INF1820: Introduksjon til språk-og kommunikasjonsteknologi

Første forelesning

Lilja Øvrelid

16 januar, 2017

Praktisk

Hvor og når

Tidspunkt

- **Forelesning:** Mandag 12:15-14, Seminarrom Caml
- **Grupper:** Onsdager
  - 10:15-12 Datastue Fortress
  - 12:15-14 Datastue Fortress
- Obligdeadline: fredag kl 23:57

Hvor og når

Tidspunkt

- **Forelesning:** Mandag 12:15-14, Seminarrom Caml
- **Grupper:** Onsdager
  - 10:15-12 Datastue Fortress
  - 12:15-14 Datastue Fortress
- Obligdeadline: fredag kl 23:57

Tidsregnskap:

- Arbeidsmengde: 37,5 / 3 = 12,5 timer
- Etter forelesning+gruppe: 9,5 timer

Obliger

- 3 obligatoriske oppgaver, hver delt i to deloppgaver (1a + 1b, 2a + 2b, 3a + 3b)
- Poengsystem:
  - 100 mulige poeng per deloppgave
  - For å gå opp til eksamen:
    - bestå alle tre obligatoriske oppgaver
    - akkumulere min 100 poeng for hver obligatoriske oppgave

Obliger

- 3 obligatoriske oppgaver, hver delt i to deloppgaver (1a + 1b, 2a + 2b, 3a + 3b)
- Poengsystem:
  - 100 mulige poeng per deloppgave
  - For å gå opp til eksamen:
    - bestå alle tre obligatoriske oppgaver
    - akkumulere min 100 poeng for hver obligatoriske oppgave

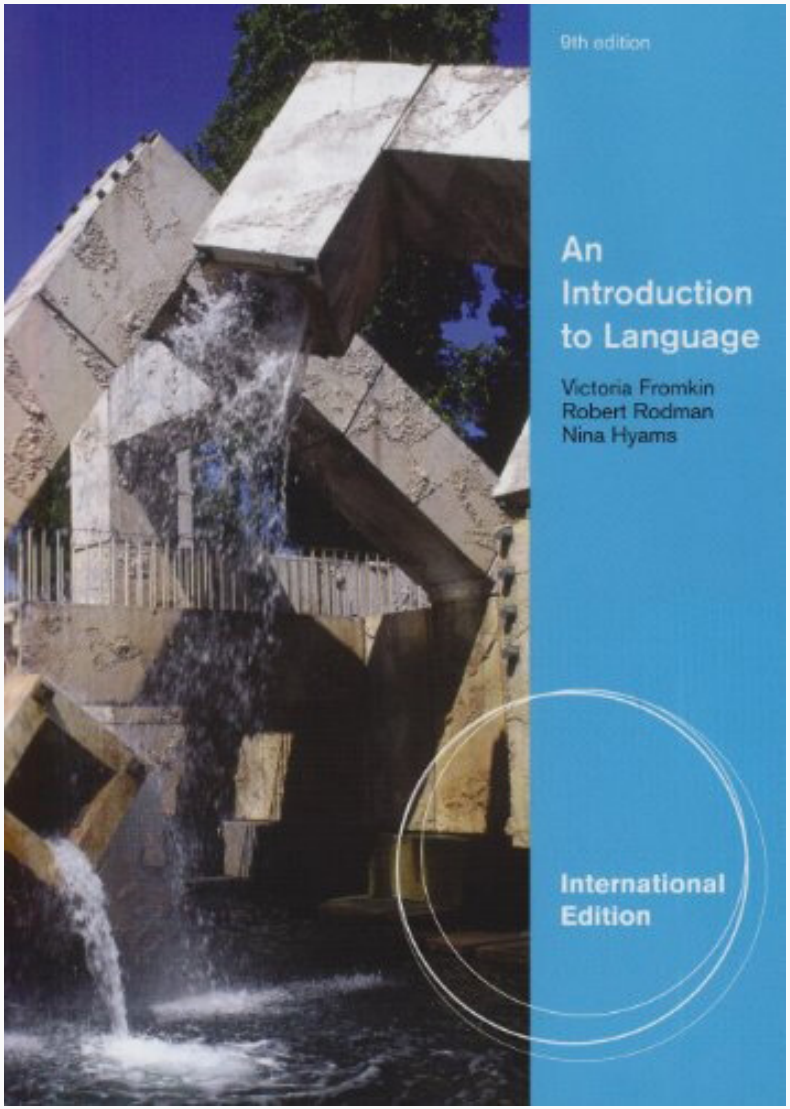
Eksempel

- Oblig 1a: 40 poeng (av 100 mulige)
- Oblig 1b: minimum 60 poeng

Obliger

- Absolutte frister
- Kopiering/plagiat
- Tidsrammer og planlegging
- Viktighet av gruppeundervisningen

