REFERAT FRA ELRC-WORKSHOP I DANMARK, 7. MARTS 2016

Program
10:00 – 10:20 **Velkomst**
Michael Vedsø, Europa-Kommissionens repræsentation i Danmark
Sabine Kirchmeier, Dansk Sprognævn

10:20 – 10:30 **Målsætninger**
Andrejs Vasiļjevs, ELRC/Tilde

10:30 – 10:40 **Europa og flersprogethed**
Derrick Kinck Olesen & Uffe Sonne Svendsen, Europa-Kommissionen, Generaldirektoratet for Oversættelse

10:40 – 11:20 **Sprog og sprogteknologi i Danmark**
Sabine Kirchmeier, Dansk Sprognævn

11:20 – 11:45 **Diskussion: Flersprogethed i den offentlige sektor – hvordan møder vi udfordringen?**
Deltagere: Sigurd Slot Jacobsen, Konkurrence- og Forbrugerstyrelsen,
Anne-Mette Olsen, Region Sønderjylland-Schleswig, Trine Engelberg (International House, Københavns Kommune)

12:00 – 12:20 **Automatisk oversættelse: Hvordan fungerer det?**
Anders Søgaard, Center for Sprogteknologi, Københavns Universitet

12:20 – 12:40 **Hvordan får de offentlige institutioner gavn af CEF.AT-platformen?**
Spyridon Pilos, Directorate General for Translation, European Commission

12:40 – 13:00 **Tilde – Maskinoversættelse til e-handel og den digitale offentlige sektor - et eksempel**
Rihards Kalniņš, ELRC/Tilde

14:00 – 14:30 **Hvilke data er der behov for? Hvorfor?**
Andrejs Vasiļjevs, ELRC/Tilde

14:30 – 15:00 **De juridiske rammer for levering af data, European Data Portal**
Cathrine Lippert, Digitaliseringsstyrelsen

15:00 – 15:30 **Diskussion: Data og sprogresurser i Danmark**
Peter Juel Henrichsen, CBS/IBC, Bolette Sandford Petersen, Københavns Universitet , Bodil Nistrup Madsen, CBS/IBC/DANTERMcentret

16:00 – 16:30 **Data- og sprogresurser: Tekniske og praktiske aspekter**
Andrejs Vasiļjevs, ELRC/Tilde (præsentation)

16:30 – 17:00 **Interaktiv session: Den bedste fremtidige praksis – offentlige institutioners data skal bruges til at forbedre systemet. Hvordan engagerer vi os?**

Andrejs Vasiļjevs, ELRC/Tilde og Sabine Kirchmeier, Dansk Sprognævn/Danish Language Council

17:00 – 17:10 **Et sprogteknologisk netværk i Danmark**
Bolette Sandford Petersen, Københavns Universitet

17:10 – 17:25 **Opsamling, konklusioner og tilsagn**

Deltagere
Der deltog ca. 70 personer i mødet, som repræsenterede et bredt udsnit af offentlige institutioner, forskningsinstitutioner og virksomheder i Danmark, heriblandt:

Region Sønderjylland, Københavns Kommune - International House, Konkurrence- og Forbrugerstyrelsen, Digitaliseringsstyrelsen, Folketingets Administration, DR, Det europæiske Miljøagentur, Region Hovedstaden - Center for It, Medico og Telefoni, Sundhedsplatformen, Arbejdsskadestyrelsen, Aarhus Kommune - Borgerservice, Ankestyrelsen, Forsvarsministeriets Ejendomsstyrelse – Udbudsafdelingen, Den Europæiske Unions Domstol, Lægemiddelstyrelsen, Udlændinge-, Integrations- og Boligministeriet, Udenrigsministeriet, Styrelsen for Forskning og Innovation, Europa-Kommissionens repræsentation i Danmark , EU-Kommissionen - Generaldirektoratet for Oversættelse Nordisk Ministerråd, Dansk Sprognævn, CBS, Københavns Universitet, Syddansk Universitet, Forbundet Kommunikation og Sprog, Translatørforeningen, IBM WWTO, IBM Danmark Aps, Mette Klingsten Advokatfirma, Gyldendal, World Translation A/S, Øresundsadvokater, KMD A/S, Confianza Translation, abcSPROG ApS, Tettinek Text & Translation, SOKKELUND & CO, Ordforsyningen, Eva-Maria Arntz ApS, Studieskolen, Tilde

