Innholdsfortegnelse

1. Hva er lingvistikk

lingvistikk er vitenskapelige studie av menneskelige språk. Kunnskap om hvordan ord settes sammen til setninger.

Lingvistiske nivåer

Fonologi: man jobber med lyd. se på kombinasjon av lyder og systemer og regelmessigheter rundt det.

Fonetikk: man jobber med lyd. beskrive fysisk hvordan lydene ytres. hva som skjer inne munnen for eksempel.

Morfologi: Dreier seg om formene til ord og hvordan ord er bygd opp.

Syntaks: studiet av hvordan fraser og setninger konstrueres og forholdet mellom ordene/ studiet av prinsipper og regler for setningsdannelse.

Semantikk: handler om språkets bokstavelige betydning ikke om hvordan mottaker eller ytrer tenker, dette tar pragmatikken for seg. Setter søkelys på å analysere betydning både til ord, fraser og setningen og hvordan man tenker. (NER

Pragmatikk: konteksten er viktig. hva betyr denne ytringen.

2. Språklige data, morfologi

Språkteknologi: systemer som generaliserer over språklige mønstre
Neurolingvistikk: lingvistisk fagområde som studerer mekanismene i den menneskelige hjerne som kontrollerer språk. Prøver å lokalisere språk i hjernen.

Brocas afasi: Ugrammatisk språk, problemer med forståelse av syntaktisk komplekse konstruksjoner

Wernickes afasi: Semantisk usammenhengende, men stort sett syntaktisk korrekt

Et korpus: er en strukturert samling tekster som er elektronisk lagret. Et korpus må konstrueres slik at det er representativt.

Annotering: manuelt fordele data og å lage treningsdata til maskinlæring.

Stemming eller lemmatisering: reduksjon til baseform.

Morfologi: hvordan ord er bygd opp, bøyes, og dannes.

Tokenisering: dele opp en tekst til løpende ord.

Tokenisering problemer:

- Punktum: F.eks.
- Apostrofe: I'II
- Bindestrek: Oslo-borgeren
- Mellomrom: New York
- Tid: 10:25

Morfemet: elementær minste enhet. Det vil si at er en del av ordet.
ord har intern struktur som er regel styrt.
Ord kan består av flere meningsbærende enheter.

Vi har 2 hovedtyper morfemer:

1. **Frie morfemer:** ord som har betydning uten å trenge å koble seg til andre ord. (ord)
2. Bundne morfemer: affikser:
    a. Prefikser: de små morfemene som kobler seg i foran i et ord.
    b. Suffikser: de små morfemene som kobler seg i bak i et ord.
    c. Infikser: morfem er inne i et ord. (i noen språk)
    d. Sirkumfikser: affiks som har to deler, en som settes i begynnelsen av ordet, og en som settes på slutten

Rot: er et ord element som ikke kan deles opp i mindre deler.

Innholdsord: Innholdsord er ord med en egen betydning og eget innhold, for eksempel substantiv, verb og adjektiv. Innholdsord er en åpen klasse og kan derfor ta inn nye ord - det kan med andre ord oppstå nye innholdsord etter hvert som språket utvikler seg. solen, fin, dame, Gunnar, morgen, nå (som verb).

Funksjonsord: Funksjonsord er ord med lite eget innhold, men som ofte har en grammatisk funksjon, for eksempel preposisjoner, artikler, pronomen og konjunksjoner. Funksjonsord er en lukket klasse som svært sjeldent(/aldri) tar inn nye ord. jeg, om, nå (som tidsadverb).

Avledning: En avledning er et ord som er dannet fra et annet ord ved hjelp av et Avledningsaffiks. Avledningsbasen kan være et rotord eller en avledning. Avledningsaffiksene er bundne morfemer med klart semantisk innhold.

Avledningsaffikser:

- U- negasjon: umulig, uvel og urolig.
- For- foran: forelese, forbokstav, fornavn.
- -er- den som utfører handlingen: fisker, baker.

