Oversettingsteknologi i Norden

Rapport med artikkelbidrag

Arbeidsseminar om oversettingsteknologi i Norden
Språkrådet, Oslo
23. og 24. oktober 2008

Arbeidsgruppen for språkteknologi i Norden
Nordens språkråd
Mars 2009

Rapporten er redigert av Torbjørg Breivik
Arbetsgruppen för språkvård och språkteknologi i Norden
Nettverket for språknemndene i Norden
mars 2009

E-post: Torbjorg.Breivik@sprakradet.no

INNHOLD

Nordisk seminar om oversettingsteknologi

Torbjørg Breivik

Innledning

Seminaret her er det fjerde i rekken av arbeidsseminar som de nordiske språknemndenes arbeidsgruppe for språkteknologi arrangerer. Vi har vært i Finland (Pargas), i Sverige (Göteborg), i Danmark (København) og i 2008 var turen kommet til Norge og Oslo. Seminarene har med hensikt vært lagt opp slik at innledningene som presenteres, skal være korte og egnet til diskusjoner. Instruksen til innlederne har vært at de skal skissere aktuelle problemstillinger innenfor et avgrenset område av språkteknologien. Deltakerne inviteres og vi prøver å finne fram til deltakere fra forskning, produktutviklingsmiljø og språknemnder for at disse kan møtes og diskutere problemstillinger av felles interesse. Et viktig mål for seminarene er også å bidra til nettverksbygging. Årets seminar gikk over to halve dager, fra lunsj 23. oktober til lunsj 24. oktober 2008. Det var 27 deltakere og seminaret ble holdt i møtelokalene til Språkrådet i Oslo.

Tema

I 2008 var temaet oversettingsteknologi.

Globalisering betyr mer kontakt med andre kulturer og andre språk. Det stiller oss overfor utfordringer når det gjelder språk og kommunikasjon. I vår del av verden har vi vært vant til å bruke engelsk som et lingua franka, og til nå klart oss bra med det. Globaliseringen medfører at det oftere og oftere oppstår situasjoner det det ikke er tilstrekkelig at man bare behersker engelsk for å kommunisere. Internasjonal handel krever tilgang til, og kunnskap om, lover og avtaleverk på ulike språk. Det kreves oversetting og tolking dersom forhandlinger om kjøp og salg skal lede fram til avtaler. Det er avgjørende med raske og korrekte oversettinger, og disse oversettingene må ivaretas på en forsvarlig måte. Oversetting og tolking er ressurskrevende, og det er mange har tenkt at her må man kunne automatisere. De første forsøkene med automatisk oversetting ble gjort på sekstitallet, men fortsatt finnes det ikke programmer som er gode nok til at man kan stole fullt ut på dem eller kalle dem automatiske. Det varierer hvor mye etterarbeid som må gjøres, men man kan spare mye tid og arbeid med å benytte seg av programmer for maskinstøttet oversettelse.

Språknemndenes oppgave er å sørge for at programmene for maskinstøttet oversettelse og annen programvare holder høg språkfaglig kvalitet, og vår arbeidsgruppe konsentrerer oss spesielt om de språkteknologiske produktene som kommer på markedene i våre respektive land.

Politiske signal

Nordens språkråd har de siste fem-seks årene prioritert språkteknologi blant de viktigste innsatsområdene på den språkpolitiske arenaen. Det har vært satt i gang flere store prosjekt der målet har vært å utnytte den språkteknologiske kompetansen vi har i Norden til felles nytte. Vi kan nevne nordisk nettordbok, støtte til elektroniske, flerspråklige ordbøker, flerspråklig søking og utredningen SpråkVis som trekker opp linjene for en nordisk språkteknologisk satsing.

I Norge har man de siste årene fått en del politiske signaler om hvordan styresmaktene ser på språkteknologien. I den siste stortingsmeldinga om språk og språkpolitikk, Mål og meining (St.mld. nr. 35, 2007-2008), argumenteres det med at språkteknologien bidrar til å øke den

demokratiske deltakelsen blant innbyggerne, og bidrar til å gjøre informasjon og tjenester tilgjengelig for alle. Forutsetningen er selvsagt at produktene gjøres tilgjengelige på norsk. I nevnte stortingsmelding framheves etableringen av en norsk språkbank som det største og viktigste språk- og kulturpolitiske tiltaket i meldinga. Språkteknologiske produkt på morsmålet er helt sentrale i en kulturpolitisk sammenheng nasjonalt og internasjonalt.
For noen år tilbake ble samme budskap gitt i en annen stortingsmelding om norsk kulturpolitikk fram mot 2014: "Språkteknologifeltet kan vera ein av dei fremste arenaene der kampen om norsk språk og kultur vil utspela seg i tida framover." (St.mld. nr. 48 Kulturpolitikk fram mot 2014 (2002-2003), s. 196.)

Status for etablering av Norsk språkbank pr oktober 2008

Stortingsmeldinga Mål og meining slo fast at Norsk språkbank skal etableres (s. 136), og kulturminister Trond Giske trakk fram denne satsingen som et av de største og viktigste tiltakene for å realisere regjeringens språkpolitiske mål. Sommeren 2008 ble det skrevet en plan for oppstart inkludert et utkast til en norsk BLARK (Basic Language Resource Kit), og en foreløpig kartlegging av eksisterende ressurser. Statsbudsjettet som ble offentliggjort i begynnelsen av oktober, hadde en bevilgning på 2,5 MNOK til "kartlegging, analyser og eventuelle kjøp av bruksretter og bearbeidelse og utvikling av aktuelle språkressurser. Avhengig av framdrift og resultater [….], vil departementet vurdere å komme tilbake til saken i revidert budsjett." Revidert budsjett legges fram i mai 2009. Planen for oppstart av språkbanken kalkulerer med en kostnadsramme på 90 – 100 MNOK over seks år til etableringen og deretter 6 – 8 MNOK årlig til drift, vedlikehold, utvikling av verktøy og frikjøp av nye ressurser. Prioriteringen skal gjøres i tett samarbeid med brukerne og her vil styret for språkbanken legge føringer på prioriteringene. Fagmiljøene i Norge er små, men de har høy kompetanse innenfor hver sine områder, og det er viktig at språkbanken støtter opp om disse og bidrar til å utvikle dem videre. Språkbanken skal være en selvstendig institusjon med eget faglig styre, og den må opptre og fungere slik at den blir oppfattet som nøytral både av forskningsinstitusjonene og av kommersielle aktører på området.

Referanser:

Prosjektplan for norsk språkbank: "Samling og tilgjengeleggjering av norske språkteknologiressursar", Norsk språkråd, 2002
Stortingsmelding nr. 48 (2002-2003) Kulturpolitikk fram mot 2014, kapittel 12.9: Ein norsk språkbank.
Stortingsmelding nr. 35 (2007-2008) Mål og meining, kapittel 7.5 Språk og teknologi, særlig 7.5.6.2 (Ein norsk språkbank skal etablerast).

Oppsummering av arbeidsseminaret

Torbjørg Breivik

De nordiske språknemndenes arbeidsgruppe for språkteknologi arrangerte sitt fjerde nordiske arbeidsseminar i oktober 2008, og denne gangen var Språkrådet i Norge vertskap. Målsettingen for disse seminarene er å fungere som et møtested for personer som arbeider med ulike deler av fagområdene i språkteknologien. Temaet for årets seminar var oversettingsteknologi, og rapporten dokumenterer seminaret gjennom en oppsummering av diskusjonene, artikler fra de ulike innlederne, program og deltakerliste. Støttearkene for presentasjonene finnes i et eget vedlegg.

I Norden som ellers i den globaliserte verden, skjer mer og mer kommunikasjon på tvers av språkgrenser, og oversetting mellom språk blir nøkkelen til å forstå. Vi kan lære oss ulike språk, men ingen kan lære seg alle, og bare i Europa har vi et stort antall små og store språk. Samarbeid krever at man presenterer problemstillinger og rammeverk for hverandre, og resultatet nedfelles gjerne i dokument som seinere blir referanser for arbeidet. Juridiske avtaler gjøres og skal etterleves. Man kommer langt om man kan bruke et ”lingua franka”, men ofte må man sørge for å gjøre dokumentene tilgjengelige på deltakernes språk. En automatisering av oversettingsprosessen vil spare tid og penger for mange. På arbeidsseminaret inviterte vi deltakerne til å innlede til diskusjoner om hvor langt vi har kommet med maskinoversetting, prosessen bak det oversatte dokumentet og hvilke utfordringer vi står overfor på dette området.

Innleggene

Innleggene på seminaret tok for seg ulike aspekter ved oversettingsprosessen: hvilken støtte man kan få fra et oversetterprogram, hva oversetteren ikke kan forvente støtte til, hva de ulike typene oversettingsprogram baserer seg på (statistikk eller regler), hvilke typer oversettingsprogram som egner seg til ulike typer oversettelser, og ikke minst hva som kan eller bør gjøres av tilrettelegging for at oversettingsprosessen kan gjennomføres smidig og raskt.

Direktør Sylfest Lomheim i Språkrådet ønsket deltakerne velkommen. Han har selv undervist ved fagoversetterstudiet på Høgskolen i Agder (nå Universitetet i Agder), og kjenner problemstillingene for oversetterne godt.

Torbjørg Breivik fra Språkrådet ga en kort intoduksjon til seminaret inkludert en status for arbeidet med etablering av en norsk språkbank. Språkbanken er av interesse i denne sammenhengen fordi den bl.a. skal inneholde parallellkorpora til bruk i utvikling av oversettingsverktøy og verktøy til bruk i utvikling av nye og bedre oversettingsverktøy.

Anna Sågvall Hein har i mange år arbeidet med maskinoversetting i teori og praksis, og gav en oversikt over hvordan situasjonen er i dag sett fra hennes ståsted. En interessant problemstilling hun trakk opp var om statistisk basert programvare egnet seg bedre til en type oversetting, mens regelbasert programvare er bedre egnet til andre typer. Noen tester kan tyde på dette, men det foreligger ingen systematiske undersøkelser som kan bekrefte eller avkrefte dette.

Barbara Gawronska pekte i sitt innlegg på behovet for å ha tilgjengelig konkordanser og store korpora, en- og flerspråklige, når man arbeider med oversettelser. Det å kunne søke i

relevante, elektronisk tilgjengelige, kilder, er en forutsetning for gode oversettelser enten man oversetter manuelt eller benytter oversettingsprogrammer. Hun understreket også nødvendigheten oppdatert og relevant IT-opplæring for studenter som skal bli oversettere.

Carol B. Eckmann drøftet i sitt innlegg om valget av dataverktøy er med og påvirker kvaliteten på teksten i det ferdige produktet. Hun viste til praktiske erfaringer som tyder på at så er tilfelle, og hun etterlyste en grundigere diskusjon og undersøkelse av den delen av oversettingsprosessen som foregår fra det foreligger en råtekst til man har et ferdig produkt. Hun viste til at om oppdragsgiver er i stand til å ta vare på relevante dokument på en strukturert måte med tanke på gjenbruk, kan det spare oversetteren for mye ekstra arbeid. Det vil gjelde faguttrykk, terminologi og måten oppdragsgiver vil bli presentert på.

Krister Lindén tok utgangspunkt i EU og kravene EU-landene har til å få tekster på eget språk. Denne situasjonen tilsier at man må ha adskillige tolker og oversettere i EU-apparatet, og behovet for å effektivisere oversettingene ved hjelp av gode oversettingsprogram, er stort.

Daniel Hardt startet med et praktisk eksempel basert på et statistisk basert program.

Thomas Bilgram viste at maskinoversettingsprogrammer virker, og stilte spørsmål om hvor innsatsen nå må settes inn for å få tatt i bruk programmene i oversettingsbransjen. Hans innlegg ble på mange måter en oppfølging av Carol B. Eckmanns diskusjon om oversetterens forutsetninger og muligheter for å kunne levere et godt sluttprodukt.

Lars Nygård presenterte et program for oversetting mellom skandinaviske språk (norsk, svensk og dansk), og kommenterte hvilke forutsetninger som lå til grunn for at det kunne virke så godt som det gjør.

