LST Prep Course: Morphology and Syntax

Manfred Pinkal
Universität des Saarlandes

10-10-2006

Units of Language – Subfields of Linguistics

	Grammar	Semantics	Pragmatics
Sound	Phonetics/ Phonology	---	---
Word	Morphology	Lexical Semantics	---
Sentence	Syntax	Compositional Semantics	Pragmatics
Text& Discourse	Text & Discourse Grammar	Discourse Semantics	Pragmatics
	Structure	Meaning	Use

Morphology

UNIVERSITÄT DES SAARLANDES

- Morphology investigates the internal structure of words: their composition out of smallest meaningful or functional units, the morphemes.

Examples

UNIVERSITÄT DES SAARLANDES

- block
- block + s
- grasp + ed
- tall + er
- tall + ness
- un + friend + ly
- mis + behav+ ior

Morphology

UNIVERSITÄT DES SAARLANDES

- Morphology investigates the internal structure of words: their composition out of smallest meaningful or functional units, the morphemes.
- Morphems are typically either stems or prefixes or suffixes.

Examples

UNIVERSITÄT DES SAARLANDES

block	stem
block + s	
grasp + ed	
tall + er	prefix
tall + ness	suffix
un + friend + ly	
mis + behav+ ior	

Examples

UNIVERSITÄT DES SAARLANDES

- block
- block + s
- grasp + ed
- tall + er
- tall + ness
- un + friend + ly
- mis + behav+ ior

stem
prefix
suffix

Examples

UNIVERSITÄT DES SAARLANDES

- block
- block + s
- grasp + ed
- tall + er
- tall + ness
- un + friend + ly
- mis + behav+ ior

stem
prefix
suffix

Examples

UNIVERSITÄT DES SAARLANDES

- block
- block + s
- grasp + ed
- tall + er
- tall + ness
- un + friend + ly
- mis + behav+ ior

stem
prefix
suffix

Morphology

UNIVERSITÄT DES SAARLANDES

- Morphology investigates the internal structure of words: their composition out of smallest meaningful or functional units, the morphemes.
- Morphems are typically either stems or prefixes or suffixes.
- Functional types of morphological operations are inflection, derivation, and compounding.

Examples

UNIVERSITÄT DES SAARLANDES

- block
- block + s
- grasp + ed
- tall + er
- tall + ness
- un + friend + ly
- mis + behav+ ior

Inflection
Derivation

Examples

UNIVERSITÄT DES SAARLANDES

- block
- block + s
- grasp + ed
- tall + er
- tall + ness
- un + friend + ly
- mis + behav+ ior

Inflection
Derivation

Examples: Compounding

UNIVERSITÄT DES SAARLANDES

- English:
  - rain + bow
  - water + proof
- German:
  - Universität+s+professor
  - Universität+s+professor+en+stelle
  - Donau+dampf+schiff+fahrt+s+gesellschaft+s+kapitän

Morphological specialties

UNIVERSITÄT DES SAARLANDES

- Infixes, e.g., Arabic inflection
- German 'Umlaut': *Mutter / Mütter*
- Circumfixes: e.g., German *ge+frag+t*

More specialties

UNIVERSITÄT DES SAARLANDES

- Morpho-phonological processes at morpheme boundaries:
  - stick / stick+s, but
  - class / class+es
- Vowel harmony (Turkish)

Language types

UNIVERSITÄT DES SAARLANDES

- Isolating:
  - Chinese, English
- Inflectional:
  - Russian, Latin
- Aggutinative:
  - Finnish, Turkish

A Turkish Example

UNIVERSITÄT DES SAARLANDES

- *Evlerinizdeyiz*
- *Ev+ler+iniz+de+yiz*
- *house+pl+your+at+we-are*
- *"We are at your houses"*

Morphological Analysis in Computational Linguistics

UNIVERSITÄT DES SAARLANDES

- Stemmer/ Lemmatiser analyses inflected forms (of nouns, verbs, adjectives) and returns
  - stem/lemma + syntactic information

  Example:
  - *grasped* → 'grasp' + Past
- Full morphological analysers reduce derivations to roots and derivational affixes, compounds to their parts.

