Stockholm University

Basic Natural Language Processing II Parsing and Generation

Hercules Dalianis

Department of Computer and Systems Sciences (DSV)

hercules@dsv.su.se

Stockholm University

Overview

- Properties of natural languages
- Grammars
- Parsing (Understand text)
    - Chunking - Shallow parsing
    - Parsing
- Generation (Generate text)
    - Deep generation
    - Surface generation

Stockholm University

Properties of natural language

- Natural language is ambiguous
- Dynamic – it changes over time.
- Syntax describes the allowed tokens in a sentence or a text
- Semantics describes the meaning of a sentence or a text
- Pragmatics describes the use of a sentence.

Stockholm University

Syntax

- Sentences have syntax. The order of symbols/words
  - *This sentence is written in the correct syntax.*
  - *This sentence written is syntax wrong.*

Semantics

- *Sentences should have – semantics – the meaning of symbols*
  - *This sentence is written in the correct syntax.*
  - *Colorless green ideas sleep furiously*

    semantically nonsense OR poetry?

Discourse

- Discourse
  - How sentences are interconnected
  - Speaker and receiver
  - Coherence between sentences and phrases.

Pragmatics

- A sentence may have pragmatics. How to use the sentence. Pragmatics of symbols
  - *Can you pass the salt?*

Natural language

- Ambiguous
  - *She saw the man on the hill with the binoculars*

  Changing over time
  - New expressions and words
- Not complete
- Humour or sarcastic or fake
- Pragmatics and creative expressions
  - Background knowledge
- Poetic

Stockholm University

Tokenisation and POS tagging on previous lecture 2 Basic Natural Language Processing I by Martin Duneld)

Tokenisation

- Read a string and tokenize it
  - Space is one delimiter
  - But also **.,?!**
  - How do you tokenize 20.4 mg?

  “The patient took 20.4 mg Prednisolon.”

  =>

  ”The” “patient” ”took “20.4” “mg” “Prednisolon” “.”

Stockholm University

Tagging

- Read the tokenised text and tag it

  “The” “patient” ”took “20.4” “mg” “Prednisolon” “.”

  ⇒POS - Part of Speech tagging

  ('The', 'DT'), ('patient', 'NN'), ('took', 'VBD'), ('20.4', 'CD'), ('mg', 'NN'), ('Prednisolon', 'NNP'), ('.', '.')

Stockholm University

Chunk and chunking

- Group of words that are together in a fixed order
  - *My name is..*
  - *How do you do?*
  - *In my opinion..*
  - *air plane*
  - *master student*
- Chunks are also called collocations or lexical phrase

Chunking - Shallow parsing – Light parsing – Partial parsing

- Chunking group or segment words based on their pos tags in chunks such as nominal phrases, verb phrases etc, or just pair wise such as
  - The patient took 20.4 mg Prednisolon
  - (The patient $_{np}$), took (20.4 mg Prednisolon $_{np}$)

Stockholm University

Chunkers

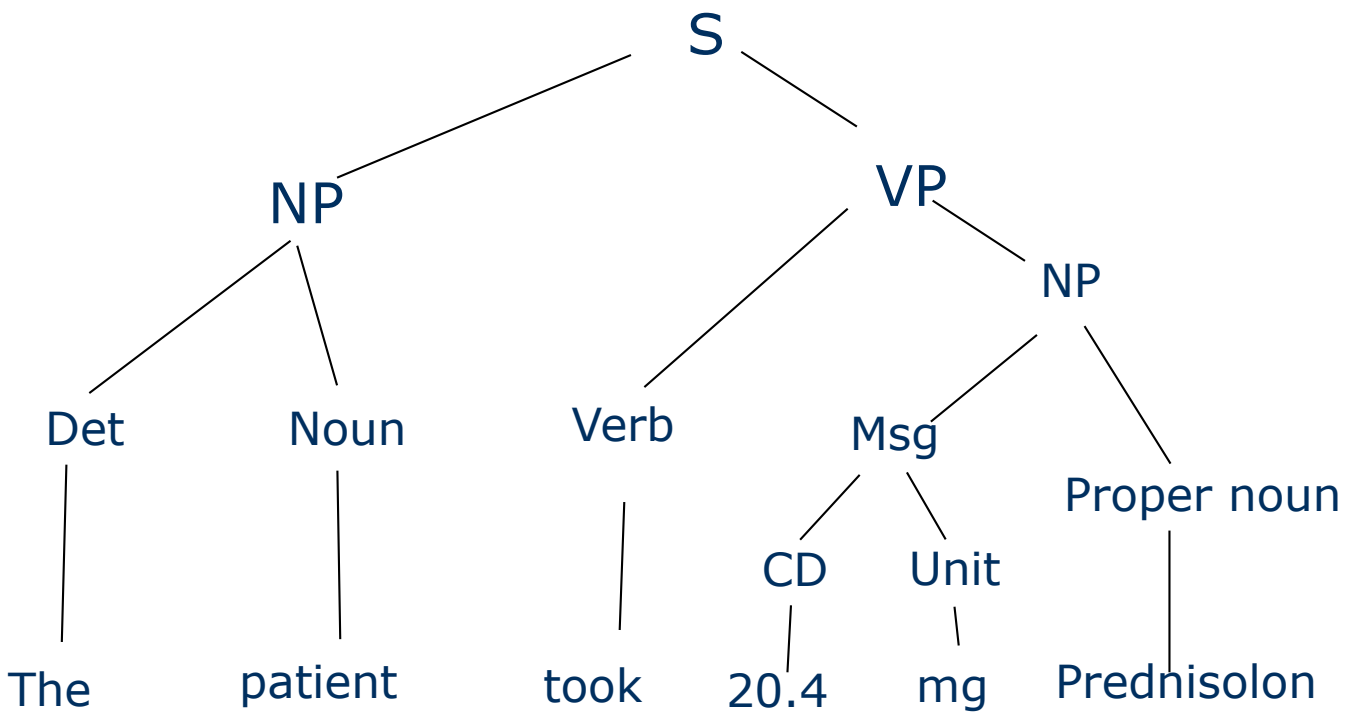
- Chunkers can be rule based
  - Lots of work to write rules
- Chunker can be trained
  - On chunked annotated corpora such as tree banks

