UNIVERSITETET I OSLO

Det matematisk-naturvitenskapelige fakultet

Eksamen i: INF2820 Datalingvistikk
Eksamensdag: 1. juni 2017
Tid for eksamen: 1430-1830
Oppgavesettet er på 4 side(r)
Vedlegg: 0
Tillatte hjelpemidler: ingen

Kontroller at oppgavesettet er komplett
før du begynner å besvare spørsmålene.

- **Du kan svare på norsk, dansk, svensk eller engelsk.**
- **Du skal svare på alle spørsmålene. Vekten på de ulike spørsmålene er oppgitt.**
- **Du bør lese gjennom hele settet slik at du kan stille spørsmål til faglærerne når de kommer i tilfelle noe er uklart.**
- **Hvis du føler noen forutsetninger mangler, lag dine egne og redegjør for dem!**

Oppgave 1 (samlet 15%)

Spørsmål 1.1 (5%)

Forklar kort hva som menes med *ordforekomst, ordform, leksem* og *lemma* med utgangspunkt i

1. *mouse, mice, rat, rats*

og teksten

2. *Rats aren't smarter than mice, and that's important. "Anything we could train a rat to do we could train a mouse to do as well", says Tony Zador. This finding is important because using mice in experiments instead of rats could open up all kinds of new research options.*

Spørsmål 1.2 (5%)

Hva er "stemming" og hva er lemmatisering? Hva er forskjellene mellom de to?

Spørsmål 1.3 (5%)

Hva mener vi med tokenisering ("tokenization") i språkteknologi? Med utgangspunkt i teksten (2) over, diskuter kort valg som kan oppstå når en skal lage en tokeniserer.

Oppgave 2 (samlet 15%)

La grammatikk G være:

```
S   -> NP VP
NP  -> ND
NP  -> DET NI
NP  -> NP PP
PP  -> P NP
DET -> NP G
VP  -> lo | smilte | danset
ND  -> faren | mora | broren | søstera | kona | mannen
NI  -> far | mor | kone | mann | søster | bror
P   -> til
G   -> sin
DET -> en | ei
NP  -> William | Kate | Elizabeth | Philip |
       Charles | Michael | Pippa | Camilla
```

Spørsmål 2.1 (8%)
Tegn opp de to trærne grammatikken gir til ordsekvensen

3. *Faren til William sin kone danset*

Spørsmål 2.2 (7%)
Hvor mange forskjellige syntaktiske analyser har sekvens (4)?

4. *Faren til William sin kone sin søster danset*

Har alle de syntaktiske analysene ulik betydning? Forklar hvilke par som evt. har samme betydning og hvilke par som evt. har forskjellig betydning og hva forskjellene i betydning består i.

Oppgave 3 (samlet 20%)

Spørsmål 3.1 (10%)
Omform grammatikk G til en grammatikk G1 på Chomsky-normalform (CNF) og vis hvordan CKY-algoritmen vil anerkjenne sekvens (4).

Spørsmål 3.2 (10%)
Vis hvordan en chart-parser for grammatikk G kan anerkjenne

5. *William sin kone danset*

Oppgave 4 (10%)
Er språket generert av G regulært? Begrunn svaret.

Oppgave 5 (20%)
Eiendomsrelasjoner og liknende relasjoner, f.eks. familierelasjoner, kan i norsk uttrykkes med *sin* som vi har sett så langt. Men de kan også uttrykkes med genitivs-s. For eksempel har setning (6) like mange analyser som setning (4) med tilsvarende betydninger.

6. *Faren til Williams kones søster danset*

Det er flere mulige måter å beskrive slike konstruksjoner. Her velger vi å betrakte *Williams* og de andre ordene med genitivs-s som vanlige ord og deretter endre de syntaktiske reglene i grammatikk G for å passe til dette. Først legger vi genetivsordene til leksikon.

```
ND  -> farens | moras | brorens | søsteras | konas | mannens
NI  -> fars | mors | kones | manns | søsters | brors
NP  -> Williams | Kates | Elizabeths | Philips |
       Charles' | Michaels | Pippas | Camillas
```

Du skal nå endre og utvide grammatikken med regler slik at den generer sekvens (6) med rett antall analyser. Genitivskonstruksjonen skal kunne gjentas vilkårlig mange ganger (*søsteras brors mors fars* ...). Du skal også utstyre grammatikkreglene med trekk slik at grammatikken ikke overgenerer. Spesielt skal den ikke generere

7. ** Moras danset*
8. ** Mora søster danset*

I tillegg skal du bruke trekk for å få riktig samsvar med den ubestemte artikkel:

ei mor, ei søster, ei kone
en mann, en far, en bror

Her skal vi ikke tillate

** ei mann, *ei far, *ei bror*

Men vi skal tillate

en mor, en søster, en kone

Deretter skal du tegne trærne – med trekk – som grammatikken gir til setningen:

9. *Faren til Williams kone danset*

Oppgave 6 (samlet 20%)

Vi skal se på klassifisering og vi bruker navnesettet fra NLTK. Det inneholder 7944 navn som hver er tilordnet en klasse, *male* eller *female*. Vi legger til side 794 navn til testing og bruker 7150 navn til trening. Av disse 7150 navnene er 4500 klassifisert som *female* og 2650 klassifisert som *male*.

Spørsmål 6.1 (10%)

Vi vil lage en Naive Bayes-klassfikator. Denne ser bare på ett trekk, siste bokstav i navnet. For bokstaven *n* er det 335 navn i treningssettet som er merket som *female* og 425 som er merket som *male*. Forklar hvordan klassifikatoren vil gå frem for å klassifisere navnet *John*. Du trenger ikke glatte verdiene.

Spørsmål 6.2 (5%)

Denne klassifikatoren vil måtte gi samme klasse til *John* og *Marilyn*. For å skille mellom dem tar vi derfor med et trekk til. Vi ser på suffiks av de to siste bokstavene. Det er 70 navn i klassen *female* som har *yn* som de to siste bokstavene, og 10 navn i klassen *male* som har *yn* som de to siste bokstavene.
Det er ingen *female*-navn som har *hn* som de to siste bokstavene, mens det er 3 *male*-navn som har det.
Hvordan vil denne klassifikatoren gå frem for å klassifiser *Marilyn* og *John*? Får noen av dem endret klasse når vi legger til dette trekket?

Spørsmål 6.3 (5%)

Trekket *hn* opptrer svært sjelden. Det samme gjelder f.eks. for *wn* som forekommer 3 ganger i klassen *female* og en gang i klassen *male* og *ln* som forekommer en gang i klassen *male* og 0 ganger i klassen *female*. Forklar hvorfor dette kan bli et problem. En måte å rette opp dette på er å bruke glatting ("smoothing"), og den enkleste formen er å "legge-til-en" glatting ("add-one"). Hvordan vil du bruke denne metoden for å glatte verdiene på suffiks-trekket?

SLUTT