Referat

1. Velkomst

Michael Vedsø (Europa-Kommissionens repræsentation i Danmark) indledte med at forklare at formålet med workshoppen er at skabe opmærksomhed om maskinoversættelsesplatformen CEF. AT, som udvikles af EU's Generaldirektorat for Oversættelse. Det er en videreudbygning af MT@EC, som er EU-Kommissionens eksisterende maskinoversættelsessystem. MT@EC bygger på de oversættelser i EU-regi der er lavet igennem de seneste 20 år. MT@EC bruges allerede i offentlige institutioner i alle EU's medlemslande og er gratis. MT@EC bygger på EU-data på 24 sprog. Der var 940 mio. sætninger i systemet ved udgangen af 2015, og det vokser med 2 mio. sætninger pr. md. CEF står for Connecting Europe Facility. AT står for Automated Translation Platform.

Oversættelse af dokumenter udgør Europa-Kommissionens største udgift. Årligt bliver der brugt 1,1 milliarder euro på oversættelse og tolkning i de forskellige EU-institutioner. Dette skal ses i lyset af at EU-borgere har ret til at henvende sig til EU på deres eget sprog og tilsvarende kan forlange at få svar tilbage på deres eget sprog, samt at oversættelse er en del af lovgivningsproceduren i EU. Erfaringen viser at oversættelse og korrekturlæsning af originaltekster bidrager til en bedre lovgivning.

Der er tekniske og juridiske udfordringer i at dele de data som offentlige myndigheder allerede har. Det er den anden grund til at afholde workshoppen.

Hvis udbygningen af MT@EC bliver realiseret, har EU-Kommissionen stor tiltro til at det kan udgøre et væsentligt bidrag til virkeliggørelsen af et digitalt indre marked.

Sabine Kirchmeier (Dansk Sprognævn) pegede indledningsvist på at en konkret dansk udfordring er at fx bilag til love m.v. ikke altid foreligger på dansk, ligesom baggrundsmateriale for beslutninger i Folketinget ikke altid bliver oversat. En praksis hvor dokumenter som har betydning for politiske

beslutninger eller lovgivningen, ikke findes på dansk, er ikke kun et sprogligt, men også et demokratisk problem.

2. Målsætninger

Andrejs Vasiljevs (Tilde/ELRC) pegede på at workshoppens hovedfokus var:

- At øge bevidstheden om værdien og vigtigheden af de data som man som offentlig myndighed ellers blot betragter som dokumenter. De kan være meget nyttige i arbejdet med at udvikle den automatiserede oversættelse så man får adgang til systemer af bedre kvalitet der kan oversætte mellem ens eget sprog og de andre europæiske sprog.
- At opfordre danske offentlige institutioner til at dele data og derved bidrage aktivt til at forbedre maskinoversættelsestjenesten for at støtte det europæiske samarbejde - og i virkeligheden også støtte det danske sprog.
- At hjælpe deltagerne med at forstå og løse praktiske og juridiske problemer ved at dele data med EU's Generaldirektorat for Oversættelse med henblik på at forbedre maskinoversættelsen.

3. Europa og flersprogethed

Derrick Kinck Olesen (EU-Kommissionen, Generaldirektoratet for Oversættelse)
nævnte i sit oplæg at maskinoversættelse allerede bruges i et vist omfang af oversættere i EU-regi med systemet MT@EC. Men kvaliteten af oversættelserne giver endnu ikke den store produktionsgevinst i dag (maks. en side pr. oversætter pr. dag i den danske afdeling). På årsbasis oversættes mellem 75 og 85.000 dokumenter fra dansk, hvilket svarer til ca. 4 % af alle Kommissionens oversættelser.
Uffe Sonne Svendsen (EU-Kommissionen, Generaldirektoratet for Oversættelse) fortalte om de erfaringer der er med maskinoversættelse, og om hvordan MT@EC hjælper med at holde en konsistens i alt der bliver oversat. Der er størst brugertilfredshed ved de analytiske sprog (dvs. sprog som ikke har mange bøjningsendelser). Systemet kan reparere ortografiske fejl, men maskinoversættelse kan ikke omkalfatre hele sætningskonstruktioner endnu.

