# Oppgave 3.2

**3.2 Ordvektorer og likhet**

Når vi skal regne ut likheten mellom to ordvektorer, kan vi bruke kosinus-likhet (cosine similarity) som et likhetsmål. Kosinus-likhet mellom to vektorer $A$ og $B$ er definert som følger:

$$sim(A,B) = \frac{A \cdot B}{\|A\|\|B\|} = \frac{\sum_{i=1}^{n} A_i B_i}{\sqrt{\sum_{i=1}^{n} A_i^2}\sqrt{\sum_{i=1}^{n} B_i^2}}$$

Telleren i formelen er prikkproduktet (dot product) til vektorene $A$ og $B$:

$$\sum_{i=1}^{n} A_i B_i$$

Nevneren her er størrelsen eller lengden til vektorene $A$ og $B$:

$$\sqrt{\sum_{i=1}^{n} A_i^2}\sqrt{\sum_{i=1}^{n} B_i^2}$$

Under finner du en matrise med termer og kontekster, som inneholder f.eks. TF-IDF eller PPMI verdier for tre **termer** og 8 *kontekst-ord*. Verdiene her er funnet på, for å gjøre utregningene lettere.

| | **eple** | **pære** | **frukt** |
|---|---|---|---|
| *anana* | 3 | | 0 |
| *eple* | 0 | | 2 |
| *pære* | 0 | | 1 |
| *skjære* | 2 | | 0 |
| *desser* | 2 | | 4 |
| *lunsj* | 3 | | 0 |
| *grill* | 1 | | 0 |
| *Eva* | 3 | | 2 |

La oss kalle ordvektorene for termene **eple**, **pære** og **frukt** for $V_{eple}$, $V_{pære}$ and $V_{frukt}$.

Hva er $\|V_{eple}\|$, dvs. lengden til $V_{eple}$? Gi svaret som et tall:

$||V_{eple}|| =$

$3^2+0^2+0^2+2^2+2^2+3^2+1^2+3^2 =$

$9+4+4+9+1+9= 36$

Kvadratrot av 36 = 6

$||V_{eple}|| = 6$

# Oppgave 3.2

### 3.2 Ordvektorer og likhet

Når vi skal regne ut likheten mellom to ordvektorer, kan vi bruke kosinus-likhet (cosine similarity) som et likhetsmål. Kosinus-likhet mellom to vektorer $A$ og $B$ er definert som følger:

$$sim(A,B) = \frac{A \cdot B}{\|A\|\|B\|} = \frac{\sum_{i=1}^{n} A_i B_i}{\sqrt{\sum_{i=1}^{n} A_i^2}\sqrt{\sum_{i=1}^{n} B_i^2}}$$

Telleren i formelen er prikkproduktet (dot product) til vektorene $A$ og $B$:

$$\sum_{i=1}^{n} A_i B_i$$

Nevneren her er størrelsen eller lengden til vektorene $A$ og $B$:

$$\sqrt{\sum_{i=1}^{n} A_i^2}\sqrt{\sum_{i=1}^{n} B_i^2}$$

Under finner du en matrise med termer og kontekster, som inneholder f.eks. TF-IDF eller PPMI verdier for tre **termer** og 8 *kontekst-ord*. Verdiene her er funnet på, for å gjøre utregningene lettere.

| | **eple** | **pære** | **frukt** |
|---|---|---|---|
| *ananas* | 3 | 1 | 0 |
| *eple* | 0 | 0 | 2 |
| *pære* | 0 | 0 | 1 |
| *skjære* | 2 | 0 | 0 |
| *dessert* | 2 | 1 | 4 |
| *lunsj* | 3 | 1 | 0 |
| *grill* | 1 | 1 | 0 |
| *Eva* | 3 | 0 | 2 |

La oss kalle ordvektorene for termene **eple**, **pære** og **frukt** for $V_{eple}$, $V_{pære}$ and $V_{frukt}$.

Videre skal vi bruke kosinus-liket som likhetsmål.

