IN1140: Introduksjon til språkteknologi

Forelesning #4

Lilja Øvrelid

Universitetet i Oslo

9. september 2024



Tema for i dag

- Korpusdata
- Morfologi

Tema for i dag

- Korpusdata
- Morfologi
  - Ord
  - Morfemet
  - Orddannelse

Språklige data

- Språkteknologi i 2023 er en data-drevet disiplin.
- Trenger språklige data
  - Introspeksjon
  - Faktisk språkbruk – Korpusdata
- Språkteknologi: systemer som generaliserer over språklige mønstre
  - Tilgjengelighet av data er avgjørende

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 språkteknologiske modeller

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:

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:

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 er representativt

Et tenkt korpus for norsk

- Vi må inkludere forskjellige typer tekster:

Et tenkt korpus for norsk

- Vi må inkludere forskjellige typer tekster:
  - Skrift og tale? [registere]

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]

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]

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? [demografi]

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? [demografi]
  - Skal vi inkludere nyhetsstoff? Hva med kronikker, romaner og e-post? Tegneserier og tekstmeldinger? [sjanger]

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? [demografi]
  - Skal vi inkludere nyhetsstoff? Hva med kronikker, romaner og e-post? Tegneserier og tekstmeldinger? [sjanger]

Eksisterende korpuser

- (Forsøk på) representative korpuser for engelsk
  - 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 (3 milliarder websider)

Eksisterende korpuser

- Korpuser for andre språk enn engelsk
  - Arabisk Gigaword
  - Chinese news
  - Norsk Aviskorpus
    - norske nyheter 1998-2016
    - ca. 1.5 millarder ord
  - Norwegian Colossal Corpus (7 milliarder ord)
    - aviser, bøker, offentlig informasjon
    - ca 7 milliarder ord
  - NoTa-korpuset
    - transkripsjoner av samtaler og intervju fra informanter født og oppvokst i Oslo-området, transkribert tekst og tale
    - søk her: `http://www.tekstlab.uio.no/nota/oslo`

Eksisterende korpuser

- Parallelle korpuser
  - EUROPARL: EU-parlamentet
  - OPUS: undertekster fra TV

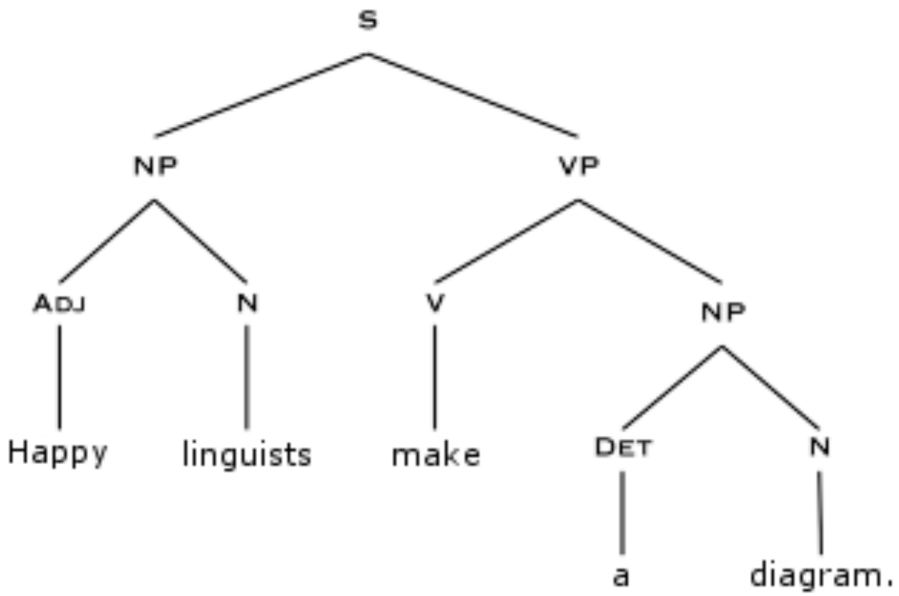
41.	**Chapter 3, Stenmarck (SV)**	**fr**	**nl**
context	That is true **as long as account** is taken of the 20 per cent of the total postal services market where , in practice , there is still a monopoly , that is where the state is the only player .	C' est exact si l' on considère la question en tenant compte des 20 pour cent du marché total des services postaux où le monopole s' est maintenu dans la pratique , c'est-à-dire là où l' État est le seul acteur .	Dat klopt als men alleen kijkt naar 20% van de totale postmarkt , waar de staat in de praktijk nog steeds het monopolie heeft .
42.	**Chapter 3, MacCormick**	**fr**	**nl**
context	The Commission should not , for example , take a stepwise jump from 350 grammes to , as some have suggested , **as low as 50 grammes** .	Par exemple , la Commission devrait éviter de passer de 350g à 50g , comme l' ont suggéré certains .	De Commissie moet bijvoorbeeld niet helemaal van 350 gram naar 50 gram gaan zakken , zoals sommigen hebben geopperd .