Eva Petterson presenterte et program for oversetting av akademiske kursplaner fra svensk til engelsk. Programmet er i bruk ved Universitetet i Uppsala og fungerer godt innenfor de fagområdene universitetet tilbyr kurs.

Til sist presenterte *Matthew McGowan* et søkeprogram som er basert på søking innenfor avgrensene temaer, og hvilke resurser som kreves for å få et flerspråklig søkeprogram til å fungere godt.

Diskusjonene

Noe av det viktigste som kom fram i diskusjonene, var at man bør se mer på hele oversettingsprosessen og tilrettelegge for oversetting fra dag en. Det å ha historikk tilgjengelig i form av elektroniske og søkbare (indekserte) dokumenter med mulighet for å hente fraser, idiomer og spesielle uttrykksmåter for et fagområde, effektiviserer prosessen betydelig.

Et annet diskusjonstema var om det er mulig å sammenligne ulike program for oversetting. Det er ikke noen reell sammenligning om man sammenligner et statistisk basert program med et regelbasert program. De er laget for og egner seg til ulike formål, noe som vil gi utslag i en test. Det burde likevel gjøres et arbeid med testing av programvare for oversetting for å vurdere hvor gode de ulike programmene er til ulike typer tekst. Det er ønskelig å få gjennomført en uavhengig evaluering av programmene, inkludert anvendelsesområde og forutsetninger for å virke godt. Arbeidsgruppa tar med seg forslaget og undersøker om det finnes muligheter for å realisere et slikt prosjekt.

Det er nødvendig å ha mer fokus på etterarbeidsfasen og kvaliteten på arbeidet som gjøres med ferdigstilling av et dokument. Dette ble understreket av flere av innlederne og tatt opp igjen i diskusjonene. Utdanningen av oversettere og translatører må inkludere opplæring i bruk av relevante dataverktøy, bruk av termbaser og anvendelse av parallellkorpus og konkordanser.

Evaluering:

Deltakerne ble bedt om å gi tilbakemelding til arbeidsgruppa på hva de syntes om seminaret, og de fleste var godt fornøyde. Deltakerne ble også bedt om å komme med forslag til forbedringer og temaer for kommende seminarer.

Når det galt forbedringer kom det fram følgende forslag:

- ønskelig at alle innlederne lager ppt-presentasjoner (lettere å forstå og følge med)
- ønskelig med en kort presentasjon av at deltakerne
- ønskelig med navneliste med institusjonstilhørighet (e-postlistene gir ikke nok informasjon om dette)
- mer brukerorientert, mer vekt på utveksling av erfaringer fra praktisk bruk av systemer
- ønskelig med navnelapper

Følgende forslag til temaer for nye seminarer kom fram:

- diskusjon om hvilke områder maskinoversetting er nyttig / brukelig vs ubrukelig
- evaluering av oversettelsesteknologi
- hvordan kombinere forskjellige typer språkteknologi
- språk og kommunikasjon
- Norden og de små språkene
- Nordens språkråd / Nordisk ministerråd som initiativtaker til deling av korpora, særlig fra og for offentlige instanser som trenger oversettinger

Sammendrag av innledningene

Maskinöversättning i teori och praktik. Var står vi i dag?

Anna Sågvall Hein

Maskinöversättning är det äldsta datorlingvistiska tillämpningsområdet med rötter tillbaka till mitten av 50-talet. Intresset för området och tron på maskinöversättningens möjligheter som ett komplement eller alternativ till mänsklig översättning har varierat under årens lopp. I början av 90-talet tog dock såväl forskning som tillämpning ny fart då man systematiskt började utveckla metoder för att på olika sätt återanvända tidigare översättningar. Det handlar både om att bygga översättningssystem som helt grundar sig på tidigare översatt text, s.k. statistiska översättningssystem, och att bygga upp och komplettera språkbeskrivningarna i språkvetenskapligt grundade översättningssystem, s.k. regelbaserade system.

De statistiska översättningssystemen visade sig kunna producera oförutsett bra översättningar, något som drivit utvecklingen framåt. En annan faktor som haft avgörande betydelse för utvecklingen av maskinöversättning är den ökande användningen av Internet och tillgången till språkliga data på nätet. Sedan något tiotal år har det funnits tillgång till fria översättningstjänster på nätet, som trots den bristande översättningskvaliteten fått stor användning. Redan i dag översätts mer text maskinellt än manuellt, och mer än 50 miljoner översättningsuppdrag körs dagligen via Internet (Jaap van der Meer 2008).

Våren 2008 inträffade en händelse av stor betydelse då Google släppte en testversion av en fri maskinöversättningstjänst, som utvecklats av Google Translation Center. Den omfattar mer än 30 olika språk och ett första allmänt intryck är att översättningarna är väsentligt mycket bättre än de fria översättningstjänster som tidigare funnits att tillgå på webben. Tjänsten är lätt att använda och man kan förvänta sig att användningen kommer att öka kraftigt. Även om översättningarna är bättre än många andra maskinella översättningar, så är de inte perfekta och kvaliteten varierar för olika språk. För att översättningarna ska uppnå publiceringskvalitet måste de efterredigeras.

Vilka förbättringar kan man då räkna med på sikt i Googles översättningar? De bygger på statistisk maskinöversättning som tränas på tidigare översättningar och som finslipas mot statistiska syntaxmodeller över målspråket. Erfarenheter från forskarvärlden visar att storleken på träningsdata är av avgörande betydelse för översättningsresultatet. Det är dock inte den enda framgångsfaktorn. Det är också viktigt att träningsdata hämtas från samma domän som den som systemet ska tillämpas på. Vidare spelar språkskillnader och språklikheter stor roll.

Den stora fördelen som företaget har gentemot andra utvecklare och forskare ligger i den närmast obegränsade tillgången till textmaterial på olika språk för träning av översättningsmodulerna. Finns det en gräns för hur långt man kan komma i fråga om att höja översättningskvaliteten i ett statistiskt system genom lägga på mer och mer träningsdata? Kan man hamna i en situation där mer data skapar fler alternativ och kvaliteten degraderar? Något som de statistiska systemen inte kommer åt är textberoenden över meningsgränserna, vilka många gånger är avgörande för korrekt översättning.

Det finns i dag publikt tillgängliga programvaror – open source – som kan användas för uppbyggnad av statistiska system. Det är en möjlighet som tagits tillvara av flera unga företag med goda resultat. Inom forskarvärlden ägnas mycken uppmärksamhet åt att förbättra de statistiska systemen genom att inkludera lingvistisk kunskap. Det är en svår uppgift, då systemen inte har några regler som kan förfinas utan består att sannolikheter som avkodas.

Framför allt försöker man finna metoder för att kompensera för brister i träningsdata genom att analysera dem lingvistiskt och därigenom utnyttja dem så effektivt som möjligt.

Hur ser det ut på marknaden? Där har hittills de regelbaserade systemen dominerat, men där kan man förvänta sig en förändring. För regelbaserade system ligger den stora svårigheten i att få språkbeskrivningarna heltäckande, en hart när omöjlig uppgift, så länge man inte använder sig av kontrollerade språk. För att kompensera för brister i språkbeskrivningen utvecklas olika upphämtningsstrategier, som utnyttjar statistiska metoder. Det är en metod som dagens Systran använder sig av och också det svenska företaget Convertus, som bland annat översätter kursplaner från svenska till engelska. En fråga man måste ställa sig är vilken roll de regelbaserade systemen spelar i dag och i framtiden.

Hur bra kan den maskinella översättningen bli? Hur bra behöver den bli för olika ändamål? Hur kan användarna bidra med sin kunskap? Ett viktigt område för forskning och utveckling är utvärdering. Det finns såväl automatiska som manuella metoder. Den manuella utvärderingen är dyrbar och utvärderarna är sällan överens i sina bedömningar. Den automatiska utvärderingen är nyttig för utvecklarna men den är inte adekvat i ett kundsammanhang. För att de automatiska måtten ska kunna användas i sådana sammanhang måste de tolkas. Är det överhuvudtaget möjligt?

I föredraget kommer jag att vidareutveckla de frågor som ställs ovan.

Korpora och konkordanser i mänsklig och automatisk översättning

Barbara Gawronska

Användning av parallellkorpora inom maskinöversättnig är nuförtiden en självklarhet. Inom översättarutblidning uppskattas också parallelkorpora, och blivande yrkesöversättare får normalt träning i att bygga egna parallella textsamlingar med hjälp av översättningsminnen, t ex TRADOS.
Betydelsen av monolingvala korpora i översättarutbildning och översättning underskattas dock fortfarande. Jag kommer däför att fokusera på hur ettspråkiga korpora - både "generella" och specialiserade - kan underlätta översättarens arbete och höja kvaliteten av automatisk översättning och automatisk sammanfattning.

Kübler (2003:27) påpekar vikten av terminologiarbete under översättningsprocessen:

"In specialized translation, translators also work as terminologists, as they have to make up a list of terms of a specific domain, as well as the list of their translations into the target language".

Specialiserade korpora utgör en oerhört värdefull källa för extrahering av domänspecifika termer. Korpora möjliggör vidare att inte enbart identifiera termerna, utan även att finna deras definitioner och kontext.
Detta är speciellt viktigt inom relativt nya forskningsområden, t ex genetik, där nya termer formas i takt med uppptäkter av nya gener, proteinmolykeler mm.

Förutom att underlätta terminologiextrahering, specialiserade korpora kan användas för identifiering av semantiskt besläktade ord, för extrahering av grammatisk information samt för jämförelse av domänspecifika stilistiska drag mm (Gawronska et al. 2002). Alla dessa funktioner är värdefulla inte enbart för mänskliga översättare, utan också för utveckling av maskinöversättningssystem och system för automatisk informationsextrahering.

En förutsättning för effektiv användning av korpora är ett välfungerande konkordansverktyg. I min presentation kommer jag att demonstrera konkordansverktyget Lexware Culler (Dura och Gawronska 2007), som baseras på lingvistiska regler och möjliggör en domänanpassad sökning.

Referenser:

Dura, E. and Gawronska, B. 2007. Novelty Extraction from Special and Parallel Corpora. In: Proceedings of 3rd Language & Technology Conference 2007, Adam

Mickiewicz University, Poznan, Poland, pp. 305-309.

Gawronska, B., Erlendsson, B. and Duczak, H. 2002. Extracting semantic classes and morphosyntactic features for English-Polish Machine Translation.

Hvordan påvirker valg av dataverktøy kvaliteten i tekstproduksjonen?

Carol B. Eckmann

Flere og flere institusjoner innfører krav om parallell publisering av sin utadrettet informasjon på flere språk. Dette igjen stiller nye krav til oversetterne, som skal produsere flere sider, ofte fortere, uten å gi avkall på kvalitet. Hva slags funksjonalitet trenger så oversetterne for å møte denne nye hverdagen? Hva slags rolle kan språkteknologiske produkter tenkes å ha i tilretteleggingen av samarbeid internt hos oppdragsgiveren, mellom oversetter og oppdragsgiver, og mellom oversetter og oversetter?

Språkteknologi och flerspråkighet inom EU

Krister Lindén

Översättning mellan de olika EU-språken görs idag inom EUs institutioner med hög och jämn kvalitet. Målet är att öka produktiviteten med bibehållen kvalitet. Ur denna synvinkel är maskinöversättning bara en av de många språkteknologier som stöder översättningsarbetet. Produktiviteten kan även ökas med olika språkvårdsredskap, rationalisering av arbetsflödet och med köpta tjänster från översättningsbyråer. Ett annat stort område för översättningsverksamhet inom EUs institutioner är tolkning. Där erbjuder internet och videokonferenstekniker nya möjligheter till kostnadseffektiva lösningar.

Dessutom är det inte bara arbetet inom EUs institutioner som fordrar översättning eller flerspråkiga lösningar. EU har en uttalad målsättning att ”alla borde beredas möjlighet att lära sig kommunicera på två språk förutom sitt modersmål”. Även EUs strävan att öka konkurrensen på en gemensam marknad och att befrämja arbetskraftens rörlighet medför ökade krav på flerspråkighet, översättning och tolkning inom de olika länderna i EU. En rörlig arbetskraft bör på lång sikt lära sig ett av det nya hemlandets nationalspråk. Detta leder till långsiktiga utmaningar för utbildningsväsendet och ger möjligheter för språkteknologi inom datorstödd språkundervisning. På kort sikt fordrar dock akuta situationer i början av en utlandsvistelse tolkning och översättning.