4. Sprog og sprogteknologi

Sabine Kirchmeier (Dansk Sprognævn) udtrykte bekymring for at dansk halter bagud på det sprogteknologiske område, og at ekspertisen i dansk sprogteknologi forsvinder. Der bør ske noget på området, men man har ikke kunnet skabe politisk enighed om en samlet indsats hidtil. Der har været mange sporadiske projekter der så er afsluttet og ikke bliver fulgt op. Der mangler penge/bevillinger til forskning og udvikling. Der har også været nogle initiativer fra privat side, fx oprettelse af sprogteknologiske virksomheder. Men det er ikke nok til at sprogteknologien kan udvikle sig i Danmark, da markedet er for lille til at bære de nødvendige udviklingsinvesteringer i starten. Der er få midler, og der mangler sammenhæng mellem initiativerne. Holland har i perioden 2006-2009 haft en national strategi der målrettet har udviklet de nødvendige basisteknologier (Basic Language Resource Kit – BLARK). Det har betydet at Holland er væsentligt længere fremme end Danmark på dette område. Også i Norge og Sverige investeres der mere i sprogteknologi og i basisteknologi for norsk og svensk.

Sprogteknologi kræver en løbende indsats da sproget stadig ændrer sig, og nye ord kommer til. Derfor handler sprogteknologi også om at sørge for at data indsamles løbende. Det er der bl.a. mulighed for med PSI-direktivet. Offentlige institutioner har derfor en vigtig rolle, idet de kan skubbe på udviklingen, kræve bedre sprogteknologi og bidrage med gode data.

5. Flersprogethed i den offentlige sektor. Hvordan imødegås udfordringen?

Erfaringer fra **International House Copenhagen** viser at der er store forskelle i Københavns Kommune med hensyn til hvordan sproglige udfordringer håndteres. De store sprog er: urdu, tyrkisk, arabisk og somalisk. Københavns Kommune har valgt at køre en ensproget model hvor kommunens medarbejdere taler engelsk til udenlandske borgere, medmindre er tale om flygtninge der har krav på en tolk. **Trine Engelberg** fortalte at ordlister bliver opdateret decentralt, og at det eneste anvendte automatiserede oversættelsesværktøj er Google Translate. Der bliver ikke systematisk gjort brug af maskinoversættelse, og sproglig viden bliver ikke koordineret. Derfor ansætter kommunen typisk "frontoffice"-medarbejdere med meget brede sprogkundskaber.

Anne Mette Olsen (Region Sønderjylland-Schleswig) kunne fortælle at man i regionen bruger mange resurser på tosproget materiale, og at meget af tiden går med at oversætte tekster i grænsependlerrådgivningen. Nogle gange er der også behov for at "oversætte" oversættelsen da en oversættelse kan være af en sådan kvalitet at den ikke kan forstås umiddelbart, fordi referencerammen først skal sættes. Oversættelse er med andre ord ikke lig forståelse, så der er stort fokus på at skrive et kort og klart sprog, og på at det er vigtigt at være præcis. I regionen er der overvejelser om at investere i et oversættelsesværktøj for at forhindre at udgifterne til oversættelse løber løbsk, og for at få mere konsekvens i eget sprogbrug. Regionen ser gerne at offentlige myndigheder oversætter noget mere – især til tysk - så de automatiske systemer kan blive bedre.

Sigurd Slot Jacobsen (Konkurrence- og Forbrugerstyrelsen) er oprettet som den ene af de to danske brugere af MT@EC (den anden er Sabine Kirchmeier), men har ikke gjort sig så mange erfaringer med systemet endnu. Engelsk er det naturlige andet sprog i styrelsen, men der kan også være behov for oversættelse fra fx processprogene tysk og fransk til dansk i en dagligdag hvor EU-retten og den danske lovgivning går hånd i hånd, og hvor løsningen på en problemstilling ofte ligger i den enkelte sætning eller endda i det enkelte ord. Der er plads til forbedring idet styrelsen ikke har noget system til vedligeholdelse af flersproglige data. Det er også en udfordring med manglende oversættelse i forbindelse med vidensdeling til resten af EU.