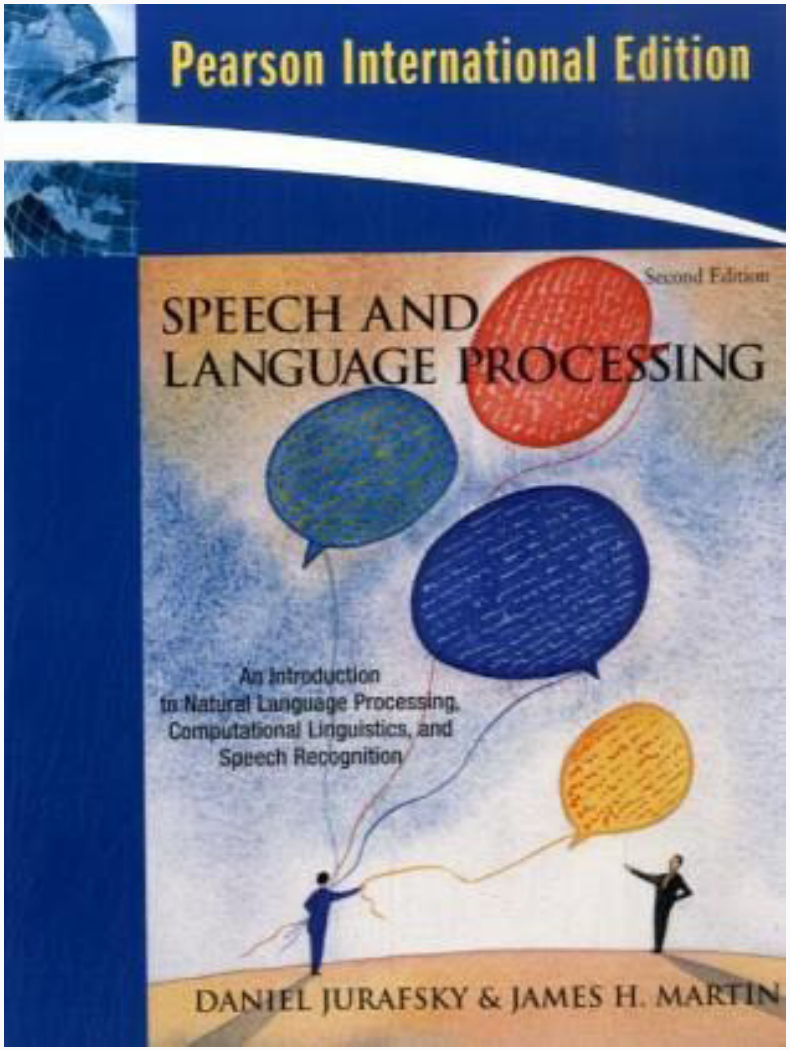
Pensumlitteratur



An Introduction to Language (8th Edition) by Victoria A. Fromkin, Robert Rodman, and Nina Hyams

- Utvalgte deler (ca 5 kapitler)

