Forkurs i Språkteknologi

Annika Willoch Olstad
annikaol@uio.no

Petter Mæhlum
pettemae@uio.no

Institutt for Informatikk
Universitetet i Oslo

August 2022

0 Introduksjon

Høsten 2022 arrangerer vi forkurs i språkteknologi for første gang. Forkurset er et valgfritt tilbud som gis på frivillig basis, med mål om å gjøre studentene bedre rustet til introduksjonsfagene i språkteknologi. Dette heftet inneholder en oppsummering av kursets to deler, lingvistikk og matematikk, og tilhørende øvingsoppgaver.

Har du kommentarer, innspill eller spørsmål til forkurset, oppsummeringen eller oppgavene, blir vi glade for at du sender oss en mail.

Lykke til!

1 Lingvistikk

Lingvistikk er det vitenskapelige studiet av naturlig språk, altså språk vi mennesker bruker. I denne delen av forkurset studerer vi kort noen viktige områder innenfor dette feltet: ord, setninger, ordklasser og setningsanalyse.

1.1 Del 1 - Ord, setninger og ordklasser

1.1.1 Ord og setninger

Hva er et ord? Det kan være vanskelig å definere helt presist hva et ord er, men en vanlig definisjon er at ord er språklige enheter med en egen betydning eller funksjon.

Merk allikevel at det er mange spørsmål som kan dukke opp når vi definerer ord slik. Er det bare ord i ordbøkene som er "ekte" ord? Hva er da egennavn som "Petter" og "Annika"? Er tegnsetting som punktum og komma ord? Dette er definisjonsspørsmål det er viktig å stille, og svaret vårt kan avhenge av hva ordene skal brukes til.

Hva er en setning? En *setning* er en sekvens av ord som minimum består av et subjekt og et verbal. Enkelt sagt må setningen minst inneholde informasjon om *noen* som *gjør noe*.

Vi skiller gjerne mellom to typer setninger - *hovedsetning* og *leddsetning*:

- **Hovedsetning** er en hel setning som kan stå alene.
- **Leddsetning** er en delsetning som må stå som del av en hovedsetning.

1.1.2 Ordklasser

Ordklasser er kategorier for ord som deler bestemte egenskaper, og sier noe om hva slags kontekst et ord forekommer i. Hvilke egenskaper dette er, avhenger av hvordan vi definerer tilhørighet til en ordklasse. Det kan gjøres på flere måter, men en vanlig gruppering er å se på hva slags funksjon det aktuelle ordet har i en setning og ordets bøyningsformer.

Under følger en beskrivelse av ordklassene for norsk. Merk at vi deler ordklassene inn i åpne og lukkede klasser, avhengig av om klassene tar inn nye ord eller ikke.

Åpne ordklasser
De åpne ordklassene tar inn nye ord i klassen, for eksempel låneord fra andre språk eller nyoppståtte ord.

Substantiv Substantiv er ord som bøyes i bestemthet og tall, og betegner ofte "ting", som for eksempel "sjokolade":

sjokolade - sjokolad**en** - sjokolad**er** - sjokolad**ene**

Et unntak er *egennavn*, som ikke bøyes.

Verb Verb er ord som bøyes i tid, og betegner ofte handlinger og tilstander. Vi skiller mellom *hovedverb* og *hjelpeverb*:

- Hovedverb står alene. For eksempel *presens*: "jeg spiser".
- Hjelpeverb står som regel aldri alene, men alltid sammen med et annet verb. For eksempel "jeg har spist".

Adjektiv Adjektiv bøyes i samsvar med substantivet det modifiserer, i bestemthet, kjønn og tall. Adjektiv kan også *gradbøyes*, for eksempel:

en smart hund - en smart**ere** hund - den smart**este** hunden

Interjeksjoner Interjeksjoner er enkeltytringer som kan stå alene, og som normalt ikke har en spesifikk funksjon i en setning. De kan heller ikke bøyes. Eksempler på interjeksjoner er *æsj*, *ja*, og *hallo*.

Adverb Ordklassen adverb er en klasse der mange forskjellige ord havner. På en måte er det en "restklasse". Hvis et ord ikke passer godt inn i en annen ordklasse, kan det godt være et adverb. Men noen ting kan vi si: adverb forteller oss ofte noe om hvordan, når, hvor og hvorfor ting skjer. De kan ofte plasseres ganske fritt i setninger.

Lukkede ordklasser
De lukkede ordklassene er ordklasser der innholdet ikke varierer så mye.

