UNIVERSITETET I OSLO

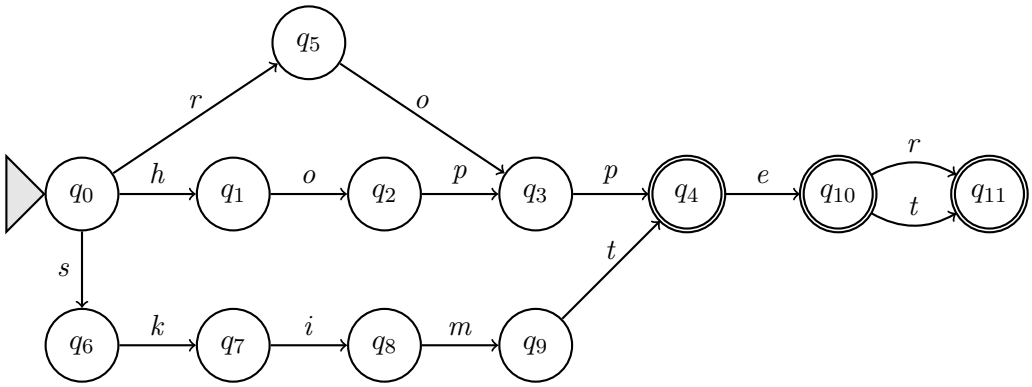
Det matematisk-naturvitenskapelige fakultet

Eksamen i:	INF1820 — Introduksjon til språk- og kommunikasjonsteknologi
Eksamensdag:	17. juni 2016
Tid for eksamen:	14.30 – 18.30
Oppgavesettet er på 6 sider.	
Vedlegg:	Ingen
Tillatte hjelpemidler:	Ingen

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

(Fortsettes på side 2.)

Oppgave 1 Tilstandsmaskiner og regulære uttrykk (vekt 25%)



1. I figuren over ser du en tilstandsautomat som gjenkjenner enkelte norske verbformer. Hvilke former av hvilke verb gjenkjenner automaten? Du trenger ikke liste opp alle formene, men beskriv alle mulighetene for automaten.
   Imperativ, infinitiv, presens, preteritum av rope, hoppe, skimte.

2. Skriv et regulært uttrykk som gjenkjenner de samme formene som automaten.

   ```
   /(rop|hopp|skimt)(e|er|et)?/
   /(rop|hopp|skimt)(e[rt]?)?/
   /(rop|hopp|skimt)(e(r|t)?)?/
   ```

3. I faget har vi snakket en del om skillet mellom deterministiske og ikke-deterministiske automater. Hva kjennetegner en deterministisk automat?
   En deterministisk automat kjennetegnes av at det kun finnes én mulig transisjon fra enhver tilstand gitt et symbol hentet fra alfabetet.

Oppgave 2 Morfologi (vekt 25%)

1. Vi skiller mellom såkalte åpne og lukkede ordklasser. Hva skiller de to typene ordklasse fra hverandre?
   De åpne ordklassene utvides ofte og får stadig nye medlemmer. De lukkede derimot består av en liten gruppe ord som svært sjelden får nye medlemmer

2. Av de ti ordklassene vi opererer med på norsk, hvilke regner vi som åpne?

   Substantiv, verb, adjektiv (adverb)

(Fortsettes på side 3.)

$$P(\vec{t}|\vec{w}) = \prod_{i=1}^{N} P(t_i|t_{i-1})P(w_i|t_i)$$

3. Ligningen over viser sannsynligheten en HMM-modell tilordner en taggsekvens gitt en ordsekvens. Hvilke sannsynligheter trenger vi for å regne ut sannsynligheten for at setningen Babyen begynte å gå tidlig har taggsekvensen subst verb sbu verb adv?
   Observasjonssannsynligheter og transisjonssannsynligheter

   ```
   P(Babyen|subst) x P(subst|<s>)  x
   P(begynte|verb) x P(verb|subst) x
   P(aa|sbu) x P(sbu|verb) x
   P(gaa|verb) x P(verb|sbu) x
   P(tidlig|adv) x P(adv|verb)
   ```

4. Et viktig problem for HMM-modeller er ukjente ord. Beskriv kort hvilket problem ukjente ord medfører for en HMM-modell og hvordan vi løser det.
   Ved MLE fra et korpus vil den estimerte sannsynligheten for et ukjent ord bli 0 og dermed blir sannsynligheten for hele setningen 0 (pga multiplikasjon). Dette kan løses ved såkalt smoothing, der man reserverer noe av sannsynlighetsmassen for ukjente ord (f.eks. add-one smoothing).

Oppgave 3 Syntaks (vekt 25%)