Hva er likheten mellom de to *likeste* ordene?

**Lengde eple = 6  Lengde pære = 2  Lengde frukt =  5**

Sim(eple, pære) =

$$\frac{3*1+0*0+0*0+2*0+2*1+3*1+1*1+3*0}{6*2} = \frac{9}{12} = 0{,}75$$

Sim (eple, frukt) =

$$\frac{3*0+0*2+0*1+2*0+2*4+3*0+1*0+3*2}{6*5} = \frac{14}{30} = 0{,}46$$

Sim (pære, frukt) =

$$\frac{1*0+0*2+0*1+0*0+1*4+1*0+1*0+0*2}{2*5} = \frac{4}{10} = 0{,}40$$

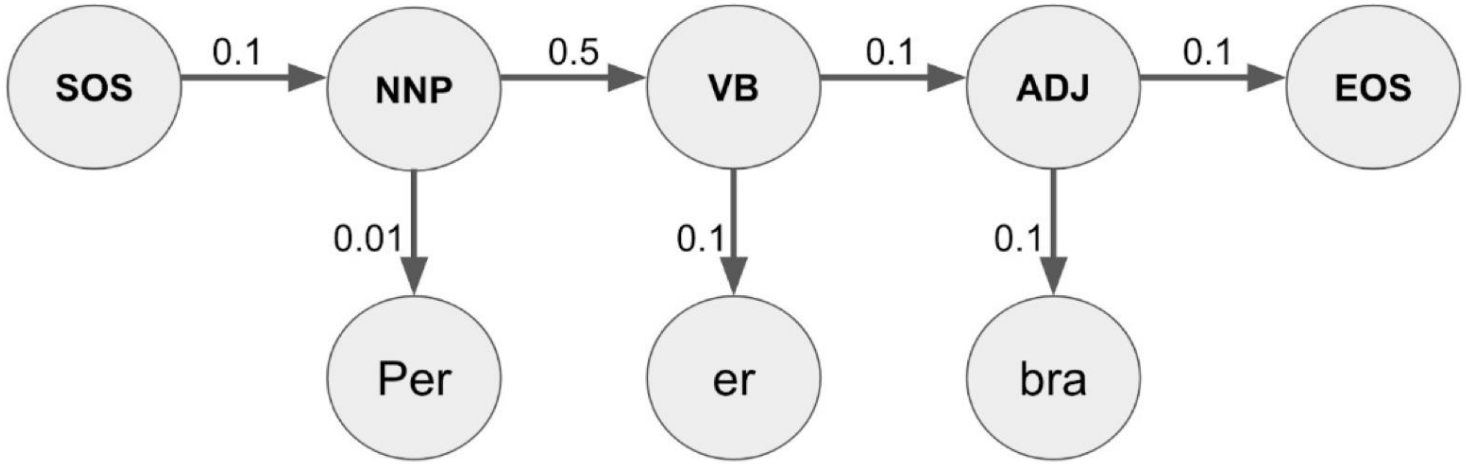
# Oppgave 4

**4**

## Hidden Markov Models

Du har følgende grafiske framstilling av observerte og skjulte tilstander, med sannsynligheter, gitt en HMM-modell.

Dette er det eneste du vet om modellen.



# Oppgave 4.1

## Komponentene i HMM

I pensum i kurset leste du om Hidden Markov Models. Slik Jurafsky og Martin definerte HMM, består den av følgende fem komponenter:

1. $Q = q_1 q_2 \dots q_n$
2. $A = a_{11} a_{12} \dots a_{n1} \dots a_{nn}$
3. $O = o_1 o_2 \dots o_T$
4. $B = b_i(o_t)$
5. $q_0, q_F$

I denne oppgaven skal du forklare hva hver av de fem komponentene er. (Du vil få en del begreper og hint i de neste oppgavene, så hvis du har problemer med å huske, kan det være lurt å lese gjennom disse først).