Konjunksjoner Konjunksjoner (tidligere kalt sideordnede konjunksjoner), sideordner to ledd eller to setninger. Leddene skal altså være like, på en eller annen måte. Eksempler på slik sideordning er "Kari og Ola" og "Jeg spiser og drikker". Hovedkonjunksjonene på norsk er *og*, *men* og *for*, *så* og *eller*. Konjunksjonene kjennetegnes ved at vi ikke får inversjon i setningen etter konjunksjonen, altså at verbet ikke bytter plass etter adverb, slik det gjør hos subjunksjonene.

Subjunksjoner Subjunksjonene (tidligere kalt underordnede konjunksjoner), underordner et ledd i en annen setning. Vi bruker subjunksjonene til å bruke setninger som subjekter, objekter, adverb og lignende. Noen av de vanligste subjunksjonene er *at, hvis, om* (som også kan være en preposisjon), *når, da* (som begge kan være adverb) og *som*. Subjunksjonene kjennetegnes ved at vi ofte får inversjon, altså at verbet bytter plass etter adverb. For eksempel:

$$\text{Jeg spiser ikke ost.} \longrightarrow \text{At jeg ikke spiser ost.}$$

Pronomener Pronomener er kort fortalt ord som står istedenfor substantiver. På norsk er det en liten gruppe ord som inneholder de typiske pronomenene *jeg, meg, hun, den, vi,* osv. I tillegg har vi *man, seg, hverandre, hvem* og *hva*.

Determinativer Determinativer er ord som sier noe om bestemthet og mengde. Ord som *mange*, *ingen* og *begge* er determinativer, sammen med possessivene (*min, din,* osv.), grunntallene (*en, to*) demonstrativer(*denne, disse,* osv.) og artikler (*en, ei, et, den,* osv.).

Preposisjoner Preposisjoner er stort sett korte ord som har mange ulike funksjoner. Betydningsmessig har de ofte med plassering å gjøre (*over, under, på, i, ved, rundt*), men de kan også uttrykke andre ting. Preposisjoner brukes også sammen med verb, og da kan de også omtales som verbalpartikler, som for eksempel i "slappe *av*".

Et vanlig kriterium er at preposisjonene kan stå foran substantiver, men de kan altså også brukes sammen med verb.

1.2 Del 2 - Setningsanalyse

Vi deler inn setingsanalysen i to bolker, av praktiske årsaker. Den første delen handler om setningsanalyse uten leddsetninger og sideordnede setninger, mens den andre inneholder underordnede og sideordnede setninger.

Del 1
I del 1 skal vi snakke om setningsanalyse uten subjunksjoner og konjunksjoner. Vi introduserer følgende ledd:

Subjekt : Den som utfører noe i en setning

SUBJEKT	VERBAL	D. OBJEKT
Jeg	klapper	katten
PRON	VERB	SUBST

Direkte objekt : Den som en handling blir utført på eller mot (men ikke mottaker)

SUBJEKT	VERBAL	**DIREKTE OBJ.**
Kua	spiser	**gress**
SUBST	VERB	**SUBST**

Indirekte objekt : Mottakeren av noe i en handling.

SUBJEKT	VERBAL	**INDIREKTE OBJ.**	DIREKTE OBJ.	
Jeg	ga	**henne**	et	brev
PRON	VERB	**PRON**	DET	SUBST

Verbal : Verbet i setningen

SUBJEKT	**VERBAL**	SUBJEKTSPRED.
Du	**er**	snill
PRON	**VERB**	ADJ

Subjektspredikativ : Det som beskriver subjektet.

SUBJEKT	VERBAL	**SUBJEKTSPRED.**
Treet	er	**grønt**
SUBST	VERB	**ADJ**

Del 2
Vi introduserer ikke flere ledd, men ser at setninger kan brukes som adverbial, subjekt og direkte objekt, samt at vi kan lage lange ledd.

Flere av subjunksjonene brukes når vi har hele setninger som inngår som ledd i andre setninger. *Hvis*, *når*, *om* og *da* introduserer adverbialer, *som* beskriver

substantiver, og *at* gjør at vi kan ta en hel setning som subjekt eller objekt i en setning.

