INF1820: Introduksjon til språk-og kommunikasjonsteknologi

Tredje forelesning

Lilja Øvrelid

30 januar, 2017

Språklige data

Vitenskapsteori

Empiricism *is a theory of knowledge which asserts that knowledge comes via the senses' experience*

Rationalism *is any view appealing to reason as a source of knowledge or justification*

(Eng Wikipedia)

Vitenskapsteori

Empiricism *is a theory of knowledge which asserts that knowledge comes via the senses' experience*

Rationalism *is any view appealing to reason as a source of knowledge or justification*

(Eng Wikipedia)

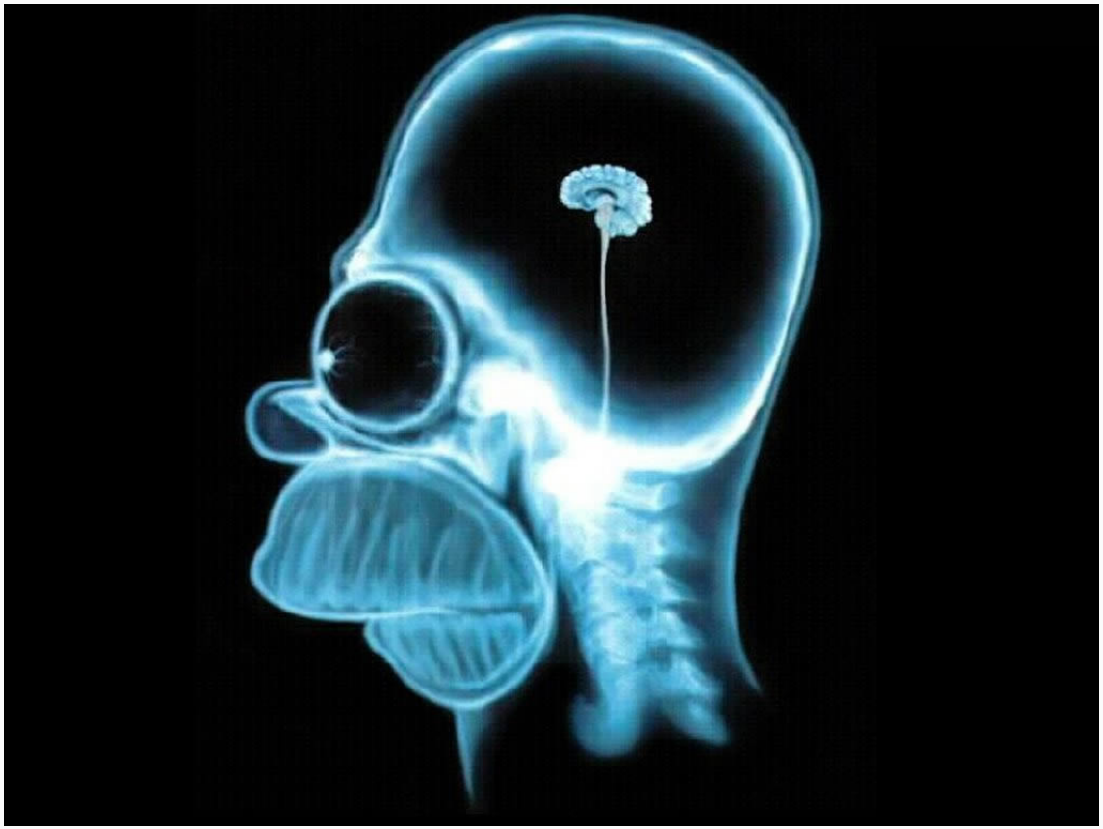
- Paradigmeskifter lingvistikk og datalingvistikk
- Rasjonalistene: teori-drevet, symbolske metoder
- Empiristene: data-drevet, statistiske metoder
  "big data"-disiplin

Språklige data

- Modellere språklig kunnskap
- Trenger språklige data
  - Introspeksjon
  - Faktisk språkbruk – korpusdata
- Språkteknologi: programmer som generaliserer over språklige mønstre
  - **Korpus**data helt sentralt
- Menneskelig språkprossessering: hvordan modelleres språk i hjernen?