I tillegg snakket vi om ordklassetagging som en anvendelse av HMM'er. Der det er relevant, gi eksempler fra ordklassetagging. For eksempel, kan du i den første delen, gi eksempler på hva $q$ kan være i kontekst av en ordklassetagger.

**Forklar hver av de fem komponenente i en HMM her...**

Q er mengden av **skjulte tilstander** i modellen. I en ordklassetagger vil dette **svare til ordklassene**.

A er **transisjons-sannsynlighetene**, dvs sannsynligheten for at man går fra **en skjult tilstand til en annen**. For ordklassetagging, kan dette være sannsynligheten for at man for **eksempel ser et verb etter et pronomen**.

O er **observasjonene** i modellen. For en ordklassetagger **vil dette være ordene**. O kan også kalles vokabularet.

# Oppgave 4.1

## Komponentene i HMM

I pensum i kurset leste du om Hidden Markov Models. Slik Jurafsky og Martin definerte HMM, består den av følgende fem komponenter:

1. $Q = q_1 q_2 \ldots q_n$
2. $A = a_{11} a_{12} \ldots a_{n1} \ldots a_{nn}$
3. $O = o_1 o_2 \ldots o_T$
4. $B = b_i(o_t)$
5. $q_0, q_F$

I denne oppgaven skal du forklare hva hver av de fem komponentene er. (Du vil få en del begreper og hint i de neste oppgavene, så hvis du har problemer med å huske, kan det være lurt å lese gjennom disse først).

I tillegg snakket vi om ordklassetagging som en anvendelse av HMM'er. Der det er relevant, gi eksempler fra ordklassetagging. For eksempel, kan du i den første delen, gi eksempler på hva $q$ kan være i kontekst av en ordklassetagger.

**Forklar hver av de fem komponenente i en HMM her...**

B er **emmisjonssannsynlighetene**. Dette betyr sannsynligheten for å se en **observasjon gitt en skjult tilstand.** For ordklassetaggeren, er det sannsynligheten for et ord gitt en ordklasse. For eksempel kan det være **sannsynligheten for at vi ser ordet "Per" gitt ordklassen egennavn**.

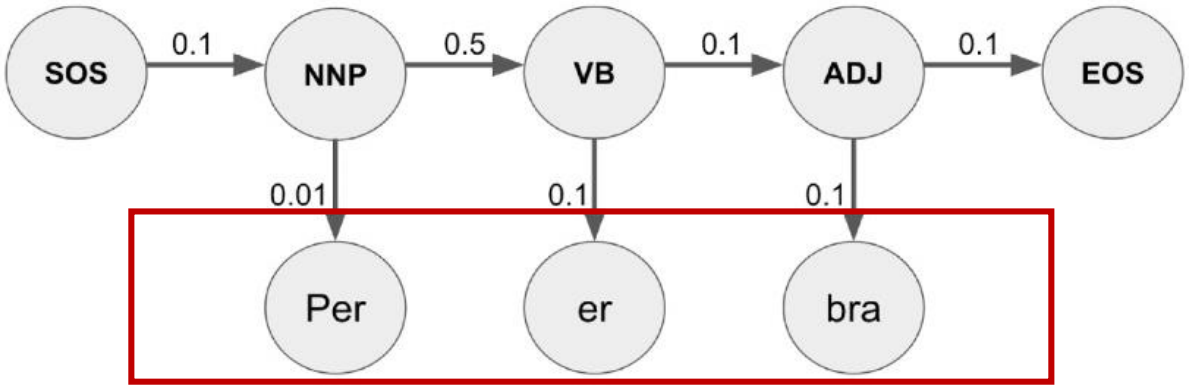
Dette er de spesielle **start- og sluttilstandene.** Disse brukes for å kunne beregne sannsynligheter for første og siste tilstand i en sekvens. Disse har ingen spesiell rolle for ordklassetagging, men kan sees på som **setningsstart og setningsslutt.**

# Oppgave 4.2

2 **Vokabular**

Hva er vokabularet i denne HMM-en?
**Velg et eller flere alternativer**

- ADJ
- Per
- NNP
- EOS
- VB
- SOS
- er
- bra




O er **observasjonene** i modellen. For en ordklassetagger **vil dette være ordene**. O kan også kalles vokabularet når vi jobber med tekst.