Subjekt	Verbal	Adverbial			
Jeg	smiler	når	du	er	her
PRON	VERB	SUBJ	PRON	VERB	ADV

Subjekt	Verbal	Adverbial		
Hun	lo	da	hunden	hoppet
PRON	VERB	SUBJ	SUBST	VERB

Subjekt	Verbal	Direkte objekt		
Jeg	vet	at	han	sover
PRON	VERB	SUBJ	PRON	VERB

Subjekt				Verbal	S.predikativ
At	du	er	her	er	fint
SUBJ	PRON	VERB	ADV	VERB	ADJ

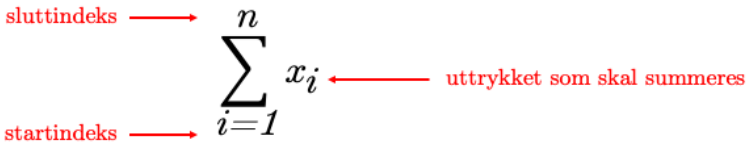
2 Matematikk

2.1 Del 1 - Summering, produkt og mengdelære

2.1.1 Sum

Summen av tall får vi ved å addere dem sammen. Addisjon markeres med plusstegnet +.

Med *summeringsnotasjon*, også kalt *sigma-notasjon*, kan vi skrive et langt summeringsuttrykk i ett enkelt uttrykk. Det gjør vi ved hjelp av symbolet $\sum$ (sigma). Et $\sum$-uttrykk består som regel av følgende deler:



$$\sum_{i=1}^{n} x_i$$

Eksempel
For å summere alle heltall fra 1 til 5, kan vi erstatte uttrykket

$$1+2+3+4+5=15 \qquad \text{med} \qquad \sum_{i=1}^{5} i = 15$$

For å summere alle partall fra 1 til 10 kan vi for eksempel erstatte uttrykket

$$2 + 4 + 6 + 8 + 10 = 30$$

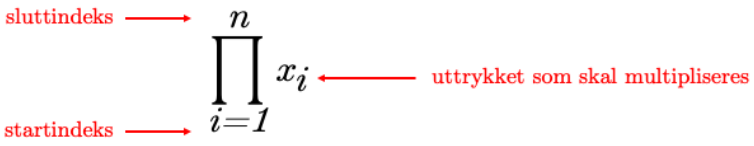
med

$$\sum_{i=1}^{5} 2i = 2 \cdot 1 + 2 \cdot 2 + 2 \cdot 3 + 2 \cdot 4 + 2 \cdot 5 = 30$$

2.1.2 Produkt

Produktet av tall får vi ved å multiplisere dem sammen. Multiplikasjon markeres med symbolet $\cdot$ eller $\times$.

Med *produktsnotasjon*, også kalt *pi-notasjon*, kan vi skrive et langt multiplikasjonssuttrykk i ett enkelt uttrykk. Det gjør vi ved hjelp av symbolet $\prod$ (pi). Et $\prod$-uttrykk består som regel av følgende deler:



Eksempel
For å multiplisere alle heltall fra 1 til 5, kan vi erstatte uttrykket

$$1 \cdot 2 \cdot 3 \cdot 4 \cdot 5 = 120 \qquad \text{med} \qquad \prod_{i=1}^{5} i = 120$$

For å multiplisere alle partall fra 1 til 10 kan vi for eksempel erstatte uttrykket

$$2 \cdot 4 \cdot 6 \cdot 8 \cdot 10 = 3840$$

med

$$\prod_{i=1}^{5} 2i = (2 \cdot 1) \cdot (2 \cdot 2) \cdot (2 \cdot 3) \cdot (2 \cdot 4) \cdot (2 \cdot 5) = 3840$$

2.1.3 Mengdelære

En *mengde* er en samling av objekter, hvor rekkefølgen og antall forekomster av hvert objekt ignoreres. Vi markerer en mengde med symbolene {}.

Eksempel

- $\{1, 2, 3, 4, 5\}$ er en mengde tall.
- $\{1, 2, 3, 4, 5\}$ og $\{4, 1, 3, 5, 2\}$ er *samme* mengde, fordi vi ignorerer rekkefølgen.
- $\{1, 2, 3, 4, 5\}$ og $\{1, 2, 3, 3, 3, 4, 5, 5\}$ er også *samme* mengde, fordi vi ignorerer antall forekomster av hvert objekt.

2.2 Del 2 - Sannsynlighet

Sannsynlighet er alltid et tall mellom 0 og 1.

