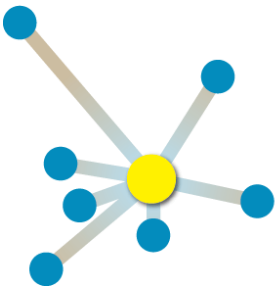
European Language Resource Coordination (ELRC) is a service contract operating under the EU's Connecting Europe Facility SMART 2014/1074 programme.



European Language Resource Coordination
Connecting Europe Facility

Deliverable Task 6

ELRC workshop report for Denmark

Author(s): Sabine Kirchmeier (Dansk Sprognævn)

Dissemination Level: Public

Version No.: <V1.1>

Date: 2016/06/10

Contents

1 Executive summary

The Danish ELRC Workshop took place in Copenhagen, on the 7th of March 2016 at Det europæiske Miljøagentur ("European Environment Agency").

Information about the workshop was disseminated through the information channels of the Danish Language Council. Invitations to the workshop were sent out to key persons in the area of public information, communication and translation at public agencies and other organizations.

The event was attended by nearly 70 participants representing a wide cross-section of public institutions, research organisations and businesses in Denmark.

The program featured presentations from the ELRC consortium, governmental, translation and language technology sectors. Both the audience and the speakers had a positive attitude towards the CEF platform and showed interest in how Danish public administration can provide language resources to the European Commission in order to develop CEF.AT.

All presentations are provided on the ELRC Denmark Workshop website in pdf format.

2 Dissemination of Workshop

Localized invitations were sent out to potential workshop participants, where the objectives of the event were explained and the importance of ELRC activities highlighted.

Invitation email (in Danish) is attached in the Appendix.

Information about the workshop was published on the Danish Language Council web site, sproget.dk and http://translatorforeningen.dk/.

