Methods in Computational Linguistics

Introduction to Language and Linguistics

Apl. Prof. Dr. Sabine Schulte im Walde

Master of Science *Computational Linguistics*
Institut für Maschinelle Sprachverarbeitung
Universität Stuttgart

Outline

Outline

Outline

Language

- What is (a) language?
  - langage vs. langue (French)
  - linguaggio vs. lingua (Italian)
  - lenguaje vs. lengua (Spanisch)
- What is (a) language?
  - reference to language in general vs. applied to particular languages
  - one cannot possess or use natural language without possessing or using some particular *natural* language
- Language: What the linguist wants to know is whether all natural languages have something in common not shared by other systems of communication (human or non-human).
- Linguistics: scientific study of language

Creativity of Language

- You cannot buy a dictionary of any language with all the words or sentences of the language.
- Memorisation of all the possible sentences in a language is impossible in principle.
- Creative aspect of language use:
  Knowing a language means being able to produce new sentences never spoken before and to understand sentences never heard before.
- You may not believe a sentence you have never heard before; you may question its logic; but you may understand it even if you never heard or read it before.

Creativity of Language: Examples

- The cat chased the mouse.
- The cat chased the mouse that ate the cheese.
- The cat chased the mouse that ate the cheese from the cow.
- The cat chased the mouse that ate the cheese from the cow that grazed in the field.

Grammar

- Grammar:
  knowledge speakers have about the units and rules of their language
  - rules for combining sounds into words (phonology)
  - rules of word formation (morphology)
  - rules for combining words into phrases and sentences (syntax)
  - rules for assigning meaning (semantics)

- Together with a mental dictionary that lists the words of the language, the grammar represents our linguistic competence.

- Universal Grammar: rules that hold in all languages;
  provides a window into the workings of the human mind

Competence and Performance

- Competence:
  - production of an infinitely large set of sentences that constitutes a language ($\rightarrow$ language-system)
  - having the knowledge necessary to produce sentences of a language
- Performance:
  - actual production of a set of sentences ($\rightarrow$ language-behaviour)
  - applying the knowledge necessary to produce sentences of a language
  - relies not only on competence but also on social conventions, beliefs, emotions, situations, etc.

Outline

Linguistics

Linguistics: scientific study of language

Linguistics: Features

- Descriptive vs. prescriptive linguistics:
  - descriptive: discover and record rules to which the members of a language-community conform (how things are)
  - prescriptive: impose rules or norms of correctness to members of a language-community (how things ought to be)
- Diachronic vs. synchronic linguistics:
  - diachronic: historical development of a language
  - synchronic: account of a language at some particular point in time