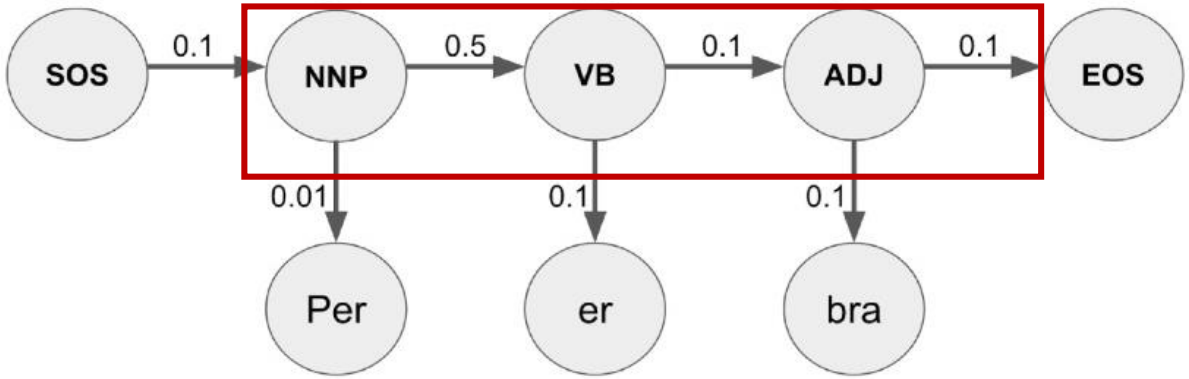
# Oppgave 4.3

## Tilstander

Hva er de skjulte tilstandene i denne HMM-en? Ta ikke med de spesielle start- og slutt-tilstandene.

**Velg et eller flere alternativer**

- NNP
- bra
- EOS
- er
- ADJ
- VB
- SOS
- Per




Q er mengden av **skjulte tilstander** i modellen. I en ordklassetagger vil dette **svare til ordklassene**.

# Oppgave 4.4

## Emmisjonssannsynligheter

Under følger en rekke påstander om emmisjonssannsynligheter, generelt og gitt grafen for "Per er bra".

Kryss av for alle påstander som er korrekt
**Velg ett eller flere alternativer**

- Emmisjonssannsynligheten for observasjonen "Per" gitt den skjulte tilstanden "NNP" er 0.01 → Ja!
- Emmisjonssannsynligheten for observasjonen "EOS" gitt den skjulte tilstanden "bra" er 0.1 → «EOS» ingen observasjon. «bra» ingen skjult tilstand.
- Emmisjonssannsynligheten for observasjonen "NNP" gitt den skjulte tilstanden "VB" er 0.5 → «NNP» ingen
- Den skjulte tilstanden "SOS" har aldri noen emmisjonssannsynligheter → Ja!

# Oppgave 5

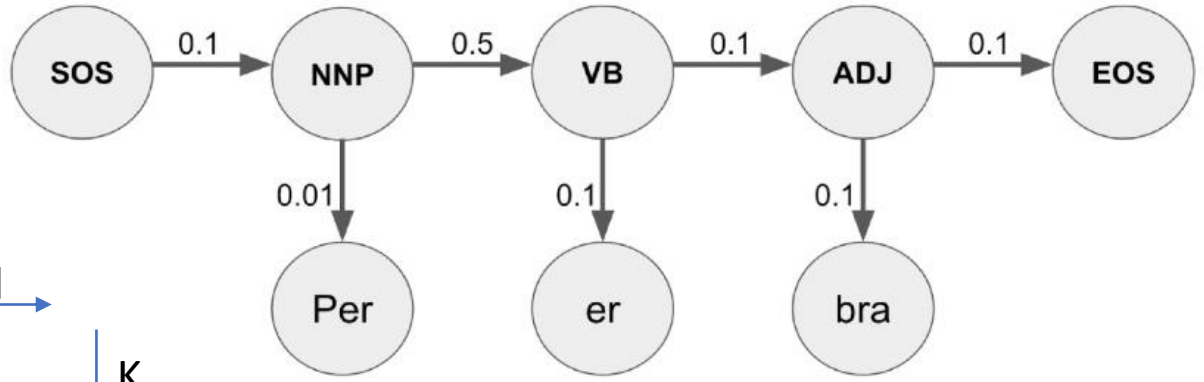
5 **Transisjonssannsynligheter**

Gitt at grafen beskrevet er alt du vet om modellen, fyll inn transisjons-sannsynlighetene i denne matrisen, der radene er forrige tilstand *t-1*, og kolonnene er nåværende tilstand *t.*

Skriv inn 0 hvis:

- sannsynligheten for transisjonen faktisk er 0
- du ikke har informasjon om sannsynligheten til denne transisjonen

Du skal (og kan) ikke fylle inn feltene med **-** (strek), bare de tomme boksene.




Transisjonsprobabilitetsmatrise

| | $SOS_t$ | $NNP_t$ | $VB_t$ | $ADJ_t$ | $EOS_t$ |
|---|---|---|---|---|---|
| $SOS_{t-1}$ | - | 0.1 | - | - | - |
| $NNP_{t-1}$ | 0 | - | 0.5 | - | - |
| $VB_{t-1}$ | - | 0 | - | 0.1 | - |
| $ADJ_{t-1}$ | - | - | 0 | - | 0.1 |
| $EOS_{t-1}$ | - | - | - | 0 | - |

Rad

Kolonne

# Oppagve 6

## Egennavn og verb

Etter egennavn ("NNP") er det i norsk ganske vanlig å finne et verb ("VB"), spesielt etter egennavn som er subjekter.

Klarer HMM-modellen å fange opp relasjonen mellom et egennavn ("NNP") og påfølgende verb ("VB")?

Forklar hvorfor/hvorfor ikke.

Ja, HMM'en fanger opp denne relasjonen. Det er akkurat det transisjonssannsynligheten melleom egennavn/NNP og verb/VB utrykker.

# Oppgave 7

**Navnet Per**

"Per" er et ganske vanlig navn i norsk. Klarer en HMM å uttrykke noe om hvor vanlig dette navnet er?

Forklar hvorfor/hvorfor ikke.

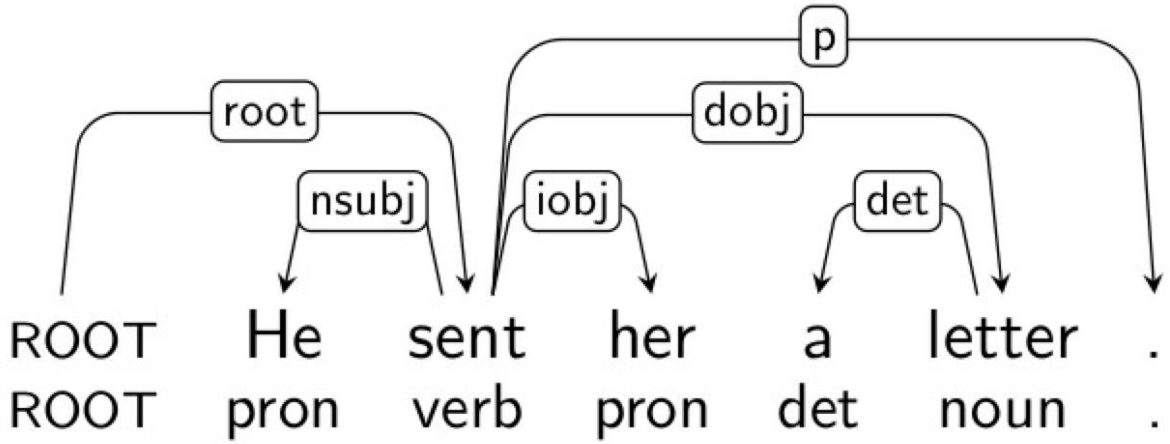
Ja, HMM'en klarer å trykke dette også. Dette svarer til emmisjonssannsynligheten for observasjonen "Per" gitt den skjulte tilstanden NNP/egennavn.