Språkteknologi och flerspråkighet inom EU

Maskinöversättning är en av flera språkteknologier som befrämjar översättningsarbetet inom EU. Det är emellertid inte bara arbetet inom EUs institutioner som fordrar översättning eller flerspråkiga lösningar, även EUs strävan att befrämja arbetskraftens rörlighet medför ökade krav på flerspråkighet, översättning och tolkning inom hela EU området.

Practical Experience with Statistical Machine Translation

Daniel Hardt

It has long been believed by many that statistical techniques alone could not produce acceptable results in MT. This has now been proven wrong. Indeed, statistical MT (SMT) is now producing impressive results, not only in the research lab but, increasingly, in the marketplace as well. In this talk, I will focus on our experience at languagelens, a Danish company we started in 2007 based on SMT technology. The languagelens system is being used to translate patent texts for Lingtech, with very positive results. I will examine the essential features of the current approach taken with the languagelens system, and show the quality of output that can be achieved with this approach. I will also offer some speculation about what the near future holds for MT.

Maskinoversættelse virker. Hvor skal der nu sættes ind, for at få større andvendelse af den i over sættelsesbranchen?

Thomas Bilgram

Lingtech har brugt maskinoversættelse som en væsentlig del af vores oversættelse af patentansøgninger siden 1993, og vores udfordring er nu at brede denne erfaring ud på andre områder. For Lingtech er oversættelse en handelsvare. For at forstå Lingtechs motivation for maskin-oversættelse (MT) er det nødvendigt at forstå vores kunders verden.

Kundens udfordringer

Tid

For vores kunder er en oversættelse altid en ulempe. Flere sprog forsinker ”time to market”, og meget ofte er teksten som skal oversættes først klar samtidig, eller ganske kort tid før, selve produktet. En øgning i hastigheden vil stille Lingtech stærkere end vores konkurrenter.

Kvalitet

Alle siger, de vil have top kvalitet, men der er forskel. For medicinalbranchen handler det om liv eller død, og selv en lidt uskarp oversættelse kan betyde katastrofe. Kvalitet koster, og sikkerhed koster. En automatisk øget konsistens ville stille Lingtech stærkere end vores konkurrenter.

Pris

Det er den manuelle del af oversættelsen, som koster. Branchen har længe brugt oversættelseshukommelse for at minimere tidsforbruget ved oversættelse. Denne metode har dog begrænsninger. Kan man skære i forbruget af mandetimer ved at give oversætterne og korrekturlæserne høj kvalitet af delvist oversat tekst som input og gode redskaber at arbejde med, så kan man spare penge.

Lingtechs udfordringer

For at lykkes med mere MT i vores oversættelsesproces har vi tre parter som vi skal have til at mødes; kunden, oversætteren og teknikeren.

Kunden skal stole på, at oversættelsen er mindst lige så god som uden MT. Ofte er kunden usikker, så budskabet skal være, at der er en fordel for kunden, fx hurtigere oversættelse, bedre kvalitet, eller lavere pris.

Oversætterne skal vide at de stadig er nødvendige, selvom deres rolle måske skifter lidt. De får stadig "samme løn for samme indsats", men med bedre input, skal de levere hurtigere, bedre og mere. En oversætter er i dag akkordlønnet målt på antal ord, og det regnestykke skal Lingtech udfordre for at få fuld fordel af den øgede kvalitet i input.

Teknikeren skal vide, at MT for Lingtech aldrig er en løsning i sig selv. En god MT-løsning er et redskab til at opnå hurtigere oversættelse, bedre kvalitet, og lavere pris. Men redskaber til færdigoversættelse kan være mindst lige så vigtige. Skal Lingtech drage fuld nytte af dette, så er det vigtigt at man ser på hele processen, og at MT-løsningen omfatter en større del af processen fra teksten kommer fra kunden, til teksten leveres tilbage til kunden.

Målet

Lingtechs mål er at tjene penge på oversættelse. Vi ved, at MT er en god hjælp til det, hvis det bliver grebet rigtigt an. Vi arbejder derfor stadig målrettet på at bruge MT på en måde, som vi mener, er lønsom for vores kunder og dermed for os.

Automatisk oversettelse mellom skandinaviske språk

Lars Nygaard

Allerede på 1980-tallet kom det første forsøket på automatisk oversettelse mellom to skandinaviske språk (PONS-systemet, for norsk og svensk). Senere har flere andre systemet kommet til, med Google som foreløpig siste ankomst. I foredraget kommer jeg til å:
- oppsummere de oversettelsessystemene jeg kjenner til
- si litt om relevante ortografiske, morfologiske og syntaktiske forskjeller
- presentere et (forholdsvis) nytt oversettelsesystem for skandinaviske språk, utviklet i samarbeid med GrammarSoft ApS.

Kursplaneöversättaren – ett system för översättning av kursplaner från svenska till engelska

Eva Pettersson

Kursplaneöversättaren är ett maskinöversättningssystem speciellt anpassat till översättning av kursplaner från svenska till engelska. Systemet är utvecklat av en forskargrupp vid institutionen för lingvistik och filologi vid Uppsala universitet. Sedan våren 2006 står avknoppningsföretaget Convertus AB för vidareutveckling och drift av Kursplaneöversättaren.

Kursplaneöversättaren har en regelbaserad kärna, med analys-, transfer- och genereringsmoduler. Därtill finns diverse upphämtningsstrategier som utnyttjas i de fall där lexikon och grammatik inte räcker till för att ge en fullgod översättning. De viktigaste upphämtningsstrategierna består i:

1. tillvaratagande av partiella analyser i de fall där analysgrammatiken inte når ända fram,
2. en statistisk språkmodell för generering när den regelbaserade genereringsmodulen inte räcker till samt
3. sammansättningsanalys av ord utanför lexikonet, och översättning av de ingående delarna.

Lexikonet kan sägas bestå av tre delar: ett svenskt lexikon för analyssteget, ett engelskt lexikon för genereringssteget samt ett svensk-engelskt lexikon för transfersteget.

Lexikonet har byggts upp utifrån de kursplaner som fanns inlagda i Uppsala universitets kursplanedatabas (Selma) i augusti 2007. Översättningsrelationerna har så långt möjligt definierats på basis av de översättningar som fanns inlagda i kursplanedatabasen. Då mindre än en tredjedel av kursplanerna hade en översättning i databasen, har övriga ord tilldelats en översättning med andra hjälpmedel, såsom ordböcker och tillgängliga resurser på webben. Det resulterande lexikonet består av totalt 35 221 översättningsrelationer, fördelade på olika ämnesområden. Därutöver tillkommer ett allmänlexikon om 10 737 ingångar och ett kärnlexikon om 2135 ingångar, vilket ger en total vokabulär om 48 093 översättningsrelationer. Därtill kommer ca 3000 lexikala transferregler för översättning av ord i kontext.

Lexikonstrukturen är hierarkisk, och i översättningsprocessen sker lexikonuppslagningen i en fördefinierad ordning, där ordet först slås upp i det mest specifika lexikonet. Ord som inte går att finna i något av lexikonen körs genom en sammansättningsanalys, och om sammansättningsanalysen lyckas översätts de ingående delarna var för sig. Om även sammansättningsanalysen misslyckas, kopieras det svenska ordet helt enkelt över till den engelska sidan.

Utöver lexikon och grammatik, innehåller Kursplaneöversättaren en minnesfunktion. Här lagras de översättningar som användarna har granskat, redigerat och godkänt. Vid översättning av nya kursplaner, konsulterar systemet översättningsminnet som ett första steg i översättningsprocessen. Om en mening återfinns i minnet, väljs den översättning som finns lagrad i minnet. Initialt innehöll minnet segment som sedan tidigare fanns manuellt översatta i kursplanedatabasen. Efter hand som systemet används, växer minnet med de översättningar som granskas och godkänns av användarna. Varje fakultet/ämnesområde har sitt eget översättningsminne.

För Uppsala universitets del är Kursplaneöversättaren integrerad i kursplaneverktyget Selma, och processen inleds med att användaren skriver in sin svenska kursplan i dess gränssnitt. Därefter trycker han/hon på knappen för stavningskontroll. Kursplaneöversättaren utför då en stavningskontroll som bygger på samma lexikon som översättningssystemet har. Ord som saknas i lexikonet rödmarkeras. Efter att användaren har genomfört eventuella korrigeringar av kursplanen i enlighet med stavningskontrollen, klickar han/hon på knappen för oversättning. Kursplanen läggs då i en översättningskö, varifrån de hämtas av översättningssystemet och bearbetas en efter en.

När översättningen är klar, skickas ett mejl till den som har utsetts till översättningsansvarig på den institution som kursplanen gäller. I mejlet finns en länk till ett efterredigeringsgränssnitt, där användaren ser den svenska kursplanen tillsammans med sin översättning. Ord i den engelska kursplanen som saknas i lexikonet är rödmarkerade.

När användaren har redigerat klart sin översättning, klickar han/hon på en knapp för godkännande, vilket gör att översättningen sparas och lagras i Selma-databasen. Dessutom lagras samtliga översättningar i översättningsminnet, så att nästa gång en användare vill översätta samma textsegment (mening, rubrik etc.) kommer de granskade och godkända översättningarna att väljas.

Driftsättning av Kursplaneöversättaren genomfördes vid Uppsala universitet i september 2007, och vid Umeå universitet i september 2008. Användarna har möjlighet att ställa frågor om systemet via Kursplaneöversättarens hemsida. Convertus har analyserat de önskemål om vidareutveckling som hittills har inkommit från användarna, och uppdaterat systemet i enlighet med detta. Just nu diskuteras frågan om vem som ska vara översättningsansvarig.

Nynodata's Temasearch - resources and techniques for multilingual search

Matthew McGowan

Nynodata's TemaSearch is a tool to enhance an existing search function with additional features, including multilingual search. Our presentation will discuss the key synonym and translation resources used in Bokmål and Nynorsk, and the way the system uses these resources to provide an enhanced search function for multilingual websites and multilingual user bases. The possibility of using TemaSearch with other European languages will be discussed.

Artikler

Maskinöversättning i teori och praktik. Var står vi i dag?

Anna Sågvall Hein

Maskinöversättning är det äldsta datorlingvistiska tillämpningsområdet med rötter tillbaka till mitten av 50-talet. Intresset för området och tron på maskinöversättningens möjligheter som ett komplement eller alternativ till mänsklig översättning har varierat under årens lopp. I början av 90-talet tog dock såväl forskning som tillämpning ny fart då man systematiskt började utveckla metoder för att på olika sätt återanvända tidigare översättningar. Det handlar både om att bygga översättningssystem som helt grundar sig på tidigare översatt text, s.k. statistiska översättningssystem, och att bygga upp och komplettera språkbeskrivningarna i språkvetenskapligt grundade översättningssystem, s.k. regelbaserade system.

De statistiska översättningssystemen visade sig kunna producera oförutsett bra översättningar, något som drivit utvecklingen framåt. En annan faktor som haft avgörande betydelse för utvecklingen av maskinöversättning är den ökande användningen av Internet och tillgången till språkliga data på nätet. Sedan något tiotal år har det funnits tillgång till fria översättningstjänster på nätet, som trots den bristande översättningskvaliteten fått stor användning. Redan i dag översätts mer text maskinellt än manuellt, och mer än 50 miljoner översättningsuppdrag körs dagligen via Internet (Jaap van der Meer 2008).

Våren 2008 inträffade en händelse av stor betydelse då Google släppte en testversion av en fri maskinöversättningstjänst, som utvecklats av Google Translation Center. Den omfattar mer än 30 olika språk och ett första allmänt intryck är att översättningarna är väsentligt mycket bättre än de fria översättningstjänster som tidigare funnits att tillgå på webben. Tjänsten är lätt att använda och man kan förvänta sig att användningen kommer att öka kraftigt. Även om översättningarna är bättre än många andra maskinella översättningar, så är de inte perfekta och kvaliteten varierar för olika språk. För att översättningarna ska uppnå publiceringskvalitet måste de efterredigeras.

