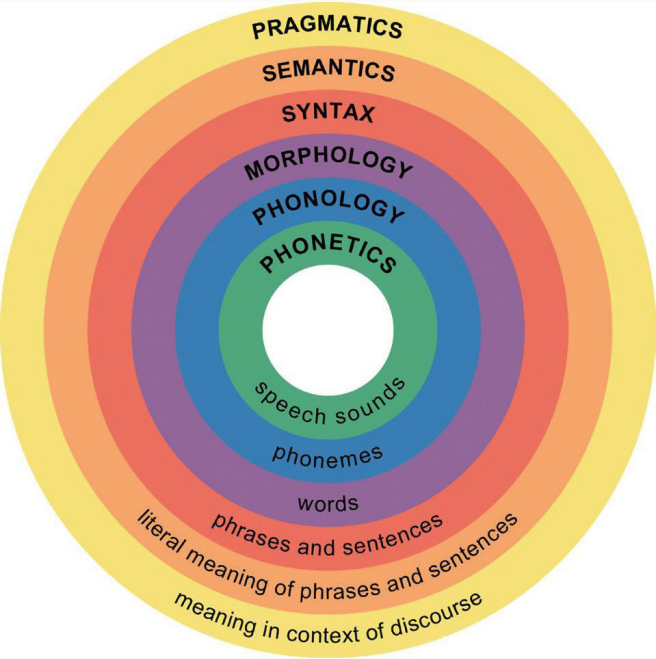
INF1820: Introduksjon til språk-og kommunikasjonsteknologi

Femte forelesning

Lilja Øvrelid

13 februar, 2017

Lingvistikk



- Bindestreksdisipliner: psykolingvistikk, neurolingvistikk, sosiolingvistikk, datalingvistikk

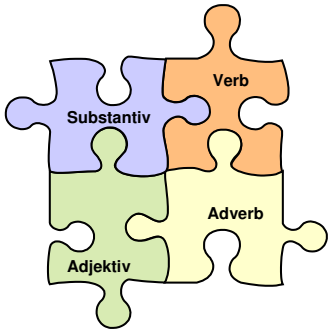
Ordklasser?

- Bindeledd mellom ordet og setningen (syntaks):
  - Sier noe om hva slags kontekster et ord forekommer i
  - Sier noe om uttale (*record*, *content*, *discount*)
- Helt essensiell i en rekke språkteknologiske applikasjoner:
  - Talesyntese
  - Morfologisk analyse
  - "Chunking", syntaktisk parsing
  - Word Sense Disambiguation
  - Informasjonsekstraksjon

Ordklasser?

`http://beta.visl.sdu.dk/visl/nor/edutainment/games/`

Ordklasser



- **Taksonomi** - uttømmende, gjensidig utelukkende, styrt av et prinsipp
- Alle ord havner i en klasse og ingen ord havner i mer enn én klasse
- Vi trenger **kriterier** for ordklasseinndeling

Ordklassekriterier

3 slags kriterier:

1. formelle eller morfologiske kriterier
    - hvilke bøyningsformer har ordet?
        - *hare* - *haren* og *redd* - *reddere*
        - **harare* och **redden*
2. funksjonelle eller syntaktiske kriterier
    - hvordan kan ordet kombineres med andre ord?
        - *en redd hare* och *redd for ilden*
        - **en redd* och **en hare for ilden*
3. betydningsmessige eller semantiske kriterier
    - hva er typiske betydninger hos ord i ordklassen?
        - *hare* - dyr, levende vesen
        - *redd* - egenskap

Ordklassekriterier

Adjektivet RØD:

- form: *rød, rødt, røde, rødere, rødest*
- funksjon: *et rødt eple* (attributiv funksjon), *Håret hennes er rødt* (predikativ funksjon)
- betydning: betegner en egenskap, typisk for adjektiv
- MEN: *De røde tapte borgerkrigen*??