# Oppgave 6.4

## Overgangssekvenser

Vi har sett på to forskjellige typer overgangssystemer (*transition systems*), dvs. *arc-eager*– og *arc-standard*–systemene. Forklar i et par setninger hva som er forskjellen mellom de to.

Gitt dependenstreet under, skriv opp overgangssekvensen som gir opphav til dette treet, både for *arc-eager*– og for *arc-standard*–varianten.



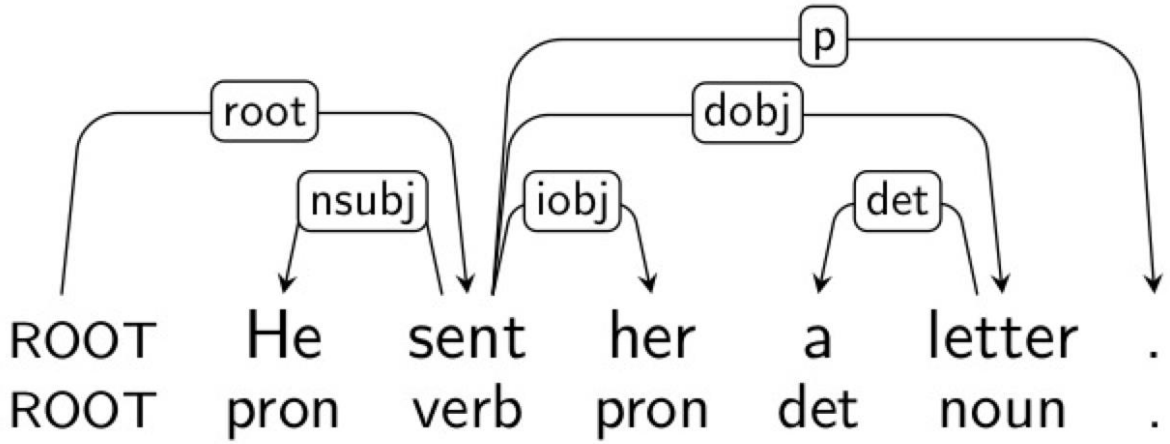

Hovedforskjellen er at *arc-standard* ligner mer på 'klassisk' shift-reduce–parsing: den **lager bare kanter mellom to elementer som ligger på stacken**, mens *arc-eager* kan allerede **lage en kant til en dependent når den står først i bufferen**. I *arc-standard* har man ikke en egen **REDUCE**, etter at både RIGHT- ARC() og LEFT-ARC() fjerner ett element fra stacken.

# Oppgave 6.4

**Overgangssekvenser**

Vi har sett på to forskjellige typer overgangssystemer (*transition systems*), dvs. *arc-eager*– og *arc-standard*–systemene. Forklar i et par setninger hva som er forskjellen mellom de to.

Gitt dependenstreet under, skriv opp overgangssekvensen som gir opphav til dette treet, både for *arc-eager*– og for *arc-standard*–varianten.




I ***arc-eager***: (SHIFT,) SHIFT, LEFT-ARC(NSUBJ), RIGHT-ARC(ROOT), RIGHT-ARC(IOBJ), REDUCE, SHIFT, LEFT-ARC(DET), RIGHT-ARC(DOBJ), REDUCE, RIGHT-ARC(P), (REDUCE, REDUCE)

I ***arc-standard***: (SHIFT,) SHIFT, SHIFT, LEFT-ARC(NSUBJ), SHIFT, RIGHT-ARC(IOBJ), SHIFT, SHIFT, LEFT-ARC(DET), RIGHT-ARC(DOBJ), SHIFT, RIGHT-ARC(P), RIGHT-ARC(ROOT).