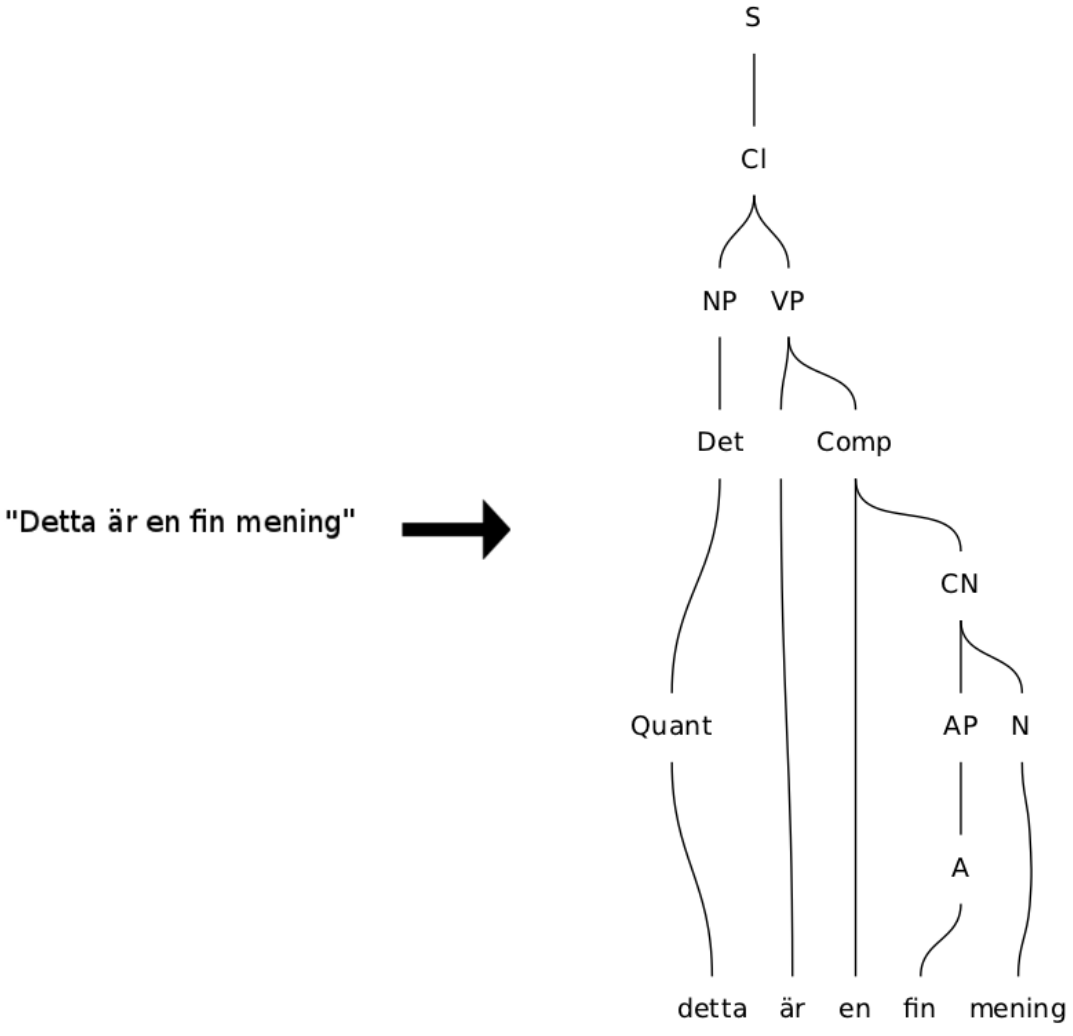
Towards a Wide-Coverage Grammar for Swedish Using GF

Malin Ahlberg

Parsing

"Detta är en fin mening"