Språk og hjerne

Språk og hjerne

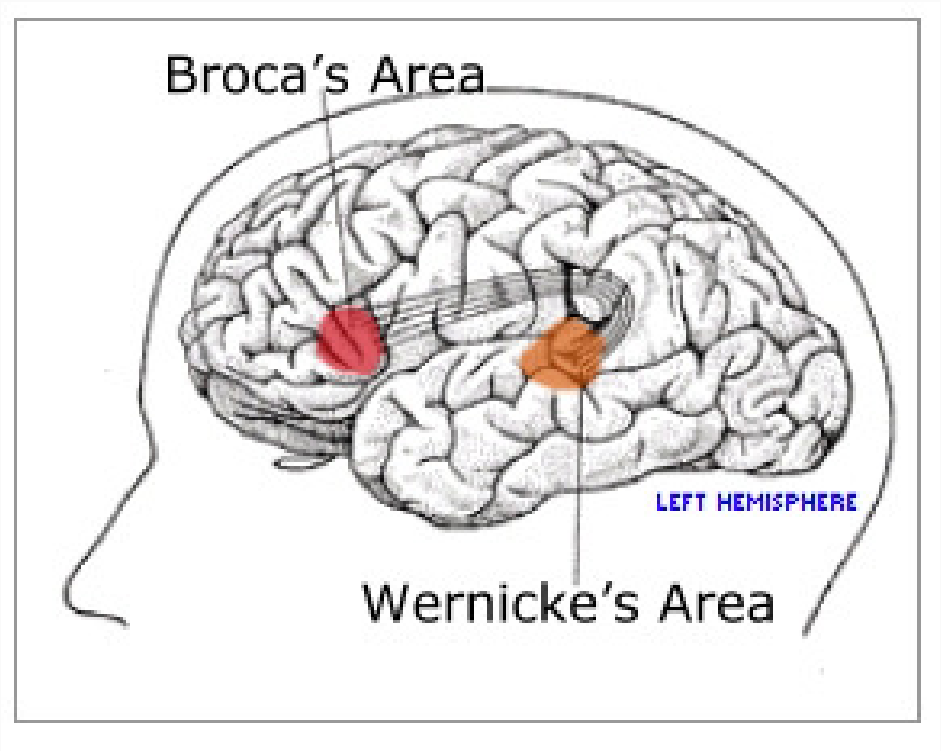


Språk og hjerne

- **Neurolingvistikk** – lingvistisk fagområde som studerer de mekanismer i den menneskelige hjerne som kontrollerer språk (-forståelse, -produksjon og - tilegnelse)
- Hjernen:
  - Storehjernen er organisert i "lapper": pannelappen, tinninglappen, isselappen, bakhodelappen.
  - Storehjernen har to halvdeler som arbeider sammen gjennom corpus collosum (over 200 millioner aktive nerveceller)

Hvor er språk lokalisert?

- Data fra atypisk språk
- Afasi
  - språkvansker etter hjerneskade
  - forskjellige typer avhengig av hvor skaden har oppstått



Hvor er språk lokalisert?

Brocas afasi

Ugrammatisk språk, problemer med forståelse av syntaktisk komplekse konstruksjoner

- *Yes... ah... Monday... er... Dad and Peter H... (his own name), and Dad.... er... hospital... and ah... Wednesday... Wednesday, nine o'clock... and oh... Thursday... ten o'clock, ah doctors... two... an' doctors... and er... teeth... yah*

Hvor er språk lokalisert?

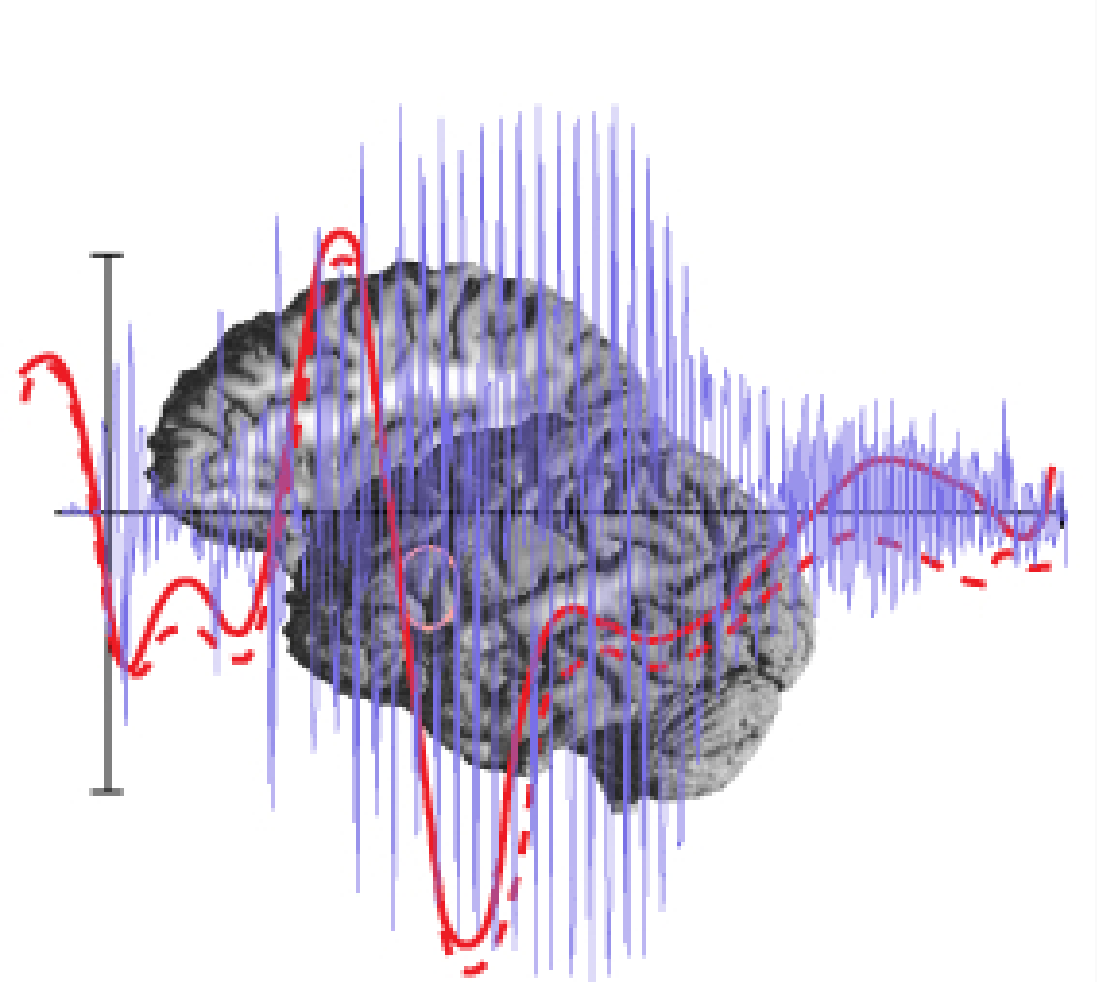
Wernickes afasi

Semantisk usammenhengende, men stort sett syntaktisk korrekt

- *I felt worse because I can no longer keep in mind from the mind of the minds to keep me from mind and up to the ear which can be to find among ourselves.*

