

Semantikk i KUNSTIs prosjekter

Diskusjonsinnledning til program møte i Bergen 18.11.03

Helge Dyvik

Betydning og fortolkning

Det er lenge siden semantikk ble identifisert som et viktig element i språkteknologien og dens forløpere. En del av den famøse ALPAC-rapportens diagnose av maskinoversettelsens bedrøvelige tilstand på 60-tallet var at maskinoversettelsesforskningen hadde støtt på en "semantisk barriere". Semantikkløs manipulasjon av ordformer og syntaktiske strukturer lot til å ha møtt veggen (eller i hvert fall en vegg); erkjennelsen var at man bare kunne komme videre hvis maskinene på en eller annen måte kunne bringes til å "forstå" de språklige uttrykkene de prosesserte. Dette tilsa systematisk arbeid med semantikk.

Utviklingen etter ALPAC på dette området kan kanskje oppsummeres i to punkter. For det første er formell semantikk for naturlige språk blitt et rikt forskningsfelt, og det er utviklet et mangfold av teoretiske redskaper for presis beskrivelse av semantiske egenskaper ved språklige uttrykk. For det annet er det blitt tydelig at den semantiske barrieren ikke var den høyeste likevel, i hvert fall ikke hvis vi definerer semantikk snevert som et studium av sannhetsbetingelser, eller eventuelt litt videre som et studium av den informasjon som ligger kodet i språkets ord og setninger ved konvensjon. Det overskyggende problemet fremstår som det Martin Kay kaller "The Resolution Problem", og som man kanskje kort kan karakterisere som gapet mellom det å forstå hva ord og uttrykk betyr, og det å oppfatte det tilsiktede budskapet i en situert ytring.

En skriftlig eller muntlig ytring er en kommunikativ handling plassert i tid og rom. Fortolkningen av en ytring bygger ikke bare på kunnskap om semantikken i de språklige uttrykkene som er valgt, men også på antagelser om intensjonen bak ytringen, på forutsetninger som ligger i konteksten og på generell kunnskap om verden. En lærdom innenfor såvel språkteknologi og datalingvistik som generell AI siden ALPAC er for det første hvor sterkt underspesifisert og flertydig den semantisk kodete informasjonen er, og for det annet i hvor begrenset grad den nødvendige ytterligere kunnskapen om kontekst og verden er formaliserbar. Dette avgrenser rommet for det prinsipielt mulige innenfor språkteknologien i betydelig grad – i hvert fall for det som er mulig med noe som ligner på de tilnæringsmåtene vi kjenner idag. Maskinoversettelse, for eksempel, vil aldri kunne utkonkurrere menneskelig oversettelse med noe som ligner på dagens metoder. Grunnen er at oversettelse utført av mennesker ikke kan *reduseres til* en formell relasjon mellom tekster, der vi bare forstår tekstene som sekvenser av gjenkjennbare symboler. Oversettelse er snarere en relasjon mellom interpretasjoner, eller fortolkninger, der selve tekstene bare er en del av det fortolkningene bygger på. I tillegg bygger fortolkningene ikke bare på språkkunnskap, men også på et stort forråd av vanskelig definerbar bakgrunnskunnskap. To tekster fungerer da mot hverandre som original og oversettelse når de interpretasjonene de gir opphav til *i den aktuelle sammenhengen*, korresponderer med hverandre, og dette synes bare i begrenset grad å være en klart komputerbar relasjon.

Språkteknologiens oppgave må derfor være å finne interseksjonen mellom det mulige og det nyttige. Til tross for de nevnte motforestillingene er grensene for det mulige tilstrekkelig utydelige til at dette er en spennende oppgave.