Grammatical Framework

Introduction

A grammar formalism for multilingual applications

Grammatical Framework

Introduction

A grammar formalism for multilingual applications

A functional programming language based on Martin-Löf type theory

Grammatical Framework

```
PredVP : NP -> VP -> S
```

Abstract syntax

```
PredVP np vp = np ++ vp
```

```
PredVP np vp = vp ++ np
```

Concrete syntax

Grammatical Framework

Abstract syntax

Categories

`NP`	"en liten katt"	(*"a small katt"*)
`VP`	"äter äpplen"	(*"eats apples"*)
`AP`	"ganska gott"	(*"rather good"*)
`V`	"sitta"	(*"sit"*)
`V2`	"gilla"	(*"like"*)
`VA`	"bli"	(*"become"*)
`VPSlash`	"gillade inte"	(*"did not like"*)
`ClSlash`	"han gillade inte"	(*"he did not like"*)

Grammatical Framework

Abstract syntax

Function types

```
SlashV2a : V2 -> VPSlash

ComplSlash : VPSlash -> NP -> VP

PredVP : NP -> VP -> Cl
```

Grammatical Framework

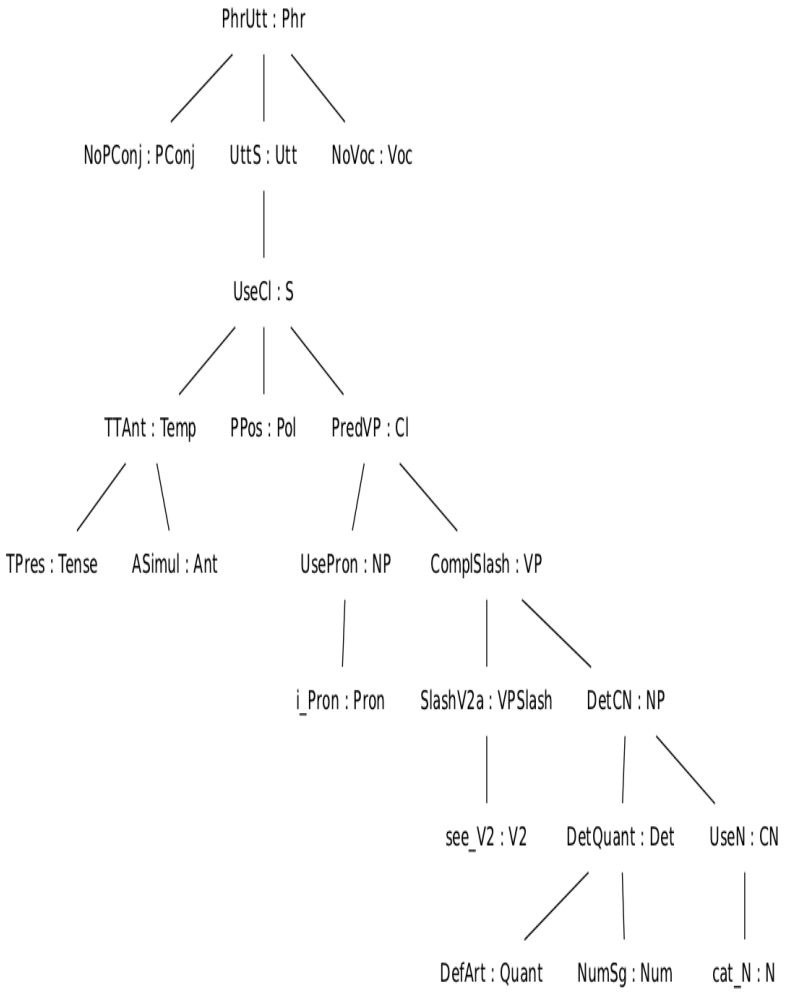
Concrete syntax

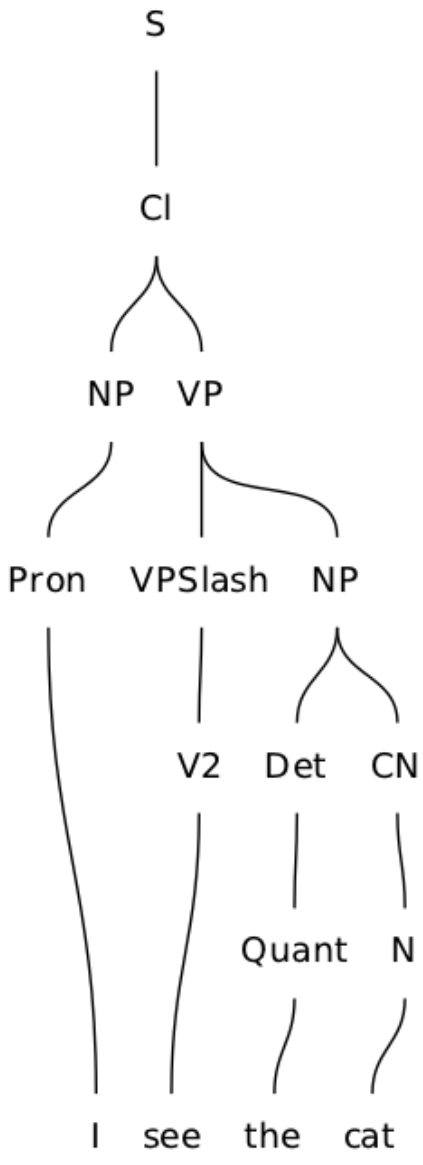
- **One concrete syntax for each grammar**
- **Linearization rules**

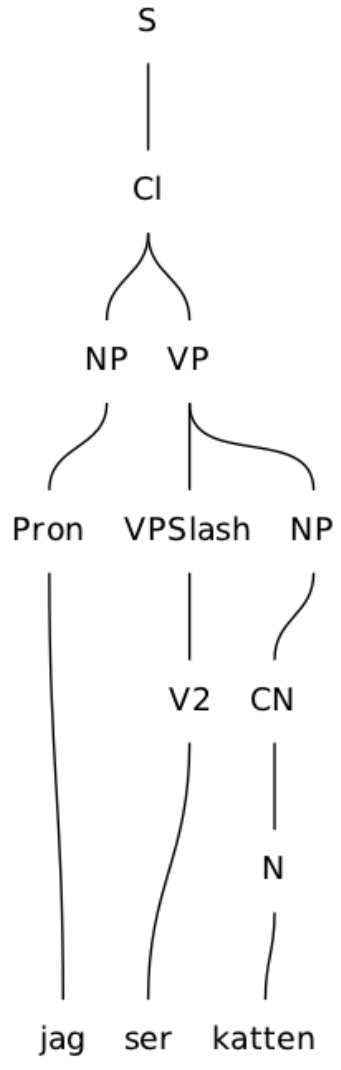
```
ComplSlash vp obj = vp ++ obj

PredVP np vp = np ++ vp
```

Grammatical Framework







Grammatical Framework

Resource grammars

The resource grammars covers about 20 languages

Grammatical Framework

Resource grammars

The resource grammars covers about 20 languages

Extra module for language specific constructions

Grammatical Framework

Resource grammars

The resource grammars covers about 20 languages

Extra module for language specific constructions

Controlled natural language

The project

Aims

- Extending the Swedish GF grammar
- Importing a large lexicon
- Creating translation between Talbanken and GF