Måling av aktivitet i hjernen

- Moderne teknologi (MRI, CT, ERP) kan gi et enda mer nøyaktig bilde
- Forandringer i hjerneaktivitet



Neurolingvistiske data:

- Neurolingvistiske data:
  - hvor språket sitter
  - empirisk støtte for teoretiske hypoteser
- Informasjon om organisering av språket: modulært organisert – separate moduler
- Ord i nettverk – semantisk, uttale
  - *table - chair*
  - *tool - pool*
- Funksjonsord adskilt fra innholdsord (Broca's afasi)

Korpusdata

Språklige data: korpusdata

- Et korpus (tekstkorpus) er en strukturert samling tekster
- Elektronisk lagret
- Kan brukes til:
  - Empiriske data for lingvistiske studier (motsetning til introspeksjon)
  - Treningsmateriale for datalingvistiske modeller av språklige fenomener

Språklige data: korpusdata

- Korpus laget for å representere et visst språk eller språklig variant
- Språklige data – to muligheter:
  1. Arkivere alle setninger i et språk: UMULIG
  2. Plukke ut et mindre utvalg ("sample") av språket: MULIG
- 2 er mulig men ikke trivielt
- Et korpus må konstrueres slik at det minimerer fordommer ("bias") og maksimerer representativitet

Et tenkt korpus for norsk

- Vi må inkludere forskjellige typer tekster:
  - Skrift og tale? [registere]
  - Fra forskjellige deler av landet? Et utvalg av dialekter? [regionale dialekter]
  - Kun fra 2000-tallet? Hva med 1990? Eller 1950? [tidsperioder]
  - Språk produsert av både menn og kvinner? Alle aldersgrupper, inkludert barn? Hva med utdanningsnivå? Sosial status? [demografi]
  - Skal vi inkludere nyhetsstoff? Hva med kronikker, romaner og e-post? Tegneserier og tekstmeldinger? [sjanger]

Et tenkt korpus for norsk

- Svarer 'JA' på alle spørsmålene
- ... men vi må også ta hensyn til fordeling, feks 50% tekst og 50% tale? 20% nyhetsstoff og 15% romaner? etc.
- Representativt korpus
  1. ekstra-lingvistiske faktorer (feks demografi)
     - 15% av innbyggerene bor i Oslo, så vi kan fordele korpuset etter dette
     - 20% av befolkningen er over 60, la oss inkludere 20% av 60+ tekster
  2. lingvistisk fordeling
     - 20% av alle norske tekster er romaner – 20% av korpuset
     - 20-åringer leser og skriver dobbelt så mye som 60+, dobbelt så mange tekster produsert av/for 20-åringer

Et tenkt korpus for norsk

- For (2): Ikke gitt at vi har denne typen informasjon om befolkningen
- For (1): Ikke tilfredsstillende heller, demografi er ikke en nøyaktig indikator for språkbruk
- Umulig å oppnå et objektivt representativt korpus. Men vi prøver allikevel ...

Eksisterende korpuser

- NLTK kommer med flere innebygde korpuser

```
>>> import nltk
>>> from nltk.book import *
*** Introductory Examples for the NLTK Book ***
Loading text1, ..., text9 and sent1, ..., sent9
Type the name of the text or sentence to view it.
Type: 'texts()' or 'sents()' to list the materials.
text1: Moby Dick by Herman Melville 1851
text2: Sense and Sensibility by Jane Austen 1811
text3: The Book of Genesis
text4: Inaugural Address Corpus
text5: Chat Corpus
text6: Monty Python and the Holy Grail
text7: Wall Street Journal
text8: Personals Corpus
text9: The Man Who Was Thursday by G . K . Chesterton 1908
```

Eksisterende korpuser

- Tilgjengelig for nedlasting: Project Gutenberg
- (Forsøk på) representative korpuser
  - Brown 1M ord
  - British National Corpus (BNC), 100M ord (register, domene, forskjellige tidsperioder, sjanger, demografi osv)
  - American National Corpus, under bygging
- Store korpuser:
  - Gigaword ($\sim$1.7 milliarder ord, nyhetstekster)
  - Common crawl (flere petabytes)

Eksisterende korpuser

- Korpuser for andre språk enn engelsk
  - Arabisk Gigaword
  - Chinese news
  - Norsk Aviskorpus
    - norske nyheter 1998-2014
    - ca. 1.5 millarder ord
  - NoWaC ("Norwegian Web as Corpus")
    - web-dokumenter fra .no-domener
    - ca 700 millioner ord
  - NoTa-korpuset
    - transkripsjoner av samtaler og intervju fra informanter født og oppvokst i Oslo-området
    - transkribert tekst og tale

