Syntactic Analysis

Daniel Zeman

http://ufal.mff.cuni.cz/course/npfl094
Level of (Surface) Syntax

• Relations between sentence parts
• Sentence part = token (word, number, punctuation)
  – Practical reasons:
    • Easily recognizable (really?)
    • Unit of previous (morphological) level of processing.
    • We don’t restore elided constituents, nor do we collapse nodes of function words; this can be done later on a deep-syntactic level.
  – On the other hand:
    • We must now also define relations between function words (prepositions, auxiliary verbs etc.), punctuation and the rest of the sentence.
Level of Surface Syntax

• Between morphology and meaning.
• Morphology provides / requires:
  – lemmas (it’s time to obtain syntactic info from the dictionary)
  – tags (part of speech and morphosyntactic features)
  – word order (now it starts to play a role)
• Typical input is ambiguous
  – ambiguous morphological analysis
• Typical output is ambiguous
  – several syntactic structures for one sentence (several readings of the sentence)
Syntactic Structure

• Different shapes in different theories
• Typically a tree
  – Phrasal (constituent) tree, parse tree
  – Dependency tree
Example of Constituent Tree

- ((Paul (gave Peter (two pears)))) .

```
S  
/  
VP  
/  
NP V NP NP Z  
/  
N  
/  
Paul  
/  
N  gave  
/  
N  Peter  
/  
C  two  
/  
N  pears  .
```
Example of Dependency Tree

- [#,0] ([gave,2] ([Paul,1], [Peter,3], [pears,5] ([two,4])), [.,6])
Words and Phrases

• Word (token)
  – smallest unit of the syntactic layer
  – grammatical (function, synsemantic) words (e.g. and in coordination *Paul and Peter, to be* in compound verb forms *he is scared, he will be scared*)
  – lexical (content, autosemantic) words (e.g. *dog; to be* in the sentence *I think, therefore I am* (René Descartes))

• Phrase
  – composed of words and/or other phrases (*immediate constituents*)
Words

• Relation to other words
  – Lexicon contains information on words and possible relations among them.
    • Subcategorization of verbs and other words (do they require an object? if so, should it be marked for a particular case?)
    • Semantic features (a noun has color, has size, can act as the subject of a particular set of verbs…)

• Idioms, multi-word expressions
  – Fixed, indivisible phrases may act as one word (e.g. compound prepositions (*in spite of*), foreign citations and named entities (*Rio de Janeiro*), compound nouns written as separate tokens (*stock exchange*))
Phrase Replaceability

• A phrase can be replaced by another phrase of the same type. Specifically, it can be replaced by its head.
  – This is related to the generation of the sentence.

⇒ The phrases $x$, $y$, $z$ can be immediate constituents of a larger phrase $f$ only if they are related to each other. This is however a matter of the particular phrase structure grammar.
  – Example: sentence "This is the man that I talked about." The part "man that I" is not a whole noun phrase because it cannot be replaced by another noun phrase, e.g. man: "*This is the man talked about."
Phrase

• Phrase
  – Sequence of immediate constituents (words or phrases).
  – May be discontinuous in some languages. cs: „Soubor se nepodařilo otevřít.“ (lit. File oneself one-was-not-able to-open) contains the phrase “open file”.

• Phrase types by their main word—head
  – Noun phrase: the new book of my grandpa
  – Adjectival phrase: brand new
  – Adverbial phrase: very well
  – Prepositional phrase: in the classroom
  – Verb phrase: to catch a ball
Noun Phrase

• A noun or a (substantive) pronoun is the head.
  – water
  – the book
  – new ideas
  – two millions of inhabitants
  – one small village
  – the greatest price movement in one year since the World War II
  – operating system that, regardless of all efforts by our admin, crashes just too often
  – he
  – whoever
Adjective Phrase

• An adjective or a determiner (attributive pronoun) is the head.

• Simple ADJPs are very frequent, complex ones are rare.
  – *old*
  – *very old*
  – *really very old*
  – *five times older than the oldest elephant in our ZOO*
  – *sure that he will arrive first*
Pronouns / Determiners

• (Substantive) pronouns: similar behavior as nouns
  – Personal pronouns (*I, you, they, oneself*).
  – Some demonstrative, interrogative, relative and negative (*who, what, somebody, something, nothing*).