Stockholm University

Parsing

- Find parts of a sentence
  - Parsing => Latin term for "part of speech" – "pars orationis.
- Parts can be verb phrases, noun phrases, prepositional phrases, that in turn is divided into determiners, nouns, adjectives, adverb, conjunctions and all this part constitutes a sentence that in turn becomes a text.

Syntactic tree



Stockholm University

Mathematics, logic, programming languages are unambiguous

- Symbols in mathematics
  - +, -, *, /, (), = , Σ, Π
- Sympols in programming languages

  print('hello') OR for n=1 do x
- Correct syntax and non-ambiguous
- Limited lexical symbols (and not growing/changing)
- Described by a formal grammar,

Stockholm University

Grammar

- Grammar that describe
  - Mathematics
  - Logic
  - Programming languages
- Described by a formal grammar, https://en.wikipedia.org/wiki/Formal_language

Compilers - parsers

- For programming languages
- Uses grammar to describe language
- Parses program code and generates machine code – so called compilers
- Produces machine code
  - C++, Java etc, (compiled)
    (Python is interpreted during runtime)
  - One can create a new programming language
- Compilers can be written using for example Prolog or Python, or C++

Stockholm University

Natural languages (NL)

- Are ambiguous
- Is a challenge for a parser / computer
- A human can deal with ambiguous NL

Stockholm University

Grammar and syntax

- A grammar is used to describe the syntax of a (natural or programming) language
- A (generative) grammar can also be used to generate correct sentences
- First grammar to describe a language was for Sanskrit by the linguist Panini in 500 B.C

Stockholm University

Grammar and parser

- A grammar describes the language
- A grammar contains rewrite rules

```
s --> np, vp.
np --> det, noun.
noun --> 'flower'.
```

- Parser uses a grammar to parse a sentence.
- For example if using Prolog DCG grammar, everything is built in you don't need to write a parser to traverse a grammar

Stockholm University

Grammars

- Definite Clause Grammar (DCG)
- Context Free Grammar (CFG)
- Dependency Grammar (DG)
- Lexical Functional Grammar (LFG)
- Look-Ahead Left to Right parser (LALR)
  - Different parsers executes grammars.

Stockholm University

Prolog and DCG grammar

- The programming language Prolog
  - Built in theorem prover
  - Facts and Theorems – proof machine
  - True - Yes or False - No answer
- DCG (Definite Clause Grammar) format can easily be used to define and execute a grammar.
- Prolog compatible

Stockholm University

```
% Grammar
s --> np, vp.
vp --> verb.
vp --> verb, np.
np --> article, noun.
np --> article, adjective, noun.
%np --> np, conj, np.

% Dictionary
verb --> [is].
article --> [the].
noun --> [flower].
adjective --> [red].
conj --> [and].
```

Parse

```
% Parse
?- s([the, flower, is, flower],[]).
false.

?- s([the, flower],[]).
false.

?- s([the, flower, is, the, flower],[]).
true .
```

Stockholm University

Generate

```
?- s(L,[]).
S = [the, flower, is] ;
S = [the, flower, is, the, flower] ;
S = [the, flower, is, the, red, flower] ;
S = [the, red, flower, is] ;
S = [the, red, flower, is, the, flower] ;
S = [the, red, flower, is, the, red, flower].
```

Stockholm University

Grammar rules

- Several thousand grammar rules to describe a natural language.
  - 40 000 and more words in a language
- Grammar rules were manually written before 1990
- After 1990 tree banks were created (similar to bio banks, seed banks)

Stockholm University

Chunking (shallow parsing) using NLTK

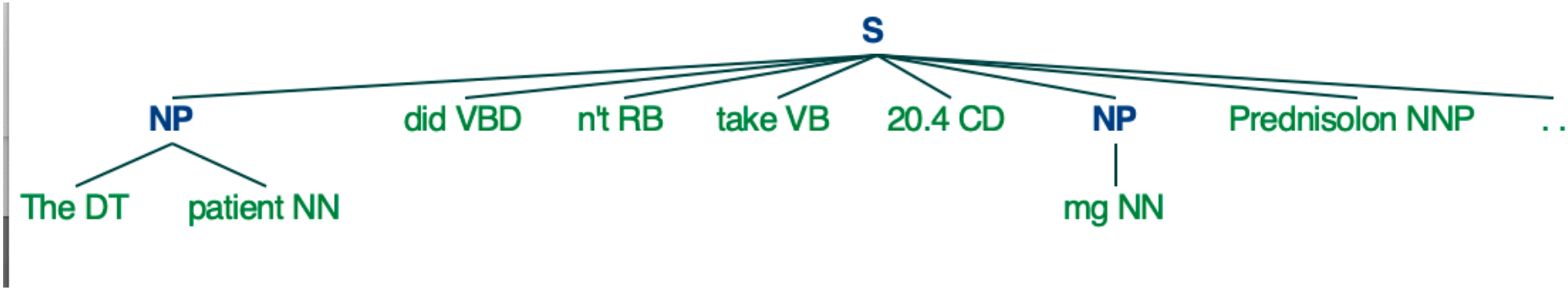
- sent = ("The patient didn't take 20.4 mg Prednisolon")
- tokenised = nltk.word_tokenize(sent)
- tagged = nltk.pos_tag(tokenised)
- chunk = nltk.ne_chunk(tagged)
- nltk.ne_chunk(tagged)

```
import nltk


tokenised = nltk.word_tokenize("The patient didn't take
20.4 mg Prednisolon. ")


print(tokenised)
['The', 'patient', 'did', "n't", 'take', '20.4', 'mg',
'Prednisolon', '.']

tagged = nltk.pos_tag(tokenised)


print(tagged)
[('The', 'DT'), ('patient', 'NN'), ('did', 'VBD'), ("n't", 'RB'),
('take', 'VB'), ('20.4', 'CD'), ('mg', 'NN'), ('Prednisolon',
'NNP'), ('.', '.')]


chunk = nltk.ne_chunk(tagged)
```

```
chunk = nltk.ne_chunk(tagged)
print(chunk)
(S
The/DT
patient/NN
did/VBD
n't/RB
take/VB
20.4/CD
mg/NN
Prednisolon/NNP
./.)

grammar = "NP: {<DT>?<JJ>*<NN>}"

NPChunker = nltk.RegexpParser(grammar)

result = NPChunker.parse(tagged)

result.draw()
```