Eksisterende korpuser

- Parallelle korpuser
  - EUROPARL, OPUS

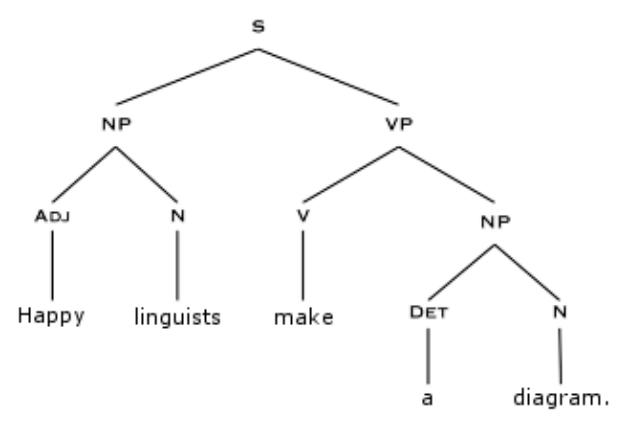
```
<?xml version="1.0"?>
<!DOCTYPE plug SYSTEM "dtd/plugXML.dtd">
<PLUG>
...
<align id="ensvfbell2" link="1-1">
  <seg lang="en">
    <s id="en2.1">Then hand luggage is opened.</s>
  </seg>
  <seg lang="sv">
    <s id="sv2.1">Sedan öppnas handbagaget.</s>
  </seg>
</align>
...
<align id="ensvfbell867" link="2-1">
  <seg lang="en">
     <s id="en867.1">You lean back with a cup of coffee to luxuriate
        in the Oriental conversation of an intelligent man.</s>
     <s id="en867.2">Immediately you are involved in a
        tormenting discussion.</s>
  </seg>
  <seg lang="sv">
     <s id="sv867.1">Man lutar sig bakåt med en kopp kaffe för att
        avnjuta en orientalisk människas intelligenta konversation,
        och omedelbart är man indragen i en plågsam diskussion.</s>
  </seg>
</align>
...
```

Annotering

- Korpuser inneholder forskjellige typer informasjon og har gjennomgått forskjellige former for (automatisk/manuell) **annotering**
- Delt opp i enheter som tilsvarer et ord, sk tokens: ord, tall, tegnsetting $\rightarrow$ tokenisering
- Stemming eller lemmatisering: reduksjon til baseform

Annotering

- Korpuser med manuell annotering
  - Mennesker merker opp lingvistisk informasjon
- Ordklasse (feks Brown)
  - The/at Fulton/np County/np Grand/jj Jury/nn said/vbd Friday/nr an/at investigation/nn ...
- Syntaks (trebanker, feks Penn Treebank)



(S (NP (ADJ Happy) (N linguists)) (VP (V make) (NP (Det a) (N diagram))))

- Ordsemantikk, diskursrelasjoner etc.

Manuelt annotert korpus for maskinlæring

Et manuelt annotert korpus

Ordbetydning

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

Trene en klassifiserer:

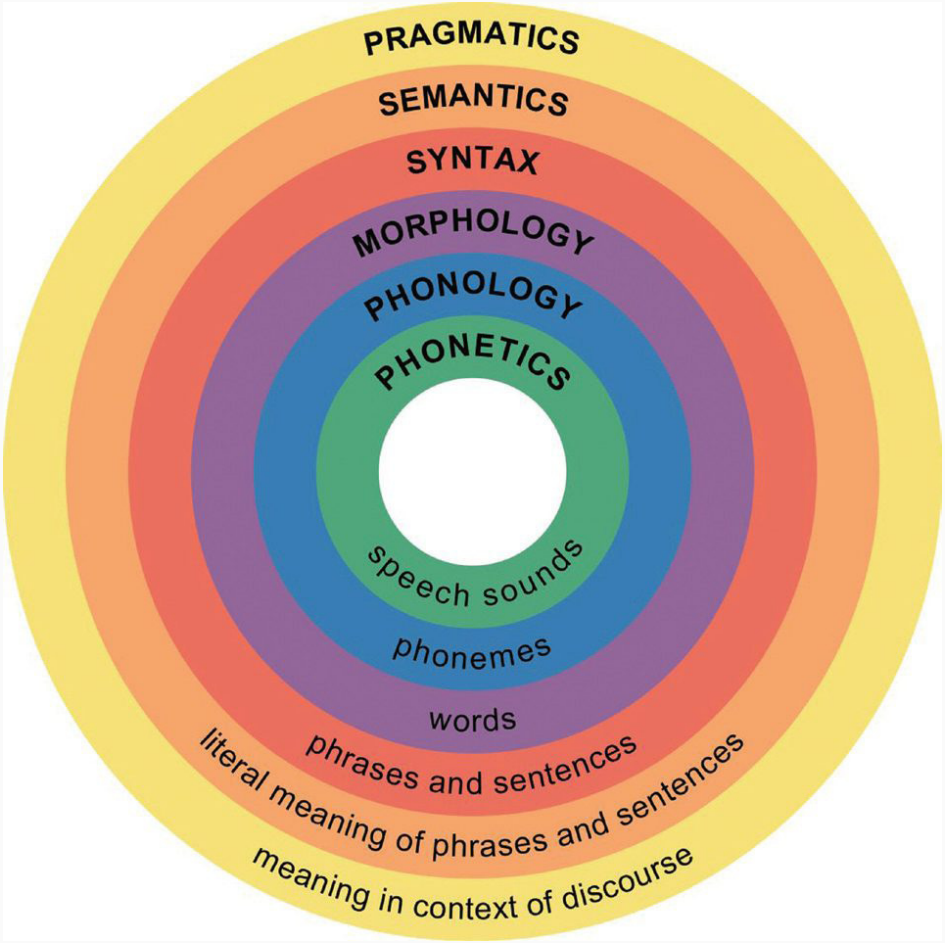
- Tren på Reading, Removing og Self_motion instanser
- Appliser på ny instans: hvilken klasse ligner den mest på?
- *A red grouse SKIMMED low over the heather: ???*