• Attributive pronouns (determiners): similar behavior as adjectives
  – Possessive pronouns (*my, your, his, whose*).
  – Articles (*the, a, an*).
  – Attributively used demonstrative, interrogative, relative and negative pronouns (*which, some, every, no*).
Numeral Phrases

• In Slavic languages not always clear what should be the head: the number, or the counted noun phrase?
  – The numeral inherits the gender of the counted noun. The noun gets its grammatical number from the numeral.
    • *jeden muž* (one man), *jedna žena* (one woman), *jedno dítě* (one child)
    • *dva muži* (two men), *dvě ženy* (two women), *dvě děti* (two children)
  – The numeral governs the case of the counted noun.
    • *pět mužů* (five men: noun in genitive, numeral in nominative, accusative or vocative)
  – Both the counted noun and the numeral have a case required by their governing preposition or verb.
    • *pěti ženami* (five women: instrumental)
Adverbiaal Phrases

- An adverb is the head.
  - *quickly*
  - *much more*
  - *how*
  - *louder than you can imagine*
  - *yesterday*
Prepositional (Postpositional) Phrase

• The preposition serves as head (because it determines the case of the rest of the phrase).
• Often have a function similar to adverbial phrases (adverbiale) or noun phrases (object of a verb).
  – *in the city center*
  – *in God*
  – *around five o’clock*
  – *to a better future*
  – *up to a situation where neither of them could back out*
  – *with respect to his nonage*
Prepositional Phrases

• Classic English example:
  – *I saw the man with a telescope.*
    1. Viděl jsem ho dalekohledem.
    2. Viděl jsem ho s dalekohledem.
Prepositional Phrases: Czech Example

• „Přišel ten pán se sousedem odnaproti.“

Lit.: Came the man with neighbor from across-the-road.
Prepositional Phrases and Syntactic Ambiguities


- *In years 1991 – 1993 I attended classes of management and marketing at Collège Bart in Canadian Québec.*

(A Czech sentence from the Prague Dependency Treebank.)
Prepositional Phrases and Syntactic Ambiguities

• *In years 1991 – 1993 I attended classes of management and marketing at Collège Bart in Canadian Québec.*

  – *attended at Collège Bart*
  – *classes at Collège Bart*
  – *management and marketing at Collège Bart*
  – *marketing at Collège Bart*
  – *Collège Bart in Québec*
  – *marketing in Québec*...
Prepositional Phrases and Syntactic Ambiguities

- **In years 1991 – 1993 I attended classes of management and marketing at Collège Bart in Canadian Québec.**

  - attended (class (of (mngmt and market))) (at Bart)
  - attended (class (of (mngmt and market)) (at Bart))
  - attended (class (of ((mngmt and market) (at Bart))))
  - attended (class (of (mngmt and (market (at Bart)))))
  - … ((at Bart) (in Québec))

- **Is Bart in Québec or Québec in Bart?**
Prepositional Phrases and Syntactic Ambiguities

• „říjnové jednání OSN o klimatických změnách v Kodani“ (Události ČT, 27.2.2009)
• “October UNO summit about climatic changes in Copenhagen” (Czech TV news, 2-27-2009)

• Question:
  Were there climatic changes in Copenhagen?
Verb Phrase

- The underlined finite verb form is the head.
- The repertory depends on the rules for analytical verb forms and varies greatly cross-linguistically.
  - *it* rains
  - *he* could at all sight *Mr. President*
  - *why we got* wet so much
  - *Go!*
  - *he* has been transported to the hospital on Sunday
  - *it began* to rain
  - *prohibits smoking* in this room
  - *give* Mary the beads that we brought from the vacation in Morocco
  - *the file* could not be opened
Clause

• Group of words with 1 predicate, e.g.:
  – *John loves Mary.*
  – *…that you are right.*

• Not necessarily same as a verb phrase (VP).
  – Nested VPs are part of the main VP.
  – Nested clauses are not parts of the main clause.
Clause and Sentence

• Clause
  – simple sentence or part of compound sentence
  – e.g. *John loves Mary.* or “*that you are right*”.