The project

Talbanken

$\sim$ 6000 manually tagged sentences

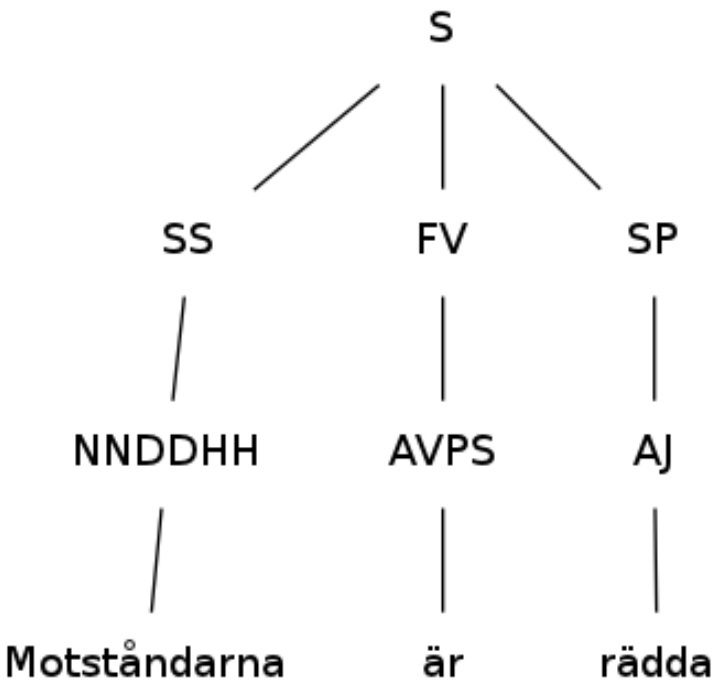
Developed at Lund university

The project

Talbanken

$\sim$ 6000 manually tagged sentences

Developed at Lund university

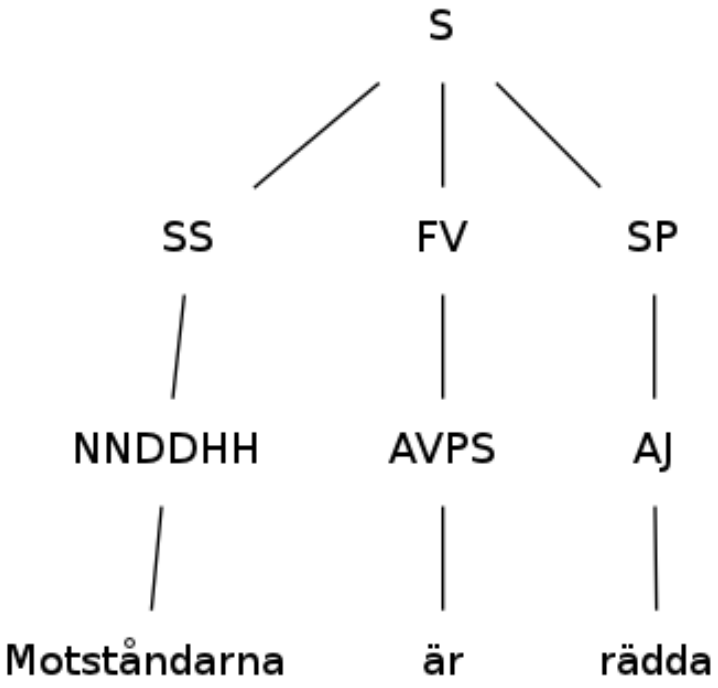


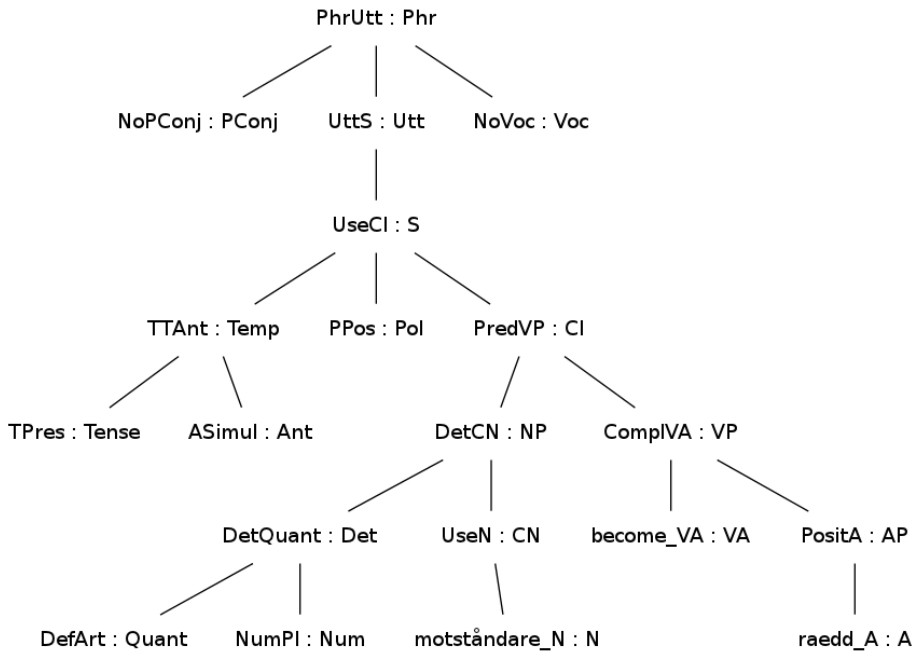
The project

Talbanken

$\sim$ 6000 manually tagged sentences

Developed at Lund university





A Treebank for GF

Translating Talbanken to GF

- Evaluate the parser

A Treebank for GF

Translating Talbanken to GF

- Evaluate the parser
- Extract grammatical and lexical information

A Treebank for GF

Translating Talbanken to GF

- Evaluate the parser
