T-(538|725)-MALV, Natural Language Processing
Partial parsing

Hrafn Loftsson[1] Hannes Högni Vilhjálmsson[1]

[1]School of Computer Science, Reykjavik University

October 2009

Outline

Outline

Full/deep parsing (í. full þáttun)

Aim

- To perform deep analysis.
- To construct "a complete" parse tree.
- Various linguistic theories have been used, e.g.:
  - CFG – **C**ontext **F**ree **G**rammar
  - PCFG – **P**robabilistic **C**ontext **F**ree **G**rammar (Collins 1996; Charniak 1997)
  - HPSG – **H**ead-Driven **P**hrase **S**tructure **G**rammar (Pollard and Sag 1994).
  - LFG – **L**exical **F**unctional **G**rammar (Kaplan 1989).
  - DG – **D**ependency **G**rammar – (Tesnière 1966)

Full parsing

The problem

- Difficult and time consuming to build a parser with large grammatical coverage.
- The solution set can grow exponentially.
  - Because often the parser tries to build all possible parse trees for a given sentence.
- The parser might also reject a correct analysis of a part of the sentence.
  - In the case when it does not fit into a global parse.

Outline

Partial/shallow parsing (í. hlutaþáttun)

Aim

- Analyse sentence parts (or chunks) without building a complete parse tree.
- "to recover syntactic information efficiently and reliably from unrestricted text, by sacrificing completeness and depth of analysis" (Abney 1996).

When is partial parsing suitable?

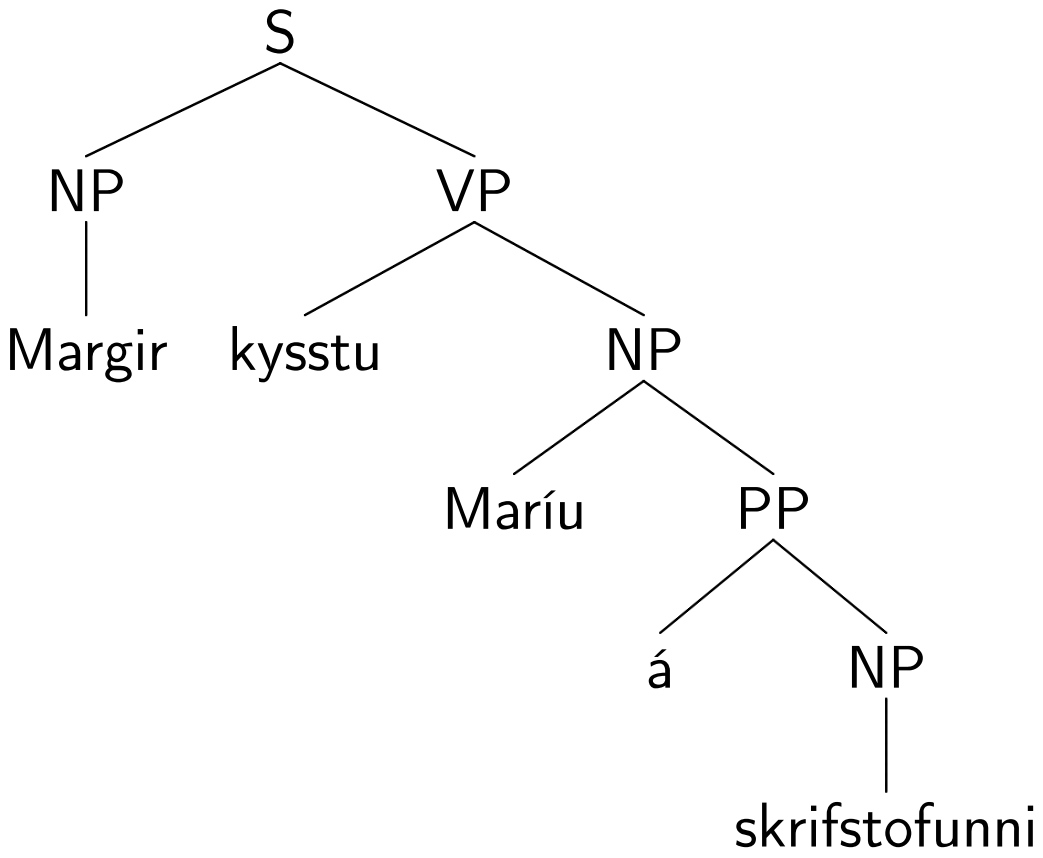
- When full analysis is not necessary.
  - For example, in information retrieval or information extraction.
- When efficiency is of prime importance.
- When the quality of the input is poor.