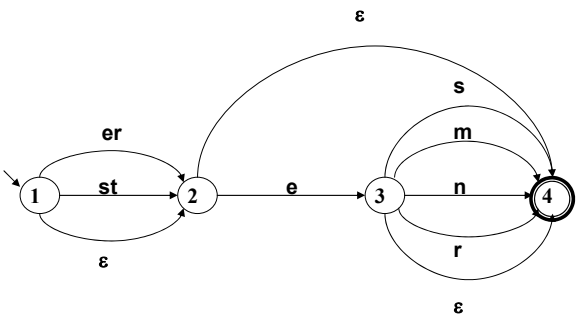
Morphological Analysers

UNIVERSITÄT DES SAARLANDES

- Morphological analysers are based on grammatical and lexical knowledge:
  - Inflectional schemata
  - Lexicon information assigning inflectional class information to the words of the language
- The best existing analysers have very good coverage, for a number of languages.
- The basic technique are finite-state automata (or finite state transducers).
- Morphological analysers are fast (linear time).

An FSA Accepting German Adjective Endings

UNIVERSITÄT DES SAARLANDES



Morphology and Syntax

UNIVERSITÄT DES SAARLANDES

- Morphology investigates the structure of words
- Syntax investigates the structure of sentences.

- In a way, syntax is the morphology of sentence, or, taken the other way round, morphology is the syntax of words.
- But: Sentence structure differs from word structure, in various respects.

Observation 1: Constituents

UNIVERSITÄT DES SAARLANDES

- A simple morphological rule of German:
  - The comparative morpheme occupies the first position of the ending (= the second position of the word)
  - *schnell*+*er*+*es* [fast+er, n, sg]
- A simple syntactic rule of English:
  - The finite verb occupies the second position of a declarative sentence
  - *John* + *gave* + *Mary* + *a* + *book*

Constituents

UNIVERSITÄT DES SAARLANDES

- Counter-examples (1)
  - Yesterday John gave Mary a book.
  - But John gave Mary a book.
- Counter-examples (2)
  - The student gave Mary a book.
  - The friendly student gave Mary a book.
  - The friendly student which I told you about yesterday gave Mary a book.

Constituents

UNIVERSITÄT DES SAARLANDES

- Counter-examples (1)
  - Yesterday John gave Mary a book.
  - But John gave Mary a book.
- Counter-examples (2)?
  - The student gave Mary a book.
  - The friendly student gave Mary a book.
  - The friendly student which I told you about yesterday gave Mary a book.
- The verb is still in second place, if we count constituents rather than words.

Arbitrarily long and complex sentences [1]

UNIVERSITÄT DES SAARLANDES

- The mouse escaped into the garden.
- The mouse that the cat chased escaped into the garden.
- The mouse that the cat which Mary owns chased escaped into the garden.

Arbitrarily long and complex sentences [2]

UNIVERSITÄT DES SAARLANDES

- *Er hat die Übungen gemacht.*
- *Der Student hat die Übungen gemacht.*
- *Der interessierte Student hat die Übungen gemacht.*
- *Der an computerlinguistischen Fragestellungen interessierte Student hat die Übungen gemacht.*
- *Der an computerlinguistischen Fragestellungen interessierte Student im ersten Semester hat die Übungen gemacht.*
- *Der an computerlinguistischen Fragestellungen interessierte Student im ersten Semester, der im Hauptfach Informatik studiert, hat die Übungen gemacht.*
- *Der an computerlinguistischen Fragestellungen interessierte Student im ersten Semester, der im Hauptfach, für das er sich nach langer Überlegung entschieden hat, Informatik studiert, hat die Übungen gemacht.*

Structural ambiguity

UNIVERSITÄT DES SAARLANDES

- Morphology talks about sequences of morphemes.
- To talk about syntactic regularities requires reference to constituent structure.
- Semantic interpretation of sentences also requires information about constituent structure:
  - *Pick up* | *a big red block.*
- *in particular, if sentences are structurally ambiguous:*
  - *John saw the man with the telescope*.

Syntactic ambiguity

UNIVERSITÄT DES SAARLANDES

- *John saw the* | *man with the telescope*
- *John saw* | *the man* | *with the telescope*
- | *Young students* | *and* | *professors* | *attended the party.*
- | *Young* | *students and professors* | *attended the party.*

Tests for constituency

UNIVERSITÄT DES SAARLANDES

- Substitution test: Word sequences that can be systematically substituted for a single word (e.g., proper name or personal pronoun) form a constituent:
  - *The student gave Mary a book.*
  - *The friendly student gave Mary a book.*
  - *The friendly student which I told you about yesterday gave Mary a book.*
  - *Mary gave John a book.*
  - *Mary gave the student a book.*
  - *Mary gave the friendly student which I told you about yesterday a book.*

  Compare with:
  - Yesterday John gave Mary a book.
  - Mary gave yesterday John a book.