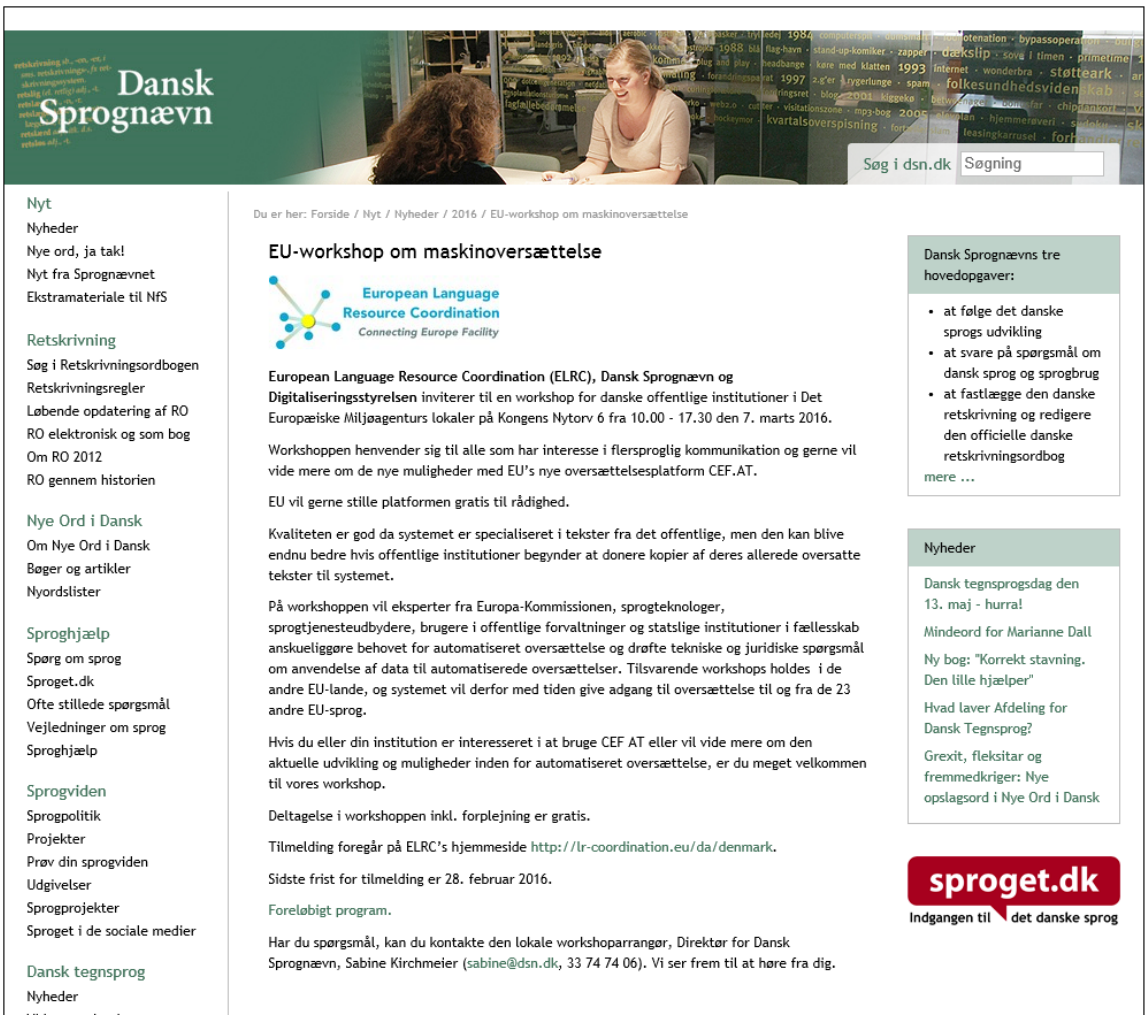


Figure 1 Snapshot of Danish Language Council Web page (http://dsn.dk/nyt/nyheder/2016/eu-konference-om-maskinoversaettelse)

European Language Resource Coordination (ELRC) is a service contract operating under the EU's Connecting Europe Facility SMART 2014/1074 programme.