Full parsing vs. Partial parsing

- *Margir kysstu Maríu á skrifstofunni (Many kissed Maria at office-the)*
  - (Höskuldur Þráinsson (1999). Íslensk setningafræði, bls. 70)
- **Full parsing**:
  - [S [NP Margir] [VP kysstu [NP Maríu [PP á [NP skrifstofunni]]]]]
  - [S [NP Margir] [VP kysstu [NP Maríu]] [PP á [NP skrifstofunni]]]
- **Partial parsing**:
  - [NP Margir] [VP kysstu] [NP Maríu] [PP á [NP skrifstofunni]]

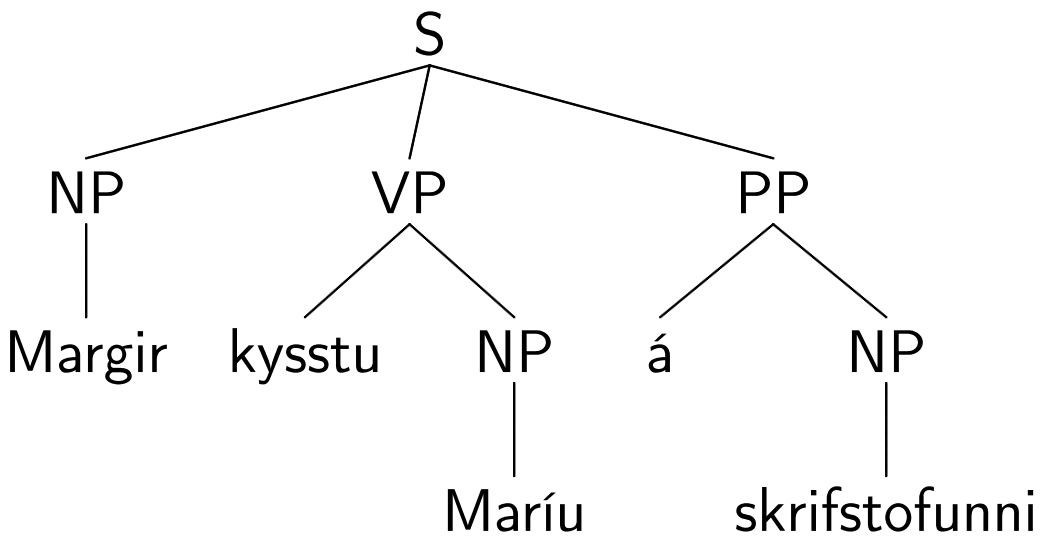
Full parsing, one interpretation

- *Margir kysstu Maríu á skrifstofunni*



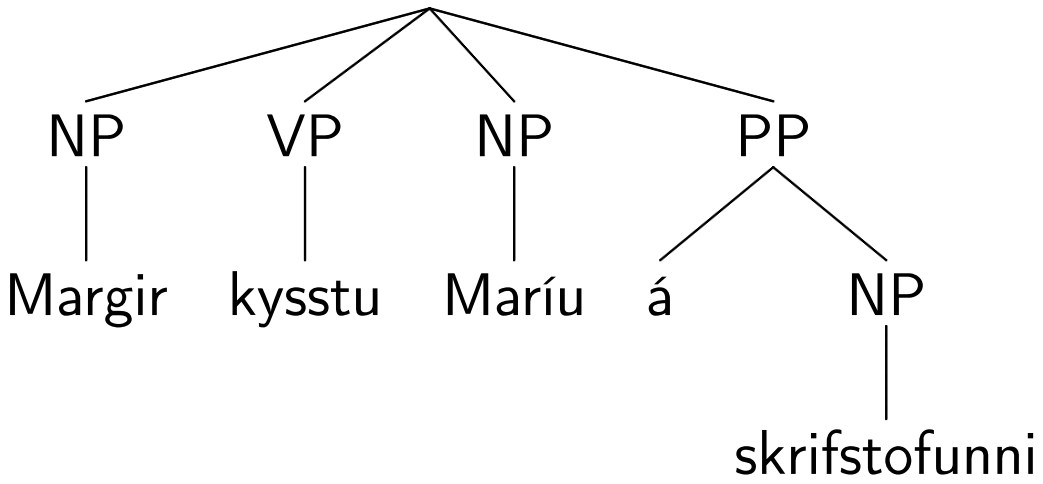
Full parsing, another interpretation

- *Margir kysstu Maríu á skrifstofunni*



Partial parsing

- *Margir kysstu Maríu á skrifstofunni*



Outline

Multiword expressions (MWE) (í. fleiryrtar segðir)