Oppsummering: språklige data

- Menneskelig språkprosessering
  - afasi-studier
  - måling av hjerneaktivitet
- Korpusdata
  - representativitet
  - størrelse
  - annotering
- Utstrakt bruk i språkteknologiske modeller

Morfologi

Morfologi


PRAGMATICS
SEMANTICS
SYNTAX
MORPHOLOGY
PHONOLOGY
PHONETICS
speech sounds
phonemes
words
phrases and sentences
literal meaning of phrases and sentences
meaning in context of discourse

Morfologi

- Hvordan ord er bygd opp
- Hvordan ord bøyes
- Hvordan ord dannes
- Hvordan ord deles i ordklasser

Ordet

- Relativ grei betydning i dagligtale
- I språkteknologi kan det derimot brukes på flere forskjellige måter

Kari og Ola gikk på tur i skogen , for de liker å gå på tur . Det var for øvrig en stor skog .

- 25 ord (**tokens**)
- men også 21 ord (**typer**)
- eller 17 ord (**leksem**)

Tokenisering

- Dele opp en tekst i løpende ord
- Første skritt i nesten alle språkteknologiske oppgaver
- Definisjon:
  a string of contiguous alphanumeric characters with space on either side; may include hyphens and apostrophes, but no other punctutation marks
  (Kucera & Francis, 1967)

Tokenisering: problemer

- Punktum
  - del av forkortelser: *f.eks.*
  - både forkortelse og setningsslutt (*Kjøper gamle møbler, bøker, klær, etc.*)
- Apostrof
  - *'the children'* vs. *the children's toys*
  - *I'll, isn't, don't*

Tokenisering: problemer

- Bindestrek
  - Ett eller flere ord?
  - *Oslo-borgeren*
  - *skrive- og leseopplæring*
- Mellomrom
  - Egennavn: *New York*
  - Faste fraser: *i fjor, blant annet*
  - Tall: *100 000*
- Annet:
  - *10,26* og *10:26*
  - URL'er

Ordet

- Kunnskap om ord viktig del av det å beherske et språk
- Kobling mellom en lydsekvens og en spesifikk betydning
- Vilkårlig kobling:
  - samme lyd - forskjellig betydning (*to*, *two*)
  - forskjellig lyd - samme betydning (*sofa*, *couch*)
- Lagret i mentalt leksikon:
  - fonologisk representasjon
  - betydning
  - (ortografi)
  - ordklasse etc.

Ordet

- Viktig skille i språk:
  - **Innholdsord:** substantiver, verb og adjektiv
    - Betegner konsepter som objekter, handlinger, egenskaper og ideer
    - *barn, skrive, spennende, anarkisme*
    - Åpen klasse: stadig nye ord, feks *hverdagsintegrering, ståhjuling*
  - **Funksjonsord:** konjunksjoner, preposisjoner, artikler og pronomen
    - Betegner grammatiske relasjoner, lite eller ingen semantisk innhold
    - *the, a* – bestemthet, *of* – eierskap
    - Lukket klasse: ikke ofte nye tilskudd, (*hen?*)

Ordet

- Skillet mellom innholdsord og funksjonsord
  - Neurolingvistikk: afasikere
  - "slips-of-tongue" hos normale språkbrukere
    - "turn to tend out" i stedet for "tend to turn out"
  - Barnespråk: overgeneralisering
    - . . . og det var mange manner der

Ordet

- Og nå: GJETTEKONKURRANSE

Ordet

- Hvilken skal ut?
  - gulest
  - gul
  - gulere
  - rød

Ordet

- Hvilken skal ut?
  - penger
  - grammatikk
  - rød
  - ere

Ordet

- Hvilken skal ut?
  - ing
  - het
  - else
  - an

Morfemet

Morfemet

- Ord har intern struktur som er regelstyrt
  - U-mulig, u-rolig, u-intelligent
  - hva betyr u-?
  - *mulig-u, *rolig-u
- Ord kan bestå av flere meningsbærende enheter
- **Morfemet** – elementær enhet (gr. 'morphe' – form)
- Morf+ologi – vitenskapen om (ord)former

Morfemet

Et ord kan bestå av ett eller flere morfemer:

- ett morfem: *boy*, *desire*, *morph*
- to morfemer: *boy+ish*, *desire+able*, *morph+ology*
- tre morfemer: *boy+ish+ness*, *desire+able+ity*
- fire morfemer: *gentle+man+li+ness*, *un+desire+able+ity*
- mer enn fire morfemer: *un+gentle+man+li+ness*, *anti+dis+establish+ment+ari+an+ism*

Morfemet

- Morfemet er den elementære (minste) lingvistiske enheten
- Kan ikke analyseres videre
- Språk består i hovedsak av diskrete enheter som kan kombineres (kreativitet)
    - et bloggbart tema