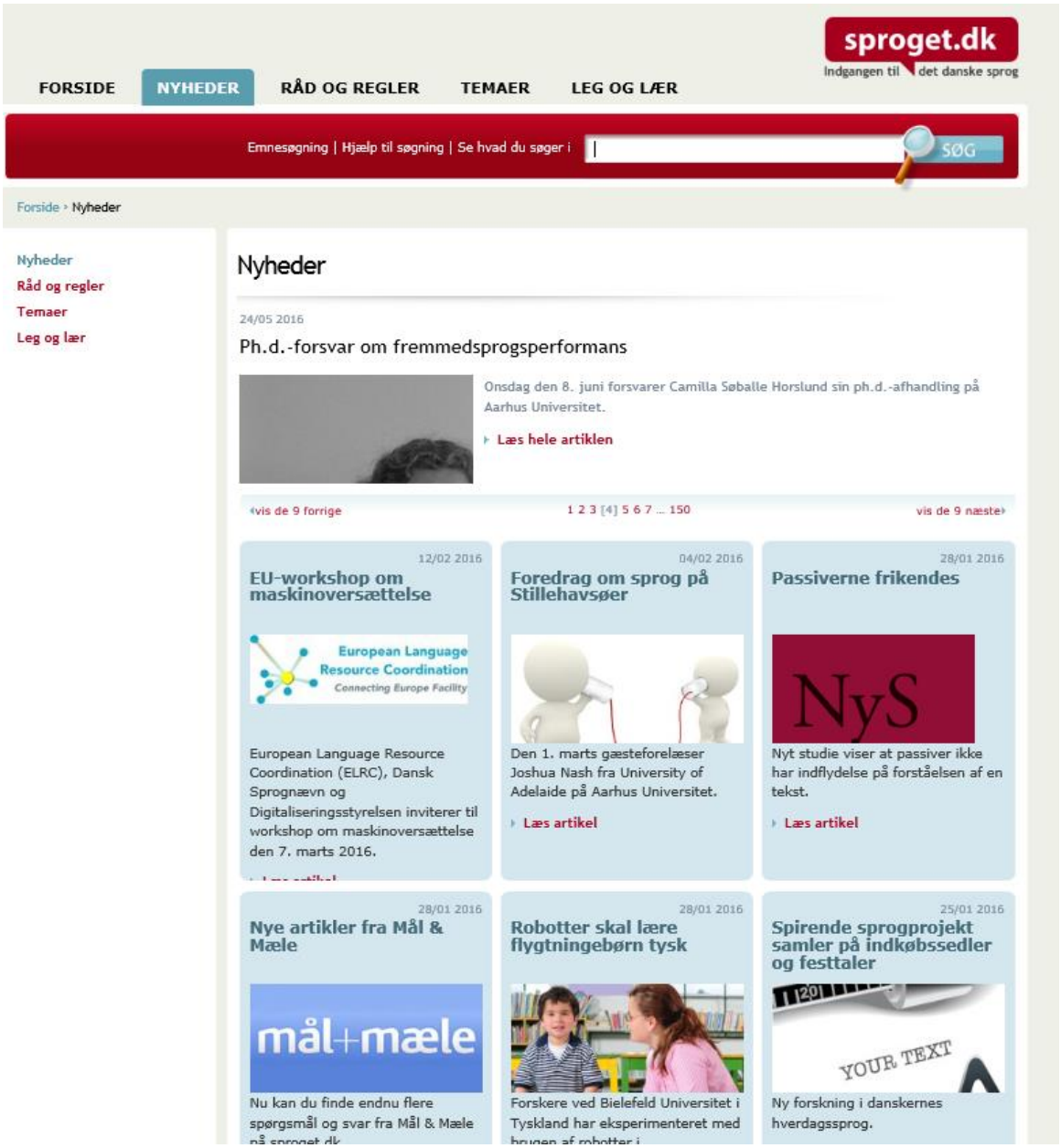


Figure 2 Snapshot of sproget.dk (http://sproget.dk/nyheder/eu-workshop-om-maskinoversaettelse)

European Language Resource Coordination (ELRC) is a service contract operating under the EU's Connecting Europe Facility SMART 2014/1074 programme.

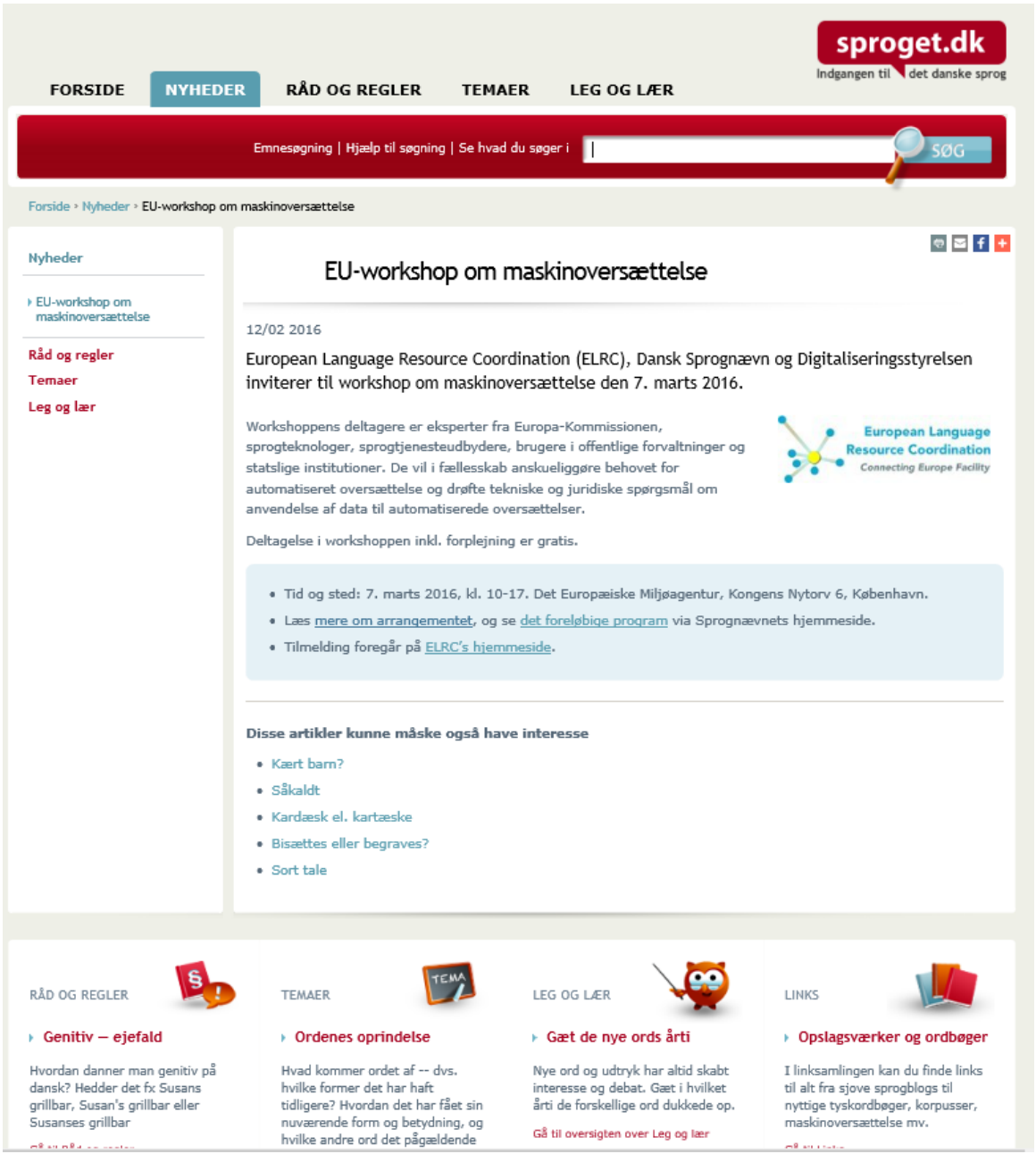


Figure 3 Snapshot of sproget.dk (http://sproget.dk/nyheder/eu-workshop-om-maskinoversaettelse)