Stockholm University

Chunk tree


S
NP
did VBD
n't RB
take VB
20.4 CD
NP
Prednisolon NNP
. .
The DT
patient NN
mg NN

Stockholm University

Tree banks

- Treebanks contains text corpora
- Corpora manually annotated with pos tags, syntactic structures and semantic roles
- Tree banks can also be automatically tagged and manually annotated
- Boot strapped on a smaller annotated corpora and then extended and corrected.
- https://en.wikipedia.org/wiki/Treebank

Stockholm University

Parsing

- Read the tokenised text and parse it

  “The” “patient” ”took “20.4” “mg” “Prednisolon” “.”

  ⇒ parsing

  ((“The” “patient”, NP),

  ((”took”, Verb),(“20.4” “mg” “Prednisolon”, NP)

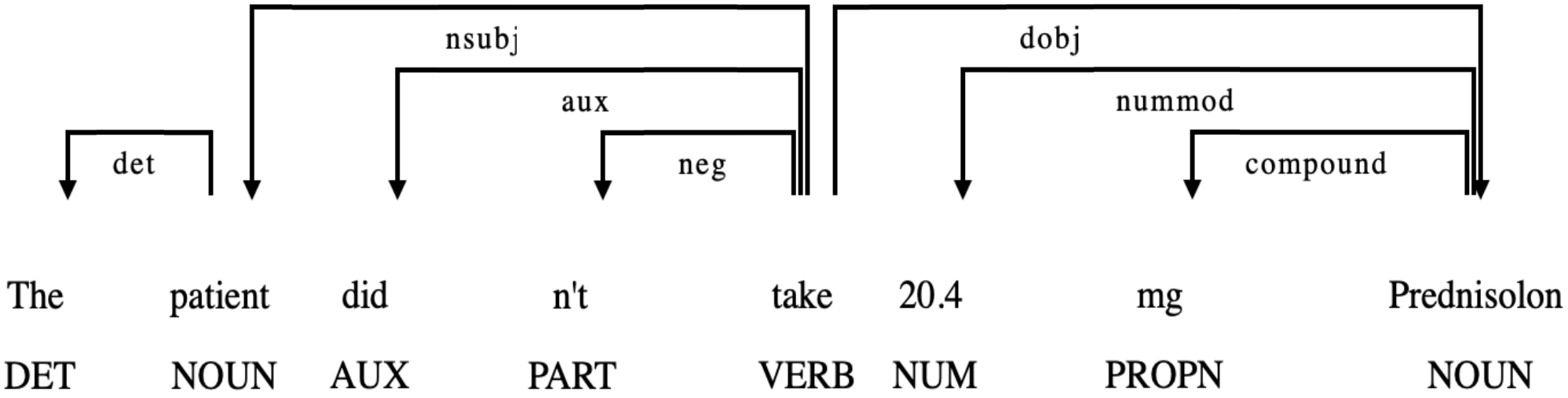
  , VP)

  , S)

Parsing using Spacy

Stockholm University

```
hercules@Python$ python
Python 3.7.1 (v3.7.1:260ec2c36a, Oct 20 2018, 03:13:28)
[Clang 6.0 (clang-600.0.57)] on darwin
Type "help", "copyright", "credits" or "license" for more information.
>>> import spacy
>>> from spacy import displacy
>>> nlp = spacy.load("en_core_web_sm")
>>> sent = ("The patient didn't take 20.4 mg Prednisolon")
>>> doc = nlp(sent)
>>> options = {"compact": True, "bg": "white","black": "white", "font":
"Source Sans Pro"}
>>> displacy.serve(doc, style="dep", options=options)


Using the 'dep' visualizer
Serving on http://0.0.0.0:5000 ...
```