Morfemet

- Vår morfologiske kunnskap har to hovedkomponenter
  - **Frie** morfemer: ord. *boy, desire, gentle, man*
  - **Bundne** morfemer: affikser.
    - prefikser: *un-, pre-, bi-*
    - suffikser: *-ing, -ish, -ness*
- Språk benytter affikser i varierende grad:
  - engelsk: dance (v.), dance (n.)
  - tyrkisk: suffiks *-ak*, *dur* 'to stop', *durak* 'stopping place'
- Noen språk har **infikser**
  - Bontov (Filippinene): *fikas* 'sterk', *fumikas* 'å være sterk'
  - *un-fuckin-believable*
- Noen språk har **sirkumfikser**
  - Tysk: *ge+lieb+t* 'har elsket'

Morfemet

- Morfologisk komplekse ord består av :
  - **Rot** + en eller flere affikser (*hus+lig*)
  - En rot er et ordelement som ikke kan deles opp i mindre (meningsbærende) deler
  - Ofte, men ikke nødvendigvis et ord som kan stå alene:
    - **ceive* som i *conceive*
    - **ling* som i *linguist*

Orddannelse

Orddannelse

- Kunnskap om morfologi innebærer kunnskap om regler for orddannelse
- Kombinerer morfemer til komplekse ord (*kjærlig-het*, *(jern+bane)+(arbeid+er)*)
- Adj + -het $\rightarrow$ Substantiv
- Verb + -er $\rightarrow$ Substantiv (en som gjør Verb)

Avledning

- En avledning er et ord som er dannet fra et annet ord ved hjelp av et avledningsaffiks (prefiks eller suffiks),
- Avledningsbasen kan være et rotord (*barn*) eller en avledning (*barnslig*)
- Avledningsaffiksene er bundne morfemer med klart semantisk innhold (som innholdsord, men er ikke ord)

Avledningsaffikser

- *u–* negasjon: *umulig, uvel, urolig*
- *for-* - foran: *forelese, forbokstav, formann*
- *-er* - den som utfører handlingen: *fisker, baker*

Avledning

- Avledningsaffikser bidrar med betydning
- Når et suffiks blir lagt til endres som regel ordklassen
- Det er siste del av ordet som bestemmer ordklasse, derfor endrer ikke prefikser ordklassen (*villig - uvillig, arbeide - bearbeide*)

Suffikser

- *-er*: Verb $\rightarrow$ Substantiv, f.eks. *fisker, baker*
- *-ing*: Verb $\rightarrow$ Substantiv, f.eks. *bading, baking, banning*
- *-lig*: Substantiv $\rightarrow$ Adjektiv, f.eks. *alvorlig, hyggelig, latterlig, vanlig*
- *-n*: Adjektiv $\rightarrow$ Verb, f.eks. *gulne, lysne, stivne*

Bøyning

markerer kategorier som tempus, numerus, kasus, etc.

Bøyningskategorier i norsk

- **Genus** (kjønn): alle substantiver har fast genus og ord som står til substantivet samsvarsbøyes (*en snill katt, et snilt beltedyr*)
- **Tall**: entall og flertall *bil-biler*
- **Bestemthet**: uttrykkes i hovedsak ved suffiks (*bilen, huset*) eller (jf. engelsk bestemt artikkel *the*)
- **Kasus**: uttrykker den funksjonen en frase har som setningsledd. To kasus i norsk: nominativ og akkusativ. I hovedsak på pronomen *hun-henne*
- ...

Bøyning

I norsk har vi følgende bøyningskategorier (forts.):

- **Grad**: tre grader uttrykkes ved bøyning, positiv, komparativ, superlativ (*fin-finere-finest*)
- **Tempus**: angir tidspunktet for handlingen eller tilstanden som setningen beskriver. I norsk uttrykkes to tempus ved bøyning: presens (nåtid) og preteritum (fortid) *spiser-spiste*
- **Diatese**: den semantiske relasjonen subjektet har til verbet i setningen, kan dannes med endelsen -s *Vi hører musikken helt hit - Musikken høres helt hit*

Bøyning

- Sammenlignet med en rekke andre språk har norsk og engelsk forholdsvis lite bøyningsmorfologi
- "morfologi konkurrerer med syntaks"

Engelsk vs russisk:

- *Victor defends Maxim* $\neq$ *Maxim defends Victor*
- *Maksim zašiščajet Viktora* $=$
- *Maksim Viktora zašiščajet* $=$
- *Viktora zašiščajet Maksim*

Bøyning vs. avledning

- Forskjeller på bøyning og avledning:
  1. Ved bøyning skifter ordet aldri ordklasse, ved avledning skifter ordet som oftest ordklasse
     - barn - barnet
     - barn - barnslig
  2. Alle prefikser er avledningsaffikser, suffikser derimot kan brukes både til bøyning og avledning
  3. Bøyning er mer produktiv

Bøyning vs. avledning

- Forskjeller på bøyning og avledning (forts.):
  4. Bøyningssuffikser i norsk har alltid svakt trykk (*bilen*, *spiste*), mens avledningssuffikser kan ha sterkt trykk (*sentral*) eller bitrykk *tenkbar*
  5. Bøyningsendelser ligger alltid i slutten av ordet, men avledningsendelsene kommer tidligere (når vi har begge deler) *galskapen*