#198 Nyt på hjemmesiden, Ja til Sprog arrangement, Kom glad i Den Gamle eller By eller i Højesteret, Deltagelse ved ELRC

Kære medlem

Bestyrelsen har forfattet en flot **forretningsorden - bestyrelseshåndbog**. Du kan læse den på hjemmesiden som en underside til siden "Om foreningen". Eller her.

*

Translatørforeningens deltagelse ved ELRC

Workshop
Dansk Sprognævn havde til mandag d. 6. marts 2016 indbudt til en daglang workshop om ELRC (European Language Resource Coordination. Tanken er at udvide brugen og anvendeligheden af maskinoversættelse inden for EU. Offentlige myndigheder i EU-landene kan i dag få adgang til at bruge EU's maskinoversættelsessystem (MT@EU), men der er ikke adgang for det private erhvervsliv. Dvs. at der heller ikke er adgang for den enkelte oversætter/translatør.....læs mere her.

*

Der er stadig ledige pladser ved arrangementerne:

Historisk tilbageblik til 1970erne - i Den Gamle By i Aarhus den 5/4.
Besøg i Højesteret - København den 18/4.
DHL 2016 i København og Aarhus - vil du med? Så følg linket og læs mere!

*

"**Den nationale sprogstrategi** - *hvordan kommer vi i mål?*"

Skal du med den 20/4? Hvis ja, vil du så ikke være så rar, at give mig besked? For bestyrelsen mener det er vigtigt, at Translatørforeningen er repræsenteret denne dag. De kan dog ikke selv, da der samme dag er bestyrelsesmøde i Odense.

Hvis du er i tvivl om hvad det er, så se i nyhedsbrev 196.

Venligst
Linda Kyllesbech
Sekretær

Translatørforeningen
Hauser Plads 20, 3. sal
DK-1127 København K
Tlf: +45 33 11 84 14
E-mail: mail@translatorforeningen.dk

Afmeld nyhedsbrev

Figure 4 Snapshot of http://translatorforeningen.dk/?id=311

3 Workshop agenda

09:30 – 10:00	**Registrering**
10:00 – 10:20	**Velkomst, målsætninger/Welcome** **Michael Vedsø,** Europa-Kommissionens repræsentation i Danmark/representative of the EU-Commission in Denmark **Lars Frelle-Petersen**, Digitaliseringsstyrelsen/Danish Agency for Digitization
10:20 – 10:30	**Målsætninger/Aims and goals** **Andrejs Vasiļjevs**, ELRC/Tilde **(præsentation)**
10:30 – 10:40	**Europa og flersprogethed/Europe and multilingualism** **Derrick Kinck Olesen** & **Uffe Sonne Svendsen,** - European Commission, Directorate-General for Translation, Danish Language Department/EU-Kommissionen, Generaldirektoratet for Oversættelse **(præsentation)**
10:40 – 11:20	**Sprog og sprogteknologi i Danmark/ Languages and Language Technologies in Denmark** **Sabine Kirchmeier,** Dansk Sprognævn/Danish Language Council (præsentation)
11:20 – 11:45	**Diskussion: Flersprogethed i den offentlige sektor – hvordan imødegås udfordringen?/ Panel: Multilingual Public Services in Denmark – how is the challenge addressed** **Deltagere:** **Sigurd Slot Jacobsen,** Specialkonsulent, Konkurrence- og Forbrugerstyrelsen/ The Danish Competition and Consumer Authority **(præsentation)** **Anne-Mette Olsen,** Konsulent, Region Sønderjylland – Schleswig **(præsentation)**