Dependency parsing

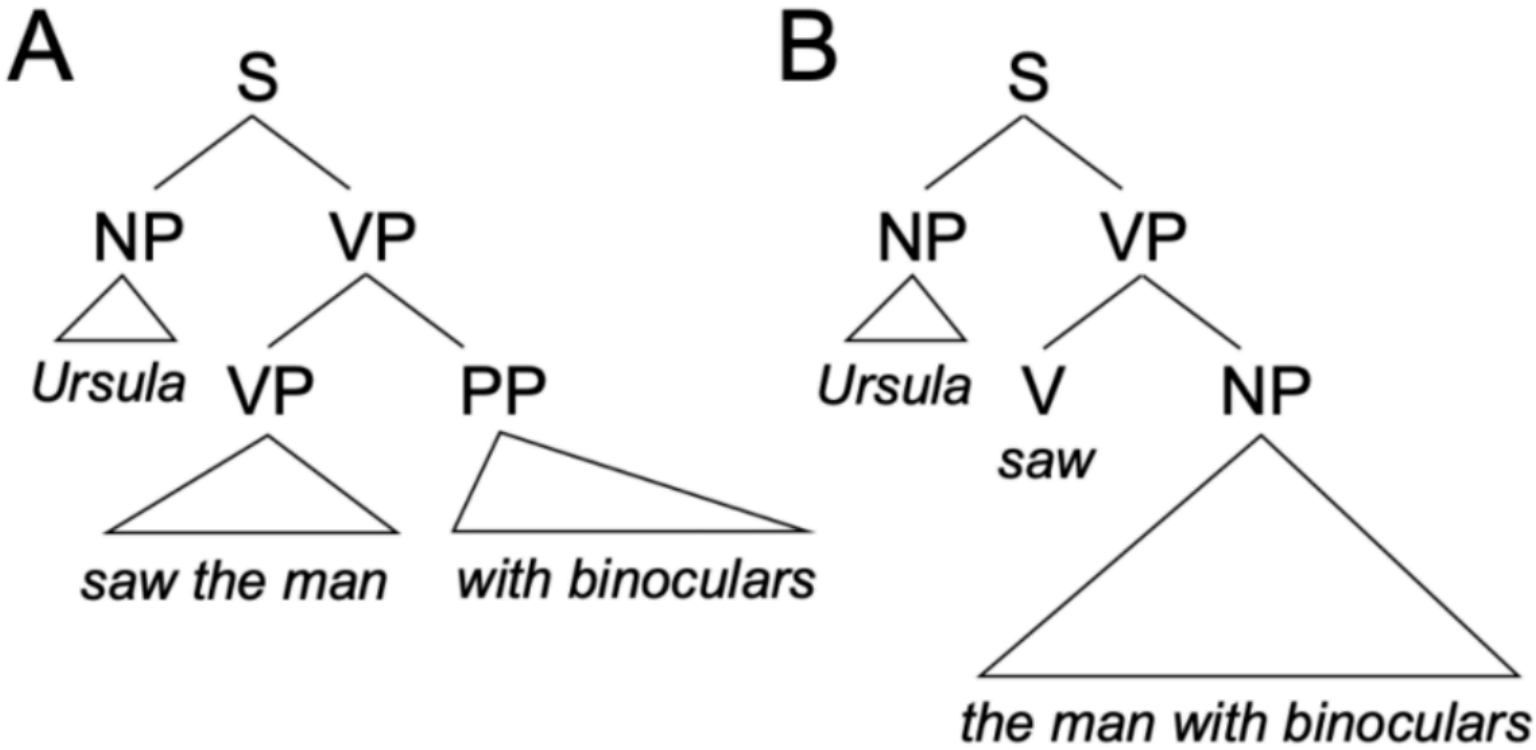
Ambiguous

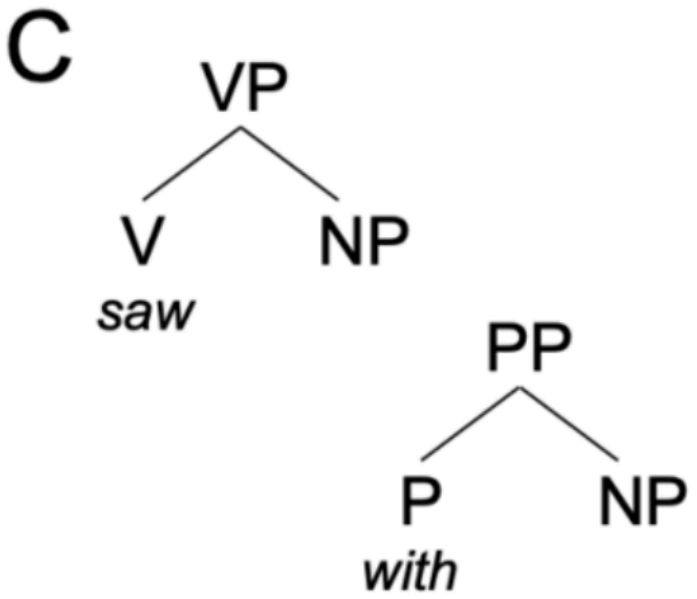
- Ursula saw the man with the binoculars

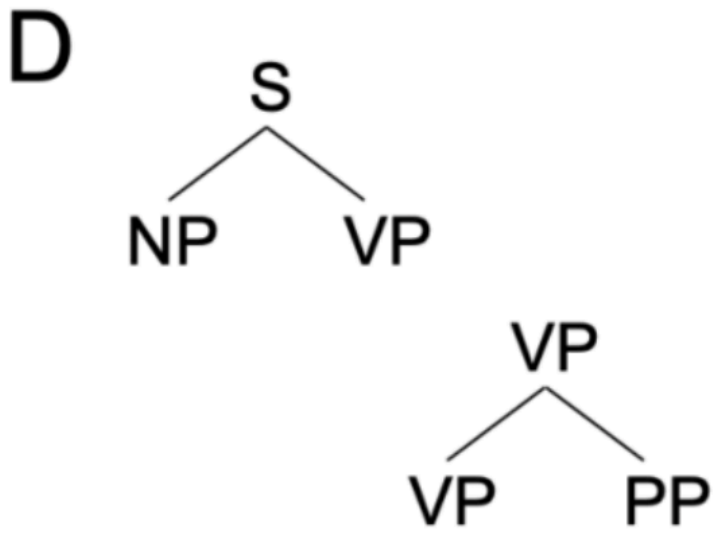
- Two meanings:

  Ursula saw the man with her binoculars

  Ursula saw the man that had binoculars


A
S
NP
VP
Ursula
VP
PP
saw the man
with binoculars
B
S
NP
VP
Ursula
V
NP
saw
the man with binoculars


C
VP
V
NP
saw
PP
P
NP
with


D
S
NP
VP
VP
VP
PP

Stockholm University

Prolog DCG grammar

- A grammar for the sentence
  - She saw the man with binoculars

Prolog with DCG grammar

```
s --> np, vp.
vp --> verb.
vp --> verb, np.
vp --> vp, pp.

np --> noun.
np --> pronoun.
np --> noun, pp.
pp --> prep, noun.

prep --> [with].
verb --> [saw].
noun --> [the, man].
noun --> [binoculars].
pronoun --> [she].
```

Prolog with DCG grammar and tree

```
s(s(NP,VP)) --> np(NP), vp(VP).
vp(verb(V)) --> verb(V).
vp(vp(verb(V),NP)) --> verb(V), np(NP).
vp(vp(VP,PP)) --> vp(VP), pp(PP).

np(noun(N)) --> noun(N).
np(pronoun(P)) --> pronoun(P).
np(np(noun(N),pp(PP))) --> noun(N), pp(PP).
pp(pp(prep(P),noun(N))) --> prep(P), noun(N).

prep(with) --> [with].
verb(saw) --> [saw].
noun(the_man) --> [the, man].
noun(binoculars) --> [binoculars].
pronoun(she) --> [she].
```

Stockholm University

Parsing

?- s(Tree, [she, saw, the, man, with, binoculars],[]).

Tree = s(pronoun(she), vp(verb(saw), np(noun(the_man),
pp(pp(prep(with), noun(binoculars)))))) ;

Tree = s(pronoun(she), vp(vp(verb(saw), noun(the_man)),
pp(prep(with), noun(binoculars)))) .

Stockholm University

Generation

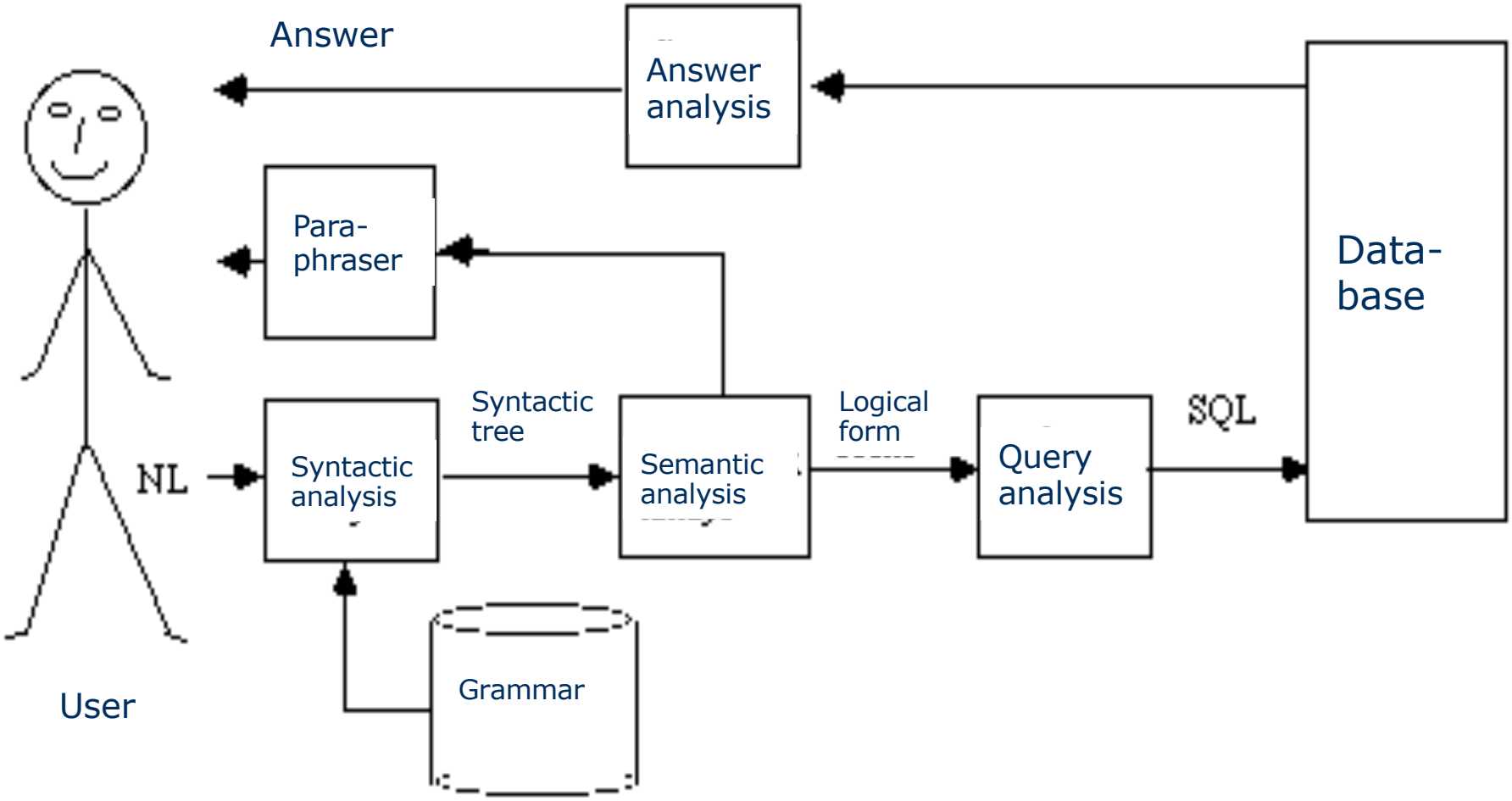
- ?- s(s(pronoun(she), vp(vp(verb(saw), noun(the_man)), pp(prep(with), noun(binoculars)))) ,L,[]).
- L = [she, saw, the, man, with, binoculars] .