Annotering

- Korpuser inneholder forskjellige typer informasjon og er ofte beriket med ulike former for (automatisk/manuell) **annotering**
- Delt opp i enheter som tilsvarer et ord, tokens: ord, tall, tegnsetting → **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, Norsk dependenstrebank)



- Ordsemantikk, sentiment 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

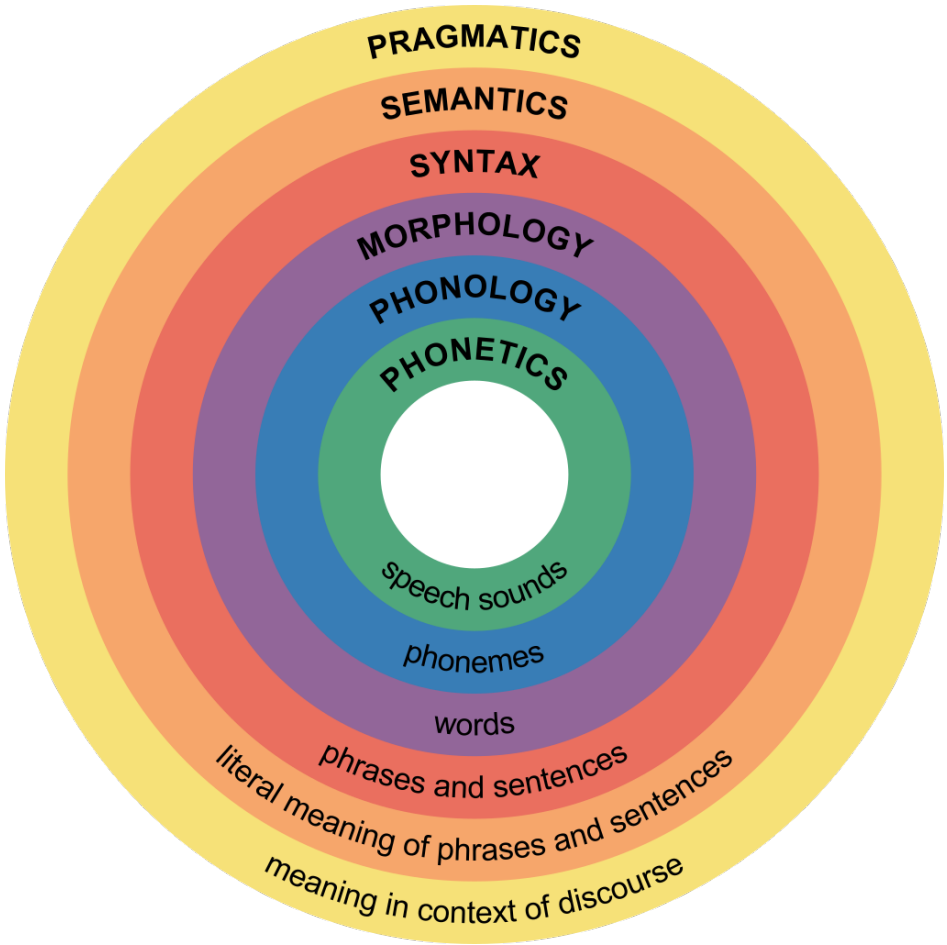
Trene en klassifiserer:

- 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

- Korpusdata
  - representativitet
  - størrelse
  - annotering
  - omfattende bruk i språkteknologiske modeller

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 gikk på tur i skogen . Hun elsker turer i skog .

Hvor mange ord?

Ordet

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

Kari gikk på tur i skogen . Hun elsker turer i skog .

Hvor mange ord?

- 13 ord (**tokens**)
- men også 11 unike ord (**typer**)
- eller 9 ord (**leksem** = leksikon oppslag)

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?

- Vi kan f.eks. ta tog!
- Kjøper gamle møbler, bøker, klær, etc.
- The children's toys
- Møtes kl.10:26? Nei, 11.26 passer bedre for meg.

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** og ofte flertydig kobling:
  - samme lyd - forskjellig betydning (*to*, *two*)
  - forskjellig lyd - samme betydning (*sofa*, *couch*)

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

Mentimeter!

Morfemet

- Ord kan bestå av flere meningsbærende enheter
- **Morfemet** – elementær, minste enhet (gr. 'morphe' – form)
- Morf+ologi – vitenskapen om (ord)former

Morfemet

Et ord kan bestå av ett eller flere morfemer:

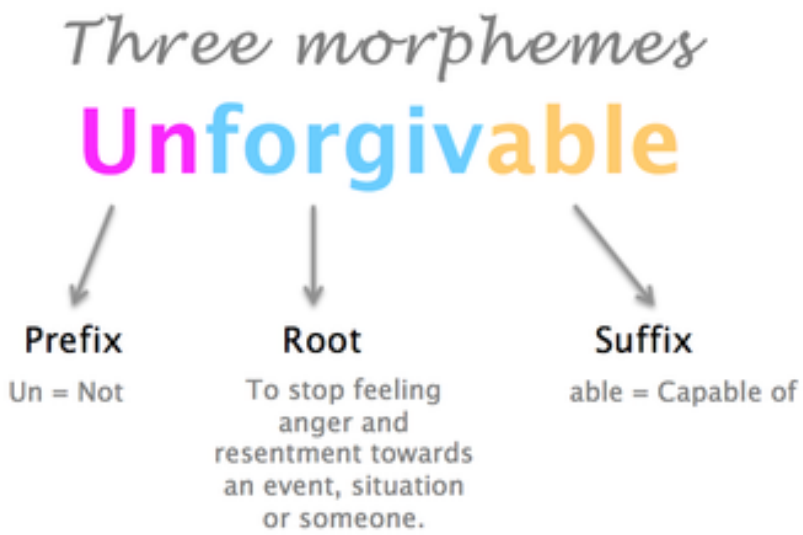
- ett morfem: *boy*, *desire*, *morph*
- to morfemer: *boy+ish*, *desire+able*, *morph+ology*
- tre morfemer: *boy+ish+ness*, *desire+able+ity*
- fire morfemer: *gentle+man+li+ness*, *un+desire+able+ity*
- mer enn fire morfemer: *un+gentle+man+li+ness*, *anti+dis+establish+ment+ari+an+ism*

Morfemet

- To hovedtyper morfemer:
  - **Frie** morfemer: ord. *boy, desire, gentle, man*
  - **Bundne** morfemer: affikser.
    - prefikser: *un-, pre-, bi-*
    - suffikser: *-ing, -ish, -ness*
- Språk benytter affikser i varierende grad
- Noen språk har **infikser**
  - Bontov (Filippinene): *fikas* 'sterk', *fumikas* 'å være sterk'
  - *un-fuckin-believable*
- Noen språk har **sirkumfikser** (affiks som har to deler, en som settes i begynnelsen av ordet, og en som settes på slutten)
  - 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



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

Bøyningsmorfemer markerer kategorier som tid (tempus), tall (numerus), kasus, etc.

Bøyningskategorier i norsk

- **Kjønn** (genus): 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 (skille subjektet fra objektet i setningen). 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*

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) *gal+skap+en*

Bøyning vs. avledning

Mentimeter!

Sammensetninger

- En tredje form for orddannelse, svært vanlig i germanske språk
- 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

Sammensetninger

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

Flere forskjellige relasjoner:

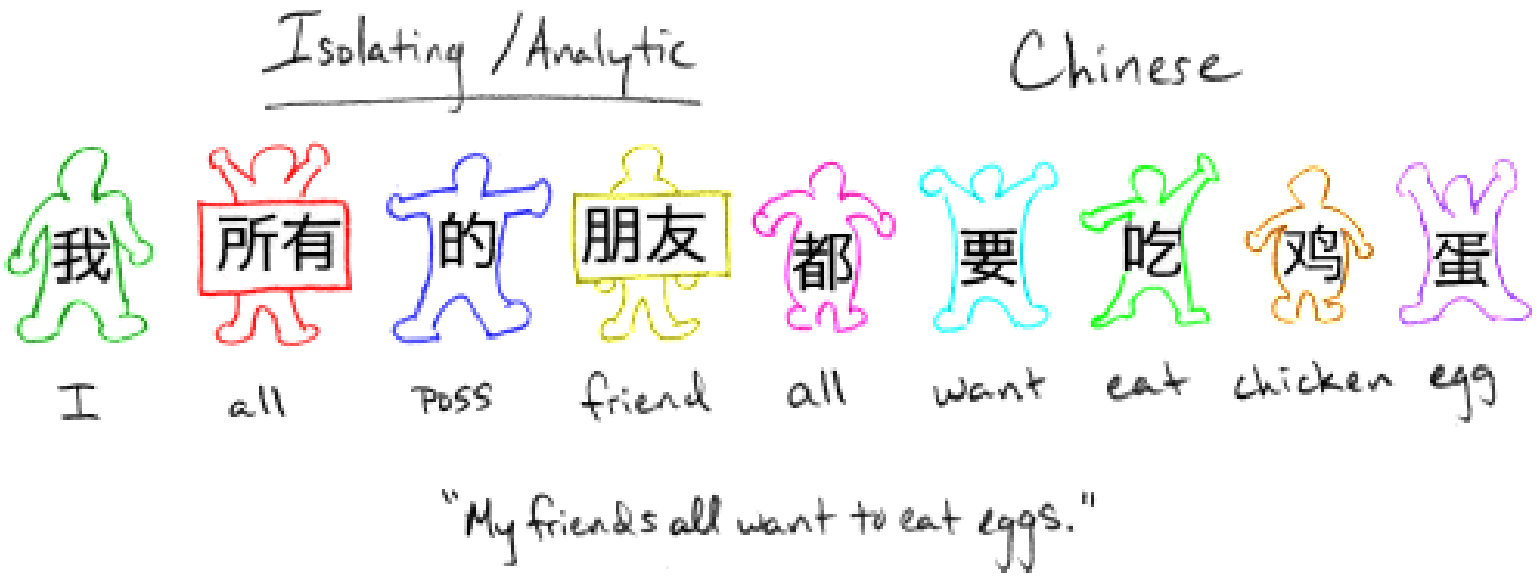
- tømmerhytte – hytte av tømmer (materiale)
- feriehytte – hytte for ferie (hensikt)
- fjellhytte – hytte på fjellet (sted)
- sommerhytte – hytte for sommerbruk (tid for bruk)
- selvbetjeningshytte – hytte med selvbetjening (måten man bruker hytten på)

Morfologisk typologi

- Typologi: delområde av lingvistikk
- Klassifiserer språk i henhold til ulike egenskaper
- I morfologisk typologi brukes to skalaer:
  - graden av **syntese** (antall morfemer i hvert ord)
  - graden av **fusjon** (antall betydninger av hvert morfem)

