

IN2110 - Obligatorisk Oppgave 2a

Stian Carlsen Swärd (stiancsw)

April 11, 2020

1 Oppgave 1

a)

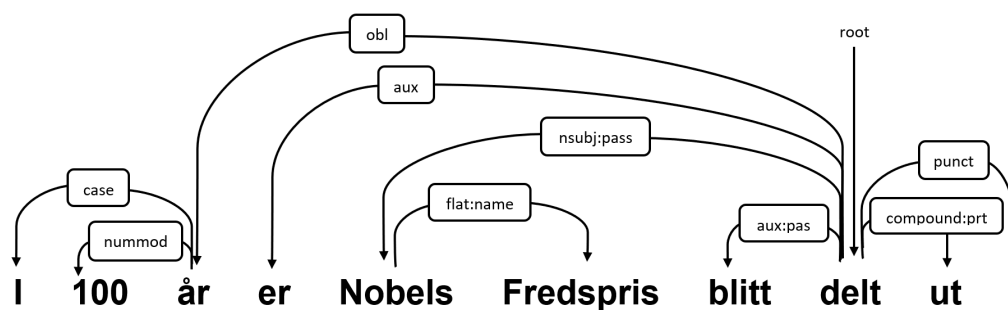


Figure 1: Dependensgraf for setningen i oppgavesettets Figur 1

b)

Transisjonssekvens for setningen i oppgave 1a) ved bruk av arc-eager algoritmen

OP	STACK	BUFFER	ARC
START	[ROOT] _S	[I, ...] _B	∅
SHIFT	[ROOT, I] _S	[100, ...] _B	∅
SHIFT	[ROOT, I, 100] _S	[år, ...] _B	∅
LA _{nummod}	[ROOT, I] _S	[år, ...] _B	$A_1 = \{(\text{år}, 100, \text{nummod})\}$
LA _{case}	[ROOT] _S	[år, ...] _B	$A_2 = A_1 \cup \{(\text{år}, I, \text{case})\}$
SHIFT	[ROOT, år] _S	[er, ...] _B	A_2
SHIFT	[ROOT, år, er] _S	[Nobels, ...] _B	A_2
SHIFT	[ROOT, ..., Nobels] _S	[Fredspris, ...] _B	A_2
RA _{flat:name}	[ROOT, ..., Fredspris] _S	[blitt, ...] _B	$A_3 = A_2 \cup \{(\text{Nobels}, \text{Fredspris}, \text{flat} : \text{name})\}$
REDUCE	[ROOT, ..., Nobels] _S	[blitt, ...] _B	A_3
SHIFT	[ROOT, ..., blitt] _S	[delt, ...] _B	A_3
LA _{aux:pass}	[ROOT, ..., Nobels] _S	[delt, ...] _B	$A_4 = A_3 \cup \{(\text{delt}, \text{blitt}, \text{aux} : \text{pass})\}$
LA _{nsubj:pass}	[ROOT, år, er] _S	[delt, ...] _B	$A_5 = A_4 \cup \{(\text{delt}, \text{Nobels}, \text{nsubj} : \text{pass})\}$
LA _{aux}	[ROOT, år] _S	[delt, ...] _B	$A_6 = A_5 \cup \{(\text{delt}, \text{er}, \text{aux})\}$
LA _{obl}	[ROOT] _S	[delt, ...] _B	$A_7 = A_6 \cup \{(\text{delt}, \text{år}, \text{obl})\}$
SHIFT	[ROOT, delt] _S	[ut, .] _B	A_7
SHIFT	[ROOT, delt, ut] _S	[.] _B	A_7
RA _{compound:prt}	[ROOT, delt] _S	[.] _B	$A_8 = A_7 \cup \{(\text{delt}, \text{ut}, \text{compound} : \text{prt})\}$
SHIFT	[ROOT, delt, .] _S	∅	A_8
RA _{punct}	[ROOT, delt] _S	∅	$A_9 = A_8 \cup \{(\text{delt}, ., \text{punct})\}$
REDUCE	[ROOT] _S	∅	A_9

Hovedforskjellene mellom arc og arc-eager er kompleksiteten ($O(n^5)$ vs $O(n^3)$), at arc-eager har en ekstra operasjon (REDUCE), samt at i arc kan ikke RIGHT-ARC-operasjonen utføres før vi har funnet alle tokens som er dependenser i en relasjon med dependensen i relasjonen vi jobber på.

2 Oppgave 3

a)

Se oblig2a.py for implementasjon av `attachment_score()`

b)

Tabell over UAS og LAS score for de ulike datasettene

Datsett	UAS	LAS
Bokmål	0.893	0.798
Nynorsk	0.682	0.563
NynorskLIA	0.485	0.324

Vi ser, ikke overraskende, at parseren scorer mye høyere på bokmålssettet enn på nynorsksettet, og høyere på nynorsksettet enn på det nynorske talespråksettet.

Vi printer ut `nnlia_dev_docs[:10]` og får følgende utskrift:

```
[vi spør først når dette her begynte for alvor og kva slags
bil du hadde å køyre med . ,
det første # det # i femogtjue . ,
og da # kj- hadde eg Forden # eg hadde Forden da au . ,
men den køyrde eg med ein månads seie så e . ,
så vart den for liten så måtte eg bytte # og eg hadde masse
bytta annakvart år # bilar . ,
ja . ,
for å få e # for å komme til noko større materiell . ,
og dette her kj- fortsette vi med og køyrde # mjølk da leste
på ein e # mellom åtti og hundre spann . ,
om om dagen # som vi bar . ,
utor mjølkekummen . ]
```

Vi ser her at setningene i datasettet ikke alltid følger konvensjonelle grammatikkregler, og at det derfor er vanskeligere for en parser som er trent på korrekt grammatikk å predikere riktig. Det er også tokens i datasettet som stammer fra påbegynte ord som har blitt avbrutt av at taleren omformulerer setningen midt i, noe som resulterer i tokens som parseren aldri har sett før.