

IN1140: Introduksjon til språkteknologi

Forelesning #6

Samia Touileb

Universitetet i Oslo

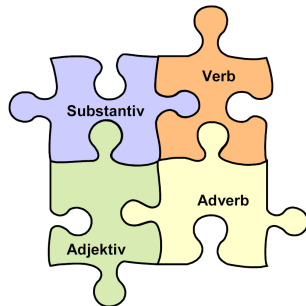
26. september 2019





- ▶ Ordklasser
- ▶ Ordklassetagede korpuser
- ▶ Ordklassetagging

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



- ▶ **Taksonomi** - et system som har kategorier som er 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

3 slags kriterier:

1. Formelle eller morfologiske kriterier

- ▶ Hvilke bøyingsformer har ordet? Har de samme ordklasse?
 - ▶ *hare* - *haren* og *redd* - *reddere*
 - ▶ **harare* og **redde*

2. Funksjonelle eller syntaktiske kriterier (hva slags funksjon har ordet i en setning, hvilken posisjoner forekommer det typisk i)

- ▶ Hvordan kan ordet kombineres med andre ord?
 - ▶ *en redd hare* og *redd for ilden*
 - ▶ **en redd* og **en hare for ilden*

3. Betydningsmessige eller semantiske kriterier

- ▶ Hva er typiske betydninger hos ord i ordklassen?
 - ▶ *hare* - dyr, levende vesen
 - ▶ *redd* - egenskap



Adjektivet RØD:

- ▶ form?
- ▶ betydning?

Adjektivet RØD:

- ▶ form: *rød, rødt (bøyning etter kjønn), røde (bestemt), rødere (komparativ form), rødest (superlativ)*
- ▶ funksjon: *et rødt eple* (attributiv funksjon – attribuer et egenskap), *Håret hennes er rødt* (predikativ funksjon – en identitesmarkør)
- ▶ betydning: betegner en egenskap, typisk for adjektiv
- ▶ MEN: *De røde tapte borgerkrigen??*
 - ▶ Adjektivet er i en substantiv kontekst: unntak.
- ▶ Fokusere på:
 - ▶ Normale (prototypiske) bruken
 - ▶ Vekting av kriteriene

Substantiv – *olje, bord, jente, sorg*

1. Bøyes i bestemthet og tall

- ▶ Bestemthet: kan knytte til seg bestemt artikkel som suffiks: *bilen, greina, huset, tanken, bordet*
- ▶ Tall: (de fleste har) forskjellige endelser for entall og flertall: *bil-biler, grein-greiner, tanke-tanker, bord-border*

2. Kjerne i substantivfraser, med modifikatorer: *en aldeles fantastisk vakker stol*

3. Betegner “ting” - mennesker, objekter, vesen, steder, fenomener og abstrakte enheter

Unntak – egennavn: bøyes ikke

Verb (hovedverb) – *sparke, sove, håpe, arbeide, bygge, leve*

1. Bøyes i tid (presens-preteritum)

inndeles i finitte vs. infinitte former

- ▶ Finitte (kan stå alene i en setning): imperativ, presens, preteritum: *spark, sparker, sparket*
- ▶ Infinitte (kan ikke stå alene, de trenger en funksjonell markør/hjelpeverb): infinitiv, perfektum partisipp (*å*) *sparke*, (*ha*) *sparket*

Transitivitet: transitiv (krever to argumenter: subjekt og objekt) -
intransitiv (tar ikke objekt, krever ikke et argument)

2. Kan stå alene som predikat

3. Betegner handlinger, aktiviteter og tilstander

Unntak – hjelpeverb (forekommer typisk med et innholdsverb): *må, skal, bli*



Hjelpeverb (fra den Store norske leksikon):

Hjelpeverb er verb som vanligvis står sammen med infinitiv eller partisipp av et annet verb, kalt hovedverbet, for å uttrykke grammatiske kategorier (perfektum, passiv, futurum og så videre).

I setningen *Hun har gått* er *har* hjelpeverb og *gått* hovedverb.

I setningen *Han må spise* er *må* hjelpeverb og *spise* hovedverb.

Noen norske hjelpeverb er *ha, bli, være, skulle, ville, kunne, måtte og burde*.



Adjektiv – *rød, snill, vanskelig, levende*

1. Samsvarsbøyes i bestemthet, kjønn og tall, *gradbøyes* (*rød, rødere, rødest, interessant, mer interessant, mest interessant*)
2. Modifikator (adledd) til substantiv
3. Betegner egenskaper
 - ▶ Gradbøyes ved bøyningsendelse eller mer–mest. Betydningen angir et punkt på en skala (feks. ung–gammel / men ikke levende–død).
 - ▶ 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*



Adverb – *her, ofte, derfor, trolig, ikke, kanskje, nå, vanligvis*

1. Ubøyelige (Engelsk: beautiful–beautifully, careful–carefully)
2. Står som modifikatorer til verb, adjektiv, adverb og setninger
3. Betegner forskjellige omstendigheter - rom, tid, måte m.m.

- ▶ **Tidsadverb** uttrykker relativ tid, dvs. et tidspunkt i forhold til et annet
 - ▶ *Han kom etterpå* (etter et tidspunkt i fortiden)
 - ▶ *Du skal komme da* (på et omtalt tidspunkt i framtiden)
- ▶ **Måtesadverb** uttrykker måten noe blir gjort på
 - ▶ *Hun gjennomgikk pensum **stykkevis***
- ▶ **Gradsadverb** uttrykker mengde, intensitet eller grad ved verbhandlingen
 - ▶ *Jeg fryser **litt***
 - ▶ *Nå har du tullet **nok***

Preposisjoner: Funksjonsord klassen – *ved, på, under, i, foran, av*

1. Ubøyelige
2. Kjerne i preposisjonsfraser, tar substantiv
3. Betegner relasjoner, f.eks.:
 - ▶ **rom** og **tid**
 - ▶ *Hytta ligger **ved** sjøen*
 - ▶ *Elevene var svake **i** engelsk*
 - ▶ *Vi drar **i** mai*
 - ▶ *Taket **på** huset ble nettopp reparert*
 - ▶ **måte** eller **middel**
 - ▶ *Hun satt **i** dype tanker*
 - ▶ *Hun åpnet døren **med** en rusten nøkkel*
 - ▶ **verbalpartikkel**
 - ▶ *De sovnet **inn***
 - ▶ *Han brøt **sammen** etter løpet*



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, får betydning fra sammenhengen (*konteksten*)
 - ▶ **Jeg** liker grammatikk
 - ▶ **Man** skal respektere hverandre
 - ▶ **Hvem** tok vesken?
-
- ▶ **Personlige pronomener:** (*jeg – meg*), (*vi – oss*)
 - ▶ **Refleksivt pronomener:** *seg*. Har antesedent i samme setning, oftest subjektet i setningen
 - ▶ **Resiproke pronomener:** *hverandre*. Uttrykker en gjensidig relasjon.
 - ▶ **Interrogative pronomener:** spørreord (*hvem, hva*).

Determinativ (artikler) – *min, din, denne, alle, noen*

1. Bøyning i kjønn og tall (min bil, mitt hus)
2. Bestemmer til substantiv
3. Bestemmer, spesifiserer substantivets referanse

3 hovedtyper:

- a) Possessiver: angir eiendom eller tilhørighet, bøyes i person. (*Det er **min** bok*)
- b) Demonstrativer: viser til eller peker på en bestemt person eller ting som kan iakttas eller er omtalt. (**Den** *hytta ligger fint til*)
- 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*)



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 og adjektivfrase)
 - ▶ *Han var på ski **og** hun var i kirken* (setning og setning)

Subjunksjoner – *å, at, om, som, før*

(en måte å danne leddsetninger = små setninger i andre setninger for å utvide setninger)

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*



Quiz!



- ▶ En **vakker** fugl suste over taket. **adjektiv**
- ▶ **Viste** du henne veien til butikken? **verb**
- ▶ De kalt katten **Sia**. **substantiv**
- ▶ **Sinte** vepser fløy etter meg. **adjektiv**
- ▶ **Jenta** sprang alt hun orket for å rekke bussen. **substantiv**
- ▶ Han **løp** opp den bratte bakken. **verb**
- ▶ Jeg **fisker** i dag. **verb**
- ▶ **Skoleklokka** ringer inn til time. **substantiv**
- ▶ **Ole** er en kjekk kar. **substantiv**
- ▶ **Greie** gutter vasket gulvene. **adjektiv**
- ▶ En hel horde med syklister raste **gjennom** tunet vårt. **preposisjon**
- ▶ **Når** du skal hente telefonen din på rektors kontor, er det nok best at jeg blir med. **subjunksjon**
- ▶ Vil du ha melk, **eller** vil du heller ha te? **konjunksjon**
- ▶ **Hva** er klokka? **pronomen**

- ▶ Reven gjemte seg **under** trappa. **preposisjon**
- ▶ Er det **meg** du vil gå tur med i dag? **pronomen**
- ▶ Ugla satt **på** greina. **preposisjon**
- ▶ Hun går i klassen **over** meg. **preposisjon**
- ▶ **Du** er jammen snill! **pronomen**
- ▶ Hun fløy av sted **som** en rakett. **preposisjon**
- ▶ **Jeg** spiser gjerne fisk til middag. **pronomen**
- ▶ Lars sang **vakkert!** **adverb**
- ▶ **Sitt!** sa Magne til hunden. **verb**
- ▶ Han er **vanligvis** en morsom taler. **adverb**
- ▶ Stolen er god **å** sitte på. **subjunksjon**
- ▶ **Man** bør være forsiktig med hva man sier. **pronomen**
- ▶ Du var tidlig oppe, **men** du hadde visst glemt at du hadde fri denne dagen. **konjunksjon**
- ▶ Den stakkars læreren så ikke annet enn **trøtte** elever. **adjektiv**



► **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 (pronomen)

► **Innholdsord vs. funksjonsord**

- **Innholdsord:** substantiv, verb, adjektiv
rikt betydningsinnhold,
 - **Funksjonsord:** mer allment betydningsinnhold. Finnes fremst i de lukkede ordklassene.
- Ikke helt én-til-én, feks hjelpeverb.



- ▶ Modellere språklig kunnskap → trenger språklige data
- ▶ Språkteknologi: programmer som generaliserer over språklige mønstre
 - ▶ Korpusdata helt sentralt
- ▶ Et korpus (tekstkorpus) er en strukturert samling av tekster
- ▶ Elektronisk lagret



- ▶ 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 (1 mill. ord), Brown-korpuset (tagget versjon), Switchboard, ATIS (sample)
 - ▶ Ordklassetagger, syntaktisk struktur (trær som representerer frasestruktur)

- ▶ Norsk dependenstrebant (2014)
 - ▶ Trebant for norsk
 - ▶ Utviklet ved Nasjonalbiblioteket
 - ▶ Manuelt tagget (lingvister, 2år)
- ▶ 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 nøyt be ent

- ▶ NoReC korpus (2018): The Norwegian Review Corpus (Velldal, Øvrelid, Bergem, Stadsnes, Touileb, Jørgensen)

▶ 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 ./.

 - ▶ DT : determiner
 - ▶ JJ : adjective
 - ▶ NN : noun, singular or mass
 - ▶ VBD : verb, past tense
 - ▶ IN : preposition or subordinating conjunction
 - ▶ NNS : noun, plural

- ▶ There/EX are/VBP 70/CD children/NNS there/RB

 - ▶ EX : existential there
 - ▶ VBP : verb, non-3rd person singular present
 - ▶ CD : cardinal number
 - ▶ RB : adverb

- ▶ Tagging følger en manual
- ▶ Noen avgjørelser er 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 substantivfrase. *Around* tagges som adverb i betydningen 'omtrent'



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



Tokenisering: 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?*.
- ▶ Tegnsetting forekommer også innad i ord: *m.p.h. cap'n, AT&T.*
- ▶ Tall kan inneholde komma: 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*)

- ▶ Input: streng av ord og en spesifisert mengde tagger
- ▶ Output: en tagg per ord

<i>Jeg</i>	<i>vil</i>	<i>drikke</i>	<i>kaffe</i>	<i>nå</i>
pron	verb	verb	subst	adv

- ▶ Flertydigheter?

<i>Jeg</i>	<i>vil</i>	<i>drikke</i>	<i>kaffe</i>	<i>nå</i>
pron	verb	verb subst	subst	verb adv

- ▶ Tall fra det engelske Brown-korpuset:
 - ▶ 12% av ordtypene er flertydige
 - ▶ 40% av tokens er flertydige
- ▶ De *fleste* engelske ord er *entydige*
- ▶ Men mange av de mest *frekvente* ordene er *flertydige*
- ▶ Heldigvis er ikke alle lesninger like sannsynlige
 - ▶ Både isolert sett og i kontekst.



- ▶ Bør alltid først definere en *baseline*:
Enklest mulige tilnærming til et problem.
- ▶ To ulike majoritets-baserte baselines for PoS-tagging:
 - 1: Tildel alle ord samme tagg; den mest frekvente (NN)
 - ▶ ca. 13% korrekt taggedde ord (Brown)
 - 2: Tildel hvert ord dets mest frekvente ordklassetagg.
 - ▶ Dersom vi lagrer de 100 mest frekvente ordene og deres tagger får vi
 - ▶ ca. 46% korrekt taggedde ord (Brown)



- ▶ To hovedkategorier:

1. Regelbaserte taggere:

- ▶ Manuelt definerte regler for å tildele ord riktig tagg i en gitt kontekst.
- ▶ Eksempel: *drikke* er substantiv, og ikke verb, dersom det følger et adjektiv.

2. Statistiske taggere:

- ▶ Bruker et (manuelt) ordklassetagget korpus ('treningskorpus') til å beregne en statistisk model for tagging.

Typisk to trinn, grovt sett:

1. Morfologisk analyse:

- ▶ Hvert ord tildeles en liste av mulige ordklasser og morfologiske trekk.
- ▶ 'Multitagging'
- ▶ To tilnærminger:
 - ▶ Fullformsleksikon: Lister med ord i alle bøyninger (løp, løper, løpt, ...), med tilhørende tagger.
 - ▶ To-nivå morfologi: morfologisk analyse med en *finite-state transducer* som mapper fra overflateform til leksem.

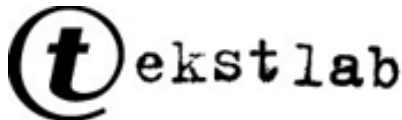
2. Entydiggjøring:

- ▶ Håndskrevne regler (gjerne mange tusen) for å disambiguere ordene.
- ▶ *Constraint Grammar* (CG) – sentral regelformalisme som har resultert i taggere for en rekke språk, deriblant engelsk og norsk.

- ▶ **Oslo-Bergen taggeren** ('OBT'): PoS-tagger for norsk.
- ▶ Constraint Grammar (CG)-regler for entydiggjøring.

```
"<som>" SELECT:3261 (prep) IF
    (1 pron-akk)
    (NOT 1 pron-nom)
;
# "Ei jente som (prep) meg"))
```

- ▶ Utviklet hos **Tekstlaboratoriet**.
- ▶ Kombinert med statistisk entydiggjøring.





- ▶ Bruker et **ordklassetagget korpus** ('treningskorpus') til å beregne den mest sannsynlige sekvensen av tagger for en gitt setning.
- ▶ En mye brukt probabilistisk model: **Hidden Markov Model (HMM)** – Tagging som klassifiseringsoppgave: Gitt en sekvens med ord, hva er den mest sannsynlige taggsekvensen?
- ▶ Ser på ordtaggene som *skjulte variabler* (eller 'tilstander') som vi ønsker å predikere basert på de *observerbare variablene*; ordene.
- ▶ Nære bånd til n -grammodeller.



- ▶ Gitt at vi har trent en modell (en PoS-tagger, språkmodell, ...),
- ▶ hvordan kan vi **evaluere** den?
- ▶ Ønsker å kvantifisere kvaliteten på testdata så vi kan sammenlikne ulike modeller.
- ▶ Generelt to stragier for å evaluere en modell:
- ▶ **Ekstrinsisk** og **intrinsisk**.



- ▶ Vi evaluerer modellen 'indirekte' utfra hvordan den påvirker resultatene for en annen oppgave.
- ▶ **Oppgavedrevet.**
- ▶ F.eks se hvordan en språkmodell påvirker maskinoversettelse, talegjennkjenning, osv.
- ▶ **Fordel:** kan teste modellen i samme kontekst som vi vil bruke den.
- ▶ **Ulempe:** ofte krevende ift tid/ressurser.



- ▶ Bruker et mer direkte mål for hvor bra modellen er på oppgaven den ble trent for.
- ▶ **PoS-tagging**: ønsker en modell som predikerer taggene for et testkorpus med høyest nøyaktighet; $accuracy = \frac{\#riktig}{\#tokens}$
- ▶ **Fordel**: ofte rask og billig.
- ▶ **Ulempe**: ikke alltid samsvar mellom ekstrinsiske og intrinsiske mål.



- ▶ Dersom vi tester på treningsdataene får vi urealistisk gode resultater sammenliknet med om vi tester på 'nye' data.
- ▶ Kalles *overfitting* dersom en model er for spesifikt tilpasset testdataene til å gi representative målinger for hvordan modellen generaliserer til usette data.
- ▶ Trenger minst to datasett: **treningsdata** og **testdata**.
- ▶ Bruker ofte også en tredje splitt: valideringsdata (*development data*).
- ▶ Viktig at datasplittene er balanserte og representative:
- ▶ F.eks samme sjanger, domene, osv.

Fra <http://www.nltk.org/book/ch05.html>

```
>>> import nltk
>>> from nltk import word_tokenize
>>> text = word_tokenize("And now for something completely
different")
>>> nltk.pos_tag(text)
[('And', 'CC'), ('now', 'RB'), ('for', 'IN'), ('something',
'NN'), ('completely', 'RB'), ('different', 'JJ')]

>>> text = word_tokenize("They refuse to permit us to
obtain the refugee permit")
>>> nltk.pos_tag(text)
[('They', 'PRP'), ('refuse', 'VBP'), ('to', 'TO'),
('permit', 'VB'), ('us', 'PRP'), ('to', 'TO'), ('obtain',
'VB'), ('the', 'DT'), ('refugee', 'NN'), ('permit', 'NN')]
```




```
>>> nltk.corpus.brown.tagged_words()  
[('The', 'AT'), ('Fulton', 'NP-TL'), ...]
```

```
>>> nltk.corpus.brown.tagged_words(tagset='universal')  
[('The', 'DET'), ('Fulton', 'NOUN'), ...]
```



```
>>> from nltk.corpus import brown
>>> brown_news_tagged =
brown.tagged_words(categories='news', tagset='universal')
>>> tag_fd = nltk.FreqDist(tag for (word, tag) in
brown_news_tagged)
>>> tag_fd.most_common()

[('NOUN', 30640), ('VERB', 14399), ('ADP', 12355), ('.',
11928), ('DET', 11389), ('ADJ', 6706), ('ADV', 3349),
('CONJ', 2717), ('PRON', 2535), ('PRT', 2264), ('NUM',
2166), ('X', 106)]
```



Ingen forelesning

<https://nettskjema.no/a/125851>

Åpent fra 26 september kl10:00 til 09 oktober kl10:00.