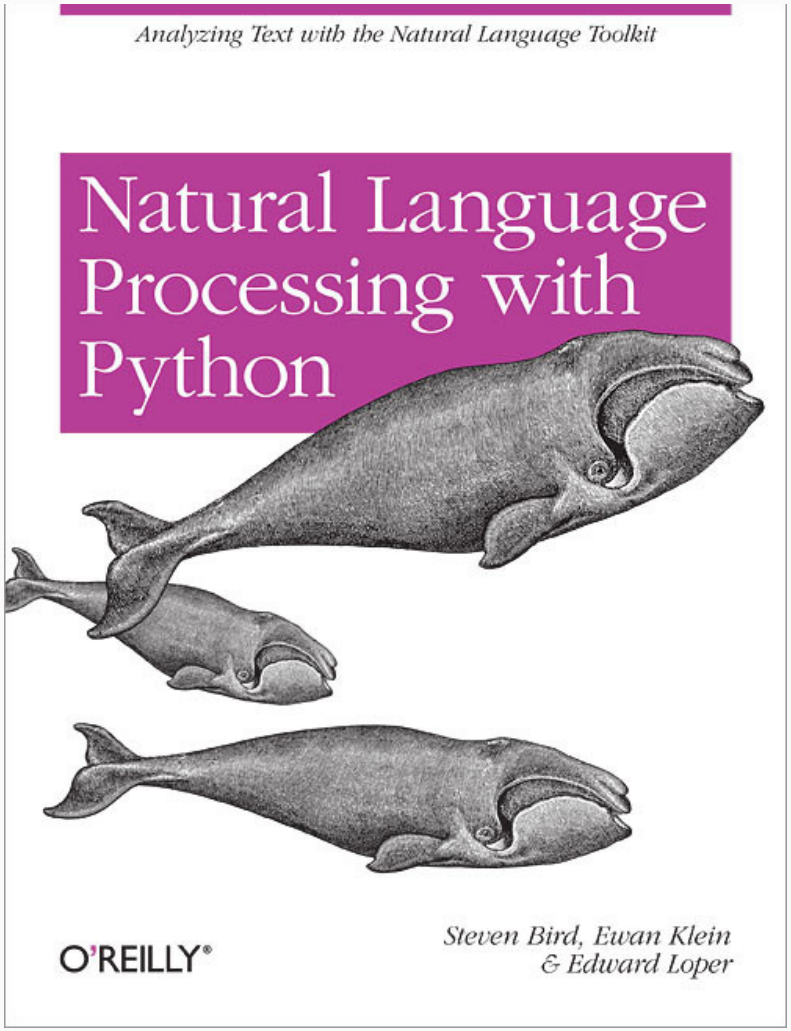
Pensumlitteratur



Speech and Language Processing: an Introduction to Natural Language Processing, Computational Linguistics, and Speech Recognition (2nd edition) by Daniel Jurafsky and James H. Martin

- Utvalgte deler

Pensumlitteratur



Natural Language Processing with Python by Steven Bird, Ewan Klein and Edward Loper (URL)

- Utvalgte deler
- Online versjon

Eksamen

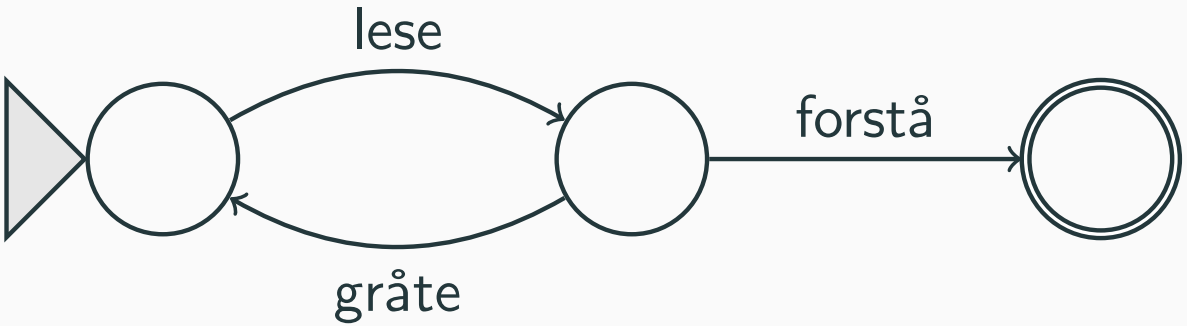
- Skriftlig (digital) eksamen på fire timer
  - 8 juni kl 14:30
- Pensumlitteratur + forelesningsnotater
- NB! Ikke en programmeringseksamen

Suksessoppskrift

- Emnesiden: timeplan, pensum, lesehenvisninger, beskjeder etc.
- Lesehenvisninger: forbered deg til forelesning
- Still spørsmål
- Gruppetimer:
  - forbered deg
  - delta aktivt
  - gjør oppgaver (også de ikke-obligatoriske)
- Benytt deg av medstudentene dine

Suksessoppskrift

- Emnesiden: timeplan, pensum, lesehenvisninger, beskjeder etc.
- Lesehenvisninger: forbered deg til forelesning
- Still spørsmål
- Gruppetimer:
  - forbered deg
  - delta aktivt
  - gjør oppgaver (også de ikke-obligatoriske)
- Benytt deg av medstudentene dine



Kursoversikt

Emnebeskrivelsen

Kurset gir en innføring i lingvistisk teori og relaterer denne til språkteknologiske problemområder, metoder og applikasjoner. Fokus er på å koble teori til praksis. Vi vil ta for oss morfologisk, syntaktisk, samt noe semantisk analyse av naturlige språk, formell språkteori og korpusbaserte metoder. Studentene vil få et første møte med noen datalingvistiske applikasjonsområder.

Tverrfaglighet

Kunnskap om språk

- oppbygning av menneskers språk
- analysere språklig materiale
- mulighetene for få datamaskiner til å forstå språk
- **flertydighet**

Tverrfaglighet

Kunnskap om teknologi

- kunne programmere
- algoritmer for språkteknologiske oppgaver
- lage små og mellomstore systemer for å løse språkteknologiske oppgaver

Introduksjon til lingvistikk

- Hva vil det si å beherske et språk?
- Hva vet vi om menneskelig språkprosessering?
- Hva mener vi med at språket er uendelig?
- Hva utgjør språkets byggeklosser?
- Hvordan settes disse sammen til meningsbærende enheter?