Vilka förbättringar kan man då räkna med på sikt i Googles översättningar? De bygger på statistisk maskinöversättning som tränas på tidigare översättningar och som finslipas mot statistiska syntaxmodeller över målspråket. Erfarenheter från forskarvärlden visar att storleken på träningsdata är av avgörande betydelse för översättningsresultatet. Det är dock inte den enda framgångsfaktorn. Det är också viktigt att träningsdata hämtas från samma domän som den som systemet ska tillämpas på. Vidare spelar språkskillnader och språklikheter stor roll.

Den stora fördelen som företaget har gentemot andra utvecklare och forskare ligger i den närmast obegränsade tillgången till textmaterial på olika språk för träning av översättningsmodulerna. Finns det en gräns för hur långt man kan komma i fråga om att höja översättningskvaliteten i ett statistiskt system genom lägga på mer och mer träningsdata? Kan man hamna i en situation där mer data skapar fler alternativ och kvaliteten degraderar? Något som de statistiska systemen inte kommer åt är textberoenden över meningsgränserna, vilka många gånger är avgörande för korrekt översättning.

Det finns i dag publikt tillgängliga programvaror – open source – som kan användas för uppbyggnad av statistiska system. Det är en möjlighet som tagits tillvara av flera unga företag med goda resultat. Inom forskarvärlden ägnas mycken uppmärksamhet åt att förbättra de statistiska systemen genom att inkludera lingvistisk kunskap. Det är en svår uppgift, då systemen inte har några regler som kan förfinas utan består att sannolikheter som avkodas. Framför allt försöker man finna metoder för att kompensera för brister i träningsdata genom att analysera dem lingvistiskt och därigenom utnyttja dem så effektivt som möjligt.

Hur ser det ut på marknaden? Där har hittills de regelbaserade systemen dominerat, men där kan man förvänta sig en förändring. För regelbaserade system ligger den stora svårigheten i att få språkbeskrivningarna heltäckande, en hart när omöjlig uppgift, så länge man inte använder sig av kontrollerade språk. För att kompensera för brister i språkbeskrivningen utvecklas olika upphämtningsstrategier, som utnyttjar statistiska metoder. Det är en metod som dagens Systran använder sig av och också det svenska företaget Convertus, som bland annat översätter kursplaner från svenska till engelska. En fråga man måste ställa sig är vilken roll de regelbaserade systemen spelar i dag och i framtiden.

Hur bra kan den maskinella översättningen bli? Hur bra behöver den bli för olika ändamål? Hur kan användarna bidra med sin kunskap? Ett viktigt område för forskning och utveckling är utvärdering. Det finns såväl automatiska som manuella metoder. Den manuella utvärderingen är dyrbar och utvärderarna är sällan överens i sina bedömningar. Den automatiska utvärderingen är nyttig för utvecklarna men den är inte adekvat i ett kundsammanhang. För att de automatiska måtten ska kunna användas i sådana sammanhang måste de tolkas. Är det överhuvudtaget möjligt?

Korpora och konkordanser i mänsklig och automatisk översättning

Barbara Gawronska

1. Introduktion

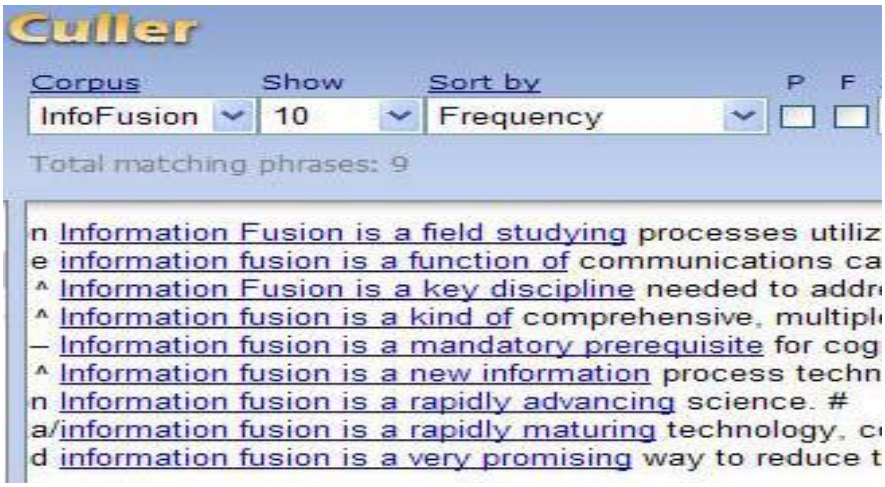
Användning av parallellkorpora inom maskinöversättnig är nuförtiden en självklarhet. Inom översättarutblidning uppskattas också parallelkorpora, och blivande yrkesöversättare får normalt träning i att bygga egna parallella textsamlingar med hjälp av översättningsminnen, t ex TRADOS. Betydelsen av monolingvala korpora i översättarutbildning och översättning underskattas dock fortfarande. I det följande kommer därför fokus att ligga på hur ettspråkiga korpora - både "generella" och specialiserade - kan underlätta översättarens arbete och höja kvaliteten av automatisk översättning och automatisk informationssökning.

2. Terminologiextrahering

Kübler (2003:27) påpekar vikten av terminologiarbete under översättningsprocessen:

> "In specialized translation, translators also work as terminologists, as they have to make up a list of terms of a specific domain, as well as the list of their translations into the target language".

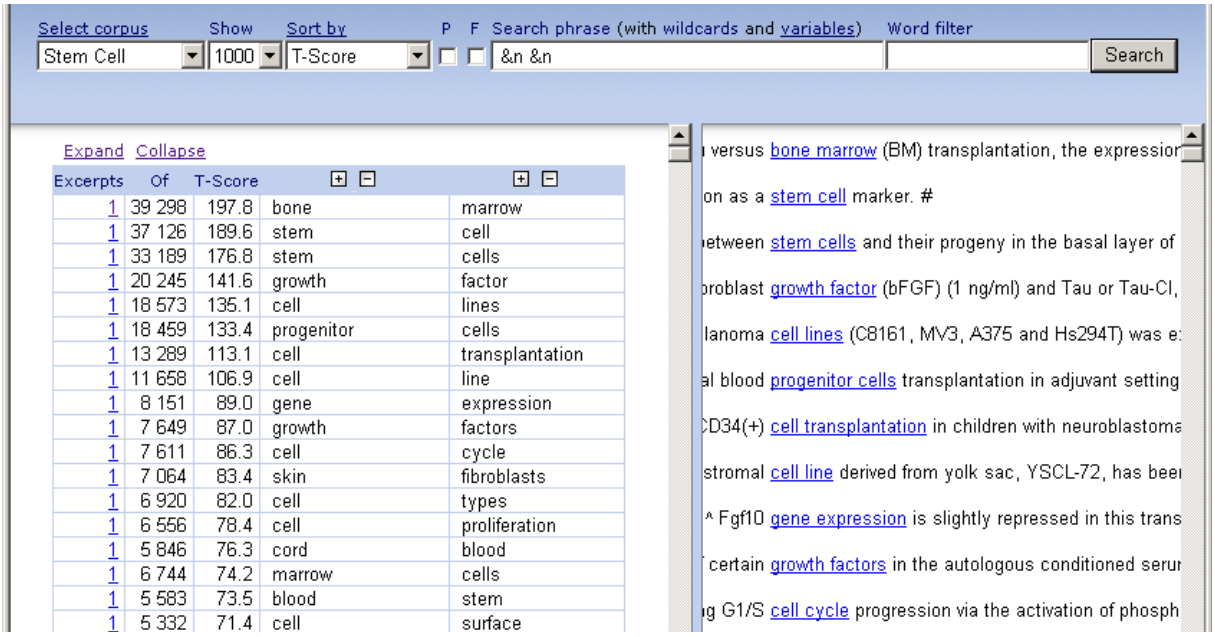
Specialiserade korpora utgör en oerhört värdefull källa för extrahering av domänspecifika termer. Med hjälp av korpora och lämpliga konkordansverktyg kan man inte enbart identifiera termerna, utan även finna deras definitioner och kontext (figur 1). Detta är speciellt viktigt inom relativt nya forskningsområden, t ex informationsfusion, där flera centrala termer får nya definitioner under forskningens gång (Dura and Gawronska 2008), eller genetik, där nya termer formas i takt med upptäckter av nya gener, proteinmolekyler mm.



Figur 1. Definitioner av termen *Information Extraction*identifierade i en omfattande samling av forskningsartiklar (Dura and Gawronska 2008)

Moderna konkordansverktyg, som t ex Lexware Culler (Dura och Gawronska 2007) ger användaren möjlighet att undersöka så kallad kollokationsstyrka (Kilgarriff 2005), dvs

sannolikheten at två ord kommer att förekomma nära varandra, samt jämföra kollokationsstyrkan mellan ordparen i olika textdomäner. Listor av fraser med högt kollokationsvärden kan integreras i översättarens eller maskinöversättningssystemets lexikon och minska risken för oidiomatisk ord-för-ord översättning. Figurer 2 och 3 visar exempel på extahering av tvåordstermer från bioinformatiska texter samt från texter rörande forskning inom informationsfusion.



Figur 2. Sammansatta termer extraherade från bioinformatisk korpus

Of	Salience		
45	34.84	uncertainty	management
20	25.27	uncertainty	measures
21	23.19	uncertainty	measure
7	19.13	uncertainty	ellipse
10	18.43	uncertainty	description
13	17.73	uncertainty	representation
3	17.26	uncertainty	calculi
13	17.18	uncertainty	area
4	17.00	uncertainty	ellipsoids

Figur 3. Sammansättningar med förleden *uncertainty* extraherade från forskningsartiklar inom Informationsfusion

3. Extrahering av semantisk, grammatisk och stilistisk information

Förutom att underlätta terminologiextrahering, korpora kan användas för identifiering av semantiskt besläktade ord, kartläggning av ordens semantiska och syntaktiska valens, extrahering av grammatisk information samt jämförelse av domänspecifika stilistiska drag mm (Nazarenko et al. 2001, Gawronska et al. 2002). Ett konkordansverktyg som integrerar statistiska matt med lingvistiska regler kan t ex visa vilka prepositioner och partiklar som kombineras med vilka verb, ge en nyanserad bild av användningen av semantiskt "lätta" verb

som *göra*, *ta*, *ställa* jämfört med t ex *do*, *make*, *take*, *put* osv., ge information om vilka substantiv som vissa adjektivattribut tenderar att förekomma tillsammans med (t ex att engelskans *wide* förekommer ofta med *eyes* och *smile*, medan *broad* kan fungera sombestämning till *smile*, men aldrig till *eyes*). Systematisk användning av kontekstuell information från korpora kan därmed minimera interferens från källspråket och öka översättningens stilistiska variation. Alla ovan exemplifierade funktioner är värdefulla inte enbart för mänskliga översättare, utan också för utveckling av maskinöversättningssystem och system för automatisk flerspråkig informationsextrahering.

4. Slutsats

För mänskliga översättare utgör ettspråkiga korpora ett ytterst värdefullt komplement till ordböcker. Möjligheten att se ordens mest frekventa kontexter underlättar valet av översättningsekvivalenter och eliminerar ”falska vänner”. Vidare kan korpora tillhandahålla stöd vid tveksamheter rörande valens, böjning och stavning. Möjligheten att sortera ordpar enligt kollokationsstyrka är mycket värdefull för den mänskliga översättarens terminologiarbete och kan öka kvalitet av automatisk översättning. En förutsättning för effektiv användning av korpora är dock välfungerande konkordansverktyg. Konkordansverktyget Lexware Culler (www.nla.se/culler, Dura 2006, Dura and Gawronska 2007) möjliggör sökning med ordklassvariabler, inkluderar morfologisk analys och tillhandahåller sortering enligt ett antal statistiska kollokationsmått. Verktyget är för närvarande tillgänglig för svenska, engelska och polska. En utveckling inriktad på norska, danska och finska skulle vara önskvärd.