Text generation

How does a human speak?

1) Obtains a question by someone and answers
2) Start to speak with a certain purpose

Stockholm University



Paraphrasing

- Other terms are rewriting, simplification,

HSQL-(Help system for structured query language) the prototype is implemented towards a database that contains information about different hospitals and their activities.

Example on generation from SQL fron NL and paraphrasing from SQL back to NL

Fråga: vilken diagnos har Amster ?

Query which diagnosis does Amster have? (Eng)

SQL:

```
SELECT DISTINCT T2.other_info, T1.name, T2.reg_no
FROM PATIENT T1, DIAGNOSIS T2
WHERE (T1.name = `Amster K.`) AND
(T2.reg_no=T1.reg_no)
```

PARAFRAS AV SQL I NATURLIGT SPRÅK

vilka diagnoser har en patient som heter Amster K.

which diagnoses does a patient have which name is Amster K? (Eng)

Paraphrase from SQL and database schema

What to Say and How to say it

- Consider the recipient of the message
  - What does the recipient already know?
- Selects what to say of everything the speaker knows.
- Decides in which order this should be said.
- Selects how to say it, language, syntactic structures and lexical objects.

Stockholm University

User modelling

- Query analysis and user modelling
  - Must analyse the question
  - Find out what the recipient already knows.

Stockholm University

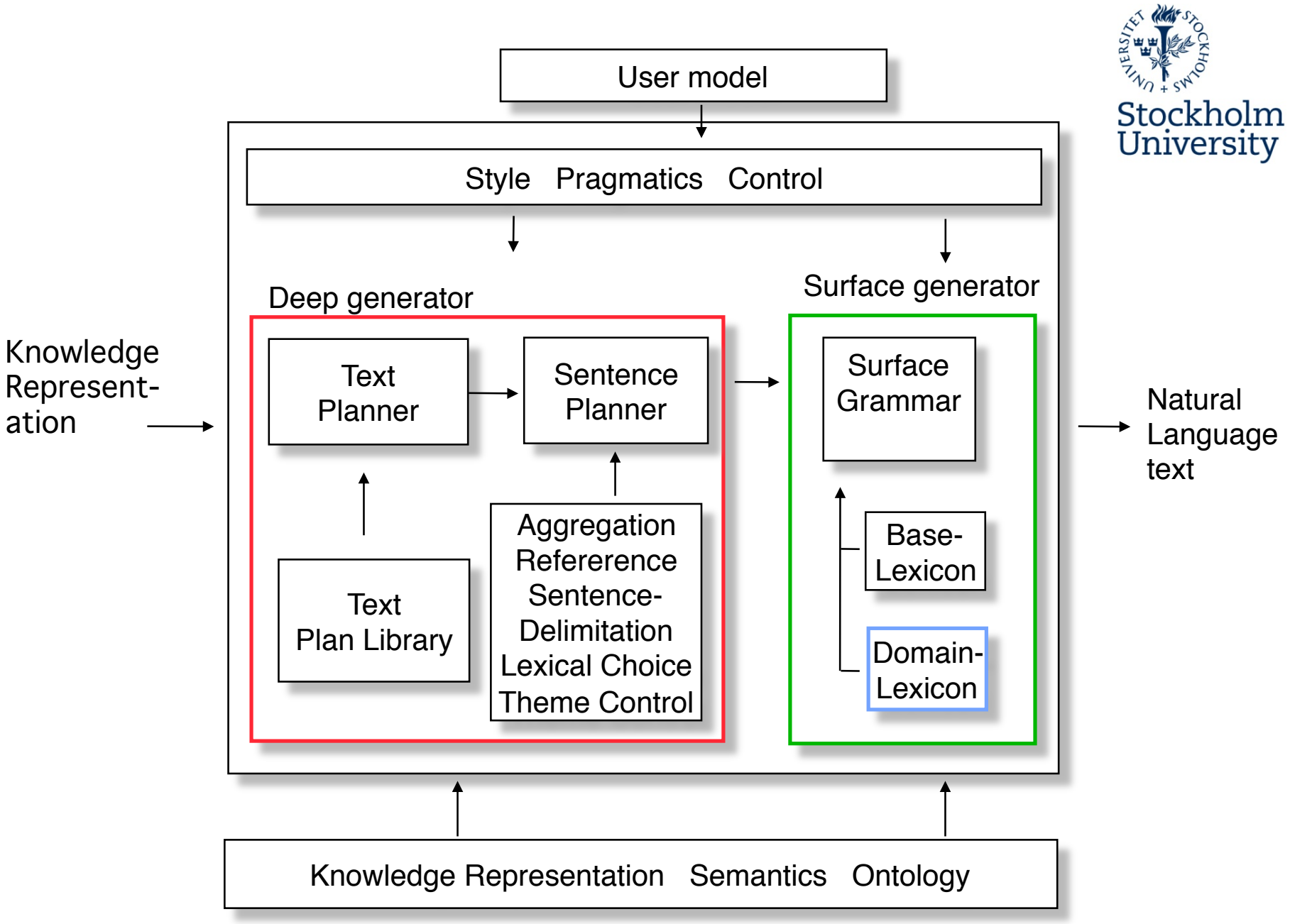
Deep generation – What to say?

- Content selection
  - Select from its rich knowledge base what to say
- Text planning
  - Creates a text plan that in turn contains sentence plans
  - The text plan must be coherent
- Sentence plan
  - Decides on the form of sentences, active, passive form, pronouns, aggregation, sentence length,

Stockholm University

Surface generation – How to say it?

- Selects language – grammar
- Lexical choices (words)
  - Base dictionary
  - Domain dictionary
- Post processing


User model
Style Pragmatics Control
Deep generator
Surface generator
Knowledge Represent-ation
Text Planner
Sentence Planner
Surface Grammar
Natural Language text
Aggregation Refererence Sentence-Delimitation Lexical Choice Theme Control
Base-Lexicon
Text Plan Library
Domain-Lexicon
Knowledge Representation Semantics Ontology

Coherence relations

- Rhetorical structure theory (RST)
  - A set or coherence relations are used to analyse a text.

    reason, elaboration, evidence, purpose. background, attribution, list, etc
  - Can also be used to generate a coherent text
  - See also Chapter 22 Discourse Coherence in Jurafsky & Martin

Stockholm University

Text generation

Deep learning models such as GPT-2 can generate text from a seed phrase.

The output maybe syntactic correct.

There is no control of the output.

Classic text generation, used templates and/or canned phrases. Fills in missing entities.

Parrot paraphrase with Python

Input_phrase: She saw the man on the hill with the binoculars

Paraphrases:

('she saw the man with the binoculars on the hill', 35)
('she saw him on the hill with the binoculars', 17)
('she saw the man on the hill with binoculars', 16)
('she saw the man on the hill with her binoculars', 14)
('she saw the man on the hill with the binoculars', 12)

https://github.com/PrithivirajDamodaran/Parrot_Paraphraser

https://colab.research.google.com/github/dataprofessor/parrot/blob/main/PARROT.ipynb#scrollTo=EXdxBG-FVae-

Text simplification

- https://rewordify.com/index.php
- Rewordify replaces difficult words with simpler wordings.

The UK has begun to withdraw staff from the British embassy in Ukraine amid warnings of a Russian invasion.

Officials say there have been no specific threats to British diplomats, but about half of the staff working in Kyiv will return to the UK.

The US has ordered relatives of its embassy staff to leave, saying an invasion could come "at any time".

Russia has denied plans for military action, but tens of thousands of troops have amassed on the border.

The embassy moves seem to be precautionary, and nothing specific is thought to have occurred in the past 24 hours to have triggered the decisions of the US and UK.

Staff working at the EU embassy will stay in place for now, with EU foreign policy chief Josep Borrell saying he would not "dramatise" the tensions.

Members of the Nato alliance, including Denmark, Spain, Bulgaria and the Netherlands, are sending more fighter jets and warships to Eastern Europe to bolster defences in the region.

With an estimated 100,000 Russian troops now at the border with Ukraine, the head of Nato has warned there is a risk of fresh conflict in Europe.

The decision by the US is one of a number of precautions the state department employs when crises could put American diplomats in harm's way, the BBC's Barbara Plett Usher reports.

The UK has begun to withdraw staff from the British government office in Ukraine in the middle of warnings of a Russian (sudden, unwanted entry into a place).
(people in charge of something) say there have been no clearly stated/particular threats to British peacekeepers, but about half of the staff working in Kyiv will return to the UK.
The US has ordered relatives of its government office staff to leave, saying a (sudden, unwanted entry into a place) could come "at any time".
Russia has denied plans for military action, but tens of thousands of troops have collected on the border.
The government office moves seem to be (related to doing things to prevent trouble or injury), and nothing specific is thought to have happened in the past 24 hours to have triggered the decisions of the US and UK.
Staff working at the EU government office will stay in place for now, with EU foreign policy chief Josep Borrell saying he would not "dramatise" the tensions.
Members of the Nato friendly partnership, including Denmark, Spain, Bulgaria and the Netherlands, are sending more fighter jets and warships to Eastern Europe to help (or increase) defences in the area.
With a guessed (number) 100,000 Russian troops now at the border with Ukraine, the head of Nato has warned there is a risk of fresh conflict in Europe.
The decision by the US is one of some (steps taken to prevent trouble or injury) the state department employs when serious problems could put American peacekeepers in harm's way, the BBC's Barbara Plett Usher reports.

