INF 2820 V2016: Obligatorisk innleveringsoppgave 4 – hele

- **Besvarelsene skal leveres i devilry innen torsdag 12.5 kl 18.00**
- **Filene det vises til finner du i /projects/nlp/inf2820/**
- **Hvis du ikke har norsk som morsmål og er usikker på de norske dataene, så bruk dine medstudenter som informanter.**
- **Hvis du har norsk som morsmål, så svar når medstudenter spør deg om de språklige dataene.**

Oppgave 1: Grammatikk med trekk (60 poeng)

Vi skal arbeide videre med grammatikken `pp.cfg` fra obligatorisk innlevering 2. Vi skal se hvordan vi ved hjelp av trekk ("features") kan utvide grammatikken på en effektiv måte. Vi vil omforme grammatikken til en fcfg-grammatikken og utvide den til å dekke flere konstruksjoner. Kopier `pp.cfg` til en fcfg-grammatikk `norsk.fcfg` og se at den virker, dvs. lag en fcfg-parser fra denne grammatikken og test den på passende eksempler.

A. Subcat (10 poeng)

I `pp.cfg` brukte vi forskjellige kategorier (metasymboler) for forskjellige typer verb, som intransitive og transitive verb. I en trekkgrammatikk kan vi i stedet ha én kategori for verb og så et trekk som sier hvilken underkategori ("subcategory") verbet tilhører. Gjør denne endringen etter mønster fra grammatikk1 for engelsk fra forelesning 12. (Finnes i mappen /projects/nlp/inf2820/fcfg/forelesn_12)

B. Bestemte og ubestemte substantiv(10 poeng)

I `pp.cfg` måtte vi innføre forskjellige kategorier for substantiv i bestemt og ubestemt form. I en trekkgrammatikk er det naturlig å ha én kategori for substantiv, og så et trekk for bestemthet. Gjør denne endringen.

Vi vil også utvide grammatikken med bestemmere som går sammen med bestemte substantiv: *det barnet*, *dette barnet*. Her er det viktig at grammatikken ikke overgenererer. (Ikke: *et barnet* eller *det barn*.) Husk også at et substantiv i bestemt form entall kan utgjøre en hel NP, mens et substantiv i ubestemt form entall ikke kan det. (Det kan være naturlig å innføre et mellomnivå – Nom eller N'-mellom N og NP.)

C. Flere substantiv (10 poeng)

I `pp.cfg` begrenset vi oss til substantiv i intetkjønn ("nøytrum") i entall. Vi skal nå også ta med substantiv i andre kjønn: hankjønn ("maskulinum"), f.eks. *bil, hund, gutt, kontrakt*, og hunkjønn ("femininum"), f.eks. *jente, hytte, and, gås*. Som for intetkjønnsordene skal vi ha med både ubestemt form og bestemt form, som i *bilen* og *hytta*.

Det er viktig å få riktig samsvar med bestemmere, f.eks. *et hus* men ikke *en hus*. Men derimot *en bil* og ikke *et bil* eller *ei bil* og det skal være *ei jente*, ikke *et jente*. Tilsvarende for *enhver*/*ethvert*.

Vi skal også ha flertallsformer av alle substantivene, både bestemte og ubestemte former og alle kjønn. For flertallssubstantiv skal vi ha med bestemmerne *mange, noen, ingen, alle, de, disse*. For hver bestemmer, tenk igjennom hvilke former av substantivet det kan forekomme sammen med. Tenk også igjennom hvilke former av substantivet som kan utgjøre en hel NP og inkluder disse i grammatikken.

D. Adjektiv(10 poeng)

I `pp.cfg` har vi med adjektiv i NP-ene. Når vi legger til substantiv av flere kjønn og tall, må vi sørge for at adjektivene samsvarer med substantivene. Det heter *et stort hus*, men *en stor gutt*, *ei stor jente* og *mange store hus*. Med de bestemte heter det *det store huset*, *den store jenta*, *de store husene*.

Vi sier at adjektiv i norsk har to bøyninger, **sterk** og **svak**. Med sterk bøyning blir det forskjell på entall (*stor*/*stort*) og flertall (*store*) og i entall blir det forskjeller mellom kjønn. Det er den sterke bøyning som modifiserer ubestemte substantiv. I den svake bøyningen blir det ingen forskjeller (*store*). Denne går sammen med de bestemte substantivene (*det store huset*, *den store jenta*, *de store husene)*.

Bruk trekk for å få en riktig behandling av adjektiv som modifiserer substantiv.

E. Predikative adjektiv (10 poeng)

Vi skal gjøre en utviding av grammatikken. Vi vil ha med VP som består av verbet *være* etterfulgt av et adjektiv. Også her er det samsvar i norsk. Det heter

Et hus er stort
Bilen er stor
Jenta er stor
Mange hus er store

En kan ikke si

Huset er store
Bilen er stort
osv.

Legg denne konstruksjonen til grammatikken og bruk trekk til å sørge for riktig samsvar.

F. Pronomen (10 poeng)

Vi skal utvide grammatikken med personlige pronomen, som *jeg, meg, du, deg, han, ham, hun, henne, den, det, vi, oss, dere, de, dem*. Disse kan utgjøre en hel NP som i setningen *Jeg liker deg*. Her må vi passe på å få riktig form på riktig plass. Det går ikke å si *Meg liker deg* eller *Jeg liker du*.