- Extract grammatical and lexical information
- Extract probabilities for GF functions

A Treebank for GF

Mapping Talbanken to GF

```
translate S = do np <- translate SS
                 vp <- translate FV
                 return (PredVP np vp)
```

A Treebank for GF

Problematic parts

XX	Unclassifiable grammatical function
NAC	Not a constituent
PU	List item

Sentence 5120:

"...

1. *förökelsen av människosläktet*
2. *motverkandet av otukt*
3. *utlevandet av genuint kristen kärlek*

..."

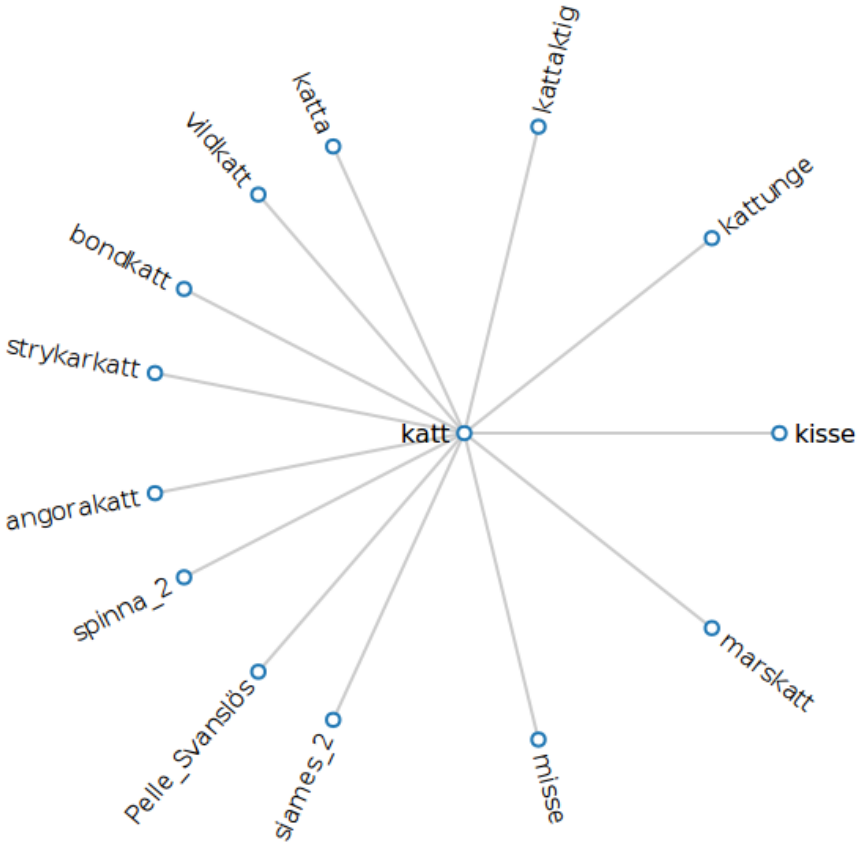
A Treebank for GF

Results

No list items	65 %
No special punctuation or bad tags	72 %
Short sentences with known words	85 %