Syntactic Categories

UNIVERSITÄT DES SAARLANDES

- Constituents that are substitutable for each other can be subdivided into larger classes that share distribution and structural properties, the Syntactic Categories, e.g.:
  - Noun phrases, consisting of a pronoun, a proper name, or a complex structure with a common noun as syntactic head element – NP
  - Prepositional phrases (*with the telescope, into the garden*) – PP
  - Adjective phrases (*friendly, very friendly, interested in linguistics*) - AP

Categories and Functions

UNIVERSITÄT DES SAARLANDES

- Syntactic categories denote classes of constituents with similar internal structure, in particular, the category /part-of-speech of their lexical head.
- Grammatical functions characterise the external role of a constituent in its syntactic context, e.g.
  - Complements: Subject, (Direct, indirect, prepositional) Object
  - Modifier / Adjunct

CFG for Syntactic Description

UNIVERSITÄT DES SAARLANDES

G = <V, Σ, P, S>, where

- V: Syntactic Categories
- $\Sigma \subseteq$ V: Parts-of-speech are terminal symbols
- P: Production rules describing constituent structure
- S: Start symbol: Category "Sentence"

A simple context-free grammar

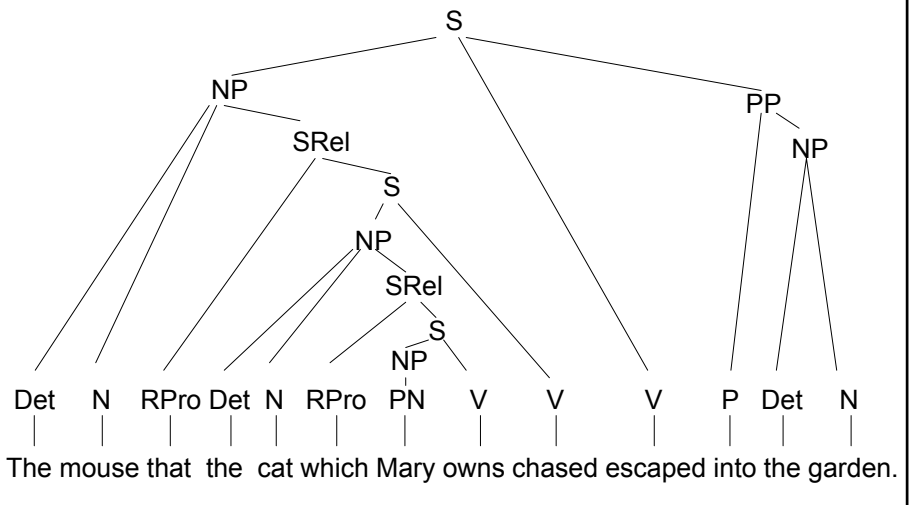
UNIVERSITÄT DES SAARLANDES

S → NP V
S → NP V NP
S → NP V NP NP
S → NP V PP
SRel → RPro S
PP → Prp N

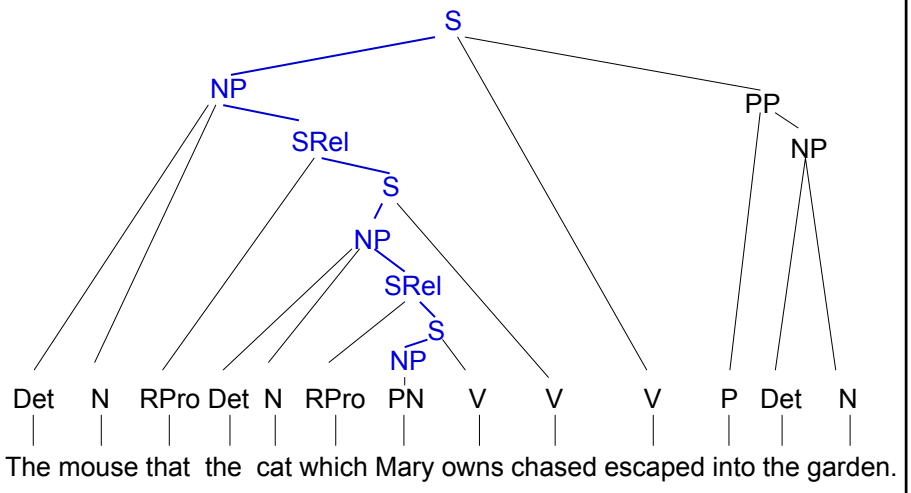
NP → Det N
NP → Det N SRel
NP → PN
NP → PPro
NP → Det N PP

A parse tree representing constituent structure

UNIVERSITÄT DES SAARLANDES



A parse tree representing constituent structure



Syntactic Description with CFGs

- CFG is a formalism that allows to model the concept for grammaticality for natural languages, by specifying the set of grammatically correct sentences, and assigning them their appropriate grammatical structures (in terms of their parse trees).
- Is it a realistic and reasonable aim to describe the set of grammatically correct sentences of a language?
  - What to do with ungrammatical input?
  - What does 'grammatical' mean after all? – Graded grammaticality!
- Is a CFG the appropriate formalism to describe the grammar of a language?