- Hvordan kan vi formalisere vår kunnskap om språk slik at den kan benyttes i automatiske systemer?
- Hvordan kan regulære uttrykk brukes til å beskrive språklige fenomener?
- Hvordan kan vi tildele ordklasser automatisk og hvordan evaluerer vi systemene våre?
- Hvordan kan vi automatisk gruppere ord til fraser?
- Hvordan kan vi automatisk skille mellom ulike betydninger av et ord?

Etter å ha tatt INF1820, kan du . . .

- . . . skrive enkle programmer for å manipulere store tekstmengder i Python

Etter å ha tatt INF1820, kan du . . .

- . . . skrive enkle programmer for å manipulere store tekstmengder i Python
- . . . trekke ut alle ord i en tekst (**oblig1**), dvs. utføre såkalt tokenisering

Etter å ha tatt INF1820, kan du . . .

- . . . skrive enkle programmer for å manipulere store tekstmengder i Python
- . . . trekke ut alle ord i en tekst (**oblig1**), dvs. utføre såkalt tokenisering
- . . . lage frekvenslister (**oblig2**)
  - Hva er "årets ord"?

Etter å ha tatt INF1820, kan du . . .

. . . beregne **sannsynligheten** for ord i en viss kontekst (**Oblig 2**)

Eksempel

ja takk, det vil jeg . . .

- *gjerne*?
- *hjerne*?

Etter å ha tatt INF1820, kan du . . .

- . . . automatisk merke opp (“tagge”) en tekst med ordklasser (**Oblig2/Oblig3**):

Eksempel

After the social browser launched two weeks earlier, talk about it exploded.

Etter å ha tatt INF1820, kan du . . .

- . . . automatisk merke opp (“tagge”) en tekst med ordklasser (**Oblig2/Oblig3**):

Eksempel

After the social browser launched two weeks earlier, talk about it exploded.

Etter å ha tatt INF1820, kan du . . .

- . . . automatisk merke opp ("tagge") en tekst med ordklasser (**Oblig2/Oblig3**):

Eksempel

After the social browser launched two weeks earlier, talk about it exploded.

```
 1   After
 2   the
 3   social
 4   browser
 5   launched
 6   two
 7   weeks
 8   earlier
 9   ,
10   talk
11   about
12   it
13   exploded
```

Etter å ha tatt INF1820, kan du . . .

- . . . automatisk merke opp (“tagge”) en tekst med ordklasser (**Oblig2/Oblig3**):

Eksempel

After the social browser launched two weeks earlier, talk about it exploded.

```
 1   After      after
 2   the        the
 3   social     social
 4   browser    browser
 5   launched   launch
 6   two        two
 7   weeks      week
 8   earlier    earlier
 9   ,          ,
10   talk       talk
11   about      about
12   it         it
13   exploded   explode
```

Etter å ha tatt INF1820, kan du . . .

- . . . automatisk merke opp (“tagge”) en tekst med ordklasser (**Oblig2/Oblig3**):

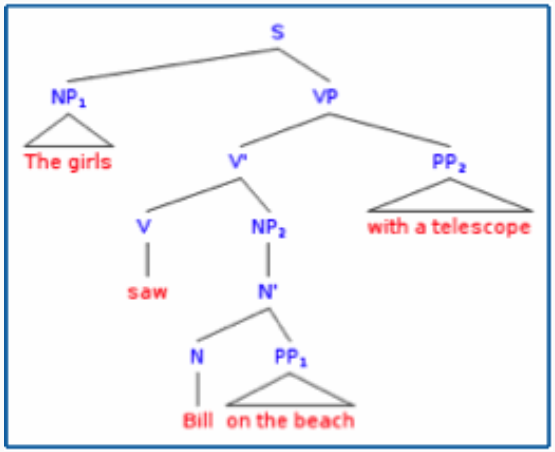
Eksempel

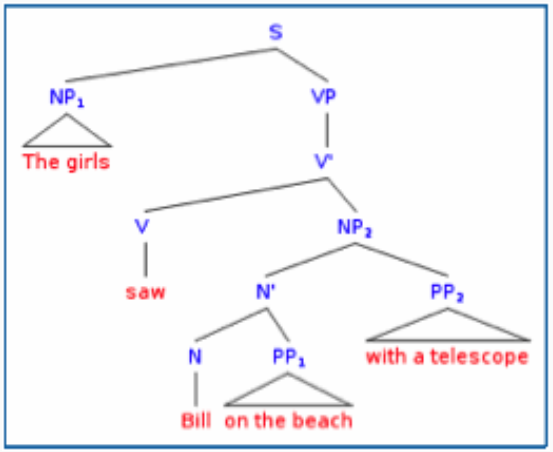
After the social browser launched two weeks earlier, talk about it exploded.