• Sentence
  – simple sentence or compound sentence
  – consists of one or more clauses
  – e.g. *John loves Mary.* or “*I realized that you were right.*”
Clause

• Predicative function
  – Certain activity of certain subjects and objects in certain time under certain conditions
• Main clause
  – Independent of other clauses in the sentence
• Nested clause, relative clause
  – Depends on another clause, carries out a function in that clause (as a dependent phrase)
• Functions of clauses:
  – Same as phrases plus some special, e.g. direct speech.
Sentence

- Consists of one or more main clauses.
- If there are more than one main clause then they are usually coordinated.
- A written sentence begins with a capital letter (if the script distinguishes case). Sometimes begins with a parenthesis or a quotation mark. An uppercase letter can occur inside of the sentence, too.
- It ends with a period, exclamation or question mark. Sometimes ends with a parenthesis or a quotation mark. A period can occur inside of the sentence, too.
- Depending on human decision, semicolons and colons may or may not terminate a sentence. It is usually possible to view them as coordinating conjunctions.
Coordination

• There is **no real head**. Technically, the conjunction, comma etc. can be proclaimed a head.

• The coordinated phrases are usually of the same type.
  – *chickens, hens, rabbits, cats and dogs*
  – *new or even newer*
  – *quickly and finely*
  – *he came to the conclusion that there is no point in hiding any more, so we might hear him here today*
  – *in the house or outside*
  – *to and from Prague*
  – *either now or later*
  – *not only on Monday and on Wednesday but also tomorrow or the day after tomorrow*
Apposition

- Similarly to coordination, joins two phrases none of which depends on the other.
- Unlike coordination, apposition has never more than two members.
- The combined meaning is also different:
  - *Charles IV, Roman Emperor and Czech King*
- Coordination: multiple different phrases carry out the same function together.
- Apposition: semantically only one entity; on surface, it is described by two different ways.
  - *and the most — 40 percent — befalls to family homes*
  - *factors, especially depreciation*
  - *caretaker — natural or legal person determined by the owner of the building*
  - *costs and increase of taxes — these are matters that...*
Elision

• A phrase omitted from the (surface of the) sentence although it is present in the underlying meaning (deep structure).
• Frequently in dialogues: the elided phrase is known from context.
  – *Whom did you see there?* — *Peter.* (Missing verb.)
• In written text often occurs in coordination.
  – *Czech and German researchers discussed*… (There was probably no researcher that was Czech and German at the same time. Instead, there were *Czech researchers and German researchers.*)
  – *The Penguins are leading 4:0, while the Colorado Avalanche only 3:2.* (verb in the second part)
• Systemic elision of subject in pro-drop languages (it is marked on the verb and can be deduced in the form of a pronoun).
  – *Sedím.* (*já*) = “*(I) sit.*”
Gaps and Discontinuous Phrases

• A constituent (phrase) was moved from the position where it is expected.
• Nothing special in free-word-order languages. The terms *gap* and *trace* are typically used in English (see the Penn Treebank).
• In Czech: *gap* is a term related to non-projective constructions and its meaning is different!
• English questions and relative clauses:
  – *Who do you work for* <gap>_{whom}?
  – *I don’t know why we have got so much rain* <gap>_{why}.
  – *On Sundays, I usually work* <gap>_{on sundays} *but I stay at home on Tuesdays."
  – *the story he never wrote* <gap>_{the story}
Summary of Phrase-Based Model

- Sentence is divided to phrases (constituents).
- Phrase may be divided to even smaller phrases.
- The largest phrase is the whole sentence.
- The smallest phrase is a word.
- Phrases are named and labeled according to their type.
Observation: Phrases Are Related to Context-Free Grammars

• Phrase structure of a sentence corresponds to the derivation tree under the grammar that generates / recognizes the sentence.

• Example:
  – $S \rightarrow NP \ VP$ (a sentence has a subject and a predicate)
  – $NP \rightarrow N$ (a noun is a noun phrase)
  – $VP \rightarrow V \ NP$ (a verb phrase consists of a verb and its object)

• Lexicon part of the grammar:
  – $N \rightarrow \text{dog} \mid \text{cat} \mid \text{man} \mid \text{car} \mid \text{John} \ldots$
  – $V \rightarrow \text{see} \mid \text{sees} \mid \text{saw} \mid \text{bring} \mid \text{brings} \mid \text{brought} \mid \ldots$
Lexicon

• In practice the lexical part can (and should) be implemented separately from the grammar.

• The nonterminals of the lowest level (immediately above the terminals) might be POS tags.
  – Then morphological analysis and tagging (disambiguation of MA) solves the lowest level of the phrase tree.
  • In fact, disambiguation is not necessary. There will be other ambiguities in the tree anyway. The parser can take care of them.
  – The grammar works only with POS tags.
  – This is why we sometimes talk about \textit{preterminals} (the nonterminals immediately above the leaf nodes).
An Extended Grammar Example for Czech (7 Cases!)

