T-(538|725)-MALV, Natural Language Processing
Syntax analysis

Hrafn Loftsson[1] Hannes Högni Vilhjálmsson[1]

[1]School of Computer Science, Reykjavik University

October 2009

Outline

Outline

Syntax analysis (í. setningagreining)

"Syntax"

- í. setningafræði, setningauppbygging.
- Derived from the Greek word "sýntaxis" = arrangement
- Refers to the grammatical arrangement of words in sentences

Relation between words

- Words are linked to each other in various ways.
- We need to consider three things:
  - **Constituents; phrases** (í. setningaliðir)
    - Noun phrases, verb phrases, etc.
  - **Syntactic functions/grammatical relations** (í. setningafræðileg hlutverk/málfræðileg vensl)
    - Subject, object, predicate nominative (í. frumlag, andlag, sagnfylling)
  - **Sub-categorisation** (í. undirflokkun)
    - For example, certain type of constituents are associated with certain kind of verbs

Slide borrowed from the course: Inngangur að tungutækni, Eiríkur Rögnvaldsson, HÍ, 2003.

Arguments for the existence of constituents

How do we know that a sequence of words constitutes a phrase?

- The sequence appears in specific places in a sentence.
- The internal word order is often fixed.
- The sequence as a whole can be move around in the sentence.
- Nothing can be inserted into the sequence.
- The sequence forms a semantic entity.

Slide borrowed from the course: Inngangur að tungutækni, Eiríkur Rögnvaldsson, HÍ, 2003.

Arguments for the existence of constituents

Placement of constituents

- Ég þekki **gamla manninn** vel en konuna hans ekki (I knew old man-the well but wife-the his not).
- **Gamla manninn** þekki ég vel en konuna hans ekki.
- Ég hitti **Jón gamla á Hóli** í morgun (I met John old on Hóli this morning)
- **Jón gamla á Hóli** hitti ég í morgun.
- Margir þekkja **gömlu konuna sem missti fölsku tennurnar á gólfið** (Many know old woman-the who dropped false teeth on floor-the)
- **Gömlu konuna sem missti fölsku tennurnar á gólfið** þekkja margir.

Höskuldur Þráinsson (1999). *Íslensk setningafræði*, bls. 67.

Constituents

Noam Chomsky

- The influential work "Syntactic structures" by Chomsky (1957) is based on the concept of constituents.
- `http://en.wikipedia.org/wiki/Syntactic_Structures`

Outline

Context-free grammar (CFG)

Purpose

- To model constituent structure
- To describe a language
- A context-free grammar consists of:
  - A set of rules or productions, each of which expresses the ways that symbols of the language can be grouped together to generate valid sentences.
  - A lexicon of words and symbols

Context-free grammar

An example

Rules	Lexicon	
S → NP VP	Det → the	Noun → day
NP → Det Noun	Noun → waiter	Verb → brought
NP → NP PP	Noun → meal	Prep → to
VP → Verb NP	Noun → table	Prep → of
VP → Verb NP PP		
PP → Prep NP		

- One symbol to the left of the arrow, one or more symbols to the right.
- S, NP, VP, PP: **non-terminals** (í. máleiningar)
- Individual words: **terminals** (í. tókar/lokatákn)

Context-free grammar

Consists of:

1. A set of **non-terminals** N
2. As set of **terminals** Σ
   - Does not overlap with N
3. A set of **productions** P, of the form $A \rightarrow \alpha$, where A is a non-terminal and α is a string of symbols from $(\Sigma \cup N)^*$
4. **A start symbol**, S

Context-free grammar

Derivation (í. afleiðsla)

- In the grammar $\rightarrow$ = "rewrite the symbol on the left with the string of symbols on the right"
- Or $\rightarrow$ = "derives"
- NP $\rightarrow$ Det Noun $\rightarrow$ the Noun $\rightarrow$ the waiter
- The string "the waiter" can thus be derived from the non-terminal NP.

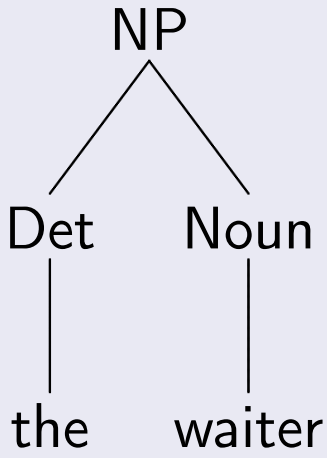
Start symbol

- The language which the CFG describes is the set of strings which can be derived from the start symbol.
- The start symbol is often named S (the sentence node).

Parsing

A parse tree

A derivation can always been shown using a parse tree:



- A leaf stands for a **terminal**, an intermediate node in the tree stands for a **non-terminal**.
- **Parsing** (í. þáttun): mapping a string of words to a parse tree.

Correct and incorrect sentences

- A CFG can be viewed as:
  - A device to **generate** sentences.
  - A device to assign **structure** to sentences.
- Grammatical sentences
  - sentences which **can** be generated by the grammar.
- Ungrammatical sentences
  - sentences which **cannot** be generated by the grammar.

Slide borrowed from the course: Inngangur að tungutækni, Eiríkur Rögnvaldsson, HÍ, 2003.

A simple CFG for Icelandic

Production	Explanation
S → NP VP (PP|AdvP)	NP=noun phrase, VP=verb phrase
VP → verb (NP|PP)	PP=preposition phrase, ADVP=adverb phrase
NP → (AP) noun|pers (PP)	AP=adjective phrase
AP → (AP) adj	pers=personal pronoun, adj=adjective
PP → prep NP	prep=preposition
AdvP → (AdvP) adverb	

A phrase inside parantheses is optional.