6. Automatisk oversættelse: Hvordan fungerer det?

Anders Søgaard (Center for Sprogteknologi, Københavns Universitet) forklarede at der er mange måder at oversætte tekster på, men at den mest anvendte metode for tiden er statistisk oversættelse. Et statistisk oversættelsessystem lærer oversættelse i to trin: Først ved at udregne sandsynligheden for at et givet ord på et sprog skal oversættes med et givet ord på et andet sprog. Dertil skal der bruges store mængder af oversatte tekster. Dernæst ved at beregne hvordan de oversatte ord skal sammensættes til sætninger for at give et korrekt resultat på målsproget. Til dette skal der også bruges store mængder af originale tekster på målsproget. Andre typer af sproglige data, fx ordbøger, begrebssystemer, tesaurusser, ontologier mv., kan også indgå i systemerne og øge deres kvalitet. Sproget er levende og ændrer sig hele tiden. Der kommer fx hele tiden nye ord og nye udtryksmåder til. Derfor handler det grundlæggende om at sørge for at data indsamles løbende. Det er desuden vigtigt at mange forskellige emneområder og teksttyper bliver repræsenteret i systemerne fordi oversættelsens kvalitet bliver bedre jo mere systemerne bliver trænet på disse tekster. Offentlige institutioner har altså store muligheder for at påvirke oversættelsens kvalitet.

Sabine Kirchmeier (Dansk Sprognævn) understregede at det ikke kun er oversatte tekster der er interessante; det er også helt almindelige tekster på et givet sprog og gerne alle mulige typer af tekster. Derfor bør offentlige institutioner gøre en indsats for at finde både oversatte og ikkeoversatte tekster samt andet sprogligt materiale frem som kan stilles til rådighed for sprogteknologi. En helt særlig

kategori er de oversættelser som myndighederne sender ud til private oversættere. De oversættes typisk med de såkaldte oversættelseshukommelser. Offentlige institutioner bør være opmærksomme på dette og lave aftaler om at disse hukommelser kommer tilbage til institutionen sammen med oversættelsen, da disse hukommelser kan integreres direkte i et maskinoversættelsessystem. Så dem kan man i princippet sende direkte videre til CEF.AT/mt@ec hvis de ikke indeholder personfølsomme data eller lign.

7. Hvordan kan offentlige institutioner få gavn af CEF.AT?

På Skype fra Bruxelles deltog **Spyridon Pilos (Head of sector "Language Applications", Directorate General for Translation, European Commission)**, der kunne præsentere en meget enkel vision: Ville det ikke være fantastisk at kunne få oplysninger på sit modersmål, ligegyldigt hvor man befinder sig henne i verden og at kunne oversætte i realtid? Borgere skal have muligheden for at blive i deres eget sprog. Det er formålet, og iflg. Spyridon Pilos er den eneste løsning på de sproglige udfordringer i EU at fremme informationsoverførsel på kryds og tværs af landegrænser. Forstår man ikke målsproget, kan man bruge maskinoversættelsen. På den måde får man hurtigt et overblik over om en tekst er relevant eller ej. Hvis man vil arbejde videre med teksten, kan man evt. vælge at sende den til menneskelig oversættelse. En maskinoversat tekst kan aldrig bruges direkte i officiel sammenhæng. Den skal altid valideres af menneskelige oversættere.

Offentlige institutioner får adgang til systemet ved at oprette sig via http://ec.europa.eu/dgs/translation/translationresources/machine_translation/index_en.htm

8. Tilde – maskinoversættelse til e-handel og til den digitale offentlige sektor. Et eksempel.

Rihards Kalnins (Tilde): Alle websteder ønsker at få flere kunder til at købe mere, men størstedelen af EU's befolkning køber ikke noget på websteder som ikke er på deres eget sprog – derfor bruges maskinoversættelse i stor grad på fx Amazon, ebay, Airbnb og Etsy. Maskinoversættelse bruges også i forbindelse med brugeranmeldelser, for det er vigtigt for de fleste at man kan læse hvad andre brugere synes om produktet, på sit eget sprog. Når man anvender maskinoversættelse i disse sammenhænge, ser man som forbruger typisk bort fra de sproglige nuancer. Man ser på helheden og træffer sin beslutning på den baggrund.

Med e-government er der større fokus på at servicere borgeren. I Letland har Tilde udviklet Hugo.lv, som kan oversætte hjemmesider og tekster fra det offentlige. Maskinoversættelse er fuldt integreret i forskellige e-services og på offentlige hjemmesider. Brugergrænsefladen er let at bruge, og den er tiltænkt både statsligt ansatte og borgere. Den er også integreret på latvia.lv – den offentlige e-governmentportal som svarer til borger.dk. Det er muligt selv at tilføje input til oversættelsen, og brugerne kan derfor være med til at forbedre systemet.

9. De juridiske rammer for levering af data

Der er en række juridiske og ophavsretlige udfordringer forbundet med at indsende data til forbedring af maskinoversættelse, og **Cathrine Lippert (Digitaliseringsstyrelsen)** kunne her bidrage med basisinformationer om både PSI-lovgivningen (adgang til offentlige data), og om hvordan private og offentlige organisationer kan bruge disse data som råstof til at udvikle andre produkter og services. Den nationale lovgivning går i spænd med PSI (Public Sector Information) idet PSI er sekundær til den nationale lovgivning, og det betyder bl.a. at national lovgivning skal respekteres, så i Danmark må der fx ikke stilles persondata til rådighed (jf. persondataloven). Det betyder at man skal være opmærksom på om de tekster man gerne vil stille til rådighed, indeholder den slags data. Der findes metoder til at

identificere og kryptere persondata, og det ville være oplagt at udvikle et generelt system for den offentlige sektor så det bliver muligt også at inddrage tekster som man normalt ikke kan udsætte for maskinoversættelse. Man skal som myndighed også være opmærksom på at man når man frigiver data, ikke længere kan bestemme over dem, ligesom man fx ikke må skele til om data skal bruges kommercielt eller ikke-kommercielt. Alle, også private virksomheder, skal have adgang.

10. Data og sprogresurser i Danmark

Peter Juel Henrichsen (International Business Communication, Copenhagen Business School) fortalte om forskning i talegenkendelse og tilblivelsen af opensource-talegenkendelse for dansk. Han understregede at der er et stort genbrugspotentiale i data til talegenkendelsessystemer hvis kommuner og andre myndigheder begynder at arbejde sammen. Man skal bl.a. være opmærksom på om hvem der ejer data når der udvikles sprogteknologiske systemer, fx talegenkendelse. Private udbydere har en tendens til at holde data for sig selv, og det gør det dyrt for kommunerne da man bliver bundet til en bestemt leverandør og ikke kan udsætte ydelsen for konkurrence. Derfor er kommunerne bl.a. blevet enige om at gå i et fælles udbud hvor de beholder retten til data og derved kan opnå lavere priser og højere kvalitet.

Bolette Sandford Petersen (Nordisk Forskningsinstitut, Københavns Universitet) talte videre om de data der er brug for. Et problem er at der ikke findes så mange oversætte tekster tilgængelige for dansk, og der arbejdes derfor i forskningen med hvordan både sammenlignelige tekster på flere sprog og ensprogede tekster kan bruges som data til træning af maskinoversættelsessystemer.

Bodil Nistrup Madsen (International Business Communication, Copenhagen Business School) talte om begrebsafklaring i samarbejdet mellem offentlige myndigheder og om behovet for struktureret information. Myndighederne taler ikke altid samme sprog som borgerne, fx i offentlige selvbetjeningsløsninger, og det er derfor meget vigtigt at offentlige institutioner er opmærksomme på deres fagudtryk og bruger dem konsistent og med præcise definitioner.

Mange offentlige institutioner har ordlister liggende som ikke kun kan have stor værdi for maskinoversættelsessystemerne og anden sprogteknologi, men også for borgerne. Det er en god ide også at være opmærksom på dem, at vedligeholde dem og at stille dem til rådighed. Hun pegede derudover på at der er et meget stort behov for at få oversat undersider til engelsk, herunder også selvbetjeningsløsninger på nettet.

11. Data- og sprogresurser: Tekniske og praktiske aspekter

Andrejs Vasiljevs (Tilde/ELRC) talte om hvilke data der er behov for. Pointen er at jo flere data offentlige myndigheder stiller til rådighed, jo bedre kvalitet er outputtet i maskinoversættelsen. Der er brug for at offentlige myndigheder i hele Europa byder ind med data – ordlister, rapporter, taler, dokumenter, brochurer, hjemmesider osv. Kun 4 % af de tekster der produceres i offentligt regi, findes ude på internettet tilgængeligt for alle. Det er de resterende 96 % - the deep web - offentlige myndigheder opfordres til at give CEF.AT adgang til. Der er brug for at den offentlige sektor udpeger synlige og usynlige data og giver CEF.AT adgang til dem.

12. Bedste fremtidige praksis – offentlige institutioners data skal bruges til at forbedre systemet

Andrejs Vasiljevs (Tilde/ELRC) talte afslutningsvis om de tekniske og praktiske aspekter ved indsamling af data- og sprogresurser. Det handler dels om at identificere de åbne kilder (altså de 4 % af

isbjerget) og om at beslutte hvordan man indsamler disse data, og dels om at få myndighederne til også at levere de resterende 96 % data fra det dybe web. Det er op til myndighederne at analysere hvilke data der kan blive tale om, men EU kan bistå offentlige myndigheder med:

- at vurdere brugbarheden af data og uddanne medarbejdere til brug af maskinoversættelse
- at vurdere dokumentation af data
- at bearbejde materialet med forskellige værktøjer, fx kan den offentlige myndighed levere rådata, og ELRC kan bearbejde disse (fx frasortering/anonymisering af persondata, frasortering af sætninger på et andet sprog i samme dokument, parallelisering af oversatte dokumenter mv.)
- at få afklaret de juridiske aspekter.

Til maskinoversættelse kan offentlige myndigheder fremsende data i alle former for digitaliserede formater. Indskannede og trykte dokumenter kan ikke bruges. Man er mest interesseret i de kildedokumenter som pdf-filer typisk er baseret på, fx Word. Det er dog også muligt, men mere besværligt at ekstrahere data af pdf-filen og tilpasse dem til maskinoversættelsessystemet.

Mange offentlige institutioner vælger at outsource oversættelsesopgaven til private. ELRC foreslår konkret at alle offentlige institutioner gennemgår deres eksisterende aftaler med oversættere og sikrer at oversætterne udover selve det oversatte dokumentet også leverer oversættelseshukommelsen tilbage så den kan genbruges i senere oversættelser. Det vil gøre det lettere at sende oversættelsesopgaver i udbud, og dermed sparer myndigheden penge, og oversættelseshukommelsen er desuden nyttig i forbindelse med. maskinoversættelse.

På www.lr-coordinatation.eu kan man finde arrangementer, workshops, teknisk og praktisk støtte. Her er både webformular, telefonnumre og e-mailadresser hvor offentlige myndigheder kan komme i kontakt med en juridisk og/eller en teknisk ekspert ift. data. På samme hjemmeside er der også en guide til hvordan man uploader sine data: "How to submit your data". Offentlige myndigheder kan:

- sende en URL, og systemet udfører datacrawling
- indsende dokumenter
- uploade data direkte
- indsende data via et fysisk medie – en form for dropbox – hvis man ikke ønsker at få data uploadet, eller hvis der er for mange eller for specifikke data.

På European Data Portal behandles rådata. Rådata bliver ikke delt – det gør kun de behandlede data.

13. Et sprogteknologisk netværk i Danmark

Bolette Sandford Petersen (Nordisk Forskningsinstitut, Københavns Universitet, formand for Fagråd for fagsprog og sprogteknologi, Dansk Sprognævn) fortalte at fagrådet gerne vil samle sprogteknologiske interessenter i et sprogteknologisk netværk. På mødet blev mødedeltagerne derfor opfordret til at deltage/henvende sig til Sabine Kirchmeier eller Bolette om at blive medlem af netværket, der er under opbygning. Det er hensigten at mødes 1-2 gange om året.

14. Spørgsmål

SPØRGSMÅL: Hvorfor giver EU ikke give private virksomheder adgang til MT@AC og på længere sigt CEF.AT?

Andrejs Vasiljevs: Det handler først og fremmest om at hjælpe offentlige institutioner til at agere på flere sprog. Hvis man frigiver MT@AC til private, risikerer man at blande sig i kommercielle interesser, idet private virksomheder har specialiseret sig i at udvikle og derfor også sælge maskinoversættelsessystemer. Umiddelbart må det handle om ikke at forvride konkurrencen, men man overvejer fortsat hvilken rolle EU-kommissionen skal spille fremover ift. private og offentlige aktører.
Derrick Kinck Olesen kunne udbygge svaret med at 25 % af alt det oversættelsesarbejde der bliver lavet i den danske sprogafdeling i EU-regi, bliver lavet af freelancere. Det svarer til ca. 16-18.000 sager om året. På denne måde deler EU allerede oplysninger med freelanceoversætterne, og begge parter har glæde af det.

SPØRGSMÅL: Bruger man aldrig skønlitteratur til maskinoversættelse, og kunne man bruge talegenkendelse som data?

