IN1140 – Uke 7

Oblig 2b

Syntaks og syntakstrær

Ukesoppgaver

Oblighjelp

Oblig 2b er ute!

- Innleveringsfrist: fredag 3. november kl. 23.59

Innlevering av oppgaven skjer i Devilry. Det er viktig at du setter deg inn i retningslinjene for obligatoriske oppgaver. Besvar oppgavene i én fil som angir brukernavnet ditt slik:

`oblig2b_brukernavn.py`

Oppgave 4 skal besvares i en egen kodefil `oppgave4.py` som du også leverer. Du trenger ikke å levere mappen `NoReC`.

Besvarelsen må kunne kjøres med Python i terminalen, slik som tidligere. Ikke pakk filene inn i en ZIP-mappe.

Poeng

En perfekt løsning av denne innleveringen er verdt 100 poeng. Du må vise koden din i oppgave 2 og 4. Koden må kunne kjøres og må inneholde kommentarer som forklarer hva koden gjør.

- Oppgave 1 —> Kontekstfrie grammatikker
- Oppgave 2 —> Chunking
- Oppgave 3 og 4 —> Naive Bayes

Kontekstfrie grammatikker

- En grammatikk som beskriver et språk fra topp til bunn
- Beskriver regler for hvordan setninger henger sammen
  - Syntaks!

S → NP VP

NP → N PP | N

VP → VP PP | V NP

PP → P NP

N → pensjonisten | gevær

P → med

V → jager

```
>>> print(final))
   File "<stdin>", line 1
     print(final))
                  ^
SyntaxError: unmatched ')'
>>>
```

Kontekstfrie grammatikker

Frasale kategorier

Leksikalske kategorier

- Disse markerer **konstituenter**
  - Et ord eller en gruppe av ord som fungerer som en enhet

S → NP VP

NP → N PP | N

VP → VP PP | V NP

PP → P NP

N → pensjonisten | gevær

P → med

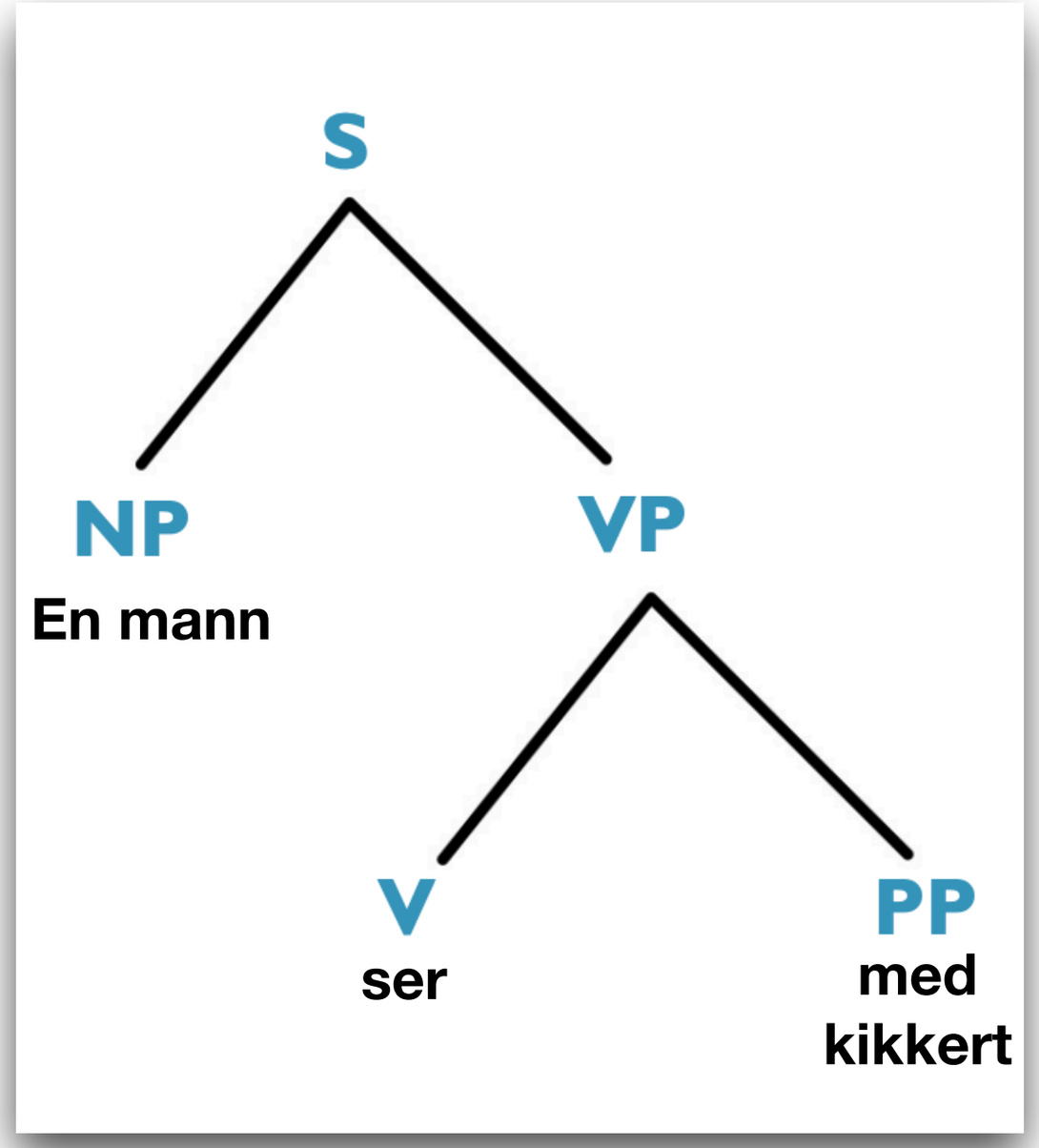
V → jager

Syntaktiske trær

- Visualisering av kontekstfrie grammatikker

- Viser oss tilhørighet i konstituenter

- Er det en konstituent?
  1. Stå alene?
  2. Erstattes av pronomen?
  3. Flyttes som enhet?



Rekursivitet i kontekstfrie grammatikker

- En naturlig del av språket
- Der en node leder tilbake til seg selv
- Må alltid ha mulighet for løsning

S → NP VP

NP → N PP | N

VP → VP PP | V NP

PP → P NP

N → pensjonisten | gevær

P → med

V → jager


S
VP
PP
NP
NP
N
V
Prep
det
N
Time flies like an arrow.

Kodeskallet til oppgave 3

```python
import nltk
# Det kan være dere må installere korpuset
nltk.download("conll2000")
from nltk.corpus import conll2000

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

# Når dere legger til mønstre er det nok lurt å gjøre noe sånt:
# grammar = r»""
#     NP: {<DT>?<JJ>*<NN>}
#         {<...>?<...>*<...>}
#     """

cp = nltk.RegexpParser(grammar)
training_chunks = conll2000.chunked_sents("train.txt", chunk_types=['NP'])

# For å inspisere dataene
# print(training_chunks[:10])

print("Evaluering av treningsettet:")
print(cp.accuracy(training_chunks))

# Når dere er ferdig kan dere teste på testsettet
test_chunks = conll2000.chunked_sents("test.txt", chunk_types=['NP'])
```

Neste uke

- Lilja tar gruppetimen
    - Naive Bayes