1	After	after	IN
2	the	the	DT
3	social	social	JJ
4	browser	browser	NN
5	launched	launch	VVD
6	two	two	JJ
7	weeks	week	NN
8	earlier	earlier	RBR
9	,	,	,
10	talk	talk	NN
11	about	about	IN
12	it	it	PP
13	exploded	explode	VVD

Etter å ha tatt INF1820, kan du . . .

- . . . forklare hva som gir opphav til flertydighet i språk og illustrere forskjeller, feks ved hjelp av syntaktiske trær (**oblig3**):
  The girls saw Bill on the beach with a telescope





Etter å ha tatt INF1820, kan du . . .

- . . . implementere (deler av) en enkel maskinlæringsalgoritme (Naive Bayes)
- . . . og bruke den til automatisk betydningsklassifisering
  - SKIM the pages for a clearer insight: Reading
  - She SKIMS through the novel which seems to fascinate them: Reading
  - Remove the vanilla pod, SKIM the jam, and let it cool: Removing
  - We SKIMMED across the surface of that sodding lake whilst all around us gathered the dark hosts of hell: Self_Motion

Hva er lingvistikk?

- Det vitenskapelige studiet av menneskelige språk
- Regler, systemer og prinsipper i språk
    - hva har ulike språk til felles? og hvordan varierer de?
    - hvordan fungerer språk?
    - hvordan forandrer språk seg over tid?
    - hvordan tilegner barn seg språk?
    - hvordan er språk representert i hjernen?

Lingvistikk

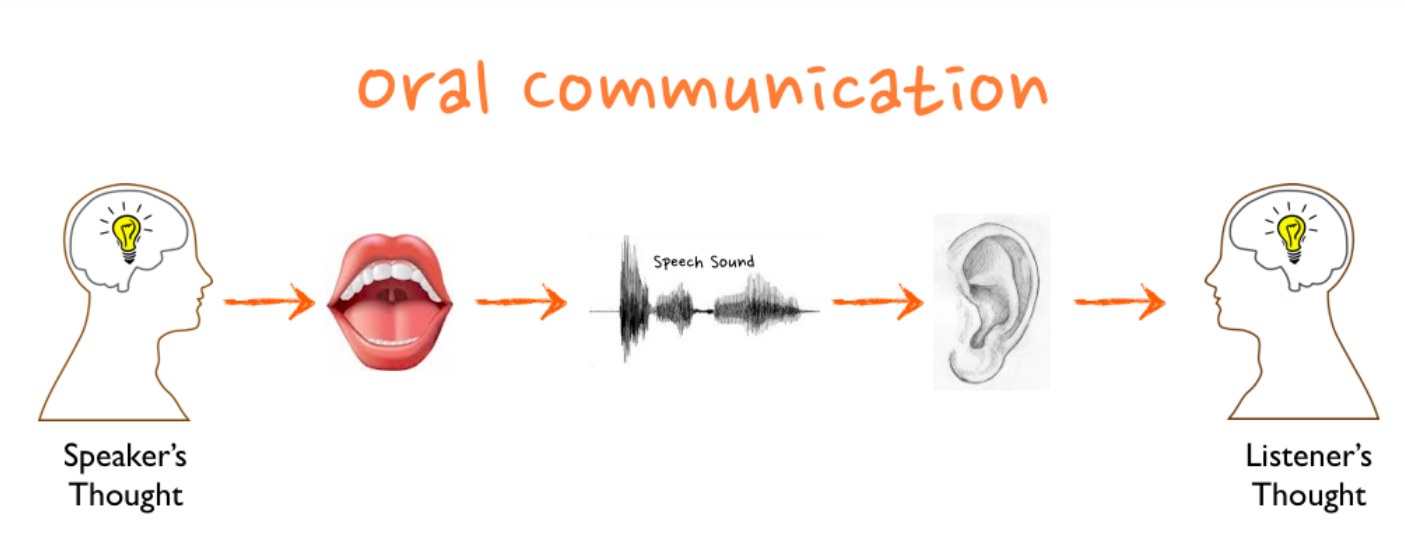
- Vitenskapelig studie av språk
- Menneskelig språk

Lingvistikk

- Vitenskapelig studie av språk
- Menneskelig språk
- Hva kjennetegner lingvistisk kunnskap?
  - Ubevisst (tacit knowledge)
  - Men det er mye kunnskap (know-how vs know-that)