- normale (prototypiske) bruken
- vekting av kriteriene

Ordklasser: substantiv

Substantiv

olje, bord, jente, sorg

1. bøyes i bestemthet og tall
   - bestemthet: kan knytte til seg bestemt artikkel som suffiks: *bilen, greina, huset, tanken*
   - tall: (de fleste har) forskjellige endelser for entall og flertall: *bil-biler, grein-greiner, tanke-tanker*
2. kjerne i substantivfraser, med modifikatorer: *en alldeles fantastisk vakker stol*
3. betegner "ting" - mennesker, objekter, vesen, steder, fenomener och abstrakte enheter

Unntak - egennavn

- bøyes ikke

Ordklasser: substantiv

- Substantivene er enten **fellesnavn** eller **egennavn**
- Fellesnavn: substantivene på forrige side
- Egennavn: ord som *Adam, Eva, Haugesund, Dagros, Norge*

Ordklasser: substantiv

Fellesnavnene har enten **konkret** eller **abstrakt** betydning

- Konkrete: en slags gjenstandsbetydning, betegner konkrete størrelser, f.eks. *blomst, bok, bord, mann, tang, tårn*
- Abstrakte: ikke gjenstandsbetydning, betegner abstrakte begreper eller forestillinger (handlinger, tilstands, egenskaper), f.eks. *fred, ro, slag, tvil, vennskap, verdighet, alderdom*
- Ikke alltid så lett å skille:
  - *en mørkhåret skjønnhet – dø i skjønnhet*
  - *spenstig som en ungdom – i min ungdom*
  - *en fin tegning – flink i tegning*

Ordklasser: substantiv

- Vi kan også skille mellom **tellelige** og **ikke-tellelige/massebetegnende** substantiver
- Tellelige: *bil, bord, okse, stang*
- Ikke-tellelige: *kjøtt, smør, støv*
- Både tellelig og ikke-tellelig betydning:
  - *et vakkert tre – bordet er laget av tre*
  - *mange høye fjell – støtt som fjell*
  - *mange steiner – mye stein*

Ordklasser: substantiv

- **Egennavn** refererer til en enkelt gjenstand eller et bestemt individ, f.eks.*Ida, Torggata, Colosseum, Apple*
- Navn på dyr, land, byer, gater, bygninger, elver, steder, planeter, foreninger, firmaer, fly, tog, kunstverk
- Vanligvis unik referanse
- Egennavn i flertallsform viser til helheten, f.eks. *De forente stater, Hebridene*
- Noen substantiver kan fungere som begge deler:
  - *Bibelen er verdens mestselgende roman – Hun leste i sin bibel*
- Egennavn som fellesnavn:
  - *Hemsedal er et skisportens mekka*
  - *. . . en quisling, et eldorado, et watergate*

Ordklasser: verb

Verb (hovedverb)

sparke, sove, håpe, arbeide, bygge, leve

1. bøyes i tid (presens-preteritum)
   inndeles i finitte vs. infinitte former
   - finitte: imperativ, presens, preteritum: *spark, sparker, sparket*
   - infinitte: infinitiv, perfektum partisipp *(å) sparke, (ha) sparket*

   transitivitet: transitiv - intransitiv
2. kan stå alene som predikat
3. betegner handlinger, aktiviteter och tilstander

Unntak - hjelpeverb: *må, skal, bli*

Ordklasser: verb

- Kan grupperes basert på semantisk og grammatisk forhold
- Tre hovedgrupper
  a) **Aktivitetsverb** uttrykker en aktiv handling, dvs. at noen gjør eller utfører noe. Forutsetter en **agens**
     - *Ida* **arbeider** *hele dagen*
     - *De* **gravde** *et hull*
     - *De fleste* **går** *til jobben*
     - *Fredrik* **spiller** *bass*

Ordklasser: verb

- Tre hovedgrupper (forts.)
  b) **Endrings-eller overgangsverb** uttrykker et forløp eller en endringsprosess, dvs subjektet er **patiens** og gjennomgår eller opplever en forandring eller overgang
    - *Faren* **døde** *like etter krigen*
    - *Bilen* **forsvant** *nedover gata*
    - *Barna* **sovnet** *fort*
    - *Plantene* **vokser** *hele året*

Ordklasser: verb

- Tre hovedgrupper (forts.)
  c) **Tilstandsverb** uttrykker en tilstand, subjektet er i gitt tilstand uten å forandres
    - *Han* **ble** *i London resten av livet*
    - *De* **bor** *nå i Mumbai*
    - *Thon* **eier** *snart hele byen*
    - *De* **lever** *et lykkelig liv sammen*
    - *Boken som* **ligger** *på bordet, tilhører meg*

Ordklasser: adjektiv

Adjektiv:

rød, snill, vanskelig, levende

1. samsvarsbøyes i bestemthet, kjønn och tall, *gradbøyes*
2. modifikator (adledd) til substantiv
3. betegner egenskaper