Syntactic Processing with CFGs

UNIVERSITÄT DES SAARLANDES

- Morphological analysers are finite-state automata (or transducers) working in linear time.
- The syntax of programming languages is recursive, and therefore described by CFGs. Because the languages typically are unambiguous, and described by deterministic CFGs, parsers for programming languages are also linear time.
- Unfortunately, grammars of natural languages are ambiguous and non-deterministic. The best algorithms (Earley Algorithm, Chart Parsing) take quadratic time to find one parse, and cubic time to find all parses.

Syntactic Processing with CFGs

UNIVERSITÄT DES SAARLANDES

- Good news: There are techniques to compile CFGs down to FSAs for many applications, without loosing much coverage (e.g., by constraining recursion depth; "finite-state technology")
- Bad news: Constituent structure is only the tip of the iceberg: More descriptive power is needed to describe syntactic structure of natural languages appropriately. Modern grammar formalisms like LFG or HPSG come in the format of typed feature structures with a context-free backbone.

Variable Word-Order in German

Peter hat der Dozentin das Übungsblatt heute ins Büro gebracht.
Peter has the lecturer the exercise-sheet today into-the office brought

Das Übungsblatt hat Peter der Dozentin heute ins Büro gebracht.

Der Dozentin hat Peter heute das Übungsblatt ins Büro gebracht.

Ins Büro hat heute Peter der Dozentin das Übungsblatt gebracht.

Heute hat Peter das Übungsblatt der Dozentin ins Büro gebracht.

Ins Büro hat das Übungsblatt der Dozentin Peter heute gebracht.

** Ins Büro heute Peter das Übungsblatt hat gebracht der Dozentin.*

** Ins heute Büro der Peter Dozentin das hat Übungsblatt gebracht.*

More syntactic phenomena

UNIVERSITÄT DES SAARLANDES

- Agreement
- Subcategorisation
- Long-distance Dependencies

Agreement

UNIVERSITÄT DES SAARLANDES

Subject-Verb agreement in English:

[The cat]$_{sg}$ chases$_{sg}$ the mouse.
[The cats]$_{pl}$ chase$_{pl}$ the mouse.

$S \rightarrow NP_{sg}\ V_{sg}$
$S \rightarrow NP_{pl}\ V_{pl}$

Agreement in German

Nominal Agreement: Gender, Number, Case

- *Der [m,sg, nom]an computerlinguistischen Fragestellungen interessierte [m,sg, nom] Student [m,sg, nom] im ersten Semester, der [m,sg, nom] im Hauptfach, für das er [m,sg, nom] sich nach langer Überlegung entschieden hat [sg], Informatik studiert [sg], hat die Übungen gemacht.*

Agreement in German

Promominal Agreement: Gender, Number

- *Der [m,sg, nom]an computerlinguistischen Fragestellungen interessierte [m,sg, nom] Student [m,sg, nom] im ersten Semester, der [m,sg, nom] im Hauptfach, für das er [m,sg, nom] sich nach langer Überlegung entschieden hat [sg], Informatik studiert [sg], hat die Übungen gemacht.*

Agreement in German

Subject-Verb Agreement

- *Der [m,sg, nom]an computerlinguistischen Fragestellungen interessierte [m,sg, nom] Student [m,3, sg, nom] im ersten Semester, der [m,3, sg, nom] im Hauptfach, für das er [m,3,sg, nom] sich nach langer Überlegung entschieden hat [3,sg], Informatik studiert [3, sg], hat [3, sg]die Übungen gemacht.*

Complements and Subcategorisation

UNIVERSITÄT DES SAARLANDES

*Mary owns a cat / *Mary owns*
*John sleeps / *John sleeps the box*

give the student a book
wait for the train
rely on the facts
put the block into the box

Long-distance Dependencies

UNIVERSITÄT DES SAARLANDES

- the cat which Mary owns _
- the cat which John believes Mary owns _
- the cat which Bill claims John believes Mary owns _

Unification Grammars

UNIVERSITÄT DES SAARLANDES

- Agreement, Subcategorisation, and Long-distance dependencies can be treated by an extension of the CFG formalism with typed feature structures and unification.

  → Berthold Crysmann's Presentation