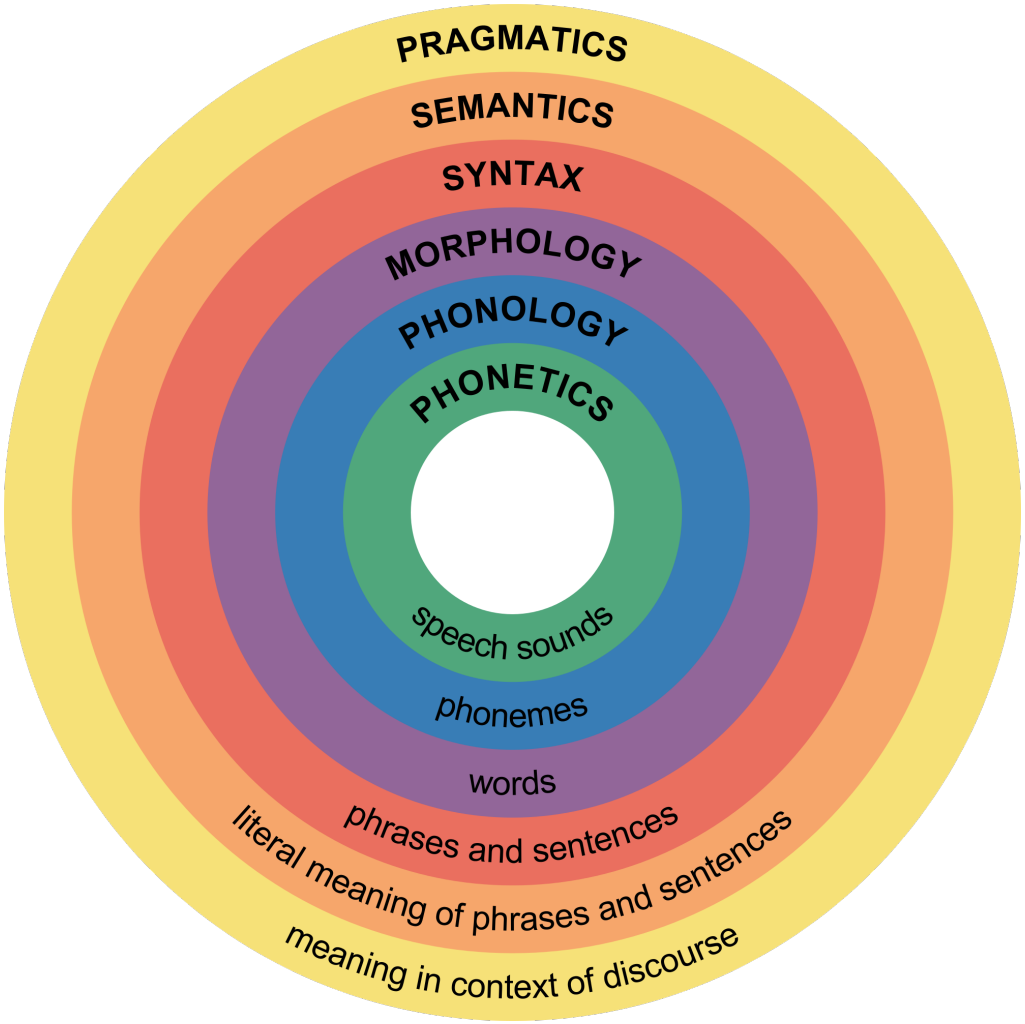
Linguistics: Features

- Theoretical vs. applied linguistics:
  - theoretical: studies of language(s) with a view to constructing a theory of their structure and functions
  - applied: application of the concepts and findings of theoretical linguistics to a variety of practical tasks

- Rational vs. empirical linguistics:
  - rational: emphasises the role of the mind in knowledge acquisition
  - empirical: all knowledge comes from experience

Outline

Linguistic Levels



accessed [Nov 7, 2020]: https://commons.wikimedia.org/wiki/File:Major_levels_of_linguistic_structure.svg

Phonetics

- The sounds of all the languages of the world together constitute a class of sounds that the human vocal tract is designed to make.
- But individual languages may use a subset of all possible sounds.
  - Languages differ in which sounds they include in their inventory.
  - example from English (but not German): *ba***th**, **th***erapy*
  - example from German (but not English): *M***ü***sli*, *T***ü***te*
- Phonetics: study of speech sounds
  - acoustic phonetics: study on the physical properties of sounds
  - auditory phonetics: study of how listeners perceive sounds
  - articulatory phonetics: study of how the vocal tract produces sounds
- Questions:
  - What types of speech sounds do we find in the languages of the world and in the individual languages?
  - How can we describe these sounds?
  - Which criteria can we use to distinguish different sounds?

Phonology

- The sounds form different patterns in different languages.
- Phonology: study of the way speech sounds form patterns;
  i.e., how they are grouped into larger phonological units, such as syllables and words
- Phonological Grammar:
  - speakers' linguistic knowledge about sound patterns of their language
  - tells you what sounds are in your language and which ones are foreign
  - tells you what combinations of sounds are legal *(blick)*,
    which ones aren't legal *(*lbick)*, and
    whether they make an actual word *(black)* or not *(blick)*
  - explains why certain phonetic features are important to the meaning of a word (such as voicing in English), while other features are not crucial to meaning (such as aspiration in English)