Det er siste del av ordet som bestemmer ordklasse, derfor endrer ikke prefikser ordklassen

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

Forskjellen mellom bøying og avledning:

- Ved bøying skifter ordet aldri ordklasse. Ved avledning skifter ordet som oftest ordklasse.
- Alle prefikser er avledningsaffikser.
- Suffikser kan brukes til bøying og avledning.
- Bøying er mer produktiv.
- Bøyningssuffikser i norsk har alltid svakt trykk (bilen, spiste), mens avledningssuffikser kan ha sterkt trykk (sentral) eller bitrykk tenkbar
- Bøyningsendelser ligger alltid i slutten av ordet, men avledningsendelsene kommer tidligere (når vi har begge deler) gal+skap+en

Avledning:
- Danner nye leksemer ved hjelp av affikser som tilfører ordet en ny betydning.
- kan danne leksemer av en annen ordklasse enn det opprinnelige leksemet.
- Eksempler: forelese [V] → foreleser [S], forelese [V] → forelesning [S]

Bøyning:
- Danner kun nye "varianter", kalt bøyninger, av leksemer, ikke nye leksemer.

For eksempel:
- Verb kan bøyes i tid (være, er, var...)
- Adjektiv kan bøyes i grad (kald, kaldere, kaldest)
- Substantiv kan bøyes i tall og bestemthet (stein, steinen, steiner, steinene)
- Endrer ikke ordklasse.
- Eksempler: forelesning → forelesningen (subs. bøyd i bestemthet), snakke → snakker (verb bøyd i tid)

Sammensetninger: ord som består av deler som hver for seg også er egne ord. Etterleddet bestemmer vanligvis ordklasse. De fleste sammensetninger er determinative: etterleddet gir hovedbetydning, mens forleddet avgrenser. bilhjul, hjulbåt.

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: ett morfem = én betydning

Bøyningsspråk: ett morfem kan ha flere betydninger.

3. Regulære uttrykk

Regulære utrykk er standard notasjon for å karakterisere tekstsekvenser. Den blir brukt for å spesifisere tekststrenger i alle slags type tekstprosessering og informasjonsutvinning.
Et regulært utrykk er en beskrivelse av en mengde strenger.

Regulære utrykk består av:

- Strenger bestående av tegn eller ord: /a/, /informatikk/.
- Disjunksjon:
    - Vanlig disjunksjon: /spise|ete/, /penge(r|ne)/
    - Tegnklasser: /[Dd]en/, /m[ae]nn/, /bec[oa]me/
    - Intervaller: [A-Z], [a-z], [0-9]
- Negasjon: alt uten om.
    - [^b]
    - [^A-Z0-9]
- Tellere: uten parenteser → siste tegn. Med parentese → hele ordet.
    - **?** fanger opp (0 eller 1)
    - ***** fanger opp (0 eller flere)
    - **+** fanger opp (minst 1 eller flere
    - **.** fange alt
    - **.*** fanger alt mellom to ord

- «wildcard». For et hvilket som helst tegn.
- Ankere: spesielle tegn som forankrer det regulære utrykke til spesifikt sted i strengen
    - Begynnelsen av linjen
    - Slutten av linjen

liste = ['6 januar 1992', '25 juni 2005', '3342 november 1988', '9 fredag 1999', '1 mars 2018', '30 juli 190221']

$regex \rightarrow [([1-9]||[12][0-9]|3[01])\backslash s$
$(januar|februar|mars|april|mai|juni|juli|august|september|$
$oktober|november|desember)\backslash s\ ((19|20)\backslash d\{2\})$

Den fanger:

- 6 januar 1992
- 25 juni 2005
- 1 mars 2018

/informatikk/	Streng	
/spise| ete/	Spise eller ete	
/[Dd]en/	Den eller den	
[A-Z], [a-z], [0-9]	Random mellom intervallet	
^	Negasjon	
?	0 eller 1	
*	0 eller flere	
+	Minst 1	
.	Hva som helst	
/^Den/	I staten av setningen	
/ der\. $ /	På slutten av linjen	
\b /\b the \b /	Ordet the	
\B /\B the \B /	Alt unntatt ordet the	
/spis(e|er|te)/		

4. Sannsynlighet og språkmodeller

Utfallsrommet er mengden Ω av mulige utfall. I For eksemplet med terningkast:

$\Omega = \{1, 2, 3, 4, 5, 6\}$

Hendelse: en delmengde av utfallsmengden, $A \subseteq \Omega$. F.eks.:

- A = {1, 2, 3}
- B = {5}

I Sannsynlighet for en hendelse: en verdi mellom 0 og 1, gitt ved P:

- $P(A) = 0.5$
- $P(B) = \frac{1}{6}$

Dersom alle utfall er like sannsynlige, har vi en **uniform distribusjon**:

- P(A) = |A| |Ω|

Sannsynlighetene for alle mulige utfall **summerer til 1**:

- $\sum_{A \in \Omega} P(A) = 1$

Markovantagelsen

De $n - 1$ siste elementene lar oss tilnærme effekten av å betrakte hele sekvensen.

Eksempel for $n = 2$: $\boldsymbol{P(w_1^k) = \sum_{i=1}^{k} P(w_i | w_{i-1})}$

Eksempel:

P (jeg, vil, drikke, kaffe, nå) = P(jeg) P (vil | jeg) P (drikke | vil) P (kaffe | drikke) P (nå | kaffe)

$$\boldsymbol{P(I| < s >) = \frac{antall forkommelse\ av\ (< s > |I)}{antall\ forkommelse\ av\ (< s >)}}$$

Glatting

Omfordeling av sannsynlighetsmassen for å unngå noen av problemene med MLE: Sørg for at alle n-grammer får frekvens > 0

$$\boldsymbol{P(w_n | w_{n-1}) = \frac{C(w_{n-1}, w_n) + 1}{C(w_{n-1}) + V}}$$

V er antall ordtyper.

5. Ordklasser og ordklassetagging

Ordklasser: bindeledd mellom ordet og setningen. Sier noe om hva slags kontekster et ord forekommer i og sier noe om uttale.
Taksonomi: et system som har kategorier som er uttømmende, gjensidig utelukkende, styrt av et prinsipp

Kriterier for ordklasseinndeling

1. Formelle eller morfologiske kriterier
    - Hvilke bøyingsformer har ordet.
        - Hare – haren. Redd – reddere
2. Funksjonelle eller syntaktiske kriterier
    - Hvordan kan ordet kombineres med andre ord
        - En hare. En redd hare
3. Betydningsmessige eller semantiske kriterier
    - Hva er typisk betydning hos ord i ordklassen
        - Redd: egenskap. Hare: dyr og levende vesen

Ordklasser

1. **Substantiv:** olje, bil, jente og gutt
2. **Verb:** sparke, sove og hjelpe
3. **Adjektiv:** rød, snill og vanskelig
4. **Adverb:** her, ofte, derfor, trolig, ikke, kanskje, nå og vanligvis.
5. **Proposisjon:** ved, på, under, i foran og av.
6. **Prononomen:** Jeg, han, meg, seg, hverandre, hvem og man.
7. **Determinativ:** min, din, denne, alle og noen.
8. **Konjunksjon:** og, eller, men, for og så.
9. **Subjunksjoner:** å, at, om, som, før.

Åpne ordklasser: substantiv, verb og adjektiv. inneholder mange tusen ord, kan enkelt fylle på med nye. Eksempel: nye bilmodeller - nye farger (brannbilrød)
Lukkede ordklasser: inneholder mange færre ord enn de åpne kan ikke fritt skape nye ord gjennom orddannelse (pronomen)
Innholdsord: substantiv, verb, adjektiv rikt betydningsinnhold
Funksjonsord: mer allment betydningsinnhold. Finnes fremst i de lukkede ordklassene.

Flertydighet

Et ord kan ha flere betydninger. Mange av de frekvente ord er flertydige.

Hovedkategorier for ordklassetaggere

Regelbasert 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 Trinn

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 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.

Statistiske taggere: Bruker et (manuelt) ordklassetagget korpus ('treningskorpus') til å beregne en statistisk model for tagging.

- 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.
- I økende grad: nevrale modeller som brukes til ordklassetagging

Evaluering

For å evaluere om modellen vi har laget er bra. Vi har 2 strategier for å evaluere en modell:

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

Datasplitter

- 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.

6. Syntaks og kontekstfrie grammatikker

Syntaks er studiet av hvordan setninger bygges opp av ord og ordkombinasjoner.
Rekkefølgen på ord er viktig.
Setninger: inneholder et verb og som regel et subjekt

Flertydighet: Flere mulige grupperinger av ord → Ulik gruppering gir ulik betydning.

Konstituenter: grupperinger av ord i en setning, fungerer som en enhet.

Konstituenttester

«Hunden lekte i hagen»

- **«Stå alene» testen:** sier noe om kan denne ord grupperingen stå alene som svar på et spørsmål. Hvor lekte hunden? I hagen.
- **«Erstattes med pronomen» testen:** et ord kan erstatte en hel gruppe med ord når de opptrer som en enhet. Hunden lekte der.
- **«flyttes som enhet» testen:** kan flyttes rundt i setningen. Kommer i ulike rekkefølge. I hagen lekte hunden.

Fraser: bygger opp setningen eller andre fraser. Et enkelt ord kan bygges ut til en gruppe ord, slik at denne nye gruppen har samme funksjon i setningen. Kan bestå av hode og modifiserende ord både foran og etter. Vi har ulike frase typer:

- **NP (substantivfraser):** hodet er et substantiv. Fungerer typisk som subjekt eller objekt i setningen.
  - Determinativ + substantiv: en hund
  - Egennavn: Barack Obama
  - Pronomen: han, hemme
- **PP (preposisjonsfraser):** hodet er en preposisjon.
  - Prep + NP: over the rooftops
  - Foranstilt et adledd: Dypest ned I skuffen
- **AdjP (adjektivfraser):** hodet er et adjektiv.
  - Foranstilt et adledd: almost crazy, pretty big
  - Crazy about dogs
- **VP (verbfraser):** hodet er et verb.

Kontekstfrie grammatikk

Formell modell som fanger inn konstituentstatus og rekkefølge. Brukes mye innenfor lingvistikk. Fungerer best for språk som engelsk, med nogenlunde fast leddstilling. De fleste moderne lingvistiske teorier inneholder en form for kontekstfri grammatikk.

CFG består av fraser og ordklasser, ord og regler

$$< N, \Sigma, R, S >$$

- N: mengde ikke terminale kategorier
- Σ: mengde terminale kategorier
- R: mengde regler på formen A → α
    - A er ikke en terminal
    - α er en streng av symboler hentet fra mengden n (Σ ∪ N). Det vil si terminale og ikke terminale kategorier.
- S: start symbol

Rekursjon: en setning inne i en setning. Mekanisme som lar oss utvide fraser. Trenger mekanisme som kan skape uendelige strukturer. Rekursive strukturer: inneholder en delstruktur av samme type som helheten.

Chunking: Dele inn setningen inn i en sekvens, chunk. En chunk inneholder et hode. Ikke-rekursive: en chunk kan ikke inneholde en chunk av samme kategori.

7. Maskinlæring og klassifisering

Statistisk klassifisering

- Sentral metode innenfor maskinlæring.
- Automatisk avgjør hvilken kategori en observasjon tilhører.
- Basert på annotert treningsdata: observasjoner der kategorien er kjent.
- Supervised klassifisering: klassifisering som bruker annotert treningsdata.

Supervised Machine learning

- **Train:** data å trene modellen
- **Dev:** data for å evaluere modellen underveis
- **Test:** data for å evaluere den beste modellen etter trening og evaluering.

Naive Bayes-klassifisering

For å finne en klasse $\hat{c}$ (hentet fra alle mulige klasser C) for en trekkvektor $\vec{f}$ må vi beregne den mest sannsynlige klassen, gitt vektoren

$$\hat{c} = \underset{c \in C}{argmax}\, P(c|\vec{f})$$

Men det er problematisk å trene direkte: "sparse data"-problemet (alltid finnes både ord og sekvenser vi ikke har sett.)