Innlevering: Den utvidete grammatikken `norsk.fcfg`

Oppgave 2: Semantikk (30 poeng)

I denne oppgaven skal vi lage en grammatikk for et lite fragment av norsk. Du kan holde leksikonet lite og f.eks. ikke ha substantiv i andre former enn neutrum (intetkjønn) entall. Grammatikken skal være utstyrt med semantiske trekk slik at vi til hele setninger får logiske formler som uttrykker det samme som setningen. Utgangspunktet er seksjonene 10.2, 10.3 og 10.4 frem til "Quantifier Ambiguity Revisited" i NLTK-boka. Det lille grammatikkfragmentet med semantiske regler presentert i NLTK-boka finnes som `simple-sem.fcfg`. Du finner den fra NLTK og i `/projects/nlp/nltk_data/grammars/book_grammars/`
Vi har laget et tilsvarende fragment for norsk, som heter `no-sem.fcfg`. Vi har endret litt på hvordan semantikken for setninger med transitive verb, blir laget. Vi gjør litt mer arbeid med semantikken i regelen VP → TV NP, mens NLTK har flyttet arbeidet til det leksikalske oppslaget for de transitive verbene. Erfaringsmessig har studentene hatt lettere for å forstå vår måte å gjøre det på.

Gjør deg kjent med grammatikken. Se hvordan grammatikken analyserer følgende setninger:

1. Ola sov
2. Kari likte Ola
3. et barn sov
4. ethvert barn beundret et dyr

Bruk som før

```
>>> import nltk
>>> semparse = nltk.load_parse("file:no-sem.fcfg")
>>> for t in semparse.parse("Kari likte Ola".split()): t.pprint()
```

For bare å se semantikken kan du skrive

```
>>> for t in semparse.parse("Kari sov".split()):
print(t.label()['SEM'])
```

A. Adjektiv(10 poeng)

For enkle adjektiv som modifiserer substantiv er det vanlig å analysere dem som predikat som er forbundet med predikatet som svarer til substantivet med &, for eksempel

- *lite hus* representeres som `\x. (lite(x) & hus(x))`

Utvid grammatikken med adjektiv med semantiske representasjoner og regler for sammensetning av adjektiv og nomen som gir dette resultatet. Se at du får riktig resultat med noen eksempler som

5. Et lite barn sov
6. Kari likte et stort pent hus

B. Setningskonjunksjon og -disjunksjon (10 poeng)

Vi ønsker at for eksempel setning (7) skal få som semantisk representasjon (8) og tilsvarende for (9) og (10). Lag syntaktiske regler med semantiske trekk som sammen med de andre reglene sørger for dette. Reglene skal være generelle og tillate gjentatt koordinering som i (11).

7. Ola sov og Kari smilte
8. `sov(ola) & smilte(kari)`
9. Ola likte et dyr eller et dyr likte Ola
10. `(exists x.(dyr(x) & likte(ola,x))) | (exists x.(dyr(x) & likte(x,ola)))`
11. Ola likte et dyr og Ola sov eller Kari smilte

C. Verbalkonjunksjon og –disjunksjon (10 poeng)

Språket kan fort bli litt stivt. I stedet for (12) er det mer naturlig å si (13)

12. Ola likte et dyr og Ola beundret Kari
13. Ola likte et dyr og beundret Kari

Lag nå regler som tillater koordinasjon av to VP-er. Vi vil ha samme semantiske representasjon (etter konverteringer) for de to setningene. Ta også med disjunksjon.

Innlevering: Grammatikkfil

Oppgave 3: Trekkgrammatikk med stort leksikon (10 poeng)

En begrensning på grammatikken og parseren vi lagde i oppgave 1 er leksikonets størrelse. Det er grense for hvor mange ord vi orker å legge inn manuelt, og derfor blir eksemplene svært ensformige. Samtidig arbeidet vi med et stort leksikon i oblig-2. Kan vi importere det i grammatikkene våre?

Filen norsk_scarrie_parser.py inneholder en del syntaktiske regler og en metode for å konstruere leksikalske regler fra scarrie-leksikonet. Legg den i samme mappe som scarrie-leksikonet fra oblig-2 og kjør den i et interaktivt vindu. Prøv så

```
>>> g = build_grammar()
>>> p = build_parser(g)
>>> for t in p.parse("en søt elefant fortalte en søvnig sebra at
hyenen sov".decode('utf-8').split()): t.pprint()
```

OBS `build_grammer()`-kommandoen tar noen minutter.

Grammatikkreglene vi har lagt inn er svært enkle. De tar utgangspunkt i pp.cfg. Vi har fjernet forskjellen mellom bestemte og ubestemte substantiv og lagt inn subkategorisering for verb, svarende til pkt. a på oppgave 1. Et problem med denne grammatikken er at den overgenerer. Den vil f.eks. akseptere «et søte elefant sov». Du kan legge inn reglene (uten leksikon) fra løsningen din til oppgave 1, i stedet for de reglene som er i denne fila. Men grammatikken vil stadig overgenerere hvis du ikke også gjør noe med hvordan de leksikalske reglene lages.

Oppgaven består altså i å modifisere `rules_from_wordid()` s.a. leksikon virker sammen med grammatikken din og gjør de samme begrensningene som grammatikken din fra oppgave 1 med hensyn til samsvar mellom substantiv, bestemmere og adjektiv, og mellom pronomen og verb.

Vær obs. på at scarrie-leksikonet tillater svært mye når det gjelder hvor verb kan forekomme, hvilke subkategoriseringsmønstre de tillater. Du vil derfor kunne oppleve at grammatikken godkjenner mere enn grammatikken din fra oppgave 1.

Løsningen på denne oppgaven har mer karakter av Python-fikling enn språkteknologi. Derfor gir den forholdsvis få poeng og kan kuttes ut av de som strever med den. Men belønningen ved å løse den er stor. En parser som kan behandle et stort leksikon.

Innlevering: Den modifiserte programfila `norsk_scarrie_parser.py`