Bokstavelig betydning

Problemene med å formalisere generell bakgrunnskunnskap medfører da at språkteknologien hovedsakelig må konsentrere seg om det språklig kodede – om den *bokstavelige betydning*. Dette er som kjent et problematisk begrep. Enkelte forkaster helt begrepet om bokstavelige meningskonstanter, og påstår at mening alltid er situasjonsavhengig og gjenstand for forhandling. Så pessimistisk på semantikkens og språkteknologiens vegne er det ikke grunn til å være. Selvsagt bidrar ordvalg og setningsvalg i en dialog til det formidlede budskap, i tillegg til trekk ved diskurssituasjonen – ellers ville jo språket være overflødig. Dette innebærer med nødvendighet at der må være betydningsegenskaper som er konvensjonelt assosiert med de valgte tegn, og som ikke er avledet av den partikulære konteksten, men er konstante fra en kontekst til en annen. Disse meningskonstantene kan vise seg å være vage og underspesifiserte, men de må være der. *Forhandling* om betydninger forekommer selvsagt, men svært omfattende forhandling om hva uttrykk skal bety, kan knapt skje hvis kommunikasjon overhodet skal finne sted. Hvis vi ikke uten videre kan stole på en felles forståelse av det store flertall av ord og uttrykk vi bruker, har vi knapt noe å forhandle *med*.

Et begrep om bokstavelig, eller situasjonsuavhengig, betydning er således uunngåelig. En ganske annen sak er det å trekke grensen mellom det bokstavelige og det situasjonsavhengige i praksis. Semantiske språkressurser som ordnett eller flerspråklige leksika forsøker å trekke denne grensen en gang for alle. Det er gode grunner til å tvile på at det kan bli helt vellykket. Riktignok klarer vi oss ikke uten et begrep om meningskonstanter, men samtidig må vi erkjenne at det neppe finnes noen objektiv og klar distinksjon som vi kan etablere en gang for alle mellom egenskaper som kan tilskrives språktegn som typer – altså konstante egenskaper – og egenskaper som bare kan tilskrives tegnforekomster i bestemte kontekster. Det er en slags skala her. Hvis vi studerer de mulige fortolkningene et språklig uttrykk har i en viss tekst i en bestemt kontekst, og så gradvis tar mer og mer generelle *typer* av tekster og kontekster i betraktning og ser på hvilke fortolkningsmuligheter som da gjenstår, så vil det gradvis virke mer og mer rimelig å se på de mulige fortolkningene av uttrykket som noe som tilhører det som *type*, sett i isolasjon, som på forhånd gitte mulige betydninger av det, snarere enn som kontekstgenererte interpretasjoner. Men vi finner neppe noe veldefinert sted der vi krysser grensen mellom de to måtene å betrakte det på.

Vi kan betrakte et eksempel: I EØS-avtalen er ordet ‘carrier’ oversatt med ‘utøver av transportvirksomhet’. En god oversetter søker å gjengi interpretasjon i kontekst, ikke bare språklige betydninger. Spørsmålet er nå om man bør si at ‘utøver av transportvirksomhet’ er en av de mulige betydningene av det engelske substantivet ‘carrier’, eller om det ikke heller er slik at konteksten gjør det klart at det er slike ‘carriers’ vi snakker om her, slik at oversettelsen her gir språklig form til noe som bare ligger i konteksten i originalen. Mitt poeng er at dette spørsmålet ikke uten videre har noe absolutt svar. Grensen mellom det bokstavelige og det kontekststilhengige er uunnværlig, men den må trekkes *relativt* til noe. Man kunne si det slik: Grensen mellom bokstavelige meningskonstanter og kontekstindusert mening