2.2.1 Utfallsrom

Utfallsrommet er en mengde med ulike utfall som kan skje når vi utfører en handling, kalt en begivenhet. Hvis vi skal regne ut sannsynligheten for en delmengde av utfallsrommet kan vi addere sannsynlighetene. Vi snakker om begivenheter. Det er handlingene vi er interessert i. Hvis vi skal kaste en terning med 6 sider én gang, har vi følgende utfallsrom:

$$U = \{1, 2, 3, 4, 5, 6\}$$

Siden det totalt er 6 mulige utfall, er sannsynligheten for hvert utfall $\frac{1}{6}$. Hvis vi adderer alle sannsynlighetene for utfallene i utfallsrommet vil vi alltid få 1.

$$\frac{1}{6} + \frac{1}{6} + \frac{1}{6} + \frac{1}{6} + \frac{1}{6} + \frac{1}{6} = \frac{6}{6} = 1$$

Vi innfører notasjonen $p(x)$ for å si "sannsynligheten for begivenhetenen x". X er da alltid en delmengde av utfallsrommet.

Utfallene i utfallsrommet kan ha lik eller ulik sannsynlighet. Hvis de har lik sannsynlighet kaller vi dem uniform eller jevn. Hvis de ikke har lik sannsynlighet kaller vi distribusjonen ikke-uniform eller ujevn. Et eksempel på et tilfelle med ikke-uniforme sannsynligheter er en ballbinge der vi skal trekke en ball, og det ikke er like mange baller av hver farge. For eksempel hvis det er 3 røde baller og 5 grønne baller i en eske, så blir sannsynligheten ulik.

Hvis vi tillater flere mulige utfall i en begivenhet, kan vi addere sannsynlighetene. Hvis vi for eksempel vil se hva sannsynligheten for å få enten 2 *eller* 3 når vi kaster en terning med seks sider, så legger vi sammen sannsynlighetene for de to utfallene "2" og "3":

$$\frac{1}{6} + \frac{1}{6} = \frac{2}{6} = \frac{1}{3}$$

2.2.2 Betinget sannsynlighet

Et viktig skille i statistikken er mellom betingede og ubetingede sannsynligheter. Hvis vi kaster en terning to ganger, har ikke det første utfallet noe å si for det andre utfallet. Utfallene er ubetingede. Hvis vi derimot trekker baller fra en eske, uten å legge ballene tilbake, endrer vi sannsynligheten hver gang, fordi antallet baller minsker. Da er utfallene betingede. Vi innfører en spesiell notasjon for betingede sannsynligheter: $p(A|B)$. Vi leser | som "gitt", og leser hele uttrykket som "sannsynligheten for A gitt B". Dette betyr at vi skal finne sannsynligheten for at A skjer, gitt at B allerede har skjedd.

2.3 Produktsetningen

Vi har sett at vi kan addere sannsynligheter for å vite sannsynligheten for delmengder når det er snakk om én begivenhet. Vi ser først på ubetingede sannsynligheter. I disse tilfellene kan vi bare gangen sammen sannsynlighetene.

Sannsynligheten for å få 1 på to terningkast etter hverandre blir $\frac{1}{6} \cdot \frac{1}{6} = \frac{1}{36}$.

Hvis sannsynligheten til en beivenhet påvirkes av tidligere begivenheter, så sier vi at de er avhengige, sånn som når vi har fargede kuler i en eske. Ellers er de uavhenige, sånn som når vi kaster terninger.

Men når vi skal se på avhengige hendelser, blir det litt mer vanskelig regne ut sansynlighetene for to hendelser. Gitt en kulebinge med 3 røde og 5 grønne kuler, hva er sannsynligheten for å trekke først en rød kule, så en grønn?

Her bruker vi det som kalles produktsetningen:

$$P(A \cap B) = P(A|B) \cdot P(B)$$

Vi kan regne ut eksempelet vårt slik:

$$P(G \cap R) = P(R|G) \cdot P(G)$$

$$\mathrm{p(G)} = \tfrac{5}{8}$$

$$p(R|G) = \tfrac{3}{7}$$

$$\frac{3}{7} * \frac{5}{8} = \frac{15}{56}$$

$P(A \cap B)$ står for at A og B skjer etter hverandre. Vi kan bruke denne formelen til å finne sannsynligheten. Det viktige her er at vi må regne ut $P(A|B)$, som er en betinget sannsynlighet. I våre tilfeller kan vi regne oss fram til svaret, men i noen tilfeller har man ikke alle tallene tilgjengelige, og da er denne setningen nyttig.

2.4 Del 3 - "Godteposen"

I "godteposen" finner vi funksjoner og operatorer vi ikke har dekket i tidligere deler, men som er sentrale i både programmering og matematikk.

2.4.1 Fakultet

Fakultetsfunksjonen er en funksjon som tar et tall n og multipliserer sammen tallene fra n ned til 1. Vi markerer fakultet med symbolet !, som leses "fakultet".

