PANORAMA
(PArallele NORdiske Annoterede Multilinguale korporA)

Summary:
PANORAMA-projektet har som hovedformål at kompilere og opmærke parallelle, bilinguale eller multilinguale, korpora for samtlige nordiske sprog: dansk, norsk, svensk, finsk, islandsk, færøsk, samisk og grønlansk. De hertil nødvendige grammatiske analyseredskaber (taggere/parsere) og søgeinstrumenter videreudvikles eller nyudvikles (alt efter sprog), samt tilpasses paralleliseringsformålet. Projektet har således et stærk sprogteknologisk islæt, og vil fremme udviklingen af IKT i Norden, herunder nordisk-flersprogede ordbøger, internetsøgemaskiner, maskinoversættelse m.m. Materialerne vil blive frit tilgængeliggjort på Internettet igennem et specialudviklet tværsprogligt søgeinterface.

PANORAMA er konciperet som et 4-årigt samarbejde mellem universiteterne i Odense (SDU), Oslo og Tromsø. Alle deltagende forskere medbringer ekspertise fra tidligere sprogteknologiske og ikke mindst korpuslinguistiske projekter.

Målsætning:
Projektet sigter overordnet på at kompilere, opmærke og tilgængeliggøre parallele korpora (flersproglige tekstdatabaser) for samtlige nordiske sprog, idet der tages særligt hensyn til også de mindre sprog i Norden.

Materialet formateres og tilgængeliggøres på en måde der vil understøtte en lang række sprogteknologiske formål, såsom elektroniske bilinguale eller multilinguale ordbøger, flersprogede nordiske internetværktøjer, flersproglig tv-tekstning samt egentlig maskinoversættelse m.m.

Forslagsstillerne finder det absolut nødvendigt at der udvikles nordisk-udrettede versioner af sprogteknologiske produkter som ovennævnte, hvis Norden skal undgå yderligere domænetab til engelsk, og hvis borgerne i Norden skal have mulighed for frit at kunne udnytte den stadig større, men ofte monolingualt publicerede, informationsmængde i det nordiske Internet.

Desuden vil parallelførte sprogdata for de nordiske sprog fremme viden om og forståelse for ligheder og forskelle på tværs af de nordiske sprog, til gavn for både undervisere og forskere i de enkelte lande. Det skal understreges at produktet således vil have es væsentlig nytteværdi ikke kun for institutionelle brugere, men for almindelige borgere. Fx vil en skolelærer, korrespondent eller professionel oversætter kunne finde ækvivalenter for et givent ord på andre nordiske sprog (endda i en naturlig sætningskontekst!)
i projektets frit tilgængelige korpussøgnings-grænseflade.

Spredning:
Projektets forskningsresultater vil blive publiceret, samt præsenteret i det relevante konferencemiljø,

og de kompilerede og opmærkede korpora vil blive tilgængeliggjort igennem et specialudviklet søgeinterface på internettet, og således stå til rådighed til forskning og undervisning i samtlige nordiske lande. Idet vi vil dække alle nordiske sprog, forventes at forskellige institutioner vil kunne bruge produktet til at øge ikke mindst den komparative viden om - og herigennem den gensidige forståelse af - de nordiske sprog.
De to for 2007 planlagte (samt efterfølgende) workshops vil være åben for eksterne deltagerne fra værtslandenes universitetsmiljø, og sprogteknologispecialister for ikke internt forsker-repræsenterede sprog vil kunne inviteres til at deltage.

Overordnet målsætning

Projektets hovedformål er at kompilere, opmærke og tilgængeliggøre parallelle korpora for de nordiske sprog: dansk, norsk, svensk, finsk, islandsk, færøsk, samisk og grønlansk. Korpusinventaret vil bestå af tekster skrevet på eet af disse sprog og oversat til et eller flere af de øvrige sprog. Der vil blive inddraget såvel bilinguale tekster (fx skøn- og fagliteratur, nationale udgivelser for minoritetssprog) og multilinguale tekster (fx visse fællesnordiske publikationer).

Med dette prosjektet vil vi særlig fokusere på parallelle tekster hvor også de mindre nordiske sprog er repræsenteret. Det vil således være målet at hvert af de "større sprog" (norsk, dansk, svensk, finsk) indgår i et bi- eller multilingualt korpus med mindst eet af de "mindre sprog" (islandsk, færøsk, samisk, grønlandsk).

Paralleliseringen vil som udgangspunkt gennemføres på sætningsniveau, men idet de enkelte tekster vil blive opmærket med morfosyntaktisk information, vil en dybere, leksikalsk eller syntaktisk, parallelisering kunne gennemføres for et eller flere af de sprogteknologisk mere veludviklede sprogpar.

Støtte til nordisk sprogteknologi

Denne IKT-ressource, der igennem et brugervenligt søgeinterface på internettet vil være gratis og nemt tilgængelig i hele det nordiske område, vil være af uvurderlig betydning for forskning og undervisning i hele NORA-området, samt ikke mindst for næringslivet, bl.a. den nordiske forlags- og IT-branche. Mulige anvendelsesområder er således:

* Tværnordiske forskningsprojekter med et nordisk sprogligt fokus
* Udvikling af flersproglige ordbøger og terminologiske arbejde
* Sammenligning og sproglig bevidstgørelse i undervisningsøjemed
* Udvikling af flersproglige, nordiske internetsøgemaskiner, spørgsmål-svar-systemer etc.
* Supportering af tolke og oversættere
* Udvikling af maskinoversættelsessystemer

Ud over de kompilerede/annoterede parallelkorpora vil også de i projektet udviklede værktøjer for grammatisk opmærkning, parallelføring og indeksering/søgning kunne finde direkte eller indirekte anvendelse i det nordiske forsknings- og virksomhedsmiljø, til gavn for nordisk sprogteknologi.

Forudsat projektet kører, som planlagt, over alle 4 år, forventer vi at kunne se, følge og supportere denne slags sekundære, applikative spin-offs fra tredjepart, herunder samarbejdende offentlige institutioner, forskningsenheder eller virksomheder.

Tværsproglighed som vehikel for nordisk integration i informationssamfundet

En af de vigtigste faktorer for at opnå et godt samarbejde mellem de nordiske lande er en god indbyrdes forståelse, uproblematisk adgang til hinandens skriftlige udgivelser og en større eksponering for andres sprog end det er tilfældet i dag. Manglende kendskab til den sproglige variation i Norden begrænser desuden udviklingen af et fælles nordisk marked, samhandel og virksomhedskultur. På det kulturelle område er det ikke mindst den passive sprogforståelse der er nøglen til andre nordiske landes kulturskat, nyhedsformidling, arbejdsmarked m.m.

I en moderne verden foregår en stor del af al sproglig udveksling ved hjælp af informationsteknoglogi, og hvis de nordiske sprog fortsat skal nyde godt af et historisk og/eller etymologisk funderet sprogfællesskab snarere end at satse ensidigt på en engelsk-international udvikling, er det ikke mindst sprogteknologien der kræver en bevidst - og aktivt støttet - indsats for at forhindre at de (i international sammenlingning små) nordiske sprog sakker agterud i sammenligning med fx engelsk, tysk eller japansk. Ud fra en IKT-synsvinkel vil især følgende ressource vinde stadig større betydning i fremtiden:

* Gode bi- eller multilinguale ordbøger og termbanker mellem de sprog der tales i området, fortrinsvis i et portabelt elektronisk format, og tilgængelig via Internettet.

* "Ordbogslignende" instrumenter, som flersprogede søgbare tekster (ovennævnte parallelkorpora), relationelle databaser (fx multilinguale ontologier), maskinoversættelsessystemer for tv-undertekster, hjemmesider etc.

* Flersproglige søgemaskiner or ekspertsystemer, der giver borgerne i norden adgang til materiale ikke alere fra deres eget land/sprog, men også fra resten af Norden. Nordisk-engelske eller nordisk-tyske systemer vil yderligere øge den tilgængelige informationsmængde.