	Thomas Jacobsen, Direktør, Kultur- og Fritidsforvaltningen, Københavns Kommune/Culture and Leisure, Copenhagen Municipality
11:45 – 12:00	*Kaffepause*
12:00 – 12:20	**Automatisk oversættelse: Hvordan fungerer det?/ Automated Translation: How does it work?** **Anders Søgaard**, Professor Center for Sprogteknologi, Københavns Universitet/ Center for Language Technology, University of Copenhagen **(præsentation)**
12:20 – 12:40	**Hvordan får de offentlige institutioner gavn af CEF.AT-platformen? How can Public Institutions benefit from the CEF.AT Platform** **Spyridon Pilos**, Head of sector "Language Applications", Directorate General for Translation, European Commission **(præsentation)**
12:40 – 13:00	**Tilde – Maskinoversættelse til den digitale offentlige sektor - et eksempel/ Tilde – MT for e-Government – a case study** **Rihards Kalniņš**, ELRC/Tilde **(præsentation)**
13:00 – 14:00	*Frokost*
14:00 – 14:30	**Hvilke data er der behov for? Hvorfor? /What Data is needed? Why?** **Andrejs Vasiļjevs**, ELRC/Tilde **(præsentation)**
14:30 – 15:00	**De juridiske rammer for levering af data, European Data Portal/ / Legal framework for Contributing Data, European Data Portal.** **Cathrine Lippert,** Specialkonsulent, Digitaliseringsstyrelsen/ Danish Agency for Digitization **(præsentation)**

15:00 – 15:30	**Diskussion: Data og sprogressourcer i Danmark/ Panel: Data and Language Resources in Denmark** **Peter Juel Henrichsen**, Lektor, Copenhagen Business School **(præsentation)** **Bolette Sandford Petersen,** Professor, Copenhagen University/Københavns Universitet (præsentation) **Asbjørn Fangel**, Danish Agency for Digitization **Bodil Nistrup Madsen,** Professor, CBS –Department of International Business Communication/DANTERMcentret (**præsentation**)
15:30 – 16:00	*Kaffepause*
16:00 – 16:30	**Data- og sprogressourcer: Tekniske og praktiske aspekter/ Data and Language Resources: Technical and Practical Aspects** **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?** **Moderatorer:** **Andrejs Vasiļjevs**, ELRC/Tilde **Sabine Kirchmeier**, Dansk Sprognævn/Danish Language Council
17:00 – 17:10	**Et sprogteknologisk netværk i Danmark** **Bolette Sandford Petersen,** Professor, Copenhagen University/Københavns Universitet, Formand for Fagråd for fagsprog og sprogteknologi, Dansk Sprognævn
17:10 – 17:25	**Opsamling, konklusioner og tilsagn/ Conclusions and commitments**

4 Summary of content of sessions

Summary of content of sessions is prepared bilingual – English and Danish.

4.1 Summary in Enlish

4.1.1 Welcome

Michael Vedsø, Head of the European Comission Representation in Denmark pointed out that the aim of the workshop is to raise awareness of the EU machine translation system platform CEF.AT developed by the European Commission's Directorate-General for Translation. It is a further development of MT@EC, which is the European Commission's existing machine translation (MT) system. MT@EC is based on the translations made at EU level over the last 20 years. MT@EC is already used in pilot projects in public institutions in all EU Member States and has no participation fees. MT@EC is based on EU data in 24 languages. There were 940 million sentences in the system at the end of 2015 and the number is growing by 2 million sentences per month. CEF.AT is a part of the Connecting Europe Facility. AT stands for Automated Translation Platform.

Translation of documents is the European Commission's most substantial expenditure. Annually, the EU spends 1.1 billion Euros on translation and interpretation in the various EU institutions. This should be seen in the light of, on the one hand, that EU citizens have a right to address the EU in their own language and correspondingly may request replies in their own language, and on the other hand that translation is part of the legislative procedure in the EU. Experience shows that translation and proofreading contributes to a better legislation.

There are technical and legal challenges in sharing the data that public authorities already have. Addressing these challenges is the second aim of the workshop. If the expansion of MT@EC takes place as planned, the European Commission has every confidence that it can make a major contribution to the realization of the digital single market.

Picture 1 Oppening by Michael Vedsø

Sabine Kirchmeier initially noted a particular Danish challenge which is that for example annexes to acts and laws are not always available in Danish, and that background material for decisions in European Parliament resolutions are not always translated. A practice where documents which have an impact on legislation are not available in Danish, is considered by the Danish Council not only to be a linguistic, but also a democratic problem.

Picture 2 Oppening by Sabine Kirchmeier

4.1.2 Aims and goals

Andrejs Vasiļjevs pointed out that the main focus of the workshop was:

- to raