

LOGON - Leksikon, Ordsemantikk, Grammatikk og Oversettelse for Norsk

Dette er et samarbeidsprosjekt mellom flere miljøer ved UiB, UiO og NTNU. Målet er å forene krefter for å nå noen av de målene som stilles i programplanen for KUNSTI. Spesielt vil vi:

1. Arbeide med et anvendelsesområde knyttet til maskinoversettelse. Her vil vi søke å utvikle en funksjonell demonstrator. Anvendelsesområdet vil tjene som en test som relevansen av det mer teoretiske arbeidet vil kunne måles opp mot.
2. Styrke forskningen og den nasjonale kompetansen på flere områder innen datalingvistik og beslektede emner hvor det så langt ikke har vært stor aktivitet i Norge. For kompetanseoppbygningen vil tildeling av doktorgrads- og andre stipendier på bestemte prosjektdeler stå sentralt.
3. Utvikle ressurser for norsk språk som vil kunne gjenbrukes og i stor grad bli fritt tilgjengelige. Spesielt vil vi utvikle en leksikalsk database som vil inneholde forskjellige leksika tilpasset ulike lingvistiske modeller og anvendelsesområder og en norsk grammatikk med gjenbrukspotensial.

Prinsipper:

- Prosjektet organiseres rundt den funksjonelle demonstratoren for oversettelse.
- Hovedstrategien for prosjektet vil være transferbasert oversettelse med unifikasjonsbaserte grammatikker.
- I første omgang vil vi oversette fra norsk til engelsk.
- Vi vil organisere prosjektet i delprosjekt. Disse vil være av to typer. Delprosjekt av den første typen er nødvendige for å få til en kjernedemonstrator. Delprosjekt av den andre typen vil bidra til å forbedre demonstratoren, men ikke være nødvendige for en minimumsdemonstrator. De vil egne seg for doktorgrads- og hovedfagsprosjekt.

Selv om målet er å utvikle ny kompetanse vil det være viktig å utnytte de produktene vi har utviklet tidligere og den kunnskapen vi har:

- Norsk ordbank (Oslo)
- Tagger basert på 'constraint grammar', CG, med sammensetningsanalysator (Oslo – Bergen)
- Norsk Komputasjonelt Leksikon (NKL) og TROLL (Trondheim)
- NorGram, LFG for norsk (Bergen)

- Arbeid med HPSG (Tr.heim)
- Kunnskaper i syntaks og unifikasjonsbaserte formalismer (alle steder)
- Semantikk og representasjoner (alle steder)
- PONS, ulike nivåer i LFG-basert maskinoversettelse (Bergen)
- "Semantiske speil", flerspråklige korpora og ordnettrelasjoner (Bergen)
- Databaser (Oslo)
- Navnegjenkjenning (Oslo)

Vi har valgt å oversette fra norsk til engelsk og ikke omvendt fordi nytteverdien vil være størst den veien siden de fleste norsktalende leser engelsk, mens svært få engelsktalende leser norsk. Vi har valgt å konsentrere oss om en av retningene for å komme lengst mulig. Vi har valgt å bruke norsk bokmål og ikke også trekke inn nynorsk av samme grunn.

1. Prosjektets innhold

1.0 Oversikt

Et transferbasert oversettelsessystem kan skjematisk illustreres ved



Figur 1

S1 er kildeprakssetning, S2 er målprakssetning. T1 er transferrepresentasjon for kildepråk, T2 for målpråk. Her må en bestemme

1. Representasjonsformat for transfer: T1 og T2
2. Analysemodul
3. Genereringsmodul
4. Transfermodul

Vi har tilgang til to mer eller mindre ferdige modeller for oversettelse delvis utprøvd for andre språkpar. For begge har vi tilgang til programvare og ressurser utviklet for andre språk.

- A. LFG-modellen der det brukes LFG-grammatikker i analyse og generering og f-strukturer som transferrepresentasjoner.
- B. HPSG-modellen der en bruker MRS (minimal recursion semantics) som basis for transfer.

I maskinoversettelseslitteraturen er det beste nivået for transfer mye diskutert: syntaks eller semantikk. Vi ønsker i utgangspunktet å bruke et semantisk nivå. Samtidig mener vi at det kan være en del å vinne ved å bruke syntaktiske nivåer og utnytte språklikhet mellom nærstående språk. Vi vil eksperimentere med dette i prosjektet.

Vi vil i prosjektet forsøke å løsrive oversettelse og transferdelen fra bestemte grammatikkformalismer. I utgangspunktet vil vi bruke en norsk LFG for analyse, en engelsk HPSG for generering og MRS for transfer. Dette har flere grunner:

- Vi vil bruke alternative analysemoduler. I tillegg til LFG vil vi prøve ut en "shallow parser" basert på CG som et alternativ til den unifikasjonsbaserte parseren i analysemodulen.
- Vi har langt på vei en ferdig "kjernegrammatikk" for norsk i LFG, og det vil være mye å tjene på å bruke denne som analysator.
- Det er interessant å studere hvordan et semantisk nivå kan legges til LFG.
- Det er interessant å betrakte transferkomponenten som en selvstendig oppgave løsrevet fra grammatikkformalismer.
- Gruppen i Trondheim har startet arbeid med en HPSG for norsk. Vi har tilgang til en HPSG for engelsk, ERG, med en MRS-semantikk og en parser og generator for denne. Her har vi også tilgang til kildekode.

1.1 Representasjoner

Valget av transferrepresentasjoner, T1 og T2 på figur 1, vil stå sentralt i prosjektet. De vil være avgjørende for hvordan modulene for analyse, transfer og generering utformes, og tjene som grensesnitt mellom disse. De vil også være avgjørende for hvor godt oversettelsen lykkes.

Vi vil bruke en form for representasjoner fra formell semantikk og underbestemte representasjoner av disse. De senere årene er det foreslått flere forskjellige slike underbestemte representasjoner, bl.a. kvasilogisk form (QLF), situasjonsskjemaer, underbestemte diskursrepresentasjonsstrukturer (UDRS), MRS, robust MRS (RMRS). De ulike representasjonene varierer både med hensyn til den underliggende semantiske modellering: førsteordens logikk, situasjonssemantikk, diskursrepresentasjonsteori (DRT), 'event semantics', og hvordan de ivaretar underspesifisering og mulighetene for trinnvis spesifisering.

I prosjektet vil vi arbeide med å videreutvikle (logisk underbestemte) semantiske representasjoner og deres egnethet for oversettelse og generering. Vi vil også se på muligheten av å utvide representasjonene med mer leksikalsk semantisk informasjon enn det som har vært vanlig (se avsnittet om flertydighet nedenfor).

Arbeidet med representasjonene vil være et kjerneprosjekt som alle delprosjektene forholder seg til og bidrar til. Det vil løpe gjennom hele prosjektperioden. Det vil først og fremst bli gjennomført ved diskusjonssesjoner på prosjektsamlingene der de ulike delprosjektene rapporterer om sine erfaringer i forhold til ulike representasjoner.

I tillegg vil vi tidlig i prosjektperioden gjennomføre et komparativt litteraturstudium av de ulike tilnærmingenes egenskaper i forhold til oversettelse.

1.2 Analyse

For å beskrive de ulike modulene kan det være hensiktsmessig å ta utgangspunkt i en tilpasset oppdeling etter Marr:

A. Språkbeskrivelse
B. Formalisme
C. Algoritme
D. Implementasjon

Figur 2

For å forstå figuren kan en tenke på kontekstfrie grammatikker. Her er formalismen, B, en formalisme for kontekstfrie grammatikker og tolkningen av den. Nivå A er en bestemt kontekstfri grammatikk. Algoritmen, C, er en algoritme som prosesserer A, for eksempel en parsingalgoritme og D er en konkret implementasjon av denne algoritmen. Prinsipielt vil hvert nivå bare forholde seg til nivået rett over og under. Språkbeskrivelsen A tar bare hensyn til formalismen B og kan derfor lages uten hensyn til C og D. Unifikasjonsbaserte grammatikker som LFG og HPSG kan beskrives tilsvarende. De er i prinsippet reversible, generering kan ses på som en annen algoritme C med samme formalisme B.

For analyse vil vi som basis bruke grammatikken NorGram (nivå A). Den er basert på LFG som formalisme (nivå B). NorGram er et resultat av et treårig prosjekt og dekker de fleste sentrale konstruksjoner i norsk. Vi har fått fri tilgang til XLE-systemet som vil ta seg av parsing (C og D) og kan i første omgang konsentrere oss om den norske språkbeskrivelsen (nivå A). Det vi vil gjøre i prosjektet er:

1. Videreutvikle grammatikken mot ”test-suite” og oversettelseskorporus.
2. Bearbeide leksikon. Kople mot database.
3. Legge en semantisk komponent til grammatikken, som blir den største oppgaven her.

1.3 Generering

Her vil vi i første omgang ta utgangspunkt i HPSG-formalismen (nivå B), LKB-systemet som implementerer denne inkludert generering fra MRS-er (C og D) og den engelske grammatikken ERG (nivå A). Det vil være to typer arbeidsoppgaver:

1. Endringer til ERG (nivå A). Dette vil gå på utviding av leksikon siden vi vil studere et domene som ikke er dekket i ERG. Det kan også gå på modifikasjoner av den semantiske komponenten hvis vi reviderer representasjonsformatet (T2) i forhold til MRS.
2. Endringer av genereringsalgoritme og implementasjon av denne (nivå C og D). Igjen vil dette bli aktualisert hvis vi reviderer T2 i forhold til MRS-formatet.

Begge disse oppgavene vil vi knytte sammen med de tilsvarende oppgavene ved transfer.

1.4 Transfer

Her vil vi arbeide på alle plan. For det første på det formelle nivået (B). Skal transferkorrespondansen (mellom T1 og T2) uttrykkes som omskrivning av hele semantiske representasjoner, være drevet av leksikalske representasjoner, eller på andre måter. Her er det flere alternativ i litteraturen som til dels har gitt navn til hele paradigmet (constraint-based, lexical-based, rule-based, transfer-based). I første fase av prosjektet vil vi bruke omskrivningsregler omtrent som i VerbMobil. Samtidig vil vi vurdere alternativ og vil kanskje revidere før fase 2. Dette vil knyttes tett opp mot utviklingen av representasjoner (pkt. 1.1 over). Nært knyttet til dette blir arbeidet med å lage algoritmer og implementasjoner av det formatet som velges (nivå C og D). Vi vil knytte dette sammen med algoritmer for generering og implementasjon av disse.

Dernest vil vi utvikle transferreglene fra norsk til engelsk (nivå A). Vi regner her med å bruke en kombinasjon av manuelle og automatiske metoder. Først vil vi ekserpere par av ord der det engelske ordet er en mulig oversettelse av det norske fra korpus og ordbøker. Vi vil konstruere transferregler manuelt for noen ordpar. Dernest vil vi generere andre transferregler for liknede par fra disse. I tillegg er det selvsagt en hel rekke av manglende direkte korrespondanse mellom norsk og engelsk som må behandles manuelt (argumentdivergens, ”head”-divergens, strukturell divergens, tempusdivergens, ...).

1.5. Flertydighet

Den største enkeltutfordringen i datalingvistikk er flertydighet. Selv om en parser basert på en dyp grammatisk analyse effektivt fjerner mange lokale flertydigheter, som å skille mellom homonymer fra forskjellige ordklasser, står vi likevel igjen med:

1. De fleste ordstrenger som har en vellykket syntaktisk analyse, vil ha flere. Dette eksploderer for en del fenomen, som PP-tilknytning.
2. Innenfor en og samme ordklasse vil det kunne finnes homonymer og polysemer.
3. Et entydig ord i kildepråket kan ha flere alternative oversettelser i målpråket.
4. En generering fra en logisk semantisk representasjon vil ha en rekke ulike realiseringer i målpråket.

Arbeidet med flertydighet vil i hovedsak bli lagt til doktorgradsprosjekt. Litt forenklet kan en si at kjerneprosjektet vil tilstrebe ”recall” for oversettelsesrelasjonen, mens ”precision” overlates til doktorgradsprosjektene. Ettersom dette vil bli lagt til doktorgradsprosjekt er det litt tidlig å si nøyaktig hvilke ideer som vil bli forfulgt og realisert, men vi vil i hvert fall arbeide med flere av følgende alternative strategier:

- Rike leksikalske representasjoner. Vi vil her bygge på bl.a. SIMPLE-leksikonet som er delvis utbygget for norsk med bl.a. qualia-strukturer (Pustejovsky), tilnæringer fra TROLL og bruk av LKB i leksikon, og tror dette bl.a. vil kunne bidra med:
 - a. En bedre syntaktisk disambiguering
 - b. En viss ordbetydningsentydiggjøring (WSD).
 - c. Valg av oversettelsesekvivalenter.
- Statistiske metoder for entydiggjøring av ord. I denne sammenheng er det naturlig å se på problem (2) og (3) under ett. Spesielt vil vi se på metoder som kombinerer en tospråklig ordbok med et stort korpus for kildepråket og et sammenliknbart korpus for målpråket (altså ikke oversettelseskorpus siden vi ikke kan regne med å ha tilgang til tilstrekkelig store slike.)
- Metoden ’semantiske speil’ (Dyvik) som tillater derivasjon av semantiske nett – inndeling i ords underbetydninger og registrering av nær-synonymi- og hyponymi-relasjoner mellom slike betydninger – på grunnlag av parallellkorpora mellom to språk. Vi vil anvende metoden på en tospråklig ordbok, der ord med mengdene av deres oversettelser kan ekserperes direkte. Resultatet av metoden vil brukes til å forbedre ordvalg ved oversettelse.
- I formell semantikk kan en skille ut to ulike tilnæringer for å videreutvikle en logisk basert semantikk til bedre å fange inn leksikalske relasjoner og semantisk finstruktur. Den ene kan kalles representasjonell. Her innfører en rikere representasjoner i det formelle språket en benytter seg av, jfr. pkt. A. Den andre kan kalles geometrisk eller modellerende. Her innfører en mer struktur på de domenene som representeres, for eksempel en geometrisk struktur for tid eller rom. Det vil være interessant å studere hvordan en slik tilnærming kan nyttiggjøres i oversettelsesprosessen.
- Når det gjelder syntaktisk disambiguering gir XLE-systemet muligheter for å legge inn preferanser for analyser. Vi vil prøve å gjøre bruk av dette. Ideelt sett burde preferanser læres automatisk, men vi har ikke ressurser til å forfølge dette i prosjektet.

1.6 Forprosessering og robusthet

I Oslo har det over noen år blitt arbeidet med ”constraint grammar” (CG) og tagging, dels i samarbeid med Bergen. Spesielt er det blitt utviklet en CG-tagger som er testet ut på mange ulike teksttyper og det er utviklet flere spesialanalyser, blant annet en sammensetningsmodul som gjør det mulig å finne delene av et ord og tilskrive tagger til ukjente ord (med kjent siste ledd). Det er arbeidet videre med syntaktiske regler i CG (dependensanalyse) og navnegjenkjenning (”named entity recognizer”). Vi ønsker å bruke dette til å forbedre oversettelsessystemet.

For det første vil vi prøve å prosessere teksten med en morfologisk tagger før LFG-analysen. Dette vil forhåpentligvis forbedre resultatet og prosesseringshastigheten.

Norsk bruker ordsammensetning, engelsk skriver sammensetninger i flere ord. Dette byr på utfordringer for oversettelse som vi vil arbeide med.

Et overraskende stort problem ved oversettelse er egennavn. Dette går både på å se at noe er et navn og grensene for det (’*Den norske stats husbank* har vedtatt å øke renten’), skille mellom navn (’Så *Jan Tore Helge*?’), og avgjøre om det skal oversettes eller forbli uoversatt.

1.7 Alternativ og utvidelser

I forhold til grunnmodellen: semantisk basert transfer + unifikasjonsbasert analyse og generering, er det mange mulige alternative veier som kan være verdt å utforske. Vi vil konsentrere oss om noen.

For det første er det spørsmål om en behøver å gå hele veien til et semantisk nivå for å utføre transfer mellom så pass nærbeslektede språk som norsk og engelsk. I et tidligere prosjekt, PONS, eksperimenterte en av oss (Dyvik) ved å legge til et (underbestemt) semantisk nivå, situasjonskjemaer, til LFG, og brukte det som basis for oversettelse mellom norsk og svensk og mellom norsk og engelsk. Samtidig ble det utforsket systematiske metoder for å utnytte syntaktisk likhet og gjøre transfer på f-strukturnivå eller c-strukturnivå der dette var mulig. Det vil være interessant også i dette prosjektet å utnytte språklig likhet.

En alternativ tilnærming til å prøve å disambiguere mest mulig (se avsnittet over om flertydighet), er å observere at ved oversettelse mellom nærstående språk vil i mange tilfeller en streng med en rekke alternative analyser kunne oversettes inn i en streng med et korresponderende sett med analyser. Typisk er en del eksempler med PP-tilordninger mellom engelsk og norsk, som

Mary saw the man in the park with the binoculars.
Mari så mannen i parken med kikkerten.

Her har begge strenger 5 korresponderende analyser. Unifikasjonsbaserte grammatikker som LFG og HPSG vil først finne alle disse 5 ulike analysene. Moderne implementasjoner av disse modellene (XLE, LKB) gir mulighet til å representere de 5 analysene i en kompakt form. En mulighet vil være å bruke disse kompakte representasjonene (i praksis en kompakt representasjon av en disjunksjon av de ulike lesningene av setningen) som basis for oversettelse, og prøve å generere strenger i kildespråket som har analyser som gir opphav til en korresponderende mengde av lesninger.

Et annet alternativ, mer i ånden som ligger under underbestemte semantiske representasjoner, ville være å gjøre en partiell, eller underbestemt syntaktisk analyse, og fra denne avlede en underbestemt semantisk representasjon som svarer til de 5 ulike lesningene. Antagelig vil en dependensanalyse utført i CG i hvert fall kunne føre et steg på veien her. Vi vil eksperimentere med en slik CG-analyse som et alternativ til LFG-analysen. Dels vil det være interessant fordi en kan beholde en del underspesifiserhet. Men det vil også være interessant fordi en slik CG analyse ("shallow parsing") vil forventes å være mer robust, å ha større dekningsgrad. På den ene siden vil vi få to oversettelsessystemer som vi kan sammenlikne. På den andre siden vil det være interessant å prøve å bygge dem sammen slik at systemet velger den optimale muligheten. Vi vil også eksperimentere med en partiell oversettelse ved hjelp av LFG. Strenger parses nedenifra og opp ("bottom-up") og de største segmentene som kan tilskrives kategorier av grammatikken ("chunks") oversettes. Disse alternativene vil ikke inkluderes i kjernedemonstratoren, men de vil typisk egne seg for doktorgradsprosjekt.

Andre alternativ og utvidelser som kan være aktuelle er:

- Bruk av HPSG i analyse av norsk som et alternativ til LFG
- Oversettelse fra engelsk til norsk
- Oversettelse mellom norsk og andre språk enn engelsk

Disse oppgavene vil vi ikke inkludere i prosjektplanen men vi håper at noen av dem vil kunne realiseres gjennom andre prosjekter vi er involvert i eller samarbeide med, eller i (fremtidige) doktorgrads- eller hovedfagsprosjekt.

1.8 Leksikalsk database

Vi vil lage en database for leksikalske ressurser. Den vil tjene to formål. Den skal representere leksikaene for de ulike grammatikkene for norsk vi vil gjøre bruk av, og indeksere dem på en slik måte at de kan operere på det samme materialet. Dernest skal den representere et leksikon for engelsk og oversettelsesrelasjoner. Her vil det være en del ulike oppgaver.

1. Opprette og drive en database etter nærmere spesifikasjoner og utvikle egnede grensesnitt for brukerne.
2. En del leksika som alt er utarbeidet, som taggerleksikonet i Oslo, NordKompLeks i Trondheim, NorGram-leksikonet fra Bergen legges inn. Leksikonene flettes, dobbeloppføringer fjernes og det sørges for en hensiktsmessig felles indeksering.
3. Det er et mål å kople databasen mot de ulike grammatikkmodellene og formalismene som benyttes på en slik måte at ulike leksika kan avledes automatisk fra databasen og oppdateringer i databasen kan gjøres uniformt for alle formalismene.
4. Det engelske leksikonet som skal brukes for generering legges inn.
5. Transferleksikonet som utvikles på grunnlag av den tospråklige ordboken skal legges inn, presenteres på en hensiktsmessig måte og lenkes til de enspråklige leksikaene.

Her har vi relativt god oversikt over hva som skal gjøres på de to første punktene, og det er viktig å komme i gang raskt. Når det gjelder de tre neste punktene vil dette være mer eksperimentelt og måtte tilpasses underveis i samarbeid med de andre delprosjektene.

1.9 Resurser og evaluering

Vi vil konsentrere oss om et domene, turistinformasjon. Vi vil forsøksvis samle inn tekster oversatt fra norsk til engelsk fra dette domenet i et omfang av 300 000 ord. Vi vil også samle inn en del mer generelle oversatte tekster. Tekster hvor det finnes alternative oversettelser er spesielt interessante. Vi vil skille ut en del av korpuset til testformål.

Det resterende korpuset vil vi sammenstille på setningsnivå ("align") med program allerede utviklet av Knut Hofland. Dette vil vi bruke som grunnlag blant annet for statistiske studier av semantiske korrespondanser mellom språkene (jfr. avsnitt over om flertydighet). Som en del av disse prosjektene vil vi også se på muligheter for ord-sammenstilling ved at vi også trekker inn en-språklige korpora for de to språkene.

Fra korpuset vil vi trekke ut et treningskorpus på 500 setninger. Demonstratoren skal minst være i stand til fullt automatisk å oversette fra norsk til engelsk 95% av setningene i treningskorpuset slik at minst én akseptabel oversettelse er å finne blant de genererte oversettelsene av hver setning. Videre er det et mål at 80% av setningene i testkorpuset av samme art dekkes på tilsvarende måte.

Videre vil vi lage en "test suite" av norske setninger for parserne. Her vil vi ta utgangspunkt i treningskorpuset på 500 setninger, testsekvensen som alt er utarbeidet for NorGram og vi vil sammenlikne med "CSLI-test suite" for engelsk. I en slik "test-suite" vil det for hver streng legges inn hvor mange analyser den er forventet å ha. Dette kan også være ingen, altså "test-suiten" inneholder negative data. Her er det et mål at minst 95% av strengene får det korrekte antall analyser.

For parsing er det mulig å få en sikker standard for evaluering. Det er mulig å oppnå intersubjektiv enighet for hvor mange analyser en streng skal ha, gitt en analysemodell, og et program kan testes ut i fra det. Det er ingen tilsvarende standard for hva som er en god eller akseptabel oversettelse. Vi vil hovedsakelig prøve å utforme passende spørreskjemaer som kan prøves ut på passende grupper som behersker begge språk, særlig på personer med engelsk som morsmål, og ta deres bedømming som utslagsgivende for kategorier som *forståelig, korrekt engelsk, idiomatisk engelsk, meningsbevarende, etc.*

Vi har ikke planlagt å legge inn automatisk evaluering. Vi vet det foregår en del eksperimentering med dette, for eksempel BLEU, og det vil være interessant som et hovedfagsprosjekt å prøve dette og liknende tilnærminger ut på resultatet av vårt prosjekt.

Det er klart at innsamling av korpora bør skje tidlig i prosjektet, evaluering etter fase 1 og særlig ved slutten av prosjektet.

1.10 Anvendelse og sammensying

De ulike delene må sys sammen til en demonstrator, et system for automatisk oversettelse av tekster fra norsk til engelsk. Det er ikke å forvente at kvaliteten på den helautomatiske oversettelsen vil være akseptabel for oversettelse av tekster eller web-sider. Vi ser i hvert fall to mulige anvendelser. Kvaliteten kan være akseptabel som et første inntrykk for lesere uten kunnskaper i norsk ("browsing quality"). Det kan derfor være interessant å lage en web-oversetter som oversetter web-sider etter modell av andre slike på veven.

Den andre muligheten er å integrere den automatiske oversetteren i en oversettelsesbenk der en menneskelig oversetter bruker den som ett av flere verktøy. I første omgang er det mest realistiske at den menneskelige oversetteren

- gis muligheten til å velge alternativ på grunnlag av de som er automatisk generert
- rette på resultatet fra den automatiske oversetteren
- oversette manuelt der det ikke foreligger noen gode alternativ

Det å bygge slike omgivelser vil vi i denne omgang gjøre på eksperimentell basis som studentarbeid.

2. Organisasjon og gjennomføring

Prosjektet vil bli organisert som et kjerneprosjekt og flere mindre delprosjekter. Kjerneprosjektet vil inneholde alt som er nødvendig for å bygge en minimumsdemonstrator. De andre delprosjektene vil være relatert til kjerneprosjektet og bidra til å forbedre demonstratoren, men de vil ikke være nødvendige for en minimumsdemonstrator.

2.1 Kjernedemonstratoren

Demonstratoren er altså organisert rundt trinnene

1. Analyse av norsk i LFG/XLE (jf. 1.2 over)
2. Transfer ved MRS eller liknende (1.4)
3. Generering av engelsk i HPSG/LKB (1.3)

I tillegg inngår

0. Forprosessering (jf. 1.6)

Og det hele skal settes sammen til ett program (1.10). Kjerneprosjektet er også delt opp i mindre delprosjekt. En stor del av kjerneprosjektet utgjøres av delprosjekt 1 Representasjoner og demonstrator. Det har ikke vært naturlig å dele det opp ytterligere. Derimot er det en del andre oppgaver som er skilt ut som egne deler fordi de er relativt vel avgrenset med vel definerte oppgaver.

DELPROSJEKT 1: REPRESENTASJONER OG DEMONSTRATOR

Oppgaver:

1. Utforming av representasjonsformat.
2. Format for transferregler.
3. Algoritmer for transfer og implementasjon av disse.
4. Genereringsalgoritme og implementasjon.
5. Forprosessering
6. Sammensying av demonstrator.

Ansvarlig: Jan Tore Lønning

Kostnad:

- 1 forskerstilling fra sommeren 2003 til format for transferregler og generering, og algoritmer og implementasjon av disse (oppgave 2-4, 6)
- ½ forskerstilling gjennom prosjektperioden til sammensying, forprosessering og lignende (oppgave 5, 6) (arbeidssted Bergen)

DELPROSJEKT 2: ANALYSE AV NORSK I LFG

Ansvarlig: Helge Dyvik

Kostnad: ½ forskerstilling gjennom prosjektperioden

DELPROSJEKT 3: TRANSFERREGLER FRA NORSK TIL ENGELSK OG ENGELSK GRAMMATIKK

Ansvarlig: Lars Hellan

Kostnad: ½ forskerstilling gjennom prosjektperioden

DELPROSJEKT 4: LEKSIKALSK DATABASE

Ansvarlig: Jan Tore Lønning

Kostnad:

- 6 månedsverk til oppretting av databasen
- 6 månedsverk til innlegging og fletting
- 1 månedsverk per år senere år til teknisk drift
- 1 forskerårsverk til faglige løsninger (punkt 3 til 5)

DELPROSJEKT 5: INNSAMLING AV KORPORA

Ansvarlig: Janne Bondi Johannessen

Kostnad: 300 000

DELPROSJEKT 6: EVALUERING

Ansvarlig: Torbjørn Nordgård

Kostnad: 300 000

2.2 Forbedringsprosjekt

Delprosjekt av den andre typen vil bidra til å forbedre demonstratoren, men ikke være nødvendige for en kjernedemonstrator. De vil egne seg for doktorgrads- og hovedfagsprosjekt. De vil gjerne være tilknyttet ett av delprosjektene av den første typen. Mulige forbedringsprosjekt er nevnt underveis i teksten. Vi samler dem her.

Mulige doktorgradsprosjekt:

- A. Rike leksikalske representasjoner (SIMPLE, Pustejovsky, LKB) (se seksj. 1.5 over)
- B. Statistiske metoder for entydiggjøring av ord (1.5)
- C. 'Semantiske speil' (1.5)
- D. Geometrisk modellteoretisk semantikk som basis for ordsemantikk og transfer. (1.5)
- E. Bruk av syntaktisk likhet i oversetting. (1.7)
- F. CG som analysator som et alternativ til LFG. Hyperunderbestemt semantikk. Valg mellom ulike parse. (1.7)
- G. Oversettelse med "chunk-parsing" og LFG (1.7)
- H. Alternative transfermekanismer. (1.7)
- I. Sammensetninger og oversettelse (1.6)

Mulige hovedfagsoppgaver og studentprosjekt i tillegg til de ovenfor:

- J. Navnegjenkjenning og oversettelse (1.6)
- K. Automatisk evaluering. (1.9)
- L. Inkorporering i web-side (1.10)
- M. Inkorporering i oversettelsesbenk (1.10)

Utvidelser som input fra andre prosjekt.

- N. HPSG som analysator for norsk (1.7)
- O. Oversettelse fra norsk til engelsk (1.7)
- P. Andre språkpar (1.7)

Vi vil minst ansette 4 doktorgradsstipendiater, helst så mye som 5. Om det vil bli rom for 5 stipendiater, vil blant annet avhenge av hvor mye ressurser vi vil trenge til utvikling av kjernedemonstratoren. Vi har gjort en foreløpig fordeling her, men vil ha en mer detaljert gjennomgang i år 2003 og ta endelig stilling til dette. Vi har heller ikke bestemt endelig hvilke av de mulige doktorgradsprosjektene vi vil prioritere.

Doktorgradsprosjektene vil veiledes av erfarne forskere knyttet til prosjektet og disse vil stå som ansvarlige for hvert delprosjekt.

2.3 Felles ressurser

HOVEDFAGSSTIPEND

Det vil styrke prosjektet å styre hovedfagsoppgaver inn mot prosjektets tema. Det vil også virke rekrutterende til fagområdet. Vi vil til enhver tid å ha 3 hovedfagsstipend. Utgift: kr 255 000 per år

GJESTEFORSKERE

Vi vil satse på å ha to utenlandske gjesteforskere per år, hver i tre måneder. Utgift: kr 190 000 per år, start i 2003.

PROGRAMMERINGSHJELP

Det meste av programmeringen vil bli utført av forskerne tilknyttet delprosjekt 1, representasjoner og demonstrator. Vi har satt av noe mer midler til programmering i 2003, og en reserve til senere år. Om dette vil være tilstrekkelig, vil vi evaluere innen utgangen av 2003 når vi tar stilling til om vi skal ansette 4 eller 5 stipendiater.

DRIFT

Det vil trenge driftsmidler. Disse vil gå til møter mellom prosjektdeltagerne, kontakt med utenlandske forskere, deltagelse på konferanser og daglig drift. Kr 400 000 per år.

PROSJEKTLEDELSE

Til ledelse av prosjektet, mulighet for sekretærhjelp og evt. redusert undervisning setter vi av kr 50 000 per år.

2.4 Infrastruktur

Vi vil ha to felles samlinger for hele prosjektet årlig. Minst en av dem skal legges til et sted bort fra alle deltagernes hjem og hjemmeinstitusjoner for best mulig konsentrasjon. Den ene årlige samlingen (vårsemesteret) vil ha preg av vi drøfter pågående aktiviteter, den andre (høstsemesteret) mer preg av rapportering om hva som er gjort i året som ligger bak. I tillegg vil vi holde kontakt ved mer uformelle møter for eksempel i forbindelse med besøk av gjester ved en av institusjonene.

Vi vil satse på å gi stipendiatene en kullfølelse selv om de er spredt på flere steder. Det vil vi bl.a. prøve å få til ved å ha egne samlinger for dem, for eksempel i forbindelse med besøk av gjesteforskere og gjesteforelesere, eller ved felles deltagelse på nordiske forskerskoler.

Det er et mål at resultatene fra prosjektet publiseres i anerkjente internasjonale kanaler og legges frem på internasjonale og nordiske konferanser innen datalingvistikk, språkteknologi og maskinoversettelse. Interne vitenskapelige rapporter vil ikke være et mål i seg selv, men tjene som et steg på veien mot publisering og som et hjelpemiddel i å lage demonstratoren. Det er å forvente at alle stipendiater bidrar med minst en rapport i året (i forbindelse med høstmøtet). Det er også å forvente at delprosjektene i kjernen bidrar med minst en rapport per ansatt årsenhet.

Alle rapportene vil bli lagt ut på prosjektets offentlige web-sider. Vi vil også ha prosjektinterne web-sider for å legge ut mer foreløpig materiale, "test-suites", korpora, ideer til drøfting etc.

2.5 Styring

Prosjektet ledes av Jan Tore Lønning, Oslo. I hver av de to andre byene er det en plassansvarlig ("site manager"), Helge Dyvik i Bergen og Lars Hellan i Trondheim. Disse vil ha ansvar for administrative og økonomiske forhold på hvert sitt sted. Sammen med de andre delprosjektlederene utgjør de styret for prosjektet.

2.6 Framdrift og milepæler

Rammen for prosjektet er at det skal avsluttes i 2006, altså det skal vare i drøyt 4 år. Siden doktorgradsprosjekt skal vare i 3 år betyr det at alle doktorgradsprosjekt skal starte i 2003, eller senest 1.1.2004. Vi regner med at ansettelsestidspunktene vil fordele seg jevnt i denne perioden, altså at det blir halve utgifter til doktorgradsstipendiater i år 2003 og 2006 og fulle utgifter i 2004 og 2005. I år 2003 vil vi gjøre en ekstra innsats for å få strukturen rundt databasen i gang, og for å etablere korpora. I år 2006 vil vi bruke mer på evaluering og på å slutføre prosjektet.

Vi vil ikke nå innarbeide tidsplaner og framdriftsplaner for de enkelte doktorgradsprosjektene utover at de skal rapportere. Stipendiater som ansettes skal sammen med sine veiledere tidlig utarbeide prosjektbeskrivelser for sine doktorgradsprosjekt.

Når det gjelder arbeidet rundt kjernedemonstratoren, er det en avveining mellom det praktiske aspektet å bygge en demonstrator og teoriutvikling. Arbeidet med å bygge en demonstrator heller i retning av tidlig å bestemme formalismer og formater, mens teoriutviklingsaspektet taler for å revidere dette underveis. For å ta hensyn til begge deler vil vi arbeide i faser:

- Fase 1. Vi bygger en første liten demonstrator basert på LFG for norsk, HPSG for engelsk, MRS som transferrepresentasjoner og VerbMobil-formatet for transfer. (Dette går frem til og med mars 2004).
- Fase 2. Evaluering av resultatet fra fase 1. Vi reviderer transferrepresentasjoner og muligens også andre deler på grunnlag av erfaringer i fase 1. (april-september 2004)
- Fase 3. Vi bygger en større demonstrator på grunnlag av dette.

Resultatet av fase 1 vil være en første demonstrator. Vi vil plukke ut et treningskorpus på 100 setninger fra treningskorpuset på 500 setninger. Vi vil så utvikle et oversettelsessystem som dekker minst 90 av disse setningene. Vi vil her bruke MRS som transferrepresentasjoner og formatet for transferregler som er utviklet for disse.

I tillegg vil vi ha følgende mål for år 2003:

- Oppretting av database med tekniske løsninger.
- Innlegging av eksisterende leksika for norsk og sammenfletting av disse.
- Innsamling av første del av et oversettelseskopus.
- Oppretting av første versjon av "test suite" for norsk parser.
- Ansetting av doktorgradsstipendiater
- Utarbeiding av detaljerte prosjektbeskrivelser for ansatte doktorgradsstipendiater.

Vi vil også i løp av 2003 ha en nøyere planlegging av resten av prosjektfasen med klarere mål og milepæler for 2004-2006.

Vedlegg: LOGON2.xls, excel-ark som viser kostnadsplan.