Phonetics & Phonology

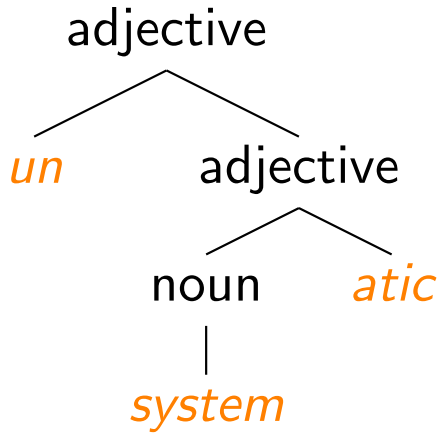
to be continued by Dmitry Nikolaev

Morphology

- Knowing a word means knowing that a particular sequence of sounds is associated with a particular meaning.
- Morphology: study of the internal structure of words, and of the rules by which words are formed
- Morpheme: most elemental linguistic unit;
  union of a sound and a meaning that cannot be further analysed
- A morpheme may be represented by a single sound (e.g., *a-moral*), or by a syllable (e.g., *boy-ish*).

Morpheme Structure

- A single word may be composed of one or more morphemes:
  - one morpheme: *boy; desire*
  - two morphemes: *boy + ish; desire + able*
  - three morphemes: *boy + ish + ness; desire + able + ity*
  - four morphemes: *gentle + man + li + ness; un + desire + able + ity*
- Complex words consist of a root and one or more affixes.
- A word is not a simple sequence of morphemes. It has internal structure.



Morphology

to be continued by Dmitry Nikolaev

Syntax

- Any speaker of any human language can produce and understand an infinite number of sentences.
  - The cat chased the mouse.
  - The cat chased the mouse that ate the cheese.
  - The cat chased the mouse that ate the cheese from the cow.
- But we cannot form sentences by adding just any word to any word.

- Sentences are composed of discrete units that are combined by rules.
- Syntax: study of the parts sentences consist of, and their connections and dependencies

Syntax

to be continued by Sabine Schulte im Walde

Semantics

- Semantics: study of the linguistic meaning of morphemes, words, phrases and sentences
- Meaning:
  - denotation: equate the meaning of a word or phrase with the entities to which it refers; e.g., the denotation of the word *dog* corresponds to the set of canines
  - connotation: set of associations that a word's use can evoke; e.g., the word *winter* evokes thoughts of snow, bitter cold, short evenings, etc.
  - extension: an expression's extension corresponds to the set of entities that it picks out in the world (i.e., its referents); e.g., the extension of *Prime Minister of the UK* is an individual
  - intension: an expression's intension corresponds to its inherent sense, the concepts that it evokes; e.g., the intension of *Prime Minister of the UK* involves the concept 'leader of the governing party'

Semantics

- Multiple senses:
  - homonymy (e.g., *bank*)
  - polysemy (e.g., *school*)
- Figurative Language:
  - literal (e.g., *grasp the bottle*)
  - metaphorical (e.g., *grasp a meaning*)
  - idiomatic (e.g., *kick the bucket*)
- Semantic relations:
  - synonymy: words or expressions with the same meanings in all contexts, such as *begin–start, vacation–holidays, big–large*
  - antonymy: words or expressions that are opposites with respect to some component of their meaning, such as *come–go, dark–light*
  - hypernymy/hyponymy: words or expressions with a super-/sub-ordinate component of their meanings, such as *bird–robin*

Semantics

to be continued by Dmitry Nikolaev

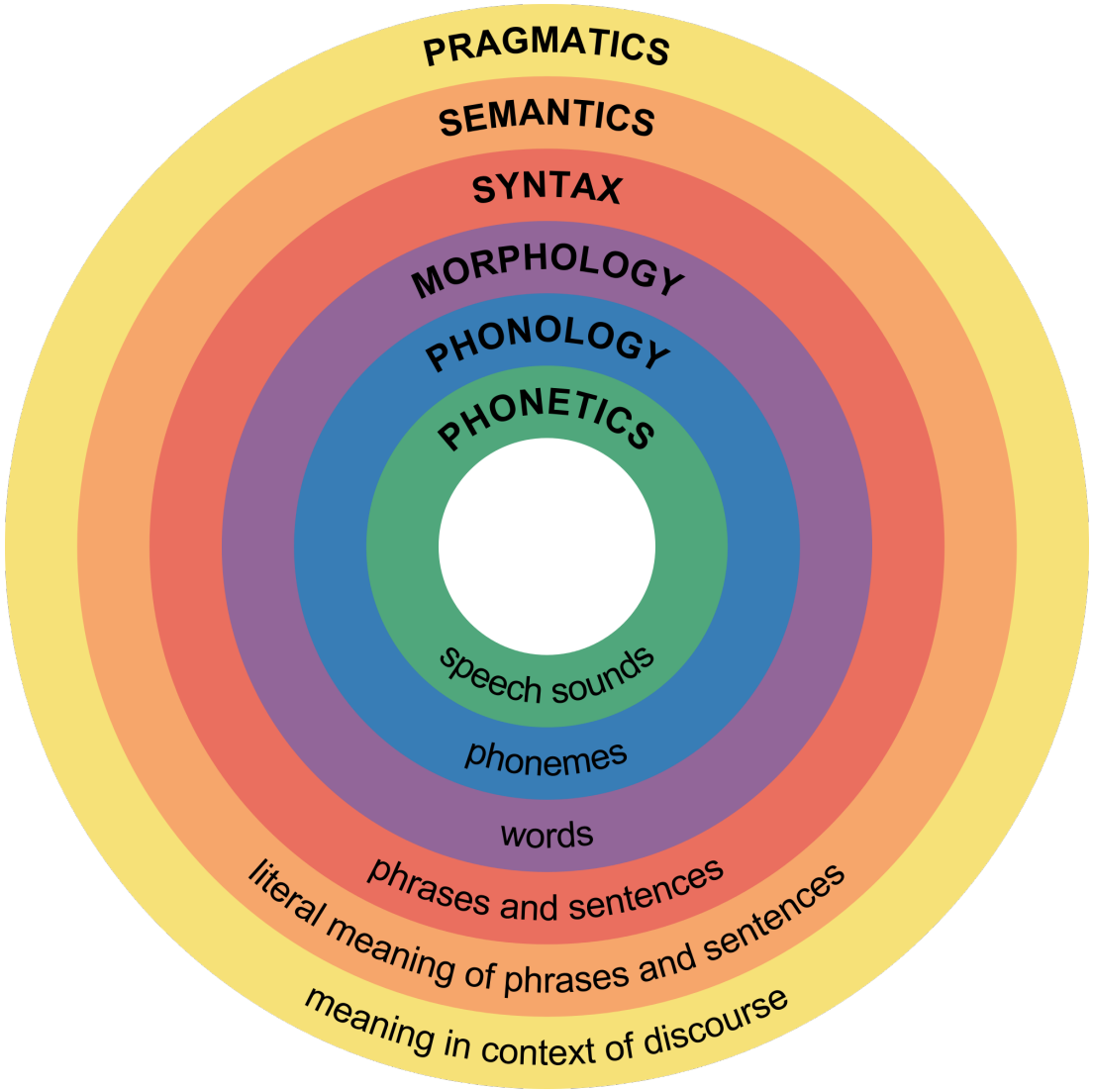
Pragmatics

- Pragmatics: interpretation of linguistic meaning in context
- Contexts:
  - linguistic context: discourse that precedes the phrase or sentence to be interpreted
  - situational context: nonlinguistic environment in which a sentence or discourse happens
    - includes the speaker, hearer, and any third parties present, along with their beliefs and their beliefs about what the others believe
    - includes the physical environment, the subject of conversation, the time of day, etc.

Pragmatics

- When people use language, the do not do this just for the sake of using it.
- Speakers have a certain intention (such as stating, promising, requesting, etc.).
- Examples:
  - friend to friend: *I'll phone you tomorrow.*
  - girl-friend to boy-friend: *On time, as always!*

Linguistic Levels



Ambiguity

- Ambiguity: a linguistic unit can be interpreted in more than one way
  - phonetic ambiguity: *realize/real ice/real eyes*
  - morphological ambiguity: *unlockable*
  - syntactic ambiguity: *I saw the man with the telescope*
  - semantic ambiguity: *bank; bright; saw*

Outline

Grammar

- Grammar (1): a kind of book with rules to obey, to use a particular language correctly; incomplete representation of language system

  → be careful to cite as evidence about (un)grammatical utterances

- Grammar (2): the knowledge speakers have about the units and rules of their language
  - rules for combining sounds into words (phonology)
  - rules of word formation (morphology)
  - rules for combining words into phrases and sentences (syntax)
  - rules for assigning meaning (semantics)

  → linguistic competence

Outline

Word Classes

- Word classes/Parts-of-Speech/Lexical or Syntactic Categories
- *John gnorbed the pirkness only twenty pripless skirps ago.*
- Words are assigned to word classes on grammatical grounds, i.e., according to their properties in phrasal and clausal structures.

Major Word Classes

- Nouns: name entities such as individuals and objects
  - common nouns: count vs. noncount/mass; concrete vs. abstract; e.g., **bun, difficulty, butter, music**
  - proper nouns/names: e.g., **Freda, Indonesia**

- Verbs: designate actions, sensations, states
  - full/lexical verbs: e.g., **believe, follow, like, see**
  - (primary) auxiliary verbs: **be, have, do**
  - modal (auxiliary) verbs: e.g., **can, may, will, must, would**

Major Word Classes

- Adjectives: designate a property or attribute of an entity
  - attributive function: premodification of a noun, e.g., *an* **ugly** *box*
  - predicative function: subject or object complements, e.g., *The box is/seems* **ugly**.
  - premodification by intensifiers: e.g., *The box is very* **ugly**.
  - comparative/superlative forms: e.g., *Children are* **happier/the happiest** *people.*
- Adverbs: denote properties/attributes of actions, sensations, states
  - clause element adverbial: e.g., *He* **almost** *forgot about it.*
  - premodifier of adjective/adverb: e.b., *They are* **quite** *happy.*
- Prepositions: closed-class items that specify a relationship between two units in a sentence
  - link to verb phrase: e.g., *runs* **to** *his office*
  - link to a noun phrase: e.g., *man* **in** *the raincoat*
  - link to an adjective phrase: e.g., *grateful* **for** *her help*

Major Word Classes

- Determiners: determine the kind of reference of a noun phrase
  - central determiners: e.g., **the/a/this** *painting*
  - predeterminers: e.g., **all** *the people*
  - postdeterminers: e.g., *the* **seven** *passengers*

- Pronouns/Pro-Forms:
  - for noun phrases: e.g., *Everyone expected that* **she** *would win.*
  - for indefinite noun phrases: e.g., *I'll buy you* **some** *soon.* $\rightarrow$ *books*
  - for clauses: e.g., *It will be funny if he* **does (so)**. $\rightarrow$ *go away*
  - for adverbials: **Here** *we stopped for lunch.*
  - for complements: *It's his parents who made him* **so**. $\rightarrow$ *a criminal*

Word Classes: Categorisations

- Open vs. closed word classes:
  - closed word classes: classes that are finite (and often small) with a membership that is relatively stable and unchanging;
    examples: *pronouns, modal verbs, prepositions, conjunctions*
  - open word classes: classes of words that are constantly changing their membership as old words drop out of the language and new ones are coined or adopted to reflect cultural changes in society;
    examples: *nouns, adjectives, full verbs, adverbs*

- Content/lexical vs. function/non-lexical word classes:
  - content word classes: nouns, verbs, adjectives, prepositions
  - function word classes: determiners, auxiliary verbs, prepositions, conjunctions

Case, Number, Gender

- Case:
  - inflectional contrast associated with nouns in many languages
  - category that encodes information about a unit's grammatical role
  - in English, functions are largely expressed through word order and preposition use
  - in German, we distinguish nominative, genitive, dative, accusative
- Number:
  - morphological category for contrasts involving countable quantities
  - simplest number contrast: singular vs. plural
  - other languages involve no (e.g., Indian Nancowry) or three (e.g., Canadian Inuktitut) contrasts
- Gender:
  - semantic distinction between kinds/sexes of nominal objects
  - most common classes: feminine, masculine, neuter
  - dinstinction through suffixes (e.g., Russian: *-/-a/-o*), determiners (e.g., French: *le/la*), etc.

Outline

Words and Phrases

- Words represent structural syntactic units.
  Combinations of words might represent structural syntactic units.
- Structural syntactic units: phrases/constituents
- Which are the words in the sentence?
  Words are separated by spaces (in many cases/languages).
- Which are the phrases/constituents in the sentence?
  $\rightarrow$ apply constituency tests

Constituency Tests

- Pronominalisation: substitution of a constituent by a pronoun

  [Many people] go to the station every morning.
  [They] go to the station every morning.

- Movement: movement of a string of words to other sentential positions

  Many people go [to the station] every morning.
  [To the station] many people go ___ every morning.

- Coordination: only constituents can be coordinated by the coordinating conjunction *and*

  [Many people] go to the station every morning.
  [[Many people] and [my friends]] go to the station every morning.

- Gapping: constituents can add a tag question that leaves a gap for which we could insert the missing string

  Many people [go to the station] every morning.
  Many people [go to the station] every morning, don't they ___?

- Sentence-Fragment: only constituents can form possible sentence fragments which speakers can use to refer to

  [Who] goes to the station every morning?
  [Many people] go to the station every morning.

Structure of Phrases

- Head: The most important element of a phrase is its head.
- How do we determine the most important element of a phrase?
  - semantic perspective:
    The head is the semantically most important element.
    [**to** the station] indicates a direction, [**at** the station] a location.
  - structural perspective: The head assigns case.
    I saw **him/*he** yesterday.
    This was a surprise for **her/*she**.
  - distributional perspective: The head of a phrase can have the same distribution as the phrase it heads.
    [The two little **kids**] meet their friends regularly at the playground.
    [**Kids**] meet their friends regularly at the playground.
  - semantico–syntactic perspective: The phrase obtains its semantic and syntactic properties from its head.
    [my older **sister**]: animate noun, feminine, singular

Structure of Phrases

- Head: The most important element of a phrase is its head.
- Phrases are named after the lexical categories of their heads: *XP*.
  - [his **sister**]$_{NP}$
  - [Maja's red-haired **brother**]$_{NP}$
  - [a **letter** to John]$_{NP}$
  - [the best **actor** in town]$_{NP}$
  - [**people** who knock at the door]$_{NP}$
  - [**to** the station]$_{PP}$
  - [**proud** of his results]$_{ADJP}$
  - [extremely **expensive**]$_{ADJP}$
  - [**drink** a glass of milk]$_{VP}$
  - [incredibly **often**]$_{ADVP}$

Outline

Grammatical Functions

- Grammarians distinguish functions that sentential constituents can fulfil.
- Subjects intuitively represent the entity which the sentence is about, the 'doer of the action'.
- Objects represent the entities that are affected by the action denoted by the verb.
- Adverbials are constituents that describe the circumstances of the event denoted by the rest of the sentence (such as time, place, manner, reason, etc.)
- Let's go for some more detail.

Grammatical Functions: Subjects

- Subjects trigger subject-verb agreement and require subject and verb to share the same person and number features.

  [Generations of students]$_{subject}$ read the textbook.
  [The graduate students]$_{subject}$ were trained in CL.

- The position of the subject in English is fixed: immediately before the verb phrase; only certain adverbials are allowed to intervene.

  [They]$_{subject}$ never came home so late.
  *[So late]$_{adverbial}$ came [they]$_{subject}$ never home.

- Subjects are obligatory in English.

- In English, there is a case marking distinction between subject and object (which is only visible with pronouns).

  [He/my brother]$_{subject}$ took [her/my friend]$_{object}$ to a concert.
  [She/my friend]$_{subject}$ took [him/my brother]$_{object}$ to a concert.

Grammatical Functions: Objects

- Objects receive object case.
- Objects do not show agreement with the verb, but instead may be strongly restricted in their distribution (e.g., in English).

  My brother often [invited]$_V$ **[her]**$_{object}$ to his parties.
  *My brother [invited]$_V$ often **[her]**$_{object}$ to his parties.
- Objects are obligatory constituents (if subcategorised).
- Direct objects denote an entity that undergoes the action/process denoted by the verb.
- Indirect objects denote the goal, the recipient or the beneficiary of the event denoted by the verb.

Passivisation

Passivisation: Sentences in the active voice can be passivised by making the object of the active sentence the subject of the passive sentence.

- She wrote [the novel]$_{object}$ at the end of the 19th century.
  [The novel]$_{subject}$ was written at the end of the 19th century.

- Next year the government will introduce [new tax laws]$_{object}$.
  Next year [new tax laws]$_{subject}$ will be introduced.

- Jill gave [him]$_{indirect\ object}$ [the book]$_{direct\ object}$ yesterday.
  [He]$_{subject}$ was given the book yesterday.
  [The book]$_{subject}$ was given to him yesterday.

Grammatical Functions: Adverbials

- Adverbials provide information about the circumstances of the action denoted by the verb and its subject and object(s).
- The circumstantial information concerns time, location, manner, cause, purpose, etc.
- They modify the clause or the verb phrase.
- Examples:

  We [often] go skiing [in the nearby mountains].
  They [never] came home [so late].
  My professor wrote two textbooks [last year].

Functions and Forms

- The sentential function can be fulfilled by different kinds of forms.
- Examples:

function	**form**	**example**
adverbial	NP	[last year]
	PP	[in the mountains]
	ADVP	[so late]
	VP	[hoping for a successful career], she applied ...
subject	NP	[a small bird], [last year]
	VP	[hoping for a successful career] helped her ...
	clause	[that you are lying] disappoints me
object	NP	[a small bird], [last year]
	PP	[to their parents]
	clause	I assume [that you are lying]

Outline

Complements

- How many phrases are in a sentence?
  - determined by the verb and its complements
  - nouns and adjectives also take complements
- Complement: structurally and semantically highly dependent sister constituent of a head
- Examples:
  - *Tom snores.* vs. **Tom snores her.*
  - **Tom is dependent.* vs. *Tom is dependent on a steady income.*

Subcategorisation/Valency

- Subcategorisation/Valency:
  potential of predicates to choose their complements
- Predicates: verbs, nouns, adjectives
- Obligatory (arguments) vs. optional complements (adjuncts)
- Argument structure: The required complements (arguments) of a word are part of the word's meaning and included in its lexical entry.
- The well-formedness of a phrase depends on at least two factors:
  - whether the phrase conforms to the structural constraints of the language as expressed in the phrase structure rules, and
  - whether the phrase obeys the selectional requirements of the head, both syntactic and semantic.

(In-)Transitivity

- Intransitivity: An intransitive verb cannot take an NP complement, e.g., **Michael* **slept** *a fish*.
- Transitivity: A transitive verb requires an NP complement (a direct object), e.g., *Harry* **hit** *the ball*.
- Some verbs are optionally transitive, e.g., *John* **ate** *(a sandwich)*.
- Some verbs select a sentence complement, e.g., *I* **think** *that Sam won the race*.
- Ditransitivity: A ditransitive verb takes an NP subject, an NP direct object, and an NP indirect object, e.g., *John* **baked** *Mary a cake*.

Subcategorisation Violation

- Syntactic subcategorisation violation:

  John found sad.
  John elapsed that Bill will come.
  John persuaded great authority to Bill.

- Semantic subcategorisation violation:

  Colorless green ideas sleep furiously.
  Golf plays John.
  Misery loves company.

(examples taken from Noam Chomsky (1965): "Aspects of the theory of syntax")

What this lecture was about.

References: Language & Linguistics


John Lyons.
Language and Linguistics – An Introduction.
Cambridge University Press, 1st edition, 1981. [Chapters 1–2]


William O'Grady, Michael Dobrovolsky, and Francis Katamba.
Contemporary Linguistics: An Introduction.
Pearson Education, Harlow, Essex, UK, 3rd edition, 1997.


Victoria Fromkin, Robert Rodman, and Nina Hyams.
An Introduction to Language.
Thomson Higher Education, Boston, MA, 8th edition, 2007.


Ingo Plag, Sabine Arndt-Lappe, Maria Braun, Mareile Schramm.
Introduction to English Linguistics.
Walter de Gruyter, Berlin/Boston, 3rd edition, 2015.

References: English/German Grammar Books


Sidney Greenbaum and Randolph Quirk.
A Student's Grammar of the English Language.
Longman, London, 1990.


Gerhard Helbig and Joachim Buscha.
Deutsche Grammatik.
Langenscheidt – Verlag Enzyklopädie, 18th edition, 1998.