Ordklasser: adjektiv

- De mest typiske egenskapene er permanente (kvaliteter):
  - Størrelse, allment: *stor, liten*, vertikalt: *høy, lav, kort*, horisontalt: *bred, smal, lang*
  - Form: *rett, krokete, rund, flat*
  - Farge/lys: *svart, gul, lys, mørk, dus*
  - Lydstyrke: *høy, lav, skarp*
  - Smak: *sur, søt, bitter*
- Noen adjektiver uttrykker midlertidige egenskaper:
  - Livstilstand: *gammel, ung, levende, frisk, syk*
  - Sinnstilstand: *sint, trøtt, glad, redd*
  - Temperatur: *varm, kald, lunken*
  - Andre egenskaper: *lat, arbeidsom, ren, skitten, rask, sein*

Ordklasser: adjektiv

- Gradbøyes ved bøyningsendelse eller mer–mest. Betydningen angir et punkt på en skala.
- Men mange adjektiv har en mer presis betydning som er vanskelig å gradere, f.eks. *død*, *gift*, *gratis*, *nybakt*, *lovlig*
- Noen av de mest sentrale adjektivene opptrer i par med motsatt betydning **antonymer**:
  - *høy – lav*
  - *stor – liten*
  - *lang – kort*

Ordklasser: adverb

Adverb:

her, ofte, derfor, trolig, ikke, kanskje, nå, vanligvis

1. ubøyelige
2. står som modifikatorer til verb, adjektiv, adverb och setninger
3. betegner forskjellige omstendigheter - rom, tid, måte m.m.

Ordklasser: adverb

- **Tidsadverb** uttrykker relativ tid, dvs. et tidspunkt i forhold til et annet
  - *Han kom etterpå* (etter et tidspunkt i fortiden)
  - *Han kom da* (på et tidspunkt i fortiden)
  - *Kom etterpå!* (etter dette tidspunktet)
  - *Kommer han nå?* (på dette tidspunktet)
  - *Du skal komme etterpå* (etter et tidspunkt i framtiden)
  - *Du skal komme da* (på et omtalt tidspunkt i framtiden)

Ordklasser: adverb

- **Måtesadverb** uttrykker måten noe blir gjort på
  - *Hun gjennomgikk pensum* **stykkevis**
  - *De lå* **andføttes**
- **Gradsadverb** uttrykker mengde, intensitet eller grad ved verbhandlingen
  - *Jeg fryser* **litt**
  - *Nå har du tullet* **nok**

Ordklasser: preposisjoner

Preposisjoner:

ved, på, under, i, foran, av

1. ubøyelige
2. kjerne i preposisjonsfraser, tar substantiv
3. betegner relasjoner, f.eks. romlige
   - *Hytta ligger* **ved** *sjøen*
   - *Elevene var svake* **i** *engelsk*
   - *Taket* **på** *huset ble nettopp reparert*

Ordklasser: preposisjoner

- Varierende semantisk innhold
- Lokalisere gjenstander og begivenheter i **rom** og **tid**
  - *Boka ligger* **på** *bordet*
  - *Den* **lå** *bak skapet*
  - *Vi drar* **i** *mai*
- Kan også uttrykke **måte** eller **middel**:
  - *Hun satt* **i** *dype tanker*
  - *Hun svarte* **med** *et lite smil*
  - *Hun åpnet døren* **med** *en rusten nøkkel*

Ordklasser: preposisjoner

- Preposisjon uten utfylling: **verbalpartikkel**
  - *De sovnet* **inn**
  - *Han brøt* **sammen** *etter løpet*
  - *Vi drakk* **opp** *all vinen*
- Danner en semantisk og syntaktisk enhet med verbet

Ordklasser: pronomen

Pronomen:

jeg, hun, dere, seg, hverandre, hvem, man

1. av svært ulik form, uregelmessig bøyning
2. som substantiv, kan fungere som setningsledd alene
3. lite eget innhold, fr betydning fra sammenhengen (*konteksten*)

- **Jeg** *liker grammatikk*
- **Man** *skal respektere hverandre*
- **Hvem** *tok vesken?*

Ordklasser: pronomen

- Pronomen får sitt innhold enten fra et element i selve talesituasjonen eller fra et nominalt ledd (typisk substantiv) i konteksten.
- Leddet som gir pronomen innhold er pronomenets **antesedent**
    - *Se* **her**
    - *Er* **du** *sulten?*
    - *Gro Harlem Brundtland er en tidligere norsk politiker.* **Hun** *var norges første kvinnelige statsminister. . .*

Ordklasser: pronomen

- **Personlige pronomen** kan bøyes, egne former for første, andre og tredje person, samt entall og flertall

	Nominativ	Akkusativ
1.pers.ent.	jeg	meg
2.pers.ent.	du	deg
3.pers.ent	han	ham
3.pers.ent	hun	henne
3.pers.ent	den/det	den/det
1.pers.flt.	vi	oss
2.pers.flt.	dere	dere
3.pers.flt.	de	dem

Ordklasser: pronomen

- **Refleksivt pronomen** er *seg* på norsk.
- Har antesedent i samme setning, oftest subjektet i setningen
- Brukes kun i tredje person. I første og andre person brukes akkusativformen
  - *Jeg vasker* **meg**
  - *Hun vasker* **seg**
  - *Vi vasker* **oss**

Ordklasser: pronomen

- **Resiproke pronomen** er *hverandre* på norsk.
- Uttrykker en gjensidig relasjon, slik at *A og B beundrer hverandre* impliserer at *A beundrer B* og *B beundrer A*
  - *Vi beundrer* **hverandre**
  - *De beundrer* **hverandre**

Ordklasser: pronomen

- **Interrogative pronomen** (spørreord)
- *Hvem* når vi spør etter et menneske, ellers er det *hva*
  - **Hvem** *er det?*
  - **Hva** *vil du ha å drikke?*
  - *Hun spurte* **hvem** *det var*

Ordklasser: determinativ

Determinativ (artikler):

min, din, denne, alle, noen

1. bøyning i kjønn og tall
2. bestemmer til substantiv
3. bestemmer, spesifiserer substantivets referanse

Ordklasser: Determinativ

3 hovedtyper:

a) Possessiver: angir eiendom eller tilhørighet, bøyes i person
   - *Det er* **min** *bok*
   - *Her har du boken* **din**

b) Demonstrativer: viser til eller peker på en bestemt person eller ting som kan iakttas eller er omtalt
   - **Den** *hytta ligger fint til*
   - **Dette** *treet er kjempestort*

c) Kvantorer: uttrykker mengde eller kvantitet, noen med bøyning (*noen, ingen, en*) og noen uten (*to, tre, visse, enkelte, utallige*)
   - *Hun har spist opp* **all** *maten*
   - *Ida har kjøpt* **noen** *bøker*

Ordklasser: Konjunksjoner

Konjunksjoner:

og, eller, men, for, så

1. ubøyelige
2. binder sammen ledd av samme slag, f.eks. ord, fraser og setninger
3. grammatisk funksjon, betegner relasjoner
   - *fullstendig ro* **og** *absolutt trygghet* (nominalfrase og nominalfrase)
   - *konkret* **og** *abstrakt betydning* (adjektivfrase och adjektivfrase)
   - *han var på ski* **og** *hun var i kirken* (setning og setning)

Ordklasser: subjunksjoner

Subjunksjoner:

å, at, om, som, før

1. ubøyelige
2. innleder leddsetninger - underordner en setning under en annen
3. grammatisk funksjon, betegner relasjoner
   - *Hun elsker* **å** *danse*
   - *Vi tror* **at** *det verste snart er over*
   - *Der er hunden* **som** *spiste kaken*

Inndeling av ordklasser

- **åpne** vs. **lukkede** ordklasser
  - åpne: substantiv, verb og adjektiv
    inneholder mange tusen ord, kan enkelt fylle på med nye
    Eksempel: nye bilmodeller - nye farger (*brannbilrød*)
  - lukkede: inneholder mange færre ord enn de åpne
    kan ikke fritt skape nye ord gjennom orddannelse
- **innholdsord** vs. **funksjonsord**
  - innholdsord: substantiv, verb, adjektiv
    rikt betydningsinnhold, refererer utenfor språket
  - funksjonsord: mer allment betydningsinnhold, refererer ikke utenfor språket. Finnes fremst i de lukkede ordklassene.
- Ikke helt én-til-én, feks hjelpeverb.

Korpuser

Korpusdata

- Modellere språklig kunnskap
- Trenger språklige data
  - Introspeksjon
  - Faktisk språkbruk – korpusdata
- Språkteknologi: programmer som generaliserer over språklige mønstre
  - Korpusdata helt sentralt

Språklige data: korpusdata

- Et korpus (tekstkorpus) er en strukturert samling tekster
- Elektronisk lagret
- Siste 30 årene innenfor lingvistikk og datalingvistikk: empirisk revolusjon
- Større og større tekstmengder tilgjengelig
- Empiriske data for lingvistiske studier (motsetning til introspeksjon)
- Treningsmateriale for datalingvistiske modeller av språklige fenomener