Bayes regel

Betinget sannsynlighet

$$P(A|B) = \frac{P(A,B)}{P(B)}$$

$$P(A,B) = P(A|B)P(B)$$

$$P(A|B) = \frac{P(B|A)P(A)}{P(B)}$$

$$\hat{c} = \underset{c \in C}{argmax}\, P(c) \prod_{j=1}^{n} P(f_j|c)$$

1. **prior-sannsynligheten for klassen p(c)**

$$p(c) = \frac{N_c}{N_{doc}}$$

N_c = antall dokumenter i treningsdataen som er i klassen c

N_{doc} = totalt antall dokumenter

2. **sannsynligheten for individuelle trekk $P(f_j|c)$**

$$P(f_j|c) = \frac{count(f_j, c)}{count(c)}$$

For å beregne $P(f_j|c)$ kan vi anta at et trekk er ett ord som finnes i dokumentets «bag of words» og kan derfor heller bergene følgene:

V hvor mange ord det er i treningsdata

Hvor mange ord det er i hver klasse

Evaluering

Lage contingency tabell

Accuracy: hvor mange ganger modellen har klassifisert riktig

Precision:

Recall:

$F_1 = \frac{2PR}{P+R}$

8. Semantikk

Semantikk: Studiet av betydning slik det uttrykkes gjennom språk. Betydningen til morfemer, ord, fraser og setninger.

Leksikal semantikk (ordsemantikk)

Semantiske trekk

Egenskaper eller deler av ords betydning som uttrykker vår kunnskap om hva ordet betyr

Leksikal relasjoner

En annen måte å beskrive et ords betydning er å beskrive hvordan det forholder seg til andre ords betydning

Leksikale relasjoner: Hvordan betydninger relaterer til hverandre

Homonymi: Urelaterte betydninger av samme fonologiske ord

- Samme ordklasse:
- Ulik ordklasse:

Polysemi: flere betydninger, men betydningene er relatert

Synonymi: Ulike ortografiske ord, men med samme, eller meget lik, betydning

Antonymi: motsatt betydning

- **Enkel antonymier:** negativ av det ene medfører positiv for det andre (død/levende)
- **Gradérbare antonymer:** negativ av ene medfører ikke nødvendigvis positiv av andre
- **Reverser:** relasjon mellom ord som betegner bevegelse i motsatt retning
- **Konverser:** relasjon mellom ord som beskriver samme situasjon fra forskjellig synspunkt.
- **Taksonomiske søstre** ord som er på samme nivå i en taksonomi (rød, blå og gul)

Hyponymi: Inkluderingsrelasjon: et hyponym inkluderer betydningen til et mer generelt ord

- Hund og katt er hyponymer av dyr

Meronymi: Relasjon mellom del og helhet

Setningssemantikk

Formell semantikk: logikk for å representere betydning til setningen.
Sannhetsverdi: hvorvidt en setning er sann eller ikke.
Sannhetsbetingelser: hvilke betingelser som må til i den virkelige verden for å gjøre en setning sann.

Entailment: Semantisk relasjon mellom setninger, uavhengig av empiriske fakta
En setning p medfører ("entails") en annen setning q dersom det er slik at når p er sann er q sann og når q er usann så er p usann

Formell semantisk analyse
Semantisk analyse ved oversettelse fra et språk til et universelt metaspråk. Førsteordenslogikk er et slikt metaspråk.

Semantisk rolle
Semantiske roller beskriver den semantiske relasjonen som argumenter har til handlingen beskrevet av verbet. Hvilke roller de forskjellige deltagerene inntar.