Sammensetninger

- En tredje form for orddannelse, svært vanlig i germanske språk, her: norsk
- Ord som består av deler som hver for seg også er egne ord
- To ledd:

Forledd	**Etterledd**
hus-	tak
etter-	prøve
fram-	på

- Etterleddet bestemmer vanligvis ordklasse (*fredlyse*, *rustfri*)

Sammensetninger

- De fleste sammensetninger er **determinative**: etterleddet gir hovedbetydning, mens forleddet avgrenser. *bilhjul, hjulbåt*

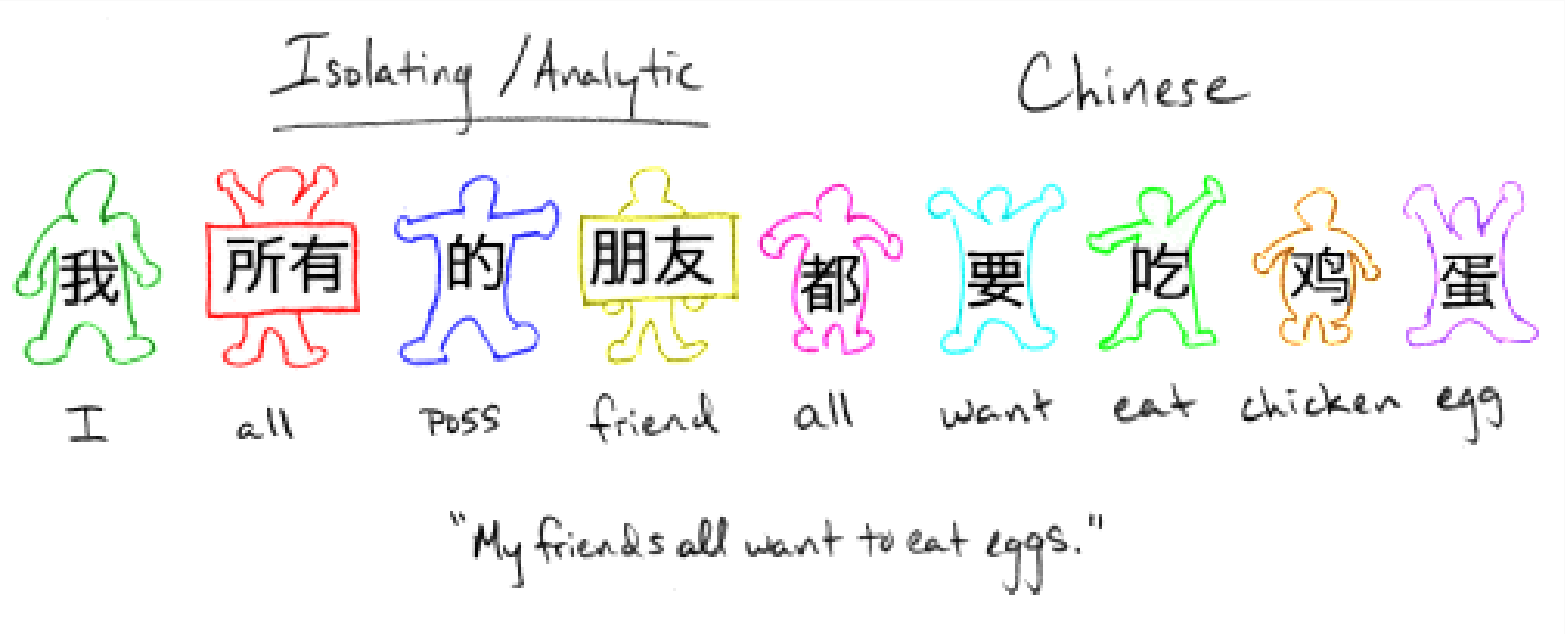
Flere forskjellige relasjoner:

- tømmerhytte – hytte av tømmer (materiale)
- feriehytte – hytte for ferie (hensikt)
- fjellhytte – hytte på fjellet (sted)
- sommerhytte – hytte for sommerbruk (tid for bruk)
- selvbetjeningshytte – hytte med selvbetjening (måten man bruker hytten på)