Ordklassetaggede korpuser

- Brown-korpuset for engelsk (1979):
  - 87 ordklassetagger
  - 1 mill. ord, utvalg fra 500 tekster hentet fra forskjellige sjangere
  - Automatisk tagget og manuelt rettet
- Penn Treebank (1993)
  - 45 ordklassetagger
  - Wall Street Journal, Brown-korpuset, Switchboard, ATIS
  - Ordklassetagger, syntaktisk struktur (frasestruktur)

Ordklassetaggede korpuser

- Norsk dependenstrebank
  - Trebank for norsk
  - Utviklet ved Nasjonalbiblioteket
  - Manuelt tagget
- ordklasser samt mye morfologisk informasjon

1	Det	det	pron	nøyt—ent—pers—3
2	er	være	verb	pres
3	hun	hun	pron	fem—ent—pers—hum—3—nom
4	som	som	sbu	_
5	eier	eie	verb	pres
6	og	og	konj	_
7	driver	drive	verb	pres
8	stedet	sted	subst	appell—nyt—be—ent

- Eksempler fra Penn (J&M):
  - The/DT Grand/JJ jury/NN commented/VBD on/IN a/DT number/NN of/IN other/JJ topics/NNS ./.
  - There/EX are/VBP 70/CD children/NNS there/RB
  - Although/IN preliminary/JJ findings/NNS were/VBD reported/VBN more/RBR than/IN a/DT year/NN ago/IN ,/, the/DT latest/JJS results/NNS appear/VBP in/IN today/NN 's/POS New/NNP England/NNP Journal/NNP of/IN Medicine/NNP

- Tagging følger en manual
- Noen avgjørelser vanskelige
- Eks: skillet mellom preposisjoner (IN), partikler (RP) og adverb (RB)
  - Mrs./NNP Shaefer/NNP never/RB got/VBD around/RP to/TO joining/VBG
  - All/DT we/PRP gotta/VBN do/VB is/VBZ go/VB around/IN the/DT corner/NN
  - Chateau/NNP Petrus/NNP costs/VBZ around/RB 250/CD
- Manualen: preposisjoner er assosiert med en etterfølgende subtstantivfrase. *Around* tagges som adverb i betydningen 'omtrent'

Ordklassetagging

Ordklassetagging

- Oppmerking av ordklasseinformasjon for hvert ord i et korpus
- Språkteknologi: automatiske systemer
- Flertydighet vanskeliggjør dette betydelig
- Ordnivå: Tokenisering

Tokenisering

- **Tokenisering**
  Tokenisering handler om å dele inn en tekst i ord og setninger. Tidligere har vi gjort det enkelt og bare splittet på mellomrom. Men dette er problematisk:
  - Tar ikke hensyn til tegnsetting og gir "ord" som *cents. said, positive." Crazy?*.
  - Her kunne vi tatt bort tegnsetting, men tegnsetting forekommer også innad i ord: *m.p.h. cap'n, AT&T*.
  - Tall inneholder komma i engelsk: 555,000
  - Det kan være ønskelig å ekspandere forkortede former som for eksempel *I,m, you're, they've* til henholdsvis *I am, you are, they have*. Da er det viktig å skille mellom slike former og genitiv 's (*Mary's*) eller anførselstegn (*'Oh no', he said*)

Ordklassetagging

- Input: streng av ord og en spesifisert mengde tagger
- Output: en tagg per ord
  - Book/VB that/DT flight/NN ./.
  - Does/VBZ that/DT flight/NN serve/VB dinner/NN ?/.
- Flertydighet: *book*, *that*

Flertydighet

- De fleste engelske ord er entydige
- Men mange av de mest frekvente ordene er flertydige
- Tall fra Brown-korpuset:
  - 11.5% av engelske ordtyper er flertydige
  - 40% av Brown tokens er flertydige
- Heldigvis er ikke alle lesninger like sannsynlige

Algoritmer for ordklassetagging

- To hovedkategorier:
  1. **Regelbaserte taggere**: stor database med håndskrevne regler. Eksempel: *book* er substantiv, og ikke verb, dersom etterfølger en determinativ
  2. **Probabilistiske taggere:** bruker et ordklassetagget korpus ("treningskorpus") til å beregne sannsynlighet for en gitt tagg i en gitt kontekst

Eksempel på en MSc-oppgave i Språkteknologi

Optimizing a PoS tag set for Norwegian Dependency Parsing

- trener ulike ordklassetaggere for norsk (NDT)
- modifiserer taggsettet for syntaktisk parsing
- viser hvordan et mer finkornet taggsett kan gi bedre resultater for parsing