Lingvistisk kunnskap

- Kunnskap om **lyd**:
  - lydsystemet for et språk
  - rekkefølgen på lyder



Lingvistisk kunnskap

- Kunnskap om **ord**:
  - Viss lydsekvens korresponderer til et visst konsept, eller **betydning**
  - **Vilkårlig** (arbitrær) kobling mellom form og betydning

- odun
- asa
- wartawan
- ciel

Lingvistisk kunnskap

- Kunnskap om **ord**:
  - Viss lydsekvens korresponderer til et visst konsept, eller **betydning**
  - **Vilkårlig** (arbitrær) kobling mellom form og betydning

- odun : "tre" – tyrkisk
- asa
- wartawan
- ciel

Lingvistisk kunnskap

- Kunnskap om **ord**:
  - Viss lydsekvens korresponderer til et visst konsept, eller **betydning**
  - **Vilkårlig** (arbitrær) kobling mellom form og betydning

- odun : "tre" – tyrkisk
- asa : "morgen" – japansk
- wartawan
- ciel

Lingvistisk kunnskap

- Kunnskap om **ord**:
  - Viss lydsekvens korresponderer til et visst konsept, eller **betydning**
  - **Vilkårlig** (arbitrær) kobling mellom form og betydning

- odun : "tre" – tyrkisk
- asa : "morgen" – japansk
- wartawan : "reporter" – indonesisk
- ciel

Lingvistisk kunnskap

- Kunnskap om **ord**:
  - Viss lydsekvens korresponderer til et visst konsept, eller **betydning**
  - **Vilkårlig** (arbitrær) kobling mellom form og betydning

- odun : "tre" – tyrkisk
- asa : "morgen" – japansk
- wartawan : "reporter" – indonesisk
- ciel : "himmel" – fransk