må trekkes relativt til identifikasjonen av det *språk* en gitt tekst er forfattet i. Språk som norsk, engelsk, norsk lovspråk, oslomål osv. er ikke veldefinerte størrelser med klare grenser som vi bare kan gå ut og oppdage. Det er et uunngåelig element av *konstruksjon* i vår isolasjon av et språk – selv om der også er åpenbare begrensninger på hva vi kan tillate oss å betrakte som et naturlig språk. For eksempel er et naturlig språk noe som nødvendigvis er felles for en gruppe individer. Konsekvensen for ‘carrier’-eksempelet er at vi må bestemme oss for om vi vil anta at EØS-avtalen er forfattet i almen engelsk eller i et spesialisert subspråk av engelsk, spesielt for en viss type lovdokumenter. Vi kan altså enten betrakte lovdomenet som en spesiell kontekst for almen engelsk, som gir opphav til visse kontekstuelle fortolkninger av ord og uttrykk, eller vi kan anse dokumentene som skrevet i et spesielt fagspråk, med spesielle betydninger assosiert med et spesialisert vokabular. Noe av det som er kontekstavhengig fortolkning under første alternativ, blir bokstavelig betydning under annet alternativ. Valget mellom disse to synsmåtene er ofte ikke et spørsmål om sant og usant, men et *spørsmål om hensikten med beskrivelsen*. I språkteknologiske applikasjoner – f.eks. dialogsystemer som skal kunne omhandle flyruteinformasjon el.l. – velger man gjerne ekstreme versjoner av annet alternativ, og koder en god del informasjon om det spesifikke bruksdomenet i den semantiske beskrivelsen av ord og setningskonstruksjoner. Dette er én strategi for å redusere "the resolution problem": Gå langt i å behandle kontekstuell informasjon og kunnskap om verden som om det var en del av semantikken i ord og uttrykk. Prisen er et språk med et svært snevert anvendelsesområde – med den medfølgende risiko for at brukeren kan bli forledet til å tro at han står overfor et system med stor almenspråklig kompetanse, og så bli urimelig skuffet.

Konsekvensen av dette er at man må ha et dynamisk syn på utviklingen av semantiske språkressurser. Almenspråklige ordnett, f.eks., bør kunne suppleres med mer domene- eller teksttypespesifikke ordnett. Dette understreker viktigheten av å utarbeide metoder og redskaper for delvis automatisk utvikling av språkressurser på grunnlag av et gitt språklig materiale, og ikke bare basere seg på statiske, manuelt utviklede ressurser.

Ulike krav til semantikk

Nå er semantikk en mangfoldig disiplin, og ikke alle aspekter av den er relevante for alle språkteknologiske applikasjoner. En hovedskillelinje går antagelig mellom informasjonssøkning i databaser og oversettelse.

Ved dialogsystemer med sikte på å fremskaffe informasjon fra en eller annen database er det viktig å vite *hva det er snakk om*. Vi må der kunne finne den intenderte referent til et uttrykk, og det intenderte påstandsinhold i en setning. Det er den intenderte referanse som står i fokus i slike systemer: Koreferanse blir en langt viktigere relasjon enn semantisk nærhet som f.eks. synonymi. (Som kjent kan ikke-synonyme uttrykk godt være koreferente, f.eks., i en gitt situasjon, ‘Kongen av Norge’ og ‘mannen i den ukledelige boblejakk’.) Dette gjør metoder fra modellteoretisk semantikk potensielt relevante: Vi trenger en modell av det omtalte domene, og en komposisjonell semantikk som tillater oss å regne ut sannhetsbetingelser. Videre trenger vi en setningsovergripende analyse av diskursanaforer som *han, hun, den* osv., stadig med sikte på å bestemme referenter. Her kan dynamiske tilnæringsmåter som f.eks. diskursrepresentasjonsteori benyttes. Den leksikalske semantikken vil på sin

side normalt kunne være snever og skreddersydd, med en granularitet og struktur som i stor grad er bestemt av granulariteten og strukturen i den aktuelle databasen.

Ved oversettelse, derimot, blir de semantiske kravene annerledes. Semantisk nærhet blir her viktigere enn koreferanse. Det kan være en interessant øvelse å lete i oversettelser etter eksempler på det motsatte, altså at kunnskap om koreferanse snarere enn semantisk nærhet var det som avgjorde valget av oversettelse. Det følgende er et eksempel fra Teknisk dokumentasjon for brannvannssystemet på Gullfaks A-plattformen, kap. 2:

E: The inhibition of *algae growth* is achieved by sodium hypochlorite injection into the pump suction.