Formelt definerer vi fakultetsfunksjonen slik, for et tall n:

$$n! = n(n-1)(n-2)...3 \cdot 2 \cdot 1$$

Eksempel
Vi kan regne ut 5! (fem fakultet), slik:

$$5! = 5 \cdot 4 \cdot 3 \cdot 2 \cdot 1 = 120$$

2.4.2 Modulo

Modulo-operasjonen gir oss resten vi sitter igjen med etter å ha dividert to heltall. Vi markerer modulo med forkortelsen *mod* eller med symbolet %.

Eksempel
Vi kan regne ut 9 mod 7, som også kan skrives 9 % 7, slik:

$$9 \mod 7 = 2$$

2.4.3 Heltallsdivisjon

Heltallsdivisjon, også kalt *divisjon med rest*, er en funksjon som returnerer største mulige heltall ved divisjon. Heltallsdivisjon markeres ofte med symbolene //.

Eksempel
Vi kan regne ut 9//4 slik:

$$9//4 = 2, \text{ med 1 i rest fordi } 2 \cdot 4 = 8 \text{ og } 9 - 8 = 1$$

3 Oppgaver

3.1 Lingvistikk

3.1.1 Del 1

1. Hva er et ord? Kan vi besvare dette spørsmålet helt presist?
2. Hvilke to elementer må vi minst ha med i en setning?
3. Hva er forskjellen på hovedverb og hjelpeverb?

4. Hva er forskjellen på åpne og lukkede ordklasser?
5. Tildel hvert ord i setningen under, riktig ordklasse:

   Jeg spiser en veldig god sjokoladekake
6. Forklar kort forskjellen på adjektiv og adverb.
7. Tildel hvert ord i setningen under, riktig ordklasse:

   Hunden er søt og snill, men den er ikke særlig smart
8. Forklar kort forskjellen på konjunksjoner og subjunksjoner.

3.1.2 Del 2

1. Nevn tre av de ulike funksjonene vi har.
2. Hva er forskjellen mellom direkte objekt og indirekte objekt?
3. Hva er det som pleier å komme etter subjunksjonene?
4. Tildel riktig ledd til setningene under.

   Jeg spiser ost.

   Jeg ga henne kaka.

   Jeg spiser og han drikker.

   Jeg vet at du spiser.

   Når du ler ler jeg.

3.2 Matematikk

3.2.1 Del 1

1. Skriv om uttrykket $5+6+7+8+9+10$ til et sigma-uttrykk, og regn ut.
2. Lag et sigma-uttrykk for å summere alle heltall fra 1 til 10. Regn ut.
3. Skriv om uttrykket $3 \cdot 6 \cdot 9 \cdot 12$ til pi-notasjon, og regn ut.
4. Lag et pi-uttrykk ved å multiplisere alle heltall fra 3 til 10. Regn ut.
5. Forklar kort hva som kjennetegner en mengde.
6. Er mengden $\{3, 2, 1\}$ og $\{1, 3, 2\}$ samme mengde? Hvorfor eller hvorfor ikke?
7. Er mengden $\{10, 9, 8, 7, 6, 5, 4, 3, 2, 1\}$ og $\{10, 9, 7, 6, 5, 4, 3, 2, 1\}$ samme mengde? Hvorfor eller hvorfor ikke?
8. Er mengden $\{1, 1, 2, 3, 4, 4, 4, 5\}$ og mengden $\{1, 2, 3, 4, 4, 4, 5\}$ samme mengde? Hvorfor eller hvorfor ikke?

3.2.2 Del 2

1. Hva er et utfallsrom?
2. I hvilke tilfeller adderer vi sannsynlighetene?
3. Hva er sannsynligheten for å få enten 2 eller 3 når man kaster en terning med 6 sider?
4. Hva er sannsynligheten for å få 8 hvis man kaster en terning med 20 sider?
5. Hva er sannsynligheten for å trekke en rød ball hvis vi har 7 røde baller og 3 grønne baller i en eske?

3.2.3 Del 3

1. Regn ut hvert uttrykk under, og vis utregningen:
   (a) $3!$
   (b) $3 \cdot 2!$
   (c) $4 \cdot 3!$
   (d) $4! \cdot 3! \cdot 2!$
2. Regn ut hvert uttrykk under, og vis utregningen:
   (a) $9 \mod 9$
   (b) $4 \mod 3$
   (c) $32 \mod 7$
   (d) $(32 \mod 7) \mod 3$
3. Regn ut hvert uttrykk under, og vis utregningen:
   (a) $9 // 8$
   (b) $8 // 9$
   (c) $32 // 7$
   (d) $(32 // 7) // 3$