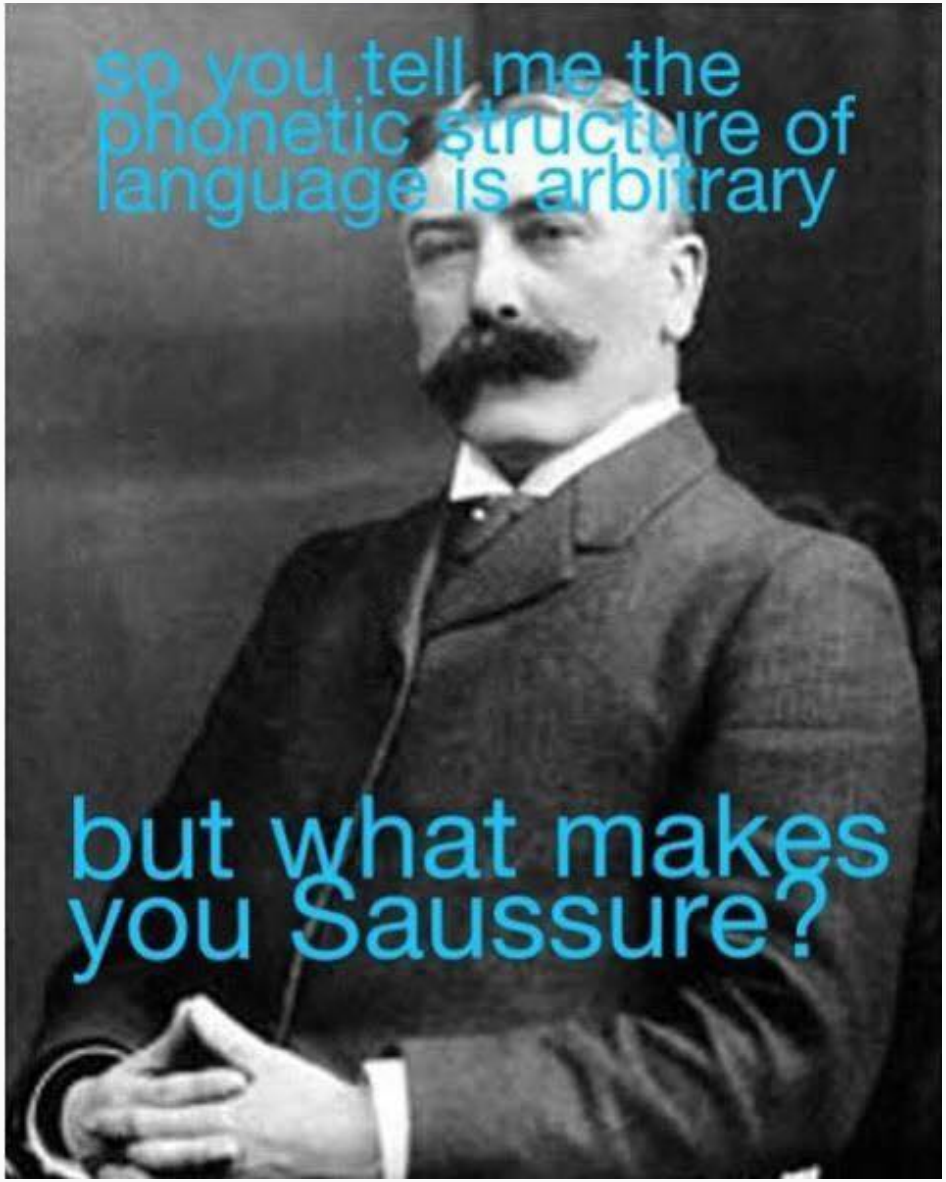
Lingvistisk kunnskap

- Kunnskap om **ord**:
  - Viss lydsekvens korresponderer til et visst konsept, eller **betydning**
  - **Vilkårlig** (arbitrær) kobling mellom form og betydning

  - odun : "tre" – tyrkisk
  - asa : "morgen" – japansk
  - wartawan : "reporter" – indonesisk
  - ciel : "himmel" – fransk

  - **Konvensjonell** sammenheng: må læres

Lingvistisk kunnskap



Lingvistisk kunnskap

- Kunnskap om hvordan ord settes sammen til setninger
- Mengden av mulige setninger er uendelig

Lingvistisk kunnskap

- Kunnskap om hvordan ord settes sammen til setninger
- Mengden av mulige setninger er uendelig

- Dette er en setning

Lingvistisk kunnskap

- Kunnskap om hvordan ord settes sammen til setninger
- Mengden av mulige setninger er uendelig

- Dette er en setning
- Dette er en setning som jeg skriver akkurat nå

Lingvistisk kunnskap

- Kunnskap om hvordan ord settes sammen til setninger
- Mengden av mulige setninger er uendelig

- Dette er en setning
- Dette er en setning som jeg skriver akkurat nå
- Dette er en setning som jeg tror at jeg skriver akkurat nå
- Dette er en setning som Fredrik mener at jeg tror at jeg skriver akkurat nå
- osv.

Lingvistisk kunnskap

- Kunnskap om hvordan ord settes sammen til setninger
- Mengden av mulige setninger er uendelig

- Dette er en setning
- Dette er en setning som jeg skriver akkurat nå
- Dette er en setning som jeg tror at jeg skriver akkurat nå
- Dette er en setning som Fredrik mener at jeg tror at jeg skriver akkurat nå
- osv.

- Dette er en kjedelig setning
- Dette er en kjedelig kjedelig setning
- Dette er en kjedelig kjedelig kjedelig setning
- osv.

Lingvistisk kunnskap

- Kunnskap om hvordan ord settes sammen til setninger
- Mengden av mulige setninger er uendelig

- Dette er en setning
- Dette er en setning som jeg skriver akkurat nå
- Dette er en setning som jeg tror at jeg skriver akkurat nå
- Dette er en setning som Fredrik mener at jeg tror at jeg skriver akkurat nå
- osv.

- Dette er en kjedelig setning
- Dette er en kjedelig kjedelig setning
- Dette er en kjedelig kjedelig kjedelig setning
- osv.

- Evne til å forstå og skape nye setninger, språkbruk er **kreativ**

Lingvistisk kunnskap

- Grammatikalitet

  - Norske sykehus bruker for mye antibiotika
  - *Sykehus norske bruker for mye antibiotika
  - *Norske sykehus for mye antibiotika bruker
  - *Norsk sykehus bruker for mye antibiotika