Gode og omfattende tværnordiske ordbøger findes imidlertid i dag kun i beskedent omfang eller slet ikke, og de termbaser der trods alt er udarbejdet, har ofte ikke den ønskede dækningsgrad. Det er her, leksikografen har brug for parallelkorpora for at kunne finde oversættelsesekvivalenter for ord, udtryk og termer i forskellige kontekster.

Brugerperspektivet: Et nordisk Internet

Et hav af relevant nordisk information er i princippet tilgængelig på Internettet for den enkelte nordiske borger: Lovtekster, rapporter, nyheder, helseinformation, undervisningsmateriale, forbrugervejledning m.m. Imidlertid er specialiserede tekster ofter kun publiceret på et enkelt sprog, og selv hvor borgeren kan forstå tekster fra beslægtede nabosprog (som fx. for dansk,svensk og norsk), er det alligevel ikke umiddelbart muligt at finde frem til teksterne på ved hjælp af en søgemaskine, ganske enkelt fordi brugeren kun har passiv fremmedsprogskendskab, og ikke nødvendigvis kan gætte sig til hvordan et givent søgeord skal staves, eller hvilket synonymt der skal benyttes, på et andet nordisk sprog. Ikke mindst for modersmålstalere fra minoritetssprog vil det ofte være væsentligt nemmere at læse videre på en tekst på matrikssproget end overhovedet at finde frem til den. Endeligt vil det være nyttigt for Internet-brugeren at kunne få løbende oversættelsessupport for ord eller tekststumper, fx. igennem en ligeledes Internet-baseret nordisk ordbog.

PANORAMA vil skabe forudsætningerne for at situationen kan forbedres - parallelkorpora og analyseværktøjer, samt nem adgang til begge dele, vil gøre det muligt at udvikle flersprogede nordiske søgemaskiner m.m., og inkluderingen af de mindre sprog vil sikre nordisk ligestilling på informationsområdet.

Projektets sproglige og tekstuelle dækningsgrad

Projektet sigter på at dække hele det nordiske område, og vil således indrage tekster og sprogteknologiske ressourcer for følgende sprog:

* Skandinavisk-germanske matriks-sprog: Dansk, Svensk, Bokmål, Nynorsk
* Andre germanske sprog: Islandsk, Færøsk
* Finsk
* "Mindre, ikke-germanske sprog": Grøndlandsk, Nordsamisk, Lulesamisk

Af disse sprog er dansk, islandsk, norsk og samisk direkte representeret i projektgruppen. Deltagerne har herudover et godt samarbejde med kolleger der arbejder med de øvrige nordiske sprog, svensk, færøsk, grønlandsk og finsk, som alle har været repræsenteret i et eller flere af de tværnordiske PaNoLa-sprogteknologi-projekter. Ud over de nævnte sprog, og for så vidt det bliver relevant for standardiseringen af det terminologiske eller komparative arbejde, vil også engelsk og tysk i nogen grad kunne inddrages som parallelsprog, ikke mindst hvor de tilgængelige flersproglige tekstkilder i forvejen rummer disse sprog i deres uopmærkede form (fx EU-tekster, manualer ...).

Projektdeltagerne: Et fælles metodologisk ståsted

Projektdeltagerne kan samlet trække på en stor ekspertise indenfor datalingvistik og IKT. Således har universiteterne i Odense (Eckhard Bick, http://beta.visl.sdu.dk/visl/about/eckhard.html), Oslo (Kristin Hagen, Anders Nøklestad) og Tromsø (Trond Trosterud, http://uit.no/humfak/tilsette/119) udviklet såkaldte taggere og parsere for henholdsvis dansk, norsk og samisk, der alle benytter sig af den sprogteknologisk særdeles robuste Constraint Grammar-teknologi, og herigennem giver den planlagte korpusopmærkning et fælles metodologisk ståsted. Eckhard Bick har herudover udviklet CG-systemer for en række andre germanske og romanske sprog, og Trond Trosterud har udviklet morfologiske analysemaskiner (finite state systems) for færøsk og grønlandsk.

ISK (Syddansk Universitet) og Tekstlaboratoriet (Universitetet i Oslo) har begge mange års erfaring med opmærkning og tilgængeliggørelse af korpora, og administrerer allerede i dag forskellige brugergrænseflader for korpussøgning til både forsknings- og undervisningsbrug (http://corp.hum.sdu.dk og http://www.hf.uio.no/tekstlab/korpus.html). Forskere fra de to sidstnævnte institutioner har tidligere samarbejdet i et fællesnordisk korpusprojekt (The Nordic Treebank Network), og Universitetet i Oslo er hjemsted for flere egentlige parallelkorpora (Sofies Verden-træbank-korpusset og det europæisk-udrettede OPUS-korpus).

Tidsplan:

Januar-Juni 2007 (alle partnere): Indsamling af tekstmateriale, afklaring af copyright, konvertering og formatering, indføjelse af metadata (sprog, kilde, periode, teksttype m.m.)

Juli-Oktober 2007 (Odense og Oslo): Indeksering, sætningsparallelisering og integration af korpora i et søgeinterface. Evalueres og tilpasses ny opmærkning efterfølgende.

April-December 2007: Opmærkning af tekster for dansk (Odense), norsk (Oslo), samisk (Tromsø). Fortsætter ind i 2008.

September - December 2007: Lemmatisering af tekster for svensk, finsk, islandsk, færøsk, grønlandsk. Fortsætter ind i 2008. Alle tre institutioner deltager, ekspertise og kontakter fra PaNoLa-projekterne benyttes.

November - December 2007: Integration af opmærkning i søgeinterfacet. Forsøg med dybere parallelføring for et udvalgt sprogpar (Forberedelse for 2008).

Forår 2007: 3-dages workshop i Norge (planlagt: Marts - Oslo)
Efterår 2007: 3-dages workshop i Danmark (planalgt: September - Odense)

Forventet resultat:

1. Parallelkorpora (flersprogede tekstdatabaser) for samtlige nordiske sprog, således at hvert sprog paralleliseres med mindst eet andet, samt at hvert "stort sprog" (dansk, svensk, norsk, finsk) har mindst en parallelisering med et "lille sprog" (islandsk, samisk, færøisk, grønlandsk) og vice versa.

2. Sætningsparallelføring og opmærkning af disse data.

3. Opmærkningsværktøjer (hvor de ikke i forvejen findes), med lemmatisering som minimum og Constraint Grammar-baseret morfosyntaktisk opmærkning som tilstræbt niveau.

4. En fællesnordisk brugerflade for søgning i de kompilerede og opmærkede korpora.

5. Dokumentation og publikationer.

Opfølgning, evaluering & dokumentation:

Siden nærværende ansøgning omhandler det første af planlagt 4 projektår, vil det være muligt at garantere et kontinuerligt arbejde med dataindsamling såvel som opmærkning, tilgængeliggørelse og brugertilpasning. Således vil erfaringerne fra de sprogteknologisk mere "velstillede" nordiske kernesprog som dansk og norsk blive fulgt op af videns- og teknologioverførsel til de mindre sprog, herunder samisk, færøisk og grønlandsk. Ligeledes vil korpusbrugerfladen løbende blive evalueret og forskellige institutionelle brugere søgt tilgodeset. Forsåvidt sprogteknologiske værktøjer skal nyudvikles til formålet, vil disse blive dokumenteret og evalueret (taggere, parsere, alignere etc.). Endeligt er det intentionen at følge op på dataindsamlingsdelen, således at indsamlingsarbejdet vil fortsætte i de efterfølgende år, bl.a. for at sikre at flestmulige nordiske sprogkombinationer dækkes af paralleltekster.
Vi kan garantere at korpusværktøjerne også vil være tilgængelige efter projetperiodens udløb, fordi IT-centrene på ISK (Syddansk Universitet) og Tekstlab (Oslo Universitet) vil hoste og supportere materialet, og fordi formatering/indeksering vil blive foretaget under hensyntagen til internationale standarder og størrest mulig portabilitet.