Referenser:

Dura, E. and Gawronska, B. 2008. Natural Language Processing in Information Fusion Terminology Management. *Proceedings of the 11th International Conference on Information Fusion*, Köln, Germany. ISIF IEEE.

Dura, E. and Gawronska, B. 2007. Novelty Extraction from Special and Parallel Corpora. In: *Proceedings of 3rd Language & Technology Conference 2007*, Adam Mickiewicz University, Poznan, Poland, pp. 305-309.

Dura, E. 2006. Culler – a User Friendly Corpus Query System. *Proceedings of the Workshop Dictionary Writing Systems at Euralex.* Turin. http://www.natcorp.ox.ac.uk/what/index.html

Gawronska, B., Erlendsson, B. and Duczak, H. 2002. Extracting semantic classes and morphosyntactic features for English-Polish Machine Translation. *Proceedings of the 9th International Conference on Theoretical and Methodological Issues in Machine Translation (TMI-2002)*, Keihanna, Japan, pp. 63-73.

Kilgarriff, A. 2005. Language is never ever ever random. *Corpus Linguistics and Linguistic Theory* 1 (2), pp. 263-276.

Kübler, N. 2003. Corpora and LSP Translation. In Zanettin, F., Bernardini, S. and Stewart, D. (eds.): *Corpora in Translator Education*. St. Jerome Publishing: Manchester, UK and Northhampton, MA, pp. 25-42.

Nazarenko, A., Zweigenbau, P., Habert, B., and Bouaud, J. 2001. Corpus-based extension of a terminological semantic lexicon. In: Bourigault, D., Jacquemin, C. and L'Homme, M-C. (eds.) *Recent Advances in Computational Terminology*. Amsterdam: John Benjamins Publishing Company, pp. 327-352.

Hvordan påvirker valg av dataverktøy kvaliteten i tekstproduksjonen?

Carol B. Eckmann

Flere og flere institusjoner innfører krav om parallell publisering av sin utadrettet informasjon på flere språk. Dette igjen stiller nye krav til oversetterne, som skal produsere flere sider, ofte fortere, uten å gi avkall på kvalitet. Hva slags funksjonalitet trenger så oversetterne for å møte denne nye hverdagen? Hva som skal til for å legge til rette for bedre samarbeid mellom oversetter og oppdragsgiver så vel som mellom oversetter og oversetter?

Jeg jobber ikke innfor noen stort konsern, ei heller et oversettelsesbyrå. Men i løpet av de siste 8 årene har jeg organisert og administrert storskala oversettelsesprosjekter for bl.a. Forskningsrådet, Fredssenteret, Slottet, og diverse departementer ved hjelp av et team uavhengige, selvstendig næringsdrivende oversettere.

Felles for disse type oppdragene er: det skal parallellpubliseres store mengder med informasjon, f. eks. på et nytt nettsted, som er produsert av et stort antall tekstforfattere; og oversetterne skal produsere høykvalitetsoversettelse av politisk viktig og faglig vanskelig tekst, som har lite innbyrdes gjentagelse og ofte krever en veldig detaljert forståelse av en virksomhets «indremedisinske» forhold;

«Høy kvalitet» i denne sammenhengen betyr å formidle et politisk/faglig/administrativt budskap på en idiomatisk, stilistisk og terminologisk riktig måte. Og det brukes selvsagt målbevisst språkteknologi som et verktøy for å sikre, og forbedre oversettelsesarbeidet.

Etter hvert er det blitt ganske klart hva slags funksjonalitet i dataverktøy som man har hatt behov for underveis og hva er det som mangler i samarbeidskonstellasjonene mellom oversettere så vel som mellom oversetter og oppdragsgivende virksomhet, som kanskje kunne løses ved utvidet bruk av slikt dataverktøy.

Altså: ***Hvordan påvirker valg av dataverktøy kvaliteten i tekstproduksjonen?***

Kvalitetsvurderingen av tekst er veldig kompleks. Jeg velger å dele kvalitetsfaktorene opp i noe jeg kaller «ikk-ene»:

Oversettelseskvalitet har å gjøre med:
grammatikk,
historikk,
stilistikk,
idiomatikk,
terminologimetodikk,
dokumentstyringsteknikk, og
«*temporikk*».
Jeg vil også påstå at kvalitet har noe med *produksjonsmetodikk* av originaldokumentene å gjøre.

Hva har så valg av dataverktøy å si for disse «ikk-ene» ?

Kvalitetsfaktor: grammatikk

Grammatikk er som alle vet vanskelig å håndtere. Det finnes stadig god rom for forbedring av grammatikkprogram. Dårlig grammatikk er ofte det første en leser legger merke til, og det kan faktisk skygge for et faglig godt budskap.

Jeg oversetter til engelsk, som er et språk alle kan, i hvert fall alle nordmenn. Mange tror at de kan skrive engelsk selv, eller de mener at de i det minste har nok innsikt i engelsk til å overprøve det som blir gjort av profesjonelle språkmedarbeidere. Ikke sjeldent med et pinlig resultat. Med adgang til et godt grammatikkprogram, kunne man kanskje ha fått fram mer korrekte tekster, f. eks. nettopp fra nordmenn som skriver på engelsk. Under grammatikk plasserer jeg også noe så enkelt som stavekontroll. Det er en svært kostnadseffektiv måte å fjerne en god del blemmer på – men det er utrolig hvor få det er som konsekvent bruker stavekontrollen i sine tekstbehandlingsprogrammer – på norsk eller engelsk.
Stavekontrollverktøyet har selvfølgelig sine klare begrensninger. Det er foreløpig forbeholdt tekstbehandlingsprogrammer. Er det kanskje en tanke å innlemme stavekontrollmuligheter på noen måte i andre dataverktøy?
Dataverktøy som kan håndtere grammatikk på en adekvat måte kan øke kvalitet ikke bare i forhold til rettskrivingen, men også ved å være tids- og kostnadsbesparende. Mange tror de kan spare penger ved å skrive noe selv på engelsk og så få det språkvasket. Hvorvidt dette medfører riktighet avhenger av hvor gode den enkeltes engelske kunnskaper er. Generelt vil jeg si at å oversette en godt gjennomtenkt og klart skrevet tekst fra norsk til engelsk gir et bedre produkt enn språkvask av en tekst skrevet på mangelfull engelsk. Men hvis den språklige kvaliteten på denne engelsken kan heves nok, kan man faktisk få fram et godt produkt og spare en del penger.

Kvalitetsfaktor: historikk

De aller færreste av de tekstene som jeg jobber med er nybrottsarbeid. Og tekstproduksjon finner ikke sted i et vakuum. Tekster som produseres er gjerne bygget på andre tekster, på tankeganger som uttrykkes andre steder, på bakgrunnsstoff som finnes i andre sammenhenger, på maler og politiske grunnlag som er opplest og vedtatt. Det er ikke alltid like lett å spore opp den riktige bakgrunnsinformasjonen. Men man trenger adgang til denne dokumentarven hvis man skal kunne føre tidligere tankeganger utvetydig framover, hvis man skal kunne hindre misforståelse og feiltolking, og hvis man skal kunne sikre samme bruk av terminologi over tid. Det er som vi alle vet ikke kvalitetsfremmende å finne opp hjulet på nytt.
For å lette adgangen til dokumentarven, trenger man å ha et verktøy som kan bygge en søkbar, fulltekst korpus av kilde- og ev. målspråksdokumenter, f. eks. tidligere utførte dokumenter, referanseinformasjon, og hvilket som helst annen tekstkilde man ønsker seg (f.eks. nettsider), og som helst kan gjøre det *fort*. Brukeren må kunne søke gjennom hele tekstmengden (og flere korpora samtidig) på samtlige språk som finnes der, for nøyaktig og *fuzzy matches* av uttrykk av varierende lengde, fra ett ord opp til tekstsegmenter som består av flere avsnitt. Forekomster av uttrykk/tekstsegmenter bør kunne gjenfinnes automatisk og lett gjenbrukes ved f. eks. at man kan velge å bevege seg gjennom uttrykk og disses ev. oversettelser i teksten på et delt skjermbilde.
Over tid bygger en oversetter opp sin egen dokumentarv også, og må også lett kunne både gjenbruke og dele den. Man må kunne hente fram/filtrere ut informasjon til et gitt formål.

Tekster bør kunne typologiseres, og det bør bære en hensiktsmessig innretning for deling av dokumentarven med andre, uansett om de benytter samme verktøy.

Kvalitetsfaktor: idiomatikk/stilistikk

Å treffe riktig stilnivå og finne de riktige idiomatiske uttrykk er en stor utfordring. Samtidig er det akkurat her hvor det er minst hjelp å få fra eksisterende dataverktøy – det er jo her det menneskelige elementet foreløpig er uovertruffent og ikke automatiserbart. Men det som hjelper veldig, er nettopp adgang til historikken, dvs. dokumentarven; altså å kunne søke etter og se hvordan terminologi og tekstsegmenter er blitt brukt i relevante kontekster. Og der kommer dataverktøy tilbake inn i bildet. Har man et dataverktøy som gir lett adgang til tidligere løsninger i kontekst, har man kommet langt. Også her vil typologisering av tekster gjøre det lettere å finne den rette stilistiske informasjonen.

For oversettere kan det faktisk ofte være mer meningsfylt å kunne se terminologisk informasjon brukt i kontekst enn å finne samme informasjon som oppslag i en ordliste/termbase. Oppslag i termbaser forteller kun det som innleggeren har valgt å ta ut om den aktuelle terminologiske informasjonen. Å kunne se ord og uttrykk i bruk i kontekster gir mange pekepinn om hvordan og i hvilke sammenhenger disse ord og uttrykk kan opptre.

Kvalitetsfaktor: terminologimetodikk

Enten man skriver eller oversetter, så er det ord det dreier seg om, og ord kan brukes på både riktig og feil måte. Har man nå først lett seg fram til de «riktige» ordene, er det en fordel å ta vare på dem for å slippe å lete dem opp igjen ved en senere anledning. Det er ikke bare nyttig for den enkelte – det er også viktig for samarbeidspartnere at de har adgang til det som andre vet eller har funnet.
Terminologisk tenkning er meget viktig i kvalitetssammenhengen. Ved å registrere terminologisk informasjon systematisk kan oversettere bedre sin ytelse, forbedre tekstkvaliteten og øke produktiviteten. En organisert samling av terminologisk informasjon gjør det mulig for medarbeidere å holde styr på, gjenbruke og dele sin ekspertise, og det forenkler samarbeidet mellom enkeltpersoner og grupper.
Oversettere og skribenter har behov for å lagre og gjenfinne et mye bredere utvalg av data enn det som tradisjonelt lagres i en terminologisk databank som lages av terminologer. I tillegg til å ivareta alle former for tradisjonell terminologi, som gjerne er knyttet til begrepssystem, trenger man et utvalg av annen terminologisk informasjon (fraseologi, tekstsammenhenger, standard tekstsegmenter av varierende størrelse, kollokasjoner, stillingsbetegnelser, institusjonsbetegnelser, foreslåtte løsninger, m.m.).

Dataverktøy har særdeles mye å si her. Verktøyet trenger å ha en integrert terminologi-behandlings komponent med tilhørende database. Det er innenfor denne komponenten at det settes rammer for hvor enkelt det blir å ekserpere, lagre og gjenfinne terminologisk informasjon, hva slags informasjon som blir lagret, og hvor fritt eller låst de enkelte postene er i sin struktur.
Vi skal selvfølgelig ikke her gå inn på alle detaljer om hva som må til for å bygge gode terminologidatabaser. For de som er veldig interessert i det, viser jeg til NS - ISO 12616 :2002 Oversettelsesrettet terminografi (Translation-oriented terminografi) og ISO 12620: 1999 Computer applications in terminology – Data categories (for tiden under revisjon). Men det er noen verktøy-kriterier som jeg mener er så vesentlige for kvaliteten i tekstproduksjonen, at jeg likevel vil nevne dem her. Oversettelsesrettet terminologiarbeid er nemlig tekstbasert, ikke begrepsbasert. Derfor:

- Man må kunne lagre fraseologi og store tekstsegmenter som hovedterm. Disse er byggesteinene i tekstproduksjon, og er ofte det som man trenger å hente ut. Derfor må de kunne lagres på hovedtermsnivået, ikke kun som tekstsammenhenger.
- Man må ha muligheten for å endre språkretning, men med klar indikasjon av hva som var kildespråket i oppslaget. Reversering kan være nyttig, men det er farlig hvis ikke man vet om det var hønset eller egget som kom først.
- Man må ha muligheten for å velge antall og type informasjonskategorier som tas med i den enkelte termposten innenfor den enkelte termbasen.
- Det må være **påbudt** med kildeangivelse. Dette gir vesentlig informasjon om påliteligheten av en term, og pålitelighet er et viktig aspekt av kvalitet.
- Det må være lett å dele ressurssene – alt eller deler av dataene.
- Det må være lett å samle output til andre filtyper, f. eks. Excel/Word-filer, for distribusjon til folk som ikke har samme programvaren.

Kvalitetsfaktor: dokumentstyringsteknikk

Med dokumentstyringsteknikk mener jeg alle de mindre cerebrale prosesser knyttet til filhåndtering ved store og små informasjonsprosjekter. For å ta et banalt men likevel ikke helt banalt eksempel: filnavn. For å automatisere innleggingsprosessen av dokumentarven i en parallelltekst-korpus, må sannsynligvis filnavnene struktureres etter en gitt protokoll. Det er de færreste som tenker på (eller forstår), og dette kan være veldig arbeidskrevende å rette på. Jeg kan for øvrig ikke si hvor mange ganger jeg har mottatt oppdrag fra en oppdragsgiver som bare heter «brosjyre» – gjerne fra flere på samme dag. Her er det snakk om å måtte ha administrative rutiner på plass som gjør det lett å håndtere mange filer uten å måtte åpne disse for å se hva innholdet er.

Andre slike viktige rutiner er versjonering, datering av dokumenter, o.l. Mens dette er ting som kanskje ikke ligger direkte til i språkteknologiske verktøy, er det likevel ting som dataverktøyet må integreres mot.

Redigeringssløyfer er ekstremt viktige. Underveis i en oversettelsesprosess vil det sikkert bli avdekket en mengde feil/svakheter/inkonsekvens i de norsk tekstene. Disse måtte rettes i en ekstra norskspråklig redigeringssløyfe, og ev. tas hensyn til ved oversettelse til andre språk som påløper samtidig. Det er avgjørende i slike situasjoner at det er siste versjon som publiseres og som gjøres tilgjengelig, f. eks. i filene som legges til dokumentarven. Her er det snakk om versjonshåndtering – den nest siste versjon bør aldri være tilgjengelig uten veldig gode grunner.

Uten god dokumentstyringsteknikk blir det veldig mye krøll. Og krøll er et kvalitetsproblem.

Kvalitetsfaktor: temporikk

Det at ting produseres fort, og gjerne fortere en «før», er et sentralt kvalitetskriterium i dag. Evnen til å levere tekster som er grammatikalsk, idiomatisk og terminologisk riktig til rett tid er unektelig et viktig aspekt av god kvalitet. Dataverktøy som tilrettelegger for rask informasjonsgjenfinning og gode kontrollrutiner er derfor av stor betydning for å kunne øke produktivitet i tid uten at det går på bekostning av kvalitet. Men under visse omstendigheter kan «time to market» virke som om det er de eneste kvalitetskriterium som gjelder. Bruk av dataverktøy kommer inn som et element i anbudsprosessen, og det er tendenser blant enkelte til å omtale disse som nærmest vidundermidler, som kan korte ned tiden til det usannsynlige. Og *det* kan gå på bekostning av andre kvalitetsperspektiver.

Kvalitetsfaktor: produksjonsmetodikk

Språkteknologiske verktøy som er i bruk er stort sett myntet på oversettelsesprosessen og parallelle tekstkorpuser, men det er et like stort behov for å kunne bistå virksomheter i sin egen primærproduksjon og vedlikehold av informasjon, som stort skjer internt utenom oversettelsessløyfen. Mitt siste poeng her i dag, er at virksomheter kunne ha stor fordel av å anvende disse verktøyene for å produsere og vedlikeholde sin egen interne og eksterne informasjon, uansett om det skal forbli ettspråklig eller oversettes. Oppbygningen av språkressursene kan tjene flere formål enn kun flerspråklige informasjonsprosesser

Mange av problemene som oversettere møter har sitt opphav i oppdragsgiverens *produksjonsprosess*. Hvis ikke oppdragsgiveren skriver klart, treffer riktig stilnivå, innlemmer riktig historikk og tenker terminologisk, blir tekstene tilsvarende vanskeligere å håndtere. Altså, oppdragsgiverens produksjonsmetodikk har mye å si for hva slags kvalitet norske tekster har, og dermed de oversatte tekstene kan få. Det er minst like viktig for *tekstprodusenter i en virksomhet* å kunne ha grammatikkprogram, å kunne gjenfinne uttrykk og tekstsegmenter i kontekst, å ha adgang til virksomhetens dokumentarv, å kunne registrere og systematisere terminologi, og å kunne produsere/oppdatere tekst rakst og sikkert. Alle mine "ikk-ene" gjelder for den primære tekstproduksjon i like stor grad som for oversettelse:

Grammatikk: Det er veldig mye rart som skrives på norsk. Adgang til dataverktøy med en god grammatikk-komponent kan gi bedre tekster, noe som er til fordel for både de som skal bruke disse på norsk, og de som ev. skal oversette disse til et annet språk.

Historikk (inkl stilistikk/idiomatikk): Adgang til dokumentarven er vel så viktig innenfor en virksomhets ettspråklige tekstproduksjon som for dens flerspråklig arbeid. Har man en lett tilgjengelig dokumentarv, kan man finne og konsultere tidligere produserte tekster for å sikre bruken av riktige formuleringer og ikke minst konsekvent terminologi. Forfattere har også et like stort behov for innsikt i riktig idiomatikk og stilistikk som oversettere. Dette vil i tillegg gjøre nye medarbeidere i stand til å sette seg fortere inn i ting, da den gjeldende ekspertisen blir gjort tilgjengelig, uten å være personavhengig.

Terminologimetodikk: Bruk av vaklende terminologi i en virksomhets intern og ekstern informasjon kan skape store forviklinger innad. Er det inkonsekvens i originalen, kan det også forplante seg til oversettelsen, med utilsiktet tvetydighet som følge. Tekstforfattere har stor bruk for å ivareta og kunne gjenfinne og gjenbruke sin egen ekspertise og få adgang til andres i den ettspråklige produksjonsprosess, uavhengig av om tekster senere skal inn i en oversettelsesprosess.
Det kan ikke gjentas for mange ganger: Konsekvent og systematisk arbeid med ettspråklig så vel som flerspråklig terminologi:

- høyner presisjon;
- forhindrer inkonsekvens;
- skaper et grunnlag for å produsere mer, fortere;
- sikrer at nye rekrutter fortere blir habile medarbeidere;
- gjør kvalitetssikringsprosesser mer hensiktsmessige
- gjør virksomheten selv bedre i stand til å etterprøve både originaltekster og oversettelser og gi god tilbakemelding.

Dokumentstyringsteknikk
Gode administrative rutiner for filhåndtering er meget viktig innad i en virksomhet. Har man et godt system på plass, blir det veldig mye enklere å organisere publisering og holde styr på

hva som befinner seg hvor i interne prosesser. Bare det å ha et klart system for hvor filer lagres ville hjelpe mange.

Temporikk
De samme tidsbesparelser man får under oversettelsesprosessen gjelder her. Og man står ofte overfor de samme kravene om å øke produktiviteten. Det er klart at bruk av språkteknologiske dataverktøy kan ha mye å si for dette.

Språkteknologi och flerspråkighet inom EU

Krister Lindén

Översättning mellan de olika EU-språken görs idag inom EUs institutioner med hög och jämn kvalitet. Målet är att öka produktiviteten med bibehållen kvalitet. Ur denna synvinkel är maskinöversättning bara en av de många språkteknologier som stöder översättningsarbetet. Produktiviteten kan även ökas med olika språkvårdsredskap, rationalisering av arbetsflödet och med köpta tjänster från översättningsbyråer. Ett annat stort område för översättningsverksamhet inom EUs institutioner är tolkning. Där erbjuder internet och videokonferenstekniker nya möjligheter till kostnadseffektiva lösningar.

Dessutom är det inte bara arbetet inom EUs institutioner som fordrar översättning eller flerspråkiga lösningar. EU har en uttalad målsättning att ”alla borde beredas möjlighet att lära sig kommunicera på två språk förutom sitt modersmål”. Även EUs strävan att öka konkurrensen på en gemensam marknad och att befrämja arbetskraftens rörlighet medför ökade krav på flerspråkighet, översättning och tolkning inom de olika länderna i EU. En rörlig arbetskraft bör på lång sikt lära sig ett av det nya hemlandets nationalspråk. Detta leder till långsiktiga utmaningar för utbildningsväsendet och ger möjligheter för språkteknologi inom datorstödd språkundervisning. På kort sikt fordrar dock akuta situationer i början av en vistelse i ett nytt land tolkning och översättning.

I ljuset av det ovanstående kan man betrakta flerspråkigheten inom EU ur två olika synvinklar, dvs. vilka utmaningar flerspråkighet ger upphov till å ena sidan för EUs institutioner och å andra sidan för EUs medlemsländer. Dessutom är det skäl att betrakta utmaningarna både för översättning, tolkning och språkundervisning, eftersom alla tre behövs i olika situationer.

För EUs institutioner är det en stor utmaning att det blev många fler språkpar vid senaste utvidgning, vilket gav en ökad översättningsmängd. Däremot har budgeten inte ökat i motsvarande grad. För att åstadkomma mera översättning med bibehållen kvalitet har det blivit nödvändigt att prioritera olika översättningsbehov och utnyttja språkteknologiska redskap så långt det är möjligt. Prioriteringen innebär att det i huvudsak översätts lagtexter och offentlig kommunikation som genomgår officiell kvalitetsgranskning, medan interna memorandum och mötesprotokoll inte översätts. Däremot finns maskinöversättning tillgängligt för alla via e-post, vilket får användas på eget ansvar.

För att effektivera översättandet används språkvårdsredskap såsom stavningskontroll och grammatikkontroll, vilket finns i textbehandlingsprogram för de flesta språk. Dessutom används i viss mån terminologikontrollredskap. Oumbärliga är de flerspråkiga terminologidatabaserna som har utvecklats under många år och som numera står som en garant för enhetlig översättning av terminologin inom EU. Den översättningsteknologi som används är översättningsminnen och i viss mån efterredigering av maskinöversättning. Speciellt maskinöversättning mellan närbesläktade språk kan framdeles komma att prioriteras, eftersom det är där maskinöversättningen har bäst möjligheter att lyckas med jämn och hög kvalitet. Eftersom textflödet är stort använder man inom EUs institutioner normal

rationalisering av arbetsflödet via versionshantering, elektroniskt dokumentflöde och resursplaneringssystem.

All översättning är det ändå inte ändamålsenligt att sköta inom EUs institutioner, utan en stor del av översättningen upphandlas av utomstående översättningsbyråer. För att kommunicera med underleverantörerna och leverera och ta emot texter används olika möjligheter som internet erbjuder. Vid ökad upphandling flyttas fokus från själva översättandet till administration av översättning såsom att förhandla fram upphandlingsavtal, att övervaka dokumentflödet och att granska kvaliteten.