- **Agent:** den som setter i gang en handling, i stand til å handle med viten og vilje.
- **patient:** entiteten som påvirkes av en handling, gjennomgår ofte en forandring
- **theme:** entiteten som blir beveget av en handling eller hvis beliggenhet beskrives
- **experiencer:** er bevisst på handlingen eller tilstanden, men er ikke i kontroll over den
- **beneficiary:** entiteten som en handling utføres for
- **instrument:** middelet som gjør at en handling kan utføres eller finner sted
- **goal:** entiteten som noe beveger seg mot
- **source:** entiteten som noe beveger seg fra

9. Semantikk i språkteknologi

Tre nivåer av betydning:

1. **Ord: Word Sense Disambiguation (WSD)**

   Aktivt felt innenfor språkteknologi. gitt en setning med et spesifikt målord og en liste med Betydninger. angi korrekt betydning for målordet i den setningen.

   Klassifisering basert på et annotert datasett.

2. **Fraser: Named Entity Recognition (NER)**

   Egennavn inneholder viktig semantisk informasjon. Automatisk gjenkjenning og kategorisering av egennavn. I Vanlige kategorier: person, organisasjon og sted.

   - Samme navn kan referere til forskjellige entiteter av samme type
   - Samme navn kan referere til entiteter av forskjellig type
   - JFK – presidenten og hans sønn → JFK – flyplass.
   - Trenger manuelt annotert korpus

   Løsning: ord-for-ord klassifisering:
   - Metoder for sekvensklassifisering
   - BIO-klassifisering: taggen indikerer om ordet befinner seg i begynnelsen (B), innenfor (I) eller utenfor (O) et egennavn, samt indikerer kategori.

3. **Setninger: Semantic Role Labeling (SRL)**

   Gitt et predikat i en setning, finn dets semantiske roller.

   Også her har vi behov for et manuelt annotert korpus.

WordNet

Manuelt konstruert database. Betydningen til ord karakteriseres gjennom relasjoner til andre ord. Semantiske konsepter karakteriseres gjennom relasjoner til andre konsepter.

Relasjoner i WordNet

- Mellom ord:
  - Synonymi (samme betydning).
  - Synonymi-relasjonen grupperer ord i synonymimengder, såkalte synsets.
- Mellom konsepter:
  - Hyponymi (mer generell, mer spesifikk).
  - Varierer noe, men antonymi og meronymi er også spesifisert for noen synsets.

Problem med semantiske roller

Ikke full enighet rundt rolleinventaret. Vanskelig å formulere formelle definisjoner av roller. generaliserte semantiske roller. Semantiske ressurser med informasjon om semantiske roller: PropBank og FrameNet

PropBank

Korpus som inneholder alle setningene i Penn Termbank. Annotert med informasjon om semantiske roller.

10. Språkteknologiske applikasjoner

Komposisjonalitet
komposisjonell
Vi forstår en frase eller setning på grunnlag av hvordan mindre deler (ord, fraser) er satt sammen. Betydningen er gitt ufra de enkelte delene og reglene som styrer hvordan de settes sammen.

https://snl.no/semantikk.

må leser før eksamen

Hva vil det si at to setninger står i en entailment-relasjon? Illustrér svaret ditt med minst ett eksempel?

En entailment-relasjon mellom to setninger innebærer at den ene setningen medfører den andre. Den er gitt ved lingvistisk informasjon (enten leksikal eller syntaktisk) og trenger ikke å avgjøres empirisk (ved å sjekke fakta i verden).

Eksempel leksikal entailment:
 Terroristen myrdet statsministeren Statsministeren er død

Eksempel syntaktisk entailment:
 Egypterne bygget pyramidene
 Pyramidene ble bygget av egypterne

Forklar kort hva BIO står for ogGi eksempler på 4 typer trekk som kan brukes for å løse denne oppgaven.

.BIO-klassifisering: taggen indikerer om ordet befinner seg i begynnelsen (B), innenfor (I) eller utenfor (O) et egennavn, samt indikerer kategori.

2. Trekk kan være:

- ordform (tokenisering): of, George, Washington, led
- lemma: of, George, Washington, lead
- shape: lower, capital, capital, lower
- affikser: of, rge, ton, ead
- ordklasse: IN, NNP, NNP, VBD
- chunk-kategori: PP, NP,
- navneliste: 0, 1, 1, 0