Saldo

A large-scale lexicon



SALDO

grundform	katt
mönster	nn_3u_film
ordklass	nn
inherenta drag	u
böjningstabell	
sg indef nom	katt
sg indef gen	katts
sg def nom	katten
sg def gen	kattens
pl indef nom	katter
pl indef gen	katters
pl def nom	katterna
pl def gen	katternas
ci	katt/katt-
cm	katts/katt/katts-/katt-
sms	katt-

A large-scale lexicon

GF lexicons are generated by smart paradigms:

Regular verb `mkV "hittar"`

Irregular verb `mkV "knyter" "knöt" "knutit"`

VF (VPres Act) : hittar
VF (VPret Act) : hittade
VF (VImper Act) : hitta
VI (VInfin Act) : hitta
VI (VSupin Act) : hittat

Importing Saldo

`mkV "hittar"`	`mkV "knyter"`
`VF (VPres Act) : hittar`	`VF (VPres Act) : knyter`
`VF (VPres Pass) : hittas`	`VF (VPres Pass) : knyts`
`VF (VPret Act) : hittade`	`VF (VPret Act) : knytte`
`VF (VPret Pass) : hittades`	`VF (VPret Pass) : knyttes`
`VI (VInfin Act) : hitta`	`VI (VInfin Act) : knyta`
`VI (VSupin Act) : hittat`	`VI (VSupin Act) : knytt`

Importing Saldo

Results

Result: Over 100 000 entries

Importing Saldo

Results

Result: Over 100 000 entries

$\sim$ 500 missing words in Talbanken

Importing Saldo

Results

Result: Over 100 000 entries

$\sim$ 500 missing words in Talbanken

$\sim$ 150 of them are used more than once

Importing Saldo

Results

Result: Over 100 000 entries

$\sim$ 500 missing words in Talbanken

$\sim$ 150 of them are used more than once

Plural tantum nouns	Irregular s-verbs
glasögon	*umgås*

Grammar

Swedish

Topicalization

Jag äter äpplet nu

Äpplet äter jag nu

“I eat the apple now”

Passive

Äpplet blir ätet av mig

Äpplet äts av mig

“The apple is eaten by me”

Grammar

Swedish

Future tense

Jag ska sova nu

"I will sleep now"

Jag kommer somna snart

"I will fall asleep soon"

Impersonal constructions

Det bor två barn i huset

"There are two children living in the house"

Det dansas på borden

"People are dancing on the tables"

Grammar

Swedish

Reflexive possessive pronous

Jag äter mitt äpple	Jag äter mitt äpple
"I eat my apple"	*"I eat my apple"*
Han äter sitt äpple	Han äter hans äpple
"He eats his apple"	*"He eats his apple"*

Grammar

Reflexive possessive pronous

sin frukt	sitt äpple	sina äpplen
"SELF'S *fruit*"	"SELF'S *apple*"	"SELF'S *apples*"

Grammar

Reflexive possessive pronous

sin frukt	sitt äpple	sina äpplen
"SELF'S *fruit*"	"SELF'S *apple*"	"SELF'S *apples*"

*Jag äter sitt äpple

"*I eat* SELF'S *apple*"

The grammar

Reflexive pronouns

```
ReflVP : VPSlash -> VP ;

         ser     -> ser sig;
```

"Han ger sina pengar till sina barn"

The grammar

Reflexive pronouns

sitt äpple

NP (OBJECT)