An example of a derivation

- S ⇒ NP VP PP ⇒ noun VP PP ⇒ noun verb PP ⇒ noun verb prep NP ⇒ noun verb prep noun ⇒ *maðurinn datt í gær* (man-the fell yesterday)
- S ⇒ NP VP AdvP ⇒ AP noun VP AdvP ⇒ adj noun VP AdvP ⇒ adj noun verb AdvP ⇒ adj noun verb adverb ⇒ *stóri maðurinn hljóp hratt* (big man-the ran fast)
- Phrases are often shown inside brackets:
  - [S [NP **maðurinn**] [VP **datt**] [PP **í** [NP **gær**]]]
  - [S [NP [AP **stóri**] **maðurinn**] [VP **hljóp**] [AdvP **hratt**]]

The Icelandic noun phrase

- NP ⇒ (AP) noun|pers (PP)
- Too simple – for example cannot handle:
  - *allir þessir þrír stóru strákar* (all these three big boys)
    - indefinite pronoun, demonstrative pronoun, numeral, AP, noun
  - *strákinn sinn* (boy-the his)
    - noun, possessive pronoun
- NP ⇒ (indefpronoun) (dempronoun) (numeral) (AP) noun|pers (posspronoun) (PP)
  - [NP Allir þessir þrír [AP stóru AP] strákar NP]
  - [NP strákinn sinn NP]

The Icelandic noun phrase

- NP ⇒ (indefpronoun) (dempronoun) (numeral) (AP) noun|pers (posspronoun) (PP)
- Better now, but cannot handle NPs like:
  - [NP Allt NP] (All)
  - [NP Jón Jónsson NP]
  - [NP Mennirnir tveir NP] (Men-the two)
  - [NP Hver NP] (Who)
  - [NP sjálfan sig NP] (himself)
  - [NP þetta allt] (this all)
  - [NP landið allt] (country all)
  - [NP hinn stóri NP] (the big)

⇒ It is not simple to develop a perfect CFG for a natural language!

Outline

Verb subcategorisation (í. undirflokkun sagna)

- Verbs demand (accept) various types of complements (í. fyllilíðir)
  - VP $\Rightarrow$ verb (VP | (NP) (NP) (PP))
- Verbs have different subcategorisation frames (í flokkunarrammar)
  - [_], [_VP], [_NP], [_NP NP], [_NP PP], ...

Slide borrowed from the course: Inngangur að tungutækni, Eiríkur Rögnvaldsson, HÍ, 2003.

Verb subcategorisation

Examples of subcategorisation frames

- Ísinn hefur bráðnað [_] (Ice-the has melted)
- Páll hefur saknað þín [_ NP] (Páll has missed you)
- Þeir hafa fjallað um málið [_ PP] (They have discussed about case-the)
- Friðrik hefur verið grannur [_ AP] (Friðrik has been slender)
- Þeir munu hafa étið útsæðið [_ VP] (They will have eaten seed-the)
- Jón hefur lánað Maríu hring [_ NP NP] (Jón has lent Maríu ring)
- Ég hef stungið klút í vasann [_ NP PP] (I have put cloth in pocket-the)
- Hann vill mála bílinn rauðan [_ NP AP] (He wants paint car-the read)

Verb subcategorisation

Transitive verbs in Icelandic

A verb which governs a case is a **transitive verb** (í. áhrifssögn) and makes the nominal stand in an oblique case (í. aukafalli).

- The nominal (í. fallorð) is called an **object** (í. andlag)
- *Ég tek hnífinn (þf.)*, *Ég mætti manni (þgf.)*, *Ég þarfnast næðis (ef.)*
  - *I take knife-the (accusative)*, *I met man (dative)*, *I need privacy (genitive)*
- *Hann gefur kettinum (þgf.) silung (þf.)* (di-transitive verb)
  - *He gives cat-the (dative) trout (accusative)*

Verb subcategorisation

Intransitive verbs in Icelandic

A verb which does not govern a case is **intransitive** (í. áhrifslaus).

- *Maðurinn hlær*
- *Man-the laughs*

Outline

Grammatical functions

Subject (í. frumlag)

- The grammatical constituent (a noun phrase) about which something is predicated.
- *[NP Gamli hundurinn NP]* beit póstinn. (Old dog bit postman-the)

Object (í. andlag)

- A constituent that is acted upon.
- [NP Gamli hundurinn NP] beit *[NP póstinn NP]*.

Grammatical functions

Predicative nominative (í. sagnfylling)

- A nominal which appears in the nominative case with an intransitive verb is called a **predicative nominative; verb complement**.
- It supplements the subject
  - Setningarliður sem stendur með áhrifslausri, ósjálfstæðri sögn (verða, verða, heita, þykja, o.s.frv.) og afmarkar eða fyllir merkingu hennar.
- [NP Hún NP] er *[NP mjög góður leikari NP]*. (She is very good actor)
- [NP Hann NP] heitir *[NP Jón NP]*. (He name Jón)

Grammatical functions

Umsögn (e. finite verb)

- A finite verb is a verb that is inflected for person, tense, number, mood.
  - Umsögn er sögn í persónuhætti (framsöguhætti, viðtengingar- eða boðhætti). Kjarni eða aðalliður setningar og allir aðrir setningarhlutar tengjast henni á einhvern hátt.
- Hann *borðar* fiskinn. (He eats fish-the).
- Ég *borða* fiskinn. (I eat fish-the).