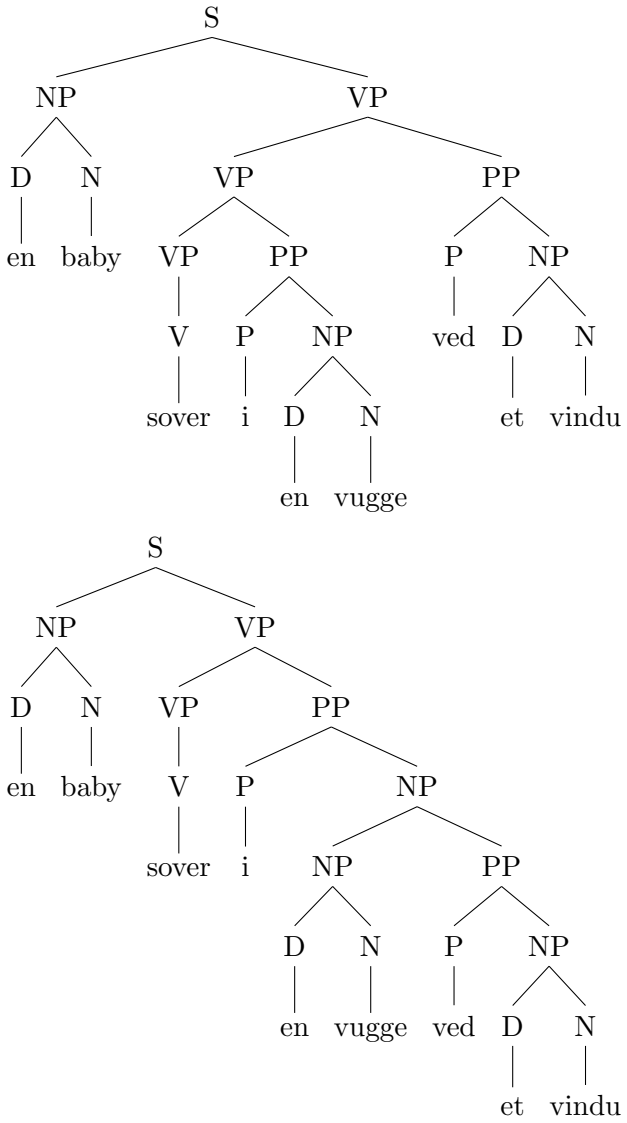
Anta følgende grammatikk for et fragment av norsk:

$S \rightarrow NP\ \ VP$	$D \rightarrow$ en | et
$NP \rightarrow D\ N$	$N \rightarrow$ baby | vindu | vugge
$NP \rightarrow NP\ \ PP$	$V \rightarrow$ sover
$VP \rightarrow V \mid V\ NP$	$P \rightarrow$ i | ved
$VP \rightarrow VP\ \ PP$	
$PP \rightarrow P\ NP$	

1. Hvor mange analyser tildeler grammatikken til setningene under?

   (a) en baby sover i en vugge ved et vindu *2 analyser*

(Fortsettes på side 4.)



(b) et barn i en vugge sover *1 analyse*

2. Tegn treet eller trærne grammatikken tilordner setning (a).

3. Utvid grammatikken slik at den tillater vilkårlig mange adjektiver foran substantivet i substantivfrasen og godtar setninger som:

   (a) en liten søt baby sover i en stor vugge

```
NP -> D Nom
Nom -> N
Nom -> Adj Nom
```

4. Hvor mange analyser tilordner den utvidede grammatikken din til setningen en rund liten søt baby sover? *1 analyse*

Oppgave 4 Semantikk (vekt 25%)

1. Vi sier at *Erna Solberg* og *statsministeren* har lik referanse, men forkjellig betydning. Hva mener vi med dette?

(Fortsettes på side 5.)

Referanse og betydning er to ulike aspekter ved mening. Der referanse peker ut objekter i verden, representerer betydning et mer varig meningsaspekt. F.eks. i 2016 er Erna Solberg statsminister, s ordene "Erna Solberg" og "statsministeren" referer til det samme, men statsminister inneholder ogs mer informasjon, f.eks. landets politiske overhode, etc. som er uavhengig av referanse.

2. Forklar kort hva en semantisk rolle er, og hvordan vi bruker semantiske roller til å analysere setninger.
   Semantiske roller beskriver de ulike rollene deltagerene i handlingen som beskrives av hovedverbet innehar. Det er i hovedsak verbets argumenter som analyseres og tildeles en rolle og dette gjøres på konstituentnivå.

3. Analyser setningen Den søte babyen kaster grøt. ved hjelp av semantiske roller. Beskriv de forskjellige rollene du har tilordnet.

   - *Den søte babyen – AGENT*
   - *grøy – THEME*

   AGENT-rollen brukes for å beskrive deltageren som utfører handlingen beskrevet av verbet med viten og vilje.
   THEME-rollen brukes for å beskrive deltageren som påvirkes av handlingen og som forflyttes som følge av at handlingen finner sted.

4. I informasjonsgjenfinning (Information Retrieval) brukes som regel den såkalte *vektorrommodellen* for å representere dokumenter. Beskriv kort hvordan dokumenter representeres i denne modellen.
   I vektorrommodellen representeres dokumenterer som en vektor der dimensjonene er ordene ("termene") som forekommer i dokumentet. Dersom termen forekommer i dokumentet er dets verdi i vektoren ikke-null, f.eks. termens frekvens i dokumentet eller en vektet frekvens (se under). La oss anta følgende dokumentsamling:

   - $d_1 = \{a,a,b,b,b\}$
   - $d_2 = \{a,a,a,b,b\}$

   For denne samlingen får vi vektorene $\overrightarrow{d_1}$ *og* $\overrightarrow{d_2}$*:*

   - $\overrightarrow{d_1}$ *= (2,3)*
   - $\overrightarrow{d_2}$ *= (3,2)*

   der dimensjonene i vektoren svarer til henholdsvis termene 'a' og 'b'.

5. I en vektorrommodell brukes ofte *tf-idf*-vekting for å forbedre modellen. Hvordan regner vi ut idf-vekten, og hva er hensikten med denne vektingen?
   tf-idf vekter brukes til vekte termers frekvens i en vektorrommodell og foretrekker ord som er vanlige i ett dokument, men sjeldne i samlingen som helhet. Vektingen regnes ut slik:

$$\mathrm{v}_{1,\mathrm{movie}} = \mathrm{tf}_{1,\mathrm{movie}} \times \mathrm{idf}_{\mathrm{movie}}$$

(Fortsettes på side 6.)

der idf-vekten regnes ut slik:

$$\text{idf}_{\text{movie}} = \log \frac{N}{n_{\text{movie}}}$$

der vi har N dokumenter i dokumentsamlingen, og n_{movie} av dem inneholder termen movie.