N: For å hindre *tilgroing av systemet* injiseres natriumhypokloritt i vannløpet til pumpen.

Algae growth ('algevekst') og *tilgroing av systemet* er åpenbart ikke synonyme uttrykk: Det er lett å tenke seg kontekster der de ikke ville kunne brukes om samme referent (sml. konge-eksempelet tidligere). Samtidig gjør den forutsatte domenekunnskap hos leseren det klart at de to uttrykkene refererer til samme fenomen i denne teksten. Eksempelet illustrerer at slike tilfeller typisk også vil være eksempler på oversettelser vi ikke vil anta at et maskinoversettelsesprogram ville klare, fordi de forutsetter kunnskap om verden – og da om en langt mer åpen verden enn den lukkede som utgjøres av en database. Et oversettelsessystem vil således først og fremst være avhengig av informasjon om semantisk nærhet, og dermed blir innhold viktigere enn referanse.

Tilsvarende blir sannhetsbetingelser mindre viktige: Flertydige eller underspesifiserte konstruksjoner behøver bare å bli disambiguert i den grad målspråket krever det. Et velkjent eksempel er kvantorrekkevidde, som ofte er flertydig. Stilt overfor en setning som "Har alle studentene tatt tre eksamener?", må et spørsmål-svar-system velge mellom tolkningen der de samme tre eksamenene tas av alle studentene ('tre eksamener' har videst rekkevidde) og tolkningen der bare antallet eksamener er felles for alle studentene ('alle studentene' har videst rekkevidde), mens et oversettelsessystem ikke behøver å bekymre seg om dette hvis heller ikke målspråket gjør det: Analysen kan der være underspesifisert. Oversettelsessystemet fordrer til gjengjeld et semantisk rammeverk som gjør det mulig å være underspesifisert på en effektiv måte: Prosessering av underspesifisert og partiell informasjon blir viktigere enn full artikulering av sannhetsbetingelser og full disambiguering.

Det er en viktig erkjennelse at den ønskede grad av underspesifisert, eller altså finkornetheten i semantikken, bestemmes av målspråkets egenskaper. Dette gjelder ikke bare på setningsnivå (som i det nevnte eksempelet), men også på ordnivå og bøyningsskategorinivå. Analyse av diskursanaforer er også nødvendig bare i den grad målspråket trekker flere distinksjoner – f.eks. genusdistinksjoner – enn kildespråket gjør i sine anaforer. Målspråkets egenskaper setter således grenser for hvor granulert den leksikalske semantikken i et oversettelsessystem behøver å være. Det betyr ikke at vi overfører målspråkets distinksjoner slavisk i vår beskrivelse av kildespråket – det ville være prinsipløst juks, med de upraktiske konsekvenser slikt gjerne har når man vil modifisere eller utvide systemene. Poenget er snarere at det ikke finnes noe entydig svar på spørsmålet om hvor mange underbetydninger et ord har. Dette er ikke noe vi kan liste opp en gang for alle – og vi ser at forsøk på å gjøre det, gjerne ender med svært ulike løsninger (sammenlign f.eks.

betydningsinndelingene i Princeton WordNet med dem i Merriam-Webster's Thesaurus). Granulariteten er også her et spørsmål om formål. Igjen er konklusjonen at vi trenger en dynamisk tilnæringsmåte til utvikling av språkressurser, f.eks. semantisk klassifiserte leksika: Snarere enn et statisk leksikon trenger vi metoder til å strukturere leksikalsk informasjon på nytt i lys av nye behov – f.eks. nye målpråk for oversettelse.

På denne bakgrunn kan vi blant annet diskutere følgende spørsmål i tilknytning til KUNSTIs prosjekter: Hvor mye har våre semantiske problemstillinger til felles? I hvilken grad har vi interesse av de samme teoretiske tilnæringsmåtene? I hvilken grad stiller vi likeartede krav til fremtidige semantiske ressurser for norsk?