Sabine Kirchmeier: Skønlitteratur er noget af det sværeste at få fat i. Forlagene og forfatterne vil meget nødig give rettighederne fra sig, hvilket er forståeligt da det er deres indtjeningsgrundlag. Men sprogteknologi er jo ikke i konkurrence med bogmarkedet. Man er jo ikke interesseret i værket som helhed, men i de enkelte ord og sætninger. Lige nu er der en dialog mellem Dansk Sprognævn og Kulturministeriet om sprogteknologi og ophavsret, ligesom der er en aftale med Det danske Akademi om at stille litterære tekster til rådighed. Der er dog også en anden udfordring ved brug af skønlitteratur, idet der er tale om helt andre tekstgenrer. Forfattere benytter fx dialog og et mere emotionelt sprog som ikke findes i tekster fra det offentlige, og ofte meget fantasifulde billeder og fremstillinger. Det er kort sagt en anden slags sprog. Hvis man anvender skønlitteratur som data til maskinoversættelse i eksempelvis EU-regi, risikerer man at få nogle ret så alternative oversættelser fordi domænet bliver for bredt.

SPØRGSMÅL: Har universiteterne adgang til MT@AC og på sigt CEF.AT?

Derrick Kinck Olesen: EMT-universiteter (universiteter som tilbyder European Master of Translation) kan bruge systemet. I Danmark er der én oversætteruddannelse (på Aarhus Universitet) som har adgang. Systemet vil med tiden blive udvidet til at andre universiteter også kan få adgang.

SPØRGSMÅL: Er der planer om at andre sprog end EU-sprog kommer med i CEF.AT, fx urdu, arabisk osv., som man gerne vil servicere borgere med?

Derrick Kinck Olesen: I vores system er der udelukkende fokus på EU-sprog, men vi tilbyder også oversættelse fra nogle af de store verdenssprog. Der er dog ikke nogen intern oversætterkapacitet. Vi bruger i stedet freelancemarkedet og kvalitetssikrer bagefter. Det drejer sig om sprogene kinesisk, russisk, urdu. Data (fx oversættelseshukommelserne) kommer med i vores database, men vi er på begyndelsesstadiet, da efterspørgslen ikke er stor nu.

SPØRGSMÅL: Kan vi i Nordisk Ministerråd bruge systemet? Vi bruger især penge på oversættelser fra finsk og islandsk

Derrick Kinck Olesen: Det bør I kunne. Problemet er blot at resultatet ikke er specielt godt for et par af de nordiske sprog på nuværende tidspunkt, fx finsk pga. sprogets karakter. Norsk og svensk giver et bedre resultat. Til dette supplerede **Sabine Kirchmeier** med forslaget om at Nordisk Ministerråd kunne påvirke kvaliteten af CEF.AT for de nordiske sprog ved at lægge de oversættelser der allerede er lavet, ind i systemet.

SPØRGSMÅL: Kan private downloade oversættelseshukommelsesbaser i Den

europæiske Dataportal?

SVAR: Portalen er helt åben og kræver ikke at man registrerer sig. Der er ingen restriktioner forbundet med Den europæiske Dataportal.

SPØRGSMÅL: Kan man ramme den rigtige sprogtone med maskinoversættelse?

Andrejs Vasiljevs: Maskinoversættelsessystemer lærer af data. Og hvis der er tilstrækkeligt mange data, så kan man også lære systemerne stil og sprogtone. Men der skal bruges store mængder data for at nå dertil. Maskinoversættelse skelner i dag ikke mellem formelt og uformelt sprog.

15. Konklusion

Sabine Kirchmeier (Dansk Sprognævn) konkluderede at det havde været en dag med mange engagerede oplægsholdere der kom godt omkring de mange facetter ved automatisk oversættelse. Der ligger mange spændende muligheder i automatisk oversættelse, og offentlige myndigheder kan spare både tid og penge ved at være mere opmærksomme på hvordan de håndterer oversættelsesopgaver.

Offentlige institutioner har stor indflydelse på hvordan fremtidens maskinoversættelse udvikler sig. Systemerne kan forbedres væsentligt til at håndtere dansk hvis de offentlige institutioner beslutter sig for at spille med og bidrage med deres data. Teknologien er endnu under udvikling, og der følger en række problemstillinger med, fx vedr. beskyttelse af persondata, som den enkelte myndighed dog kan få hjælp til at håndtere. Perspektiverne er store, og sandsynligheden for at offentlige institutioner får bedre værktøjer til at håndtere den mangesproglige virkelighed de oplever, er langt større hvis de spiller aktivt med og leverer data, end hvis de sidder med hænderne i skødet.