Skilgreining

- Sequences of two or more words that act as a single lexical unit, for example:
  - Proper nouns (names): persons, companies, institutions
  - Temporal expressions: times, dates
  - Numerical expressions: numbers and amounts
  - Sequences of words that function like a conjunction, adverb, adjective, or a preposition.

Examples of MWEs

MWE_AdvP = MWE functioning as an **adverb**
MWE_PP = MWE functioning as a **preposition**
MWE_CP = MWE functioning as a **conjunction**
MWE_AP = MWE functioning as an **adjective**

```
[CP en c CP] [MWE_AdvP einhvern fokeo veginn nkeog MWE_AdvP]
[CP but CP] [MWE_AdvP somehow MWE_AdvP] ...

[PP [MWE_PP uppi aa á aþ MWE_PP] [NP bakkanum nkeþg NP] PP]
[PP [MWE_PP up on MWE_PP] [NP bank-the NP] PP]

[MWE_CP án ae þess fphee að cn MWE_CP] [VPi hafa sng VPi]
[MWE_CP without MWE_CP] [VPi having VPi] ...
```

Examples of MWEs

```
[MWE_AdvP við aa og c við aa MWE_AdvP] [VP gægðist sfm3eþ VP]
[MWE_AdvP now and then MWE_AdvP] [VP looked VP] ...

[CP og c CP] [NP [MWE_AP hvers fohee kyns nhee MWE_AP] drasl nhen NP]
[CP and CP] [NP [MWE_AP various MWE_AP] stuff NP]
```

Examples of MWEs

- The expressions above can be solved by using some kind of word lists.
- Proper nouns, temporal expressions and numerical expressions are a different story.
  - Nevertheless, wordlists, **gazetteers**, are sometimes used for proper nouns.
- But note that it may be necessary to recognise expressions that are not matched using hard-coded lists.
- $\Rightarrow$ Regular expressions!

Outline

Chunks (í. klumpar)

- A chunk = Group of words
- The difference between a chunk and a constituent is that the latter can contain nested constituents of the same type.
  - See for example slide no. 21 in the lecture "CFG and Prolog"
- A CFG is used to allow recursion in constituents
- Chunks do not contain nested chunks of the same type
- Regular expressions are thus sufficient to describe chunks.

A description of an Icelandic noun chunk

- Let us simplify the matter and only assume that a noun chunk can contain:
  - an adverb, an adjective, a noun
  - for example, “saga” (a story)
  - or “skemmtileg saga” (a fun story)
  - or “mjög skemmtileg saga” (a very fun story)

In JFlex:

```
%% A finite-state automaton recognising simple noun chunks
%public
%class NounChunk
%standalone
%unicode

%{
    String Open="[NP " ;
    String Close="NP] ";
%}

WhiteSpace = [ \t\f]
WordChar = [^\r\n\t\f ]
Word = {WordChar}+
WordSpaces = {Word}{WhiteSpace}+

Gender = [kvhx] /* k=masculine, v=feminine, h=neuter, x=unspec */
Number = [ef]   /* e=singular, f=plural */
Case = [noþe]   /* n=nominative, o=accusative, þ=dative, e=genitive */
```

In JFlex (cont.):

```
AdverbTag = aa[me]?
AdjectiveTag = l{Gender}{Number}{Case}[sv][fme]
NounTag = n{Gender}{Number}{Case}[g\-]?[mös]?

Adverb = {WordSpaces}{AdverbTag}{WhiteSpace}+
Adjective = {WordSpaces}{AdjectiveTag}{WhiteSpace}+
Noun = {WordSpaces}{NounTag}{WhiteSpace}+

NounChunk = {Adverb}?{Adjective}?{Noun}

%%
{NounChunk} { System.out.print(Open + yytext() + Close);}
.           { System.out.print(yytext());}
```