Tolkning kan ju inte i nämnvärd grad försnabbas, men åtgärder för en jämn kvalitet till lägre pris kan vidtas. Professionell tolkning kräver vanligen två tolkar per språk under varje arbetspass. I mån av möjlighet försöker tolken få talet på förhand, men oftast finns endast rubriken och ämnesområdet tillgängligt. Under tolkningen har tolken hela tiden tillgång till EUs flerspråkiga terminologidatabaser. Genom videolösningar via internet med delat arbetsbord på datorn bör man i framtiden även kunna ordna tolkning på distans och därmed spara resekostnader och öka tillgången på tolkar genom att utvidga möjligheten till köpta tolktjänster i medlemsländerna. Utmaningarna inom översättning och tolkning för EUs medlemsländer är en följd av bl.a. ökad konkurrens på den gemensamma marknaden som uppstår via ett större utbud av varor och tjänster. Deras dokumentation måste översättas för att varorna och tjänsterna skall kunna marknadsföras och säljas inom hela EU. Utmaningar uppstår även via ökad turism och ökade kundströmmar mellan EU-länderna. För att i praktiken åstadkomma ökad rörlighet på den europeiska arbetsmarknaden både i form av rörlig arbetskraft och en ökad användning av distansarbete via olika typer av call centers, borde alla som bor i EU beredas möjlighet att få språkundervisning.

Den fria rörligheten för varor och människor inom EU leder till utmaningar för medlemsländernas utbildningsinfrastruktur, där språkundervisningen i grundskolor, gymnasier och kvällsgymnasier utgör en viktig komponent. Samtidigt leder arbetskraftsinvandring till behov av varierande kunskaper i främmande språk hos lokala myndigheter, så som skattemyndigheter och tillståndsmyndigheter, där det måste avvägas vad som kan skötas med översättning av vägledande dokument och vad som fordrar tolkning. En stor utmaning är även hur man skall ordna de enskilda invandrarnas kontakter med det lokala rättsväsendet och de lokala vårdinstitutionerna under olika skeden i livet.

I ett multikulturellt och flerspråkigt Europa behöver alla goda språkkunskaper för att kommunicera med varandra. Exempel på tillfälliga eller krisartade kommunikationsbehov i flerspråkiga situationer är t.ex. ett nödsamtal till 112, ett läkarbesök eller en polisanmälning under en semesterresa i något EU-land vars nationalspråk man inte behärskar. Liknande situationer uppstår även i början av en invandrares vistelse i ett nytt hemland. Generellt kan man säga, att långsiktiga och permanenta behov av information på ett visst språk kan skötas med språkundervisning och utbildning. Däremot sköts olika organisationers upprepade behov att sprida information på flera språk kostnadseffektivt med översättning av dokument, medan tillfälliga eller krisartade flerspråkiga situationer alltjämt fordrar tolkning.

Lindén är specialrådgivare i flerspråkig språkteknologi för EU kommissionär Leonard Orban.

Practical Experience with Statistical Machine Translation

Daniel Hardt

It has long been believed by many that statistical techniques alone could not produce acceptable results in MT. This has now been proven wrong. Indeed, statistical MT (SMT) is producing impressive results, not only in the research lab but, increasingly, in the marketplace as well. In this brief note, I will focus on our experience at languagelens, a Danish company we started in 2007 based on SMT technology. The languagelens system is being used to translate patent texts for Lingtech, with very positive results. I will examine the essential features of the current approach taken with the languagelens system, and show the quality of output that can be achieved with this approach. I will also offer some speculation about what the near future holds for MT.

The Past: Why MT "Should be" Impossible

Performing a correct translation would appear to require complex knowledge in many different areas, including grammar, meaning, terminology, as well as general world knowledge. From this point of view, the machine translation project seems daunting indeed, since none of these areas of knowledge have been successfully automated. The key to the success of statistical MT is that it redefines the problem: we bypass these sophisticated knowledge sources and instead, relatively simple statistical techniques are used to derive translation systems by relying on massive amounts of data.

The Present: How it Works

The statistical system searches for the most likely translation, based on previous translation data it has examined. The system builds a Translation Table and a Context Model from that data.

- **Translation Table:** This table is normally built from a large collection of domain-specific translation data. It includes dozens or hundreds of translation options for each entry. Entries can be individual words or multi-word units. Each translation option has an associated probability.

- **Context Model:** This guides the translation system in its choices among all the possible translation options available on the translation table. This is illustrated by the example below:

In	conclusion	,	we	have	several	points	.
I	konklusion	,	vi	har	**flere**	**punkter**	.
Afslutningsvis			vi har		adskillige	bemærkninger	.
På	afslutning		**har vi**		flere af	point	.

The Translation Model gives several options for each position – *In* can be *I* or *På*, *In conclusion* can be *I konklusion*, and so on. The context model examines each such choice, and evalutes them based on what comes before and after. In the above example, the blue text represents the selection with the help of the context model.

The process involves a straightfoward search among simple options derived from the translation data – the power of this approach is that it can accomodate massive amounts of data – data that is now widely available, in ever-increasing quantity.

The Future: the Coming Breakthrough

The languagelens statistical MT system has been in use at Lingtech since September 2007, with great success – Lingtech reports a savings of 80% in producing translation with the languagelens system. This system is built with over 30 million words of domain specific data. The same high quality can be achieved in other domains, with other language pairs, given the availability of relevant data. With the explosion of the amount of data available, there will be a rapid growth in the availability of very high quality MT systems, based on statistical systems such as the languagelens system. Furthermore, as such systems are deployed more widely, the post-edited data produced will lead to a continuous improvement of the quality of the systems deployed, which will in turn lead to the system being deployed more widely. This feedback will very likely lead to an explosive growth in the deployment of statistical MT systems in the very near future.

Maskinoversættelse virker - eller - Automatiseret oversættelse virker[1]

Thomas Bilgram

Abstract: Formålet med denne artikel er, ligesom det var tilfældet med mit indlæg på seminaret, at argumentere for at automatiseret oversættelse virker. Med dette mener jeg, at dagens niveau på automatiseret oversættelse er så godt, at der ganske enkelt ville kunne bygges løsninger, der har kommerciel værdi. Lingtech er selv beviset for, at dette er tilfældet, og i lang tid har været tilfældet. Samtidig vil jeg komme med et par tanker over, hvorfor det ikke er mere i brug, end det er i dag.

Baggrund

Lingtech A/S har oversat tekst med automatiseret oversættelse i over 15 år. Vi blev stiftet i 1993 som et samarbejde mellem de to patentkontorer, Hofman-Bang & Boutard og Lehmann & Ree, med det ene formål at udnytte AT i oversættelsesprocessen. Center for Sprogteknologi (CST) lavede i samarbejde med ejerne PaTrans til dette formål. Det tog ca. 5 år, fra idéen blev undfanget, til Lingtech oversatte størstedelen af teksterne ved brug af PaTrans. PaTrans oversatte derefter i mere end 14 år mange millioner ord patenttekst.

Skiftet til Lingtech Sunbeam

I forbindelse med at PaTrans stod foran en større opdatering, undersøgte Lingtech andre muligheder. Lingtech gik med som partner i et projekt, finansieret af Ministeriet for Forskning og Udvikling. Undervejs i dette projekt bestemte Lingtech sig for at skifte fra regelbaseret til statistisk baseret, automatiseret oversættelse. Baggrunden for dette skift var, at det var enklere og billigere at udvikle nye AT-systemer.

Udgangspunktet

Det er klart, at en automatiseret oversættelse ikke har en kvalitet som den, en oversætter udfører, og det forudsætter derfor et par ændringer af skift i fokus for at evaluere dette på en fornuftig måde, og dermed åbne for at komme videre i en kommerciel sammenhæng:

- Automatiseret oversættelse medfører ikke dårligere kvalitet for kunden, netop fordi den ikke står alene, men fuldføres af en oversætter.
- Det er forretningsmæssigt uinteressant alene at sammenligne resultatet af en automatisk oversættelse med menneskeskabte oversættelser gennem automatiserede målemetoder, som Bleu Score eller lign.[2] Den forretningsmæssige målestok handler om at være i stand til at få en besparelse i omkostningerne og dermed en forretningsmæssig fordel. Dette opnås ved at forvente, at oversætteren kan levere samme kvalitet på kortere tid.
- Det handler derfor ikke om sprog, det handler om effektivitet. For Lingtech er automatiske oversættelsessystemer ikke interessante set ud fra hverken et sprogligt synspunkt, eller et forskningsmæssigt synspunkt.

Hindringer

[1] Denne artikel handler om automatiseret oversættelse. Det er et begreb, som ikke adskiller sig på nogen måde fra maskinoversættelse. Jeg har udelukkende valgt at skifte begrebet for at følge og støtte en udvikling i branchen. Jeg indrømmer at der udelukkende er tale om et skift, som handler om marketing. Ligeledes er MT skiftet til AT.

[2] Det kan dog være interessant at bruge målemetoder af denne type som en indikation af kvaliteten, inden man påbegynder test eller oversættelser.

Når man tager i betragtning, hvilke resultater der er opnået inden for AT, er der overraskende få kommercielle AT-systemer i brug.

Jeg mener at dette skyldes en række faktorer, hvoraf disse er de væsentligste:

- Det, som er forskningsmæssigt interessant, er ikke nødvendigvis kommercielt interessant og modsat. Det er grundlæggende en forskers opgave at løse problemer. Har man løst et problem, kommer der et nyt, heldigvis, som man kan gå videre med. Afledt af dette kommer så, at så længe opgaven med automatiseret oversættelse er defineret som, at man skal oversætte perfekt, lige så længe dukker der nye problemer op, og lige så længe bliver man ikke færdig.
- Ligesom mange andre faggrupper har bremset automatiseringsprocesser inden for deres fagområde, så bremser også oversættere denne proces. Dette forstærkes af, at mange oversættelsesfirmaer ledes af oversættere, og at mange kunder har oversættere som indkøbere.

Dertil kommer to misforståelser:

- Som oversættelsesfirma oversætter vi ikke tekst, vi oversætter dokumenter. Et godt AT-system kan aflaste dele af den proces, men for at få succes skal den kunne bygges ind i arbejdsprocesser, som bruges således, at der ikke dukker nye arbejdsopgaver op, som tager den tid som er sparet ved oversættelsen.
- Ikke alle oversættelsesopgaver er velegnede til AT-systmer. En stor del af de opgaver, som oversættes af bureauer, er tekster, som ikke er velegnet til automatiserede processer. Der skal ændres for meget andet end ord. En del tekster skal tilpasses den kultur, som findes i landet, og det betyder, at oversætterens opgave ikke kun er sproglig oversættelse, men også kulturel tilpasning. Ligeledes skal teksten ofte genetableres i en grafik eller et design i et software, som ikke let kan håndteres af AT-systemer.

Et AT-system skal derfor have et API, som gør at det kan integreres i andre systemer, eller det skal bygges ind i systemer fra gang til gang. Nøgternt set handler det om, at god software ikke kun fokuserer på selve opgaven, der skal løses, men også på design, brugervenlighed og funktion.

Hvor er så mulighederne?

Jeg tror, at nogen skal se tingene fra den anden ende. Starte med oversætteren, dvs. starte med den egentlige bruger, og så få bygget en løsning, som virker. Det er sket rundt omkring, og mange af de løsninger, som fungerer i dag, er bygget i større firmaer med egen udviklingsafdeling. Disse firmaer har ofte store oversættelsesopgaver, som retfærdiggør investeringer i denne størrelse. Desværre er disse firmaer sjældne, eller oversættelse har ikke det fokus, der skal til for at få truffet beslutninger om dette på tilstrækkeligt højt niveau.

Hvor er der så andre muligheder?

Jeg tror, der skal et skift til – måske endda et stort skift, hos udviklerne af AT-systemer. Et større samarbejde med oversætterbranchen, et samarbejde med softwaredesignere og folk med erfaring i forandring af arbejdsprocesser. Dette kunne føre til, at der kom AT-systemer, som ikke kun oversætter tekst i god kvalitet, men også hele dokumenter og også på en måde som branchens brugere kan se fornuften i og ikke mindst på en måde, som gør det lettere for oversætteren.

Det, jeg ikke mener

Under den mundtlige fremstilling af disse problemstillinger på konferencen kunne jeg fornemme en tøven over for mine synspunkter. Jeg vil derfor gerne understrege, at jeg har stor respekt for den akademiske arbejdsmetode og de resultater, den fører med sig. Jeg har

gjort det til mit levebrød at kommercialisere disse resultater, og jeg kan se, hvor meget arbejde det har krævet at levere AT-systemer i den kvalitet, vi har i dag. Jeg ved derfor også, hvor vigtigt det er, at der fremadrettet forskes og udvikles inden for dette fagområde.