*Sitt äpple är stort

The grammar

Reflexive pronouns

sitt äpple *Sitt äpple är stort

NP (OBJECT)

sitt äpple **och** **en banan**

NP (OBJECT) + NP → NP (OBJECT)

The grammar

Reflexive pronouns

sitt äpple		*Sitt äpple är stort		
NP (OBJECT)				
sitt äpple	**och**	**en banan**		
NP (OBJECT)	+	NP	→	NP (OBJECT)
sina äpplen	+	**alla**	→	**alla sina äpplen**
NP (OBJECT)				NP (OBJECT)

The grammar

NP Subject/Object

The grammar

NP Subject/Object

```
ConjunctionNP : (a : NPType) -> NP a -> NP a -> NP a ;

PredVP : NP Subject -> VP -> Cl ;

ComplSlash : VPSlash -> NP Object -> Cl ;
```

The grammar

Reflexive pronouns

- The new grammar can be made compatible with the old one
- The separation of noun phrases needing an antecedent is also be needed for other languages

Evaluation

Results

- a study of how dependent types can be used in the resource grammars
- a large-scale GF lexicon and a program to redo the importation when needed
- a comparison between GF and another annotation
- a deeper testing of the Swedish resource grammar and an estimation of how well GF can be used to describe larger parts of a language

Future work

Lexicon

- Idioms
- Speed up
- Valencies

Future work

Lexicon

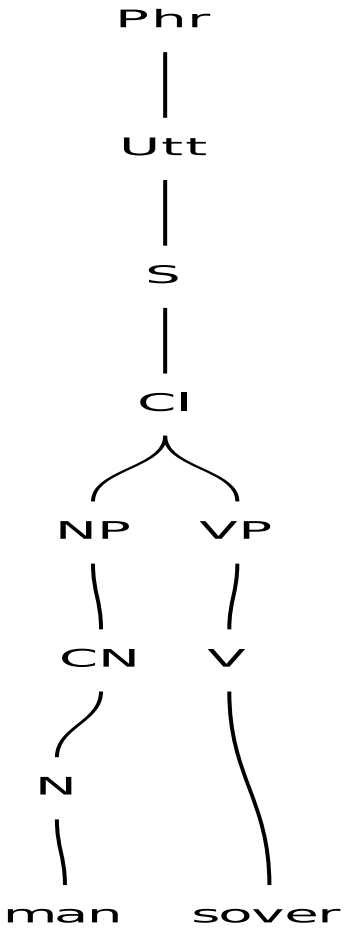
A verb may have type V2 if it is followed by:

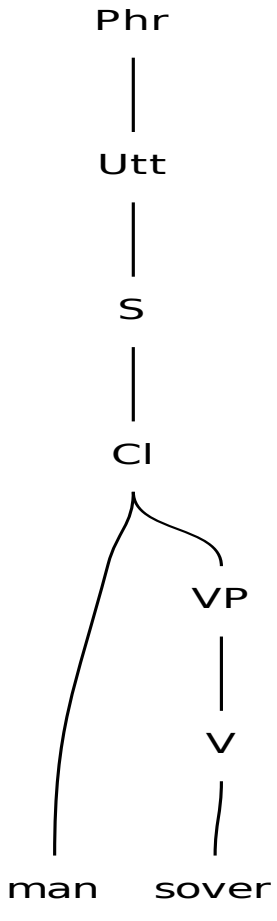
- `OO` (other object)
- `SS` (subjective predicative complement)
- `IO` (indirect object)
- `OA (PP ..)` (other adverbial with a prepositional phrase)

Future work

Probabilities

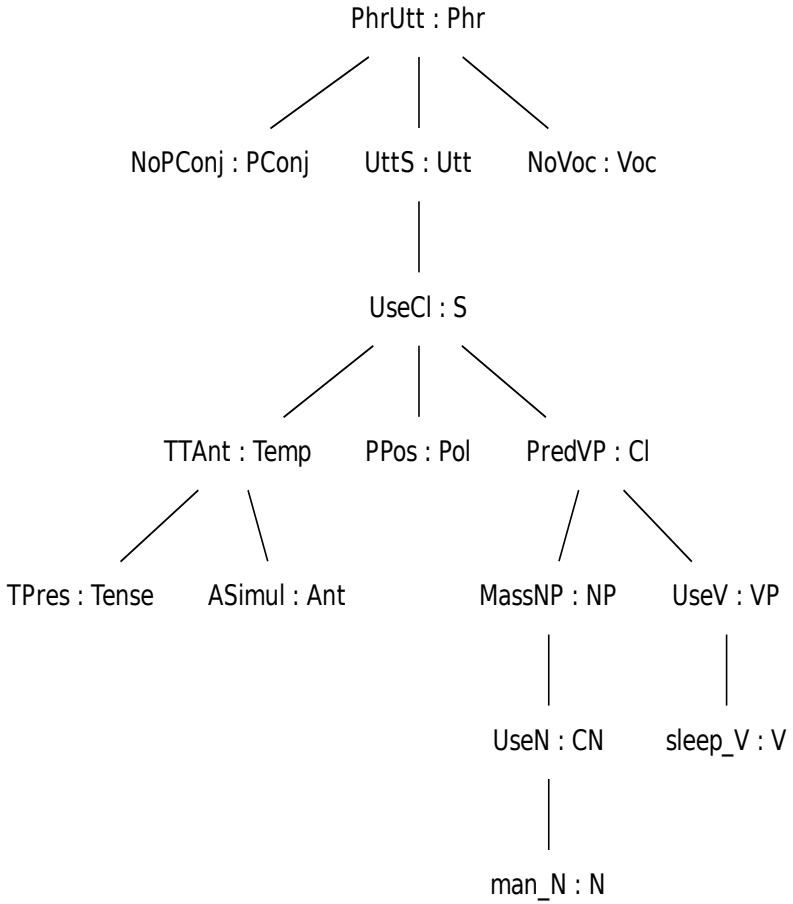
"Jag såg en"

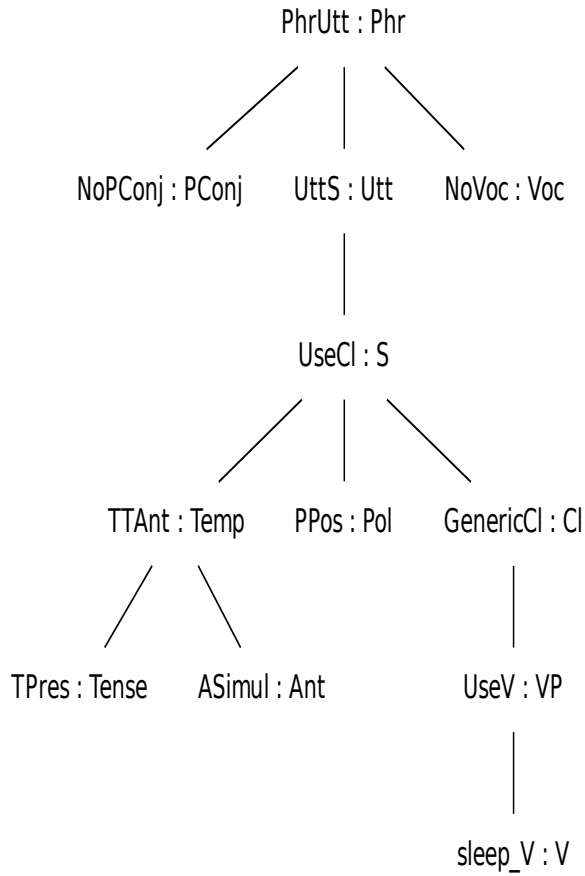




Future work

Probabilities





Future work

Robustness

- Named entity recognition
- Chunk parsing

The end

Thanks for listening

clt.gu.se/seminar/2012-01-09/masters-seminar-malin-ahlberg