Oversikt på tavla

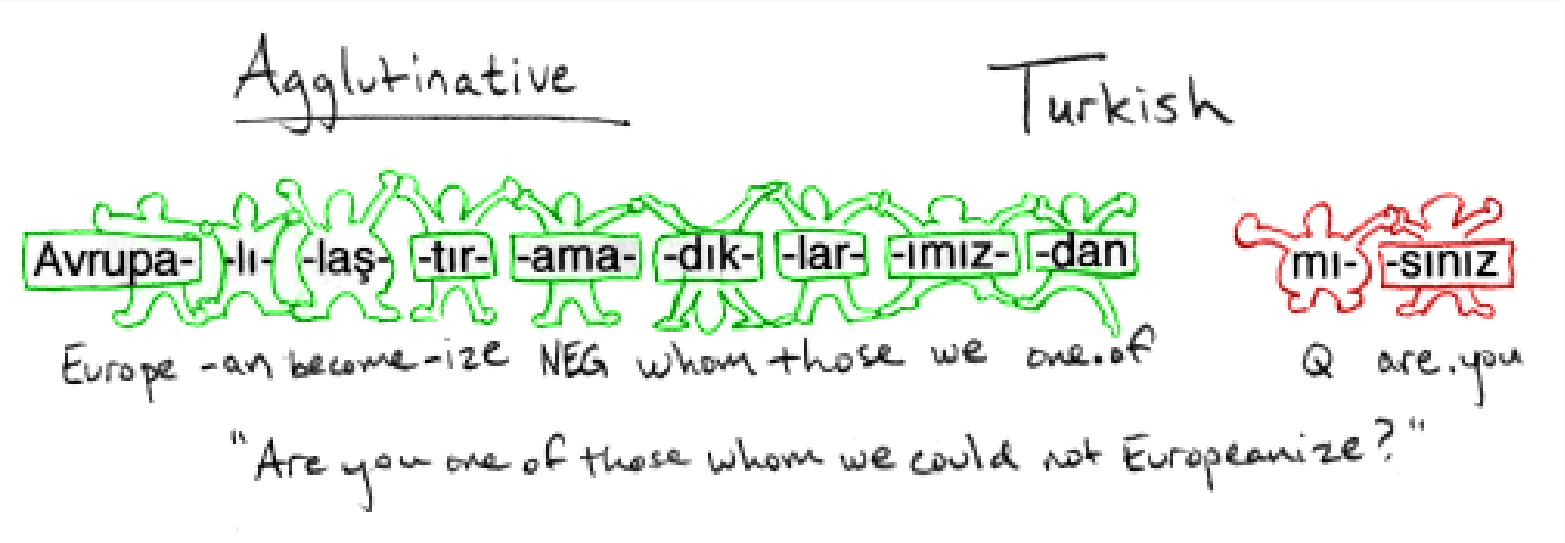
Morfologisk typologi

Isolerende språk

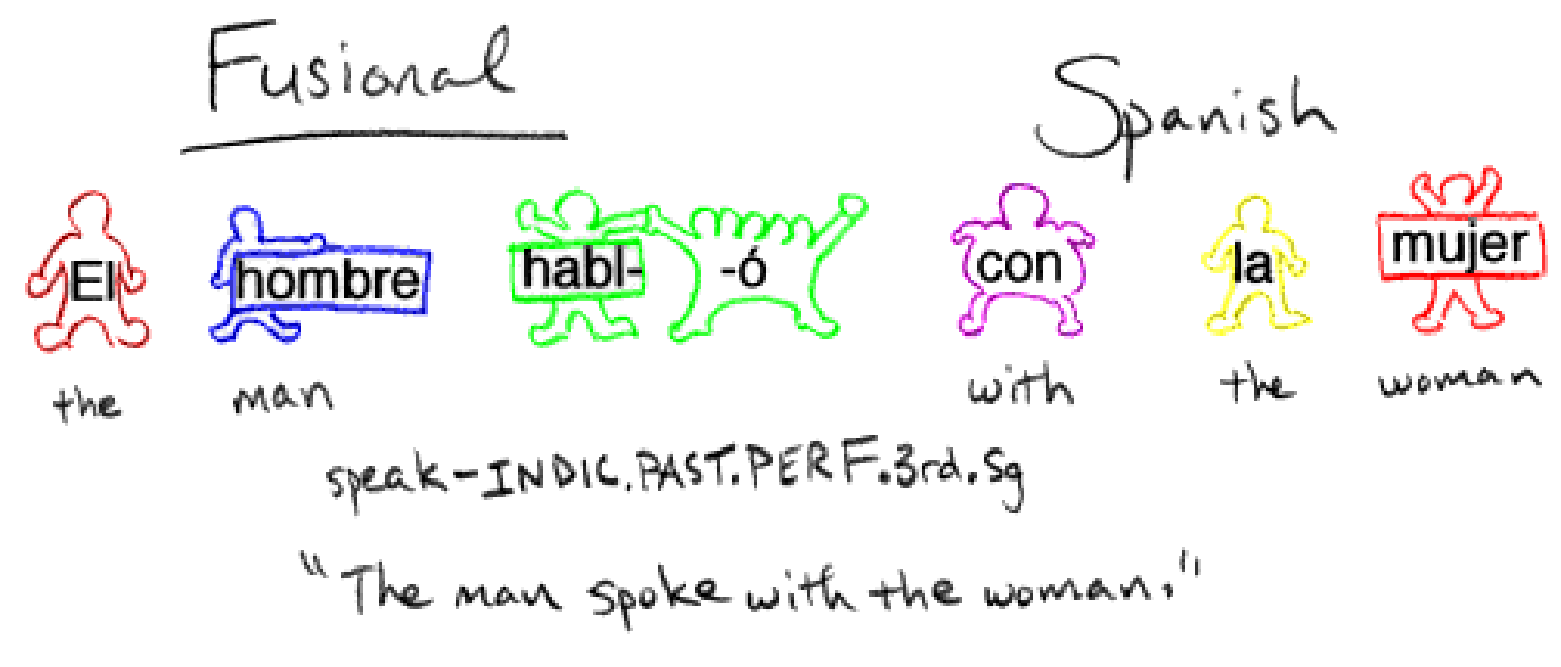


- **Syntetiske språk**, (feks de fleste indo-europeiske): de aller fleste ord formes ved affiksering til en rot.
  - Agglutinerende: Ethvert affiks representerer et distinkt trekk (feks fortid, flertall) – ethvert trekk korresponderer til ett affiks
  - Bøyningsspråk ("Inflectional") (feks romanske språk): flere grammatiske kategorier kan være representert i ett affiks

Agglutinerende språk



Bøyningsspråk



Polysyntetiske språk