, t \, _{c} \, e \)
Morphology

Input:
- sequence of phonemes (~ (lexical) letters)

Output:
- sequence of pairs (lemma, (morphological) tag)

Deals with:
- composition of phonemes (word forms) into 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”, …) (derivation vs. inflection)

- Categories: non-lexical
  - small number of possible values (< 100, often < 5-10)
Morphology Level: The Mapping

Formally: \( A^+ \rightarrow 2^{(L, C_1, C_2, \ldots, C_n)} \)

- \( A \) 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...)
- \( 2^{(L, C_1, C_2, \ldots, C_n)} \) denotes the power set of \((L, C_1, C_2, \ldots, C_n)\)
- \( A, L \) 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
Typology of Languages

By morphological features

▶ Analytical: using (function) words to express categories (1 morpheme almost 1 word)
  ▶ English, also French, Italian, ..., Chinese
    ▶ I would have been going ~ (Pol.) szłabym

▶ (Synthetic) Inflective (fusional): using prefix/suffix/infix, combines several categ.
  In one morpheme (morpheme boundary is not clear)
  ▶ Slavic: Czech, Russian, Polish,... (not Bulgarian); also French, German; Arabic
    ▶ (Cz. new(acc.)) novou (Adj, Fem., Sg., Acc., Non-neg., Pos.)

▶ (synthetic) Agglutinative: one category per (non-lexical) morpheme (morpheme boundary is clear)
  ▶ Finnish, Turkish, Hungarian; Korean/Japanese
    ▶ (Korean) meg+hi+go+it+neun+jung+i+da (have been being eaten)
Categories & Tags

- **Tagset:**
  - list of all possible combinations of category values for a given language
  - $\mathcal{T} \subseteq C_1 \times C_2 \times \ldots \times C_n$
  - 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], …
(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
Words, Phrases, Clauses, Sentences

- **Words**
  - smallest units on the syntax level
    - function/autosemantic

- **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
    - autosemantic words ("lexical" words)
  - idioms
    - fixed phrases (non-compositional) -&gt; "words"

- 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: a new book
    - Adjective: brand new
    - Adverbial: so much
    - Prepositional: in a class
    - Verb: catch a ball
Ellipsis

- Word or Phrase missing where one would normally expect one; often happens in dialogues
  - Whom did you see there?
  - Peter. ?? verb ??

- Most common in coordination (written text)
  - Pittsburgh leads 4-0 but Detroit only 3-1. ??verb in 2\textsuperscript{nd} part??

- Systematic in many languages: pro-drop (leave out a pers. pronoun in the Subject position)
  - [She] Passed the exam easily.
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)

- **Function of a Clause**
  - same as for phrase, plus some (direct speech/discourse etc.)
Gaps (Non-Continuous Constituents)

- Constituent moves from the expected position:
  - happens in questions and relative clauses
    - Who(m) do you work for <gap> whom?
      - strictly speaking, **do you work** should be **you (do work)**
    - I don’t know why we have got so much rain <gap> why?
    - On Sundays, I usually work <gap> On Sundays but I stay home on Tuesdays.
    - The story he never wrote <gap> the story
    - And finally the car she was supposed to use <gap> the car for her trip to New York broke.
      - The last two: also could be considered ellipsis (which) plus a gap.
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!
- Sometimes even semicolon (;) might be a sentence break (...vague)
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 (but 1:n mapping Dep→PS)
Phrase Structure Tree

Example:

```
((DaimlerChrysler's shares) NP (rose (three eights) NUMP (to 22) PP-NUM ) VP ) S
```

DaimlerChrysler's shares rose three eights to 22
Example:

DaimlerChrysler's shares rose three eights to 22

rose_{Pred}(shares_{Sb}(DaimlerChrysler's_{Atr}), eights_{Adv}(three_{Atr}), 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 (autosemantic - has meaning in isolation - lemmas, (morphosyntactic) tags, deep semantic 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]
```