- $NP \rightarrow N \mid AP\ N$
- $AP \rightarrow A \mid AdvP\ A$
- $AdvP \rightarrow Adv \mid AdvP\ Adv$
- $NP_{nom} \rightarrow N_{nom}$
- $NP_{nom} \rightarrow AP_{nom}\ N_{nom}$
- $NP_{nom} \rightarrow N_{nom}\ NP_{gen}$
- $NP_{gen} \rightarrow N_{gen}$
- $NP_{gen} \rightarrow AP_{gen}\ N_{gen}$
- $NP_{gen} \rightarrow N_{gen}\ NP_{gen}$
- $N \rightarrow pán \mid hrad \mid muž \mid stroj \ldots$
- $A \rightarrow mladý \mid velký \mid zelený \ldots$
- $Adv \rightarrow velmi \mid včera \mid zeleně \ldots$
- $N_{nom} \rightarrow pán \mid hrad \mid muž \ldots$
- $N_{gen} \rightarrow pána \mid hradu \mid muže \ldots$
- $N_{dat} \rightarrow pánovi \mid hradu \mid muži \ldots$
- $N_{acc} \rightarrow pána \mid hrad \mid muže \ldots$
- $N_{voc} \rightarrow pane \mid hrade \mid muži \ldots$
- $N_{loc} \rightarrow pánovi \mid hradu \mid muži \ldots$
- $N_{ins} \rightarrow pánem \mid hradem \ldots$
An Extended Grammar Example for Czech (Verbs)

- $VP \rightarrow VP_{obligatory}$
- $VP \rightarrow VP_{obligatory} \ VP_{optional}$

- $VP_{obligatory} \rightarrow V_{intr}$
- $VP_{obligatory} \rightarrow V_{trans} \ NP_{acc}$
- $VP_{obligatory} \rightarrow V_{bitr} \ NP_{dat} \ NP_{acc}$
- $VP_{obligatory} \rightarrow V_{mod} \ VINF$

- $VP_{optional} \rightarrow AdvP_{location} \mid AdvP_{time}$ …

- $V_{intr} \rightarrow \text{šedivět} \mid \text{brzdit} \ldots$
- $V_{trans} \rightarrow \text{koupit} \mid \text{ukrást} \ldots$
- $V_{bitr} \rightarrow \text{dát} \mid \text{půjčit} \mid \text{poslat} \ldots$
- $V_{mod} \rightarrow \text{moci} \mid \text{smět} \mid \text{muset} \ldots$

- … (tens to hundreds of frames)
Unification Grammar

• An alternative to nonterminal splitting
• Instead of seven context-free rules:
  - $\text{NP}_{\text{nom}} \rightarrow \text{AP}_{\text{nom}} \text{N}_{\text{nom}}$
  - $\text{NP}_{\text{gen}} \rightarrow \text{AP}_{\text{gen}} \text{N}_{\text{gen}}$
  - $\text{NP}_{\text{dat}} \rightarrow \text{AP}_{\text{dat}} \text{N}_{\text{dat}}$
  - $\text{NP}_{\text{acc}} \rightarrow \text{AP}_{\text{acc}} \text{N}_{\text{acc}}$
  - $\text{NP}_{\text{voc}} \rightarrow \text{AP}_{\text{voc}} \text{N}_{\text{voc}}$
  - $\text{NP}_{\text{loc}} \rightarrow \text{AP}_{\text{loc}} \text{N}_{\text{loc}}$
  - $\text{NP}_{\text{ins}} \rightarrow \text{AP}_{\text{ins}} \text{N}_{\text{ins}}$
• One unification rule:
  - $\text{NP} \rightarrow \text{AP} \text{N} := [\text{case} = \text{AP}^{\text{case}} \# \text{N}^{\text{case}}]$
Syntactic Analysis (Parsing)

- Automatic methods of finding the syntactic structure for a sentence
  - Symbolic methods: a phrase grammar or another description of the structure of language is required. Then: the chart parser.
  - Statistical methods: a text corpus with syntactic structures is needed (a treebank).
  - Hybrid methods: a simple grammar, ambiguities solved statistically with a corpus.
    - Chunking / shallow parsing
Parsing with a Context-Free Grammar

- Hierarchy of grammars:
  - Noam Chomsky (1957): *Syntactic Structures*
- Couple of classical algorithms.
  - CYK (Cocke-Younger-Kasami) … complexity $O(n^3)$
    - John Cocke (“inventor”)
    - Tadao Kasami (1965), Bedford, MA, USA (another independent “inventor”)
    - Daniel H. Younger (1967) (computational complexity analysis)

- Constraint of CYK: grammar is in CNF (Chomsky Normal Form), i.e. the right-hand side of every rule consists of either two nonterminals or one terminal. (CFGs can be easily transformed to CNF.)
Parsing with a Context-Free Grammar

- **Chart parser**: CYK requires a data structure to hold information about partially processed possibilities. Turn of 1960s and 1970s: the *chart* structure proposed for this purpose.

- Jay Earley (1968), PhD thesis, Pittsburgh, PA, USA
  - A somewhat different version of chart parsing.

- For details on chart parser, see the earlier lecture about morphology and context-free grammars.
Practical Phrase-Based Parsing

• Rule-based parsers, e.g. Fidditch (Donald Hindle, 1983)
• Collins parser (Michael Collins, 1996–1999)
  – Probabilistic context-free grammars, lexical heads
  – Labeled precision & recall on Penn Treebank / Wall Street Journal data / Section 23 = 85%
  – Reimplemented in Java by Dan Bikel ("Bikel parser"), freely available
• Charniak parser (Eugene Charniak, NAACL 2000)
  – Maximum entropy inspired parser
  – P ~ R ~ 89.5%
  – Mark Johnson: reranker => over 90%
• Stanford parser (Chris Manning et al., 2002–2010)
  – Produces dependencies, too. Initial P ~ R ~ 86.4%
Probabilistic Context-Free Grammars

- PCFG (probabilistic context-free grammars)
- If there are several possible parses we want to weigh them.
- Competing parses are caused by competing rules with the same left-hand side.
- The idea: probabilistic distribution for rules with the same left-hand side.
  - Example: grammar has $\text{VP} \rightarrow \text{V NP}$ and $\text{VP} \rightarrow \text{V NP PP}$.
  - The input sentence allows both these readings, too.
  - But we know (e.g.) that the second way of building a $\text{VP}$ is more frequent:
    - $p(\text{V NP} \mid \text{VP}) = 0.3$
    - $p(\text{V NP PP} \mid \text{VP}) = 0.7$
Ambiguous Parse

- $S \rightarrow NP \ VP$
- $VP \rightarrow V \ NP \ PP$
- $VP \rightarrow V \ NP$
- $NP \rightarrow N$
- $NP \rightarrow N \ PP$
- $PP \rightarrow PREP \ N$
- $N \rightarrow \text{man}$
- $N \rightarrow \text{woman}$
- $N \rightarrow \text{car}$
- $V \rightarrow \text{saw}$
- $\text{PREP} \rightarrow \text{in}$

`man saw woman in car`
Probability of Parse Tree

• Both phrases / parses are “grammatical”.
• Different readings. Which one is better in this context?
• Probabilistic context-free grammar:
  – Relations between parent and child nodes.
  – Probability of derivation, use of a rule.
  – Probability of the whole parse tree ($r_i$ are grammar rules used to generate the sentence $S$ whose parse is $T$):

$$p(T) = \prod_{i=1}^{n} p(r_i)$$
Assumptions

• Application of a rule is independent of application of other rules in the sentence (very strong and improbable assumption).

• Independence of context of other subtrees.

• Independence of context of ancestors (higher levels).

• Independence of location in the sentence (word order) or in the tree.
Rule Probability

- Rule $r_i$: $A \rightarrow \alpha$.
- Let’s denote $R_A$ the set of all rules $r_j$ whose left-hand side is the nonterminal $A$.
- Let’s define a probability distribution on $R_A$:
  \[
  \sum_{r \in R_A} p(r) = 1 \quad \text{and} \quad 0 \leq p(r) \leq 1
  \]
- In other words:
  \[
  p(r) = p(\alpha | A) \quad r = A \rightarrow \alpha \quad \alpha \in (N \cup T)^+
  \]
Estimation of Rule Probability

- A treebank based on a context-free grammar (i.e. not a dependency treebank).

\[ r = A \rightarrow \alpha_1 \alpha_2 \ldots \alpha_k \]

\[ p(r) = \frac{c(r)}{c(A)} \]

- Frequency of rule application: how often is there this subtree in the treebank

A
  / \  / \  / \
\( \alpha_1 \) \( \alpha_2 \) ... \( \alpha_k \)