Kursplaneöversättaren – ett system för översättning av kursplaner från svenska till engelska

Eva Pettersson

Kursplaneöversättaren är ett maskinöversättningssystem speciellt anpassat till översättning av kursplaner från svenska till engelska. Systemet är utvecklat av en forskargrupp vid institutionen för lingvistik och filologi vid Uppsala universitet. Sedan våren 2006 står avknoppningsföretaget Convertus AB för vidareutveckling och drift av Kursplaneöversättaren.

Kursplaneöversättaren har en regelbaserad kärna, med analys-, transfer- och genereringsmoduler. Därtill finns diverse upphämtningsstrategier som utnyttjas i de fall där lexikon och grammatik inte räcker till för att ge en fullgod översättning. De viktigaste upphämtningsstrategierna består i:

1. tillvaratagande av partiella analyser i de fall där analysgrammatiken inte når ända fram,
2. en statistisk språkmodell för generering när den regelbaserade genereringsmodulen inte räcker till samt
3. sammansättningsanalys av ord utanför lexikonet, och översättning av de ingående delarna.

Lexikonet kan sägas bestå av tre delar: ett svenskt lexikon för analyssteget, ett engelskt lexikon för genereringssteget samt ett svensk-engelskt lexikon för transfersteget.

Lexikonet har byggts upp utifrån de kursplaner som fanns inlagda i Uppsala universitets kursplanedatabas (Selma) i augusti 2007. Översättningsrelationerna har så långt möjligt definierats på basis av de översättningar som fanns inlagda i kursplanedatabasen. Då mindre än en tredjedel av kursplanerna hade en översättning i databasen, har övriga ord tilldelats en översättning med andra hjälpmedel, såsom ordböcker och tillgängliga resurser på webben. Det resulterande lexikonet består av totalt 35 221 översättningsrelationer, fördelade på olika ämnesområden. Därutöver tillkommer ett allmänlexikon om 10 737 ingångar och ett kärnlexikon om 2135 ingångar, vilket ger en total vokabulär om 48 093 översättningsrelationer. Därtill kommer ca 3000 lexikala transferregler för översättning av ord i kontext.

Lexikonstrukturen är hierarkisk, och i översättningsprocessen sker lexikonuppslagningen i en fördefinierad ordning, där ordet först slås upp i det mest specifika lexikonet. Ord som inte går att finna i något av lexikonen körs genom en sammansättningsanalys, och om sammansättningsanalysen lyckas översätts de ingående delarna var för sig. Om även sammansättningsanalysen misslyckas, kopieras det svenska ordet helt enkelt över till den engelska sidan.

Utöver lexikon och grammatik, innehåller Kursplaneöversättaren en minnesfunktion. Här lagras de översättningar som användarna har granskat, redigerat och godkänt. Vid översättning av nya kursplaner, konsulterar systemet översättningsminnet som ett första steg i översättningsprocessen. Om en mening återfinns i minnet, väljs den översättning som finns lagrad i minnet. Initialt innehöll minnet segment som sedan tidigare fanns manuellt översatta i kursplanedatabasen. Efter hand som systemet används, växer minnet med de översättningar

som granskas och godkänns av användarna. Varje fakultet/ämnesområde har sitt eget översättningsminne.

För Uppsala universitets del är Kursplaneöversättaren integrerad i kursplaneverktyget Selma, och processen inleds med att användaren skriver in sin svenska kursplan i dess gränssnitt. Därefter trycker han/hon på knappen för stavningskontroll. Kursplaneöversättaren utför då en stavningskontroll som bygger på samma lexikon som översättningssystemet har. Ord som saknas i lexikonet rödmarkeras. Efter att användaren har genomfört eventuella korrigeringar av kursplanen i enlighet med stavningskontrollen, klickar han/hon på knappen för oversättning. Kursplanen läggs då i en översättningskö, varifrån de hämtas av översättningssystemet och bearbetas en efter en.

När översättningen är klar, skickas ett mejl till den som har utsetts till översättningsansvarig på den institution som kursplanen gäller. I mejlet finns en länk till ett efterredigeringsgränssnitt, där användaren ser den svenska kursplanen tillsammans med sin översättning. Ord i den engelska kursplanen som saknas i lexikonet är rödmarkerade.

När användaren har redigerat klart sin översättning, klickar han/hon på en knapp för godkännande, vilket gör att översättningen sparas och lagras i Selma-databasen. Dessutom lagras samtliga översättningar i översättningsminnet, så att nästa gång en användare vill översätta samma textsegment (mening, rubrik etc.) kommer de granskade och godkända översättningarna att väljas.

Driftsättning av Kursplaneöversättaren genomfördes vid Uppsala universitet i september 2007, och vid Umeå universitet i september 2008. Användarna har möjlighet att ställa frågor om systemet via Kursplaneöversättarens hemsida. Convertus har analyserat de önskemål om vidareutveckling som hittills har inkommit från användarna, och uppdaterat systemet i enlighet med detta. Just nu diskuteras frågan om vem som ska vara översättningsansvarig.

Invitasjon

Nordisk arbeidsseminar oversettingsteknologi
23.-24. oktober 2008, Oslo, Norge

Sted: Språkrådet, Observatoriegt. 1 b, Oslo, Norge

En viktig politisk oppgave for språknemndene i Norden er å medvirke til at nye teknologiske produkter og tjenester blir tilgjengelige på de respektive språkene. Språknemndene har en arbeidsgruppe for språkteknologi som årlig arrangerer et arbeidsseminar for spesielt inviterte deltakere. Årets tema er ***oversettingsteknologi.*** På seminaret vil ulike oversettingsprogrammer og -teknologier bli presentert, det vil bli diskusjoner om relevante problemstillinger på området og hvilke utfordringer vi står overfor. Det er et uttalt mål at deltakerne skal representerer sentrale aktører på området: forskere, produktutviklere og språkrøktere, og gi innledere og deltakere mulighet for utveksling av ideer og erfaringer. Korte sammendrag av innledningene vil bli distribuert til deltakerne før seminaret, en sammenfattende oppsummering og presentasjonene fra seminaret vil bli gjort tilgjengelig på nettet i etterkant.

Globaliseringen er en utfordring for små språksamfunn, og desto viktigere blir gode oversettingsprodukter som en del av realiseringen av en overordnet strategi for å styrke nasjonalspråkene parallelt med at man benytter andre språk på den internasjonale arenaen. Universitets- og forskningsmiljøene velger ofte engelsk som publiseringsspråk. En strategi for parallellspråklig publisering er et incitament til å øke innsatsen for bedre oversetterverktøy. De fleste programmene som er tilgjengelige for oversetting til/fra de nordiske språkene, gir ikke fullgode oversettelser slik en kvalifisert oversetter vil levere. Dette seminaret håper vi kan bidra med innspill til å bedre denne situasjonen.

Arbeidsseminaret har begrenset antall plasser. Det er derfor nødvendig at den lokale arrangøren får tilbakemelding om du ønsker å delta eller ikke. Invitasjonen sendes til et visst antall personer fordelt på alle de nordiske landene.

Om du selv ikke kan eller ønsker å delta, må du gjerne gi arrangørene navn på andre personer som kan være aktuelle, og da er det fint om du opplyser om e-post-adressen. Invitasjonen sendes *bare* elektronisk. Vedlagt finner du programmet for konferansen. Etter påmelding vil deltakerne få tilsendt korte sammendrag av innleggene.

Arbeidsgruppens arbeid er støttet økonomisk via midler fra Nordens språkråd, og overnatting en natt i Oslo dekkes for deltakerne.

Påmelding til Torbjørg Breivik, e-post: Torbjorg.Breivik @ sprakradet.no innen 10. oktober 2008.

Språkrådet

Nordisk arbeidsseminar i oversettingsteknologi
23.-24. oktober 2008, Språkrådet, Oslo, Norge

Program

12:00	**Ankomst, registrering og lunsj**
13:00	Velkommen v/Sylfest Lomheim, Språkrådet
13:05	Introduksjon til seminaret v/Torbjørg Breivik, Språkrådet
13:15	Anna Sågvall Hein: Maskinöversättning i teori och praktik. Var står vi i dag?
13:45	Barbara Gawronska: Korpora och konkordanser i mänsklig och automatisk översättning
14:15	Diskusjon
14:45	Kaffe
15:15	Carol B. Eckmann: Hvordan påvirker valg av dataverktøy kvaliteten i tekstproduksjonen?
15:45	Krister Lindén: Språkteknologi och flerspråkighet innom EU
16:15 – 16:45	Diskusjon
19:00	**Middag på Al Chouf**
09:00	Daniel Hardt: Practical Experience with Statistical Machine Translation
09:30	Thomas Bilgram: Maskinoversættelse virker. Hvor skal der nu sættes ind, for at få en større andvendelse af den i oversættelsesbranchen?
10:00	Diskusjon
10:30	Kaffe
11:00	Lars Nygård: Automatisk oversettelse mellom skandinaviske språk
11:30	Eva Petterson: Kursplaneöversättaren – ett system för översättning av kursplaner från svenska till engelska
12:00	Matthew McGowan: Nynodata's Temasearch – resources and techniques for multilingual search
12:30	Lunsj
13:30	Oppsummerende diskusjon
14.30	**Avslutning og avreise**

Språkrådet

Deltakerliste, nordisk arbeidsseminar om oversettingsteknologi, 23. – 24. oktober 2008, Oslo

Navn	e-post	
Andersen, Gisle	Gisle.andersen@nhh.no	Norges handelshøyskole, Norge
Bilgram, Thomas	Thomas.bilgram@lingtech.com	Lingtech AS, Danmark
Breivik, Torbjørg	Torbjorg.Breivik@sprakradet.no	Språkrådet, Norge
Domeij, Rickard	Rickard.Domeij@sprakradet.se	Språkrådet, Sverige
Dysvik, Sylvi	Sylvi.dysvik@mfa.no	Utenriksdepartementet, Norge
Dyvik, Helge	Helge.dyvik@lili.uib.no	Universitetet i Bergen
Eckmann, Carol B.	ceckmann@online.no	Selvstendig næringsdrivende
Gambäck, Björn	gamback@sics.se	
Gawronska, Barbara	Barbara.gawronska@uia.no	Universitetet i Agder, Norge
Hagen, Hege	Hege.Hagen@mfa.no	Utenriksdepartementet, Norge
Hagensen, Lene	Lene.hagensen@lingtech.com	Lingtech AS, Danmark
Hardt, Daniel	Dh.isv@cbs.dk	Handelshøjskolen i København, Danmark
Kirchmeier-Andersen, Sabine	sabine@dsn.dk	Dansk Sprognævn, Danmark
Lenchow, Tove	tovel@no.ibm.com	IBM Norge
Lindén, Krister	Krister.Linden@helsinki.fi	Helsingfors universitet, Finland
McGowan, Matthew	mat@nynodata.no	Nynodata AS, Norge
Møller, Margrethe H.	mhm@sitkom.sdu.dk	Syddansk universitet, Danmark
Nordli, Finn Christian	Finn.Christian.Nordli@mfa.no	Utenriksdepartementet, Norge
Nygård, Lars	Lars.nygaard@iln.uio.no	Universitetet i Oslo, Norge
Offersgaard, Lene	leneo@hum.ku.dk	Københavns universitet, Danmark
Pettersson, Eva	Eva.Pettersson@convertus.se	Convertus AB, Sverige
Reuter, Mikael	Mikael.Reuter@focis.fi	Forskningscentralen för de inhemska språken, Finland
Seljebotn, Bjørn	bjorn@nynodata.no	Nynodata AS, Norge
Selänniemi, Juhani	juhani@selanniemi@lingsoft.fi	Lingsoft AB, Finland
Sågvall-Hein, Anna	Anna.Sagvall-Hein@convertus.se	Convertus AB, Sverige
Vries, Johan de	johannes@devries.as	Devries AS, Norge
Widenius, Risto	Risto.widenius@kotus.fi	Forskningscentralen för de inhemska språken, Finland