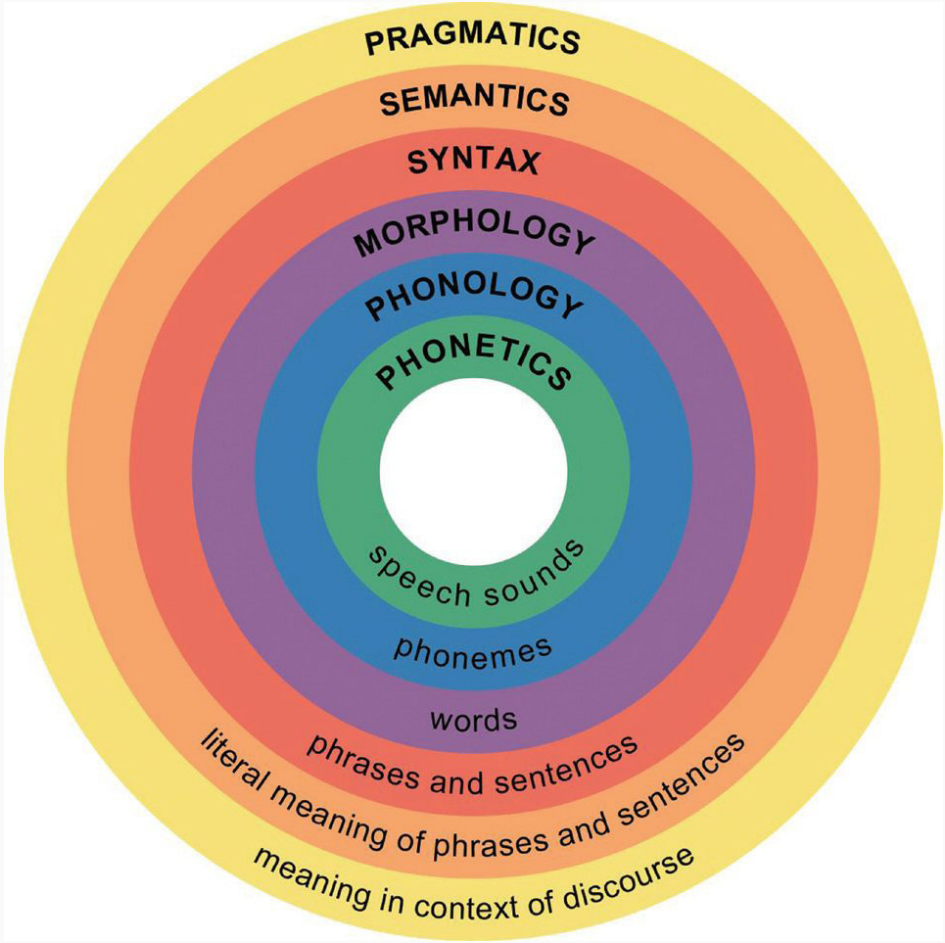
Lingvistisk kunnskap

- Grammatikalitet

  - Norske sykehus bruker for mye antibiotika
  - *Sykehus norske bruker for mye antibiotika
  - *Norske sykehus for mye antibiotika bruker
  - *Norsk sykehus bruker for mye antibiotika

- Kunnskap om *regler* for hvordan man danner setninger i et språk
  - en endelig mengde regler, med et endelig vokabular $\Rightarrow$ en uendelig mengde setninger
- Grammatikalitetsbedømninger

Lingvistiske nivåer



Fonetikk/fonologi: lyder $\Rightarrow$ ord

Morfologi: morfemer $\Rightarrow$ ord

Syntaks: ord $\Rightarrow$ fraser, fraser $\Rightarrow$ setninger

Semantikk: ord $\Rightarrow$ betydning, setninger $\Rightarrow$ betydning

Hva vi vet om språk

- Der vi finner mennesker, finner vi språk

Hva vi vet om språk

- Der vi finner mennesker, finner vi språk
- Det finnes ingen "primitive" språk

Hva vi vet om språk

- Der vi finner mennesker, finner vi språk
- Det finnes ingen "primitive" språk
- Alle språk forandrer seg over tid

Hva vi vet om språk

- Der vi finner mennesker, finner vi språk
- Det finnes ingen "primitive" språk
- Alle språk forandrer seg over tid
- Forholdet mellom lyd og betydning er (stort sett) vilkårlig

Hva vi vet om språk

- Der vi finner mennesker, finner vi språk
- Det finnes ingen "primitive" språk
- Alle språk forandrer seg over tid
- Forholdet mellom lyd og betydning er (stort sett) vilkårlig
- Alle menneskelige språk bruker endelig (finitt) mengde lyder og ord til å danne uendelig mengde mulige setninger

Hva vi vet om språk

- Der vi finner mennesker, finner vi språk
- Det finnes ingen "primitive" språk
- Alle språk forandrer seg over tid
- Forholdet mellom lyd og betydning er (stort sett) vilkårlig
- Alle menneskelige språk bruker endelig (finitt) mengde lyder og ord til å danne uendelig mengde mulige setninger
- Alle språk kan uttrykke negasjon, spørsmål, gi kommandoer, snakke om fortid/framtid, hypotetiske situasjoner

Hva vi vet om språk

- Der vi finner mennesker, finner vi språk
- Det finnes ingen "primitive" språk
- Alle språk forandrer seg over tid
- Forholdet mellom lyd og betydning er (stort sett) vilkårlig
- Alle menneskelige språk bruker endelig (finitt) mengde lyder og ord til å danne uendelig mengde mulige setninger
- Alle språk kan uttrykke negasjon, spørsmål, gi kommandoer, snakke om fortid/framtid, hypotetiske situasjoner
- Ethvert normalt barn er i stand til å lære morsmålet sitt

- Lingvistiske disipliner