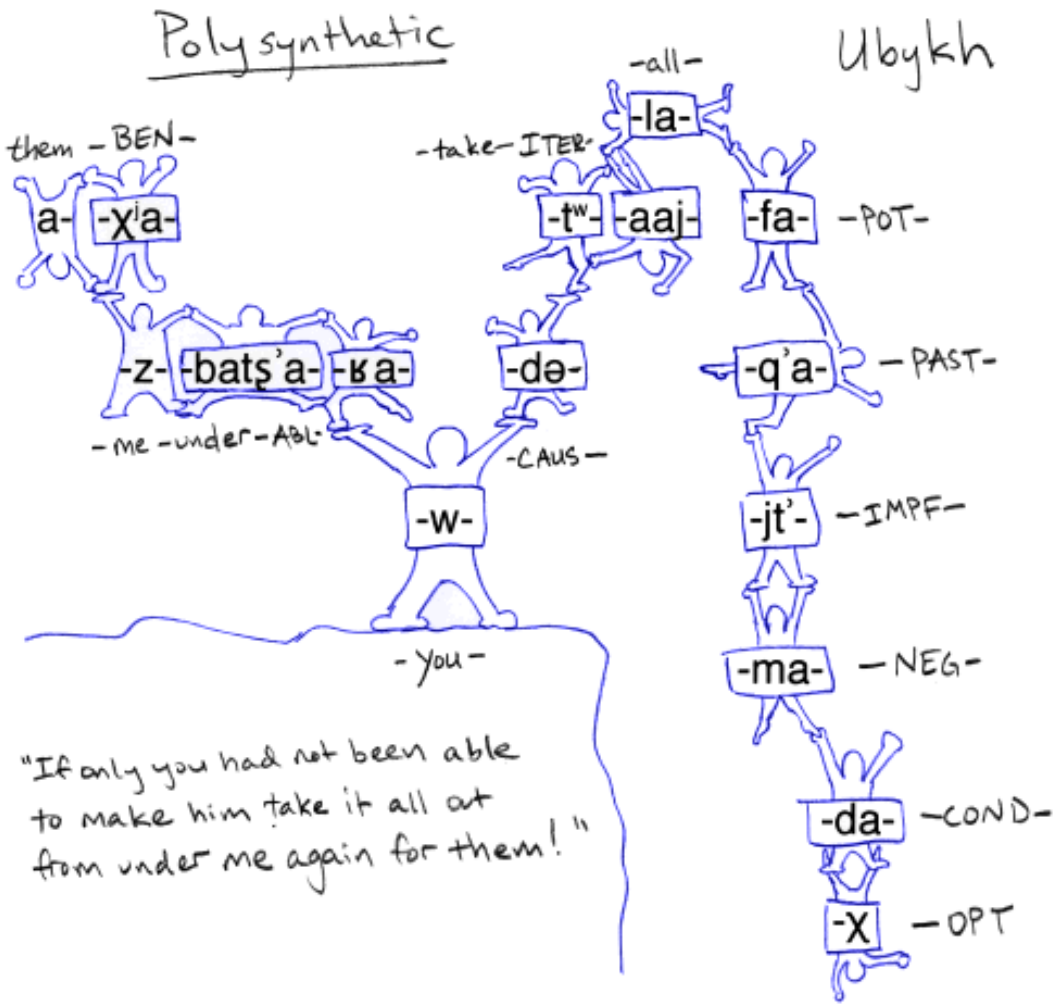
Isolerende språk

Syntese: ett ord = ett morfem



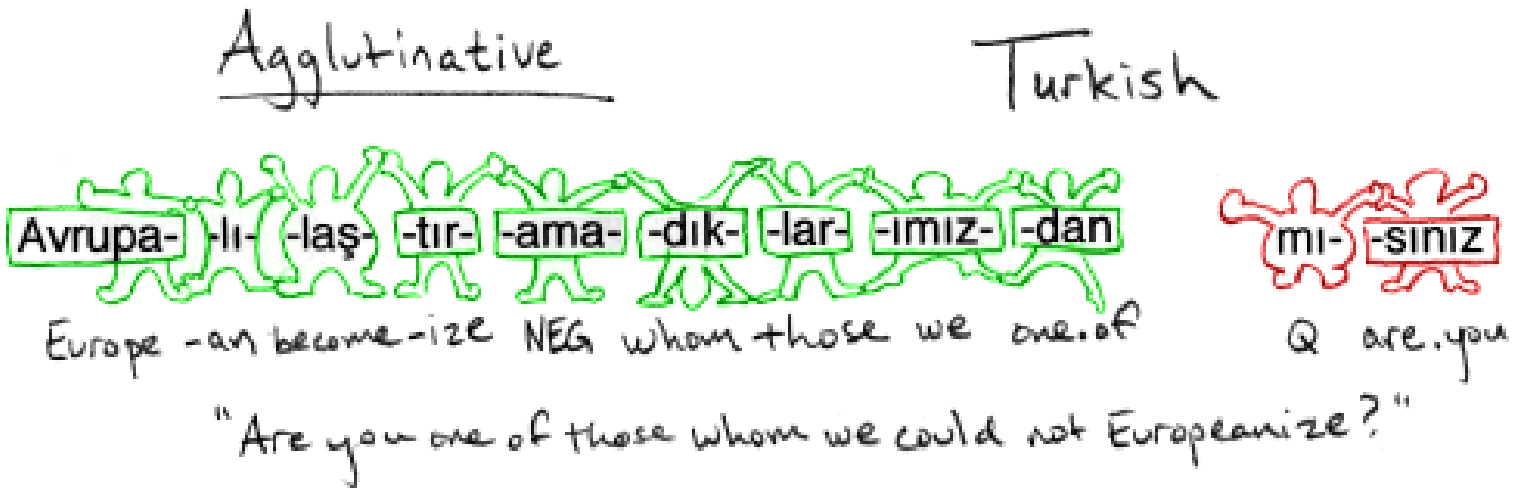
Polysyntetiske språk

Syntese: høy morfem-til-ord fordeling



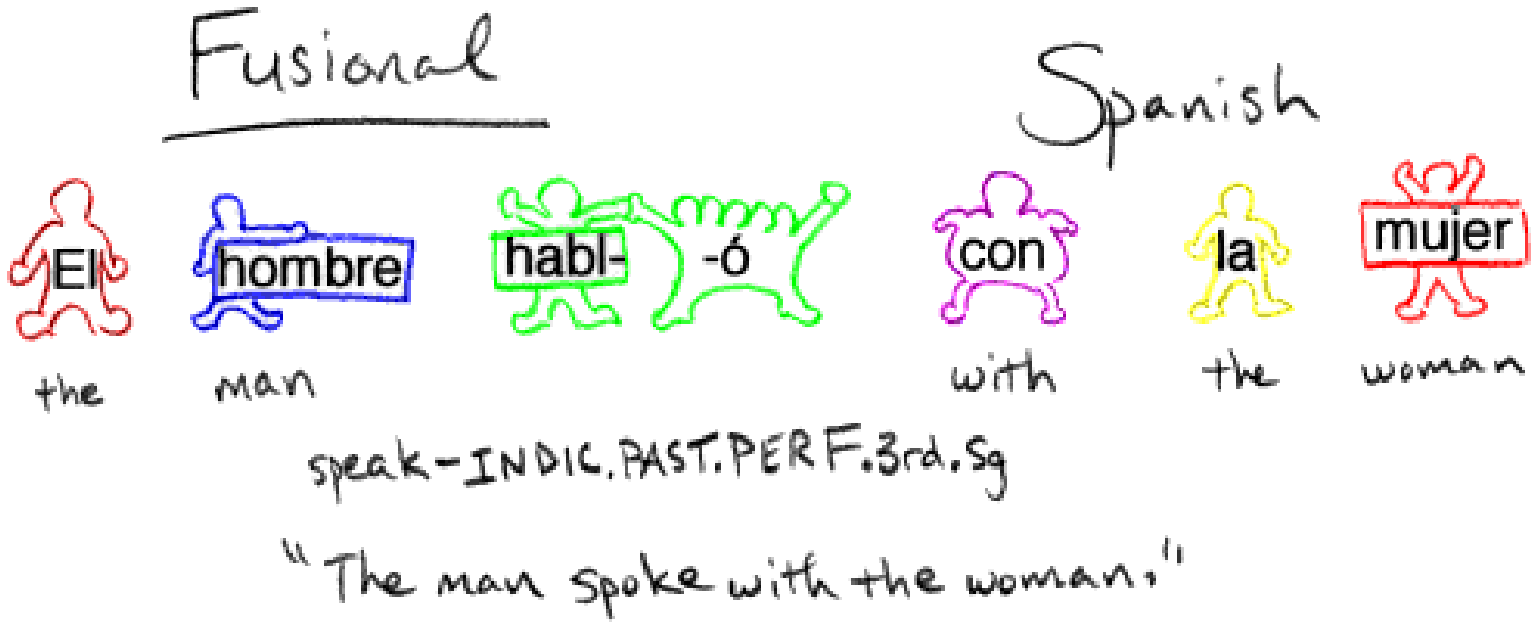
Agglutinerende språk

Fusjon: ett morfem = én betydning



Bøyningsspråk

Fusjon: ett morfem kan ha flere betydninger



Oppsummering morfologi

- Handler om **ord**:
  - hvordan ord er bygd opp (morfemer)
  - hvordan nye ord dannes (avledning, sammensetning)
  - hvordan ord bøyes
- Skiller mellom frie og bundne morfemer (affikser)
- Morfologisk komplekse ord består av
  - **Rot** + en eller flere affikser (hus+lig)
- Morfologi er noe som skiller verdens språk: syntese og fusjon

Oblig 1a

- Teoretisk: morfologi og lingvistiske nivåer
- Praktisk:
  - Tekst i Python
    - lese og skrive til fil
    - telle forekomster i tekst
    - tekst som streng og liste
  - Tokenisering av tekst (første forsøk)
    - Enkel tokenisering
    - Feilanalyse
- Legges ut i dag
- Frist: Fredag 20/9 kl. 23.59
- Innlevering via Devilry
- Bruk gruppetimene! Ikke bare i innspurten
- Husk at ukesoppgavene ofte gir deg byggeklossene som du skal anvende i obligene.