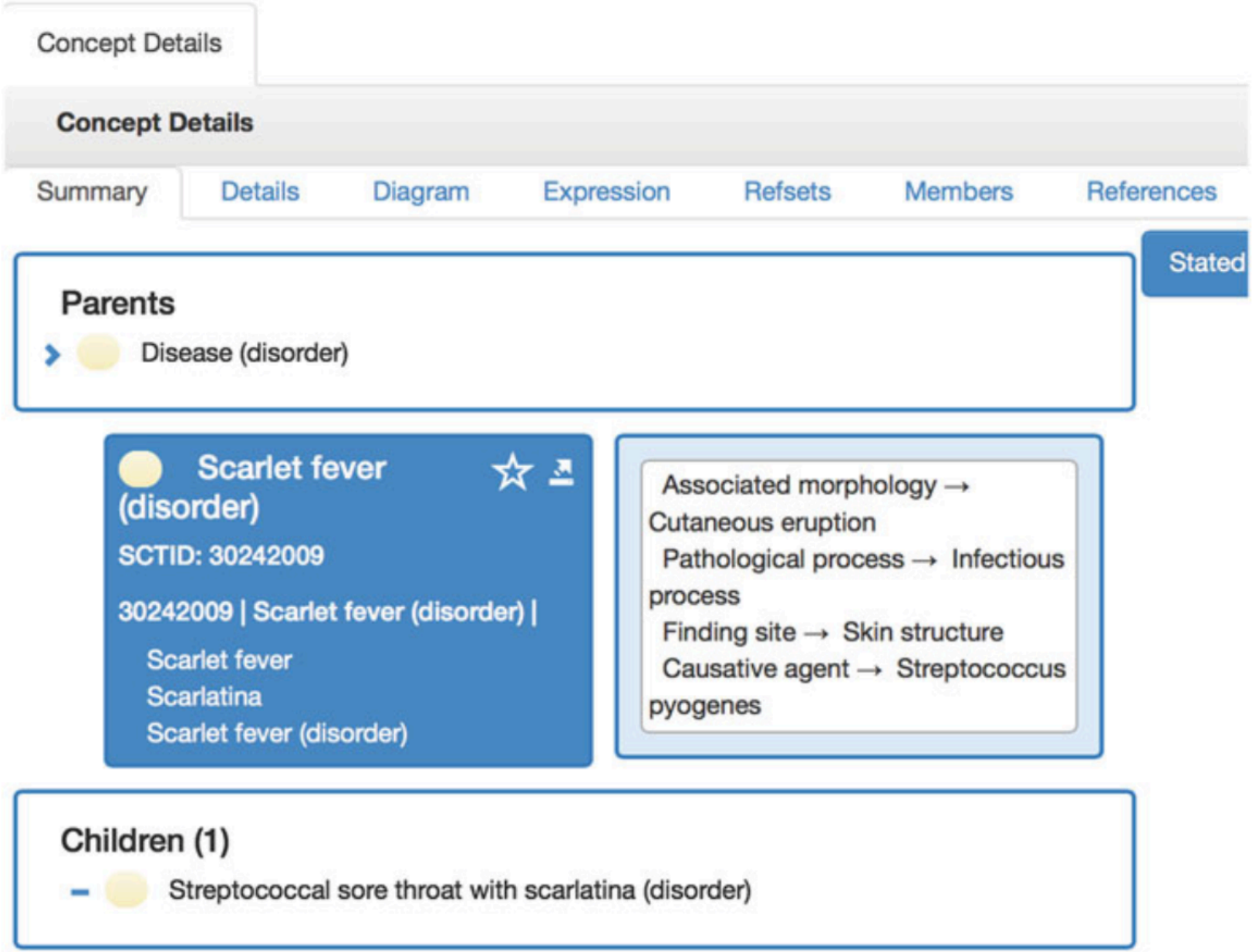


Fig. 10.9 The IHTSDO SNOMED CT Browser and its description of scarlet fever, the browser is described in Sect. 5.2

Stockholm University

Input: *Scharlakansfeber*
Output: *Scharlakansfeber är en eruption och hudsjukdom orsakad av streptokocker. Orsaken till sjukdomen är Streptococcus pyogenes. Sjukdomen finns i hud och hudstruktur och hud.*
(Translated with Google translate to English:
Input: *Scarlet fever*
Output: *Scarlet fever is an eruption and skin disease caused by streptococci. The cause of the disease is Streptococcus pyogenes. The disease is found in skin and skin structure and skin.*

Fig. 10.10 Example of natural language output from the SNOgen system, when entering the disorder *scharlakansfeber* (in Eng: scarlet fever) in SNOMED CT format and obtaining the Swedish natural language text output. Below is the corresponding machine translated English text (© 2014 The authors—reprinted with permission from the authors. Published in Kanhov (2014))

Machine learning

- Supervised
  - Annotated data sets
    - Named Entity Tagging
    - Classification
- Unsupervised
    - Clustering
- Semi supervised
  - Some annotated data and lots of un-annotated data.
  - Active learning
  - BERT Deep Learning

Machine learning

- Supervised
  - Annotated data sets
    - Named Entity Tagging
    - Classification
- Unsupervised
    - Clustering
- Semi supervised
  - Some annotated data and lots of un-annotated data.
  - Active learning
  - BERT Deep Learning

Stockholm University

Annotated data sets

- To train pos-taggers for different languages
- To train chunkers for different languages
- To train parsers for different languages

Stockholm University

NLP data sets

- English
  - ELRA, http://catalogue.elra.info/
  - GitHub, https://github.com/niderhoff/nlp-datasets
  - And lots on different places
- Swedish
  - Språkbanken, https://spraakbanken.gu.se

Health Bank – Swedish Health Records Research Bank

- Health Bank, http://dsv.su.se/healthbank
- 2 million patients, Karolinska University Hospital
- 500 clinical units
- Text and structured data
- 7 Ethical permissions
- 20 manually annotated data sets
- https://dsv.su.se/healthbank/annotated

Stockholm University

Tree banks

- https://en.wikipedia.org/wiki/Treebank

Formats for data SGML and ConLL

- Text format
- SGML format Standard Generalized Markup Language (similar to HTML)

  My name is <First_Name>Hercules</First_Name>
- ConLL format

My	0
name	0
is	0
Hercules	First_Name

Stockholm University

Quizzes for the lecture II

- What is the difference between a natural language and a formal language?
- What is a grammar?
- What is chunker?
- What is tree bank?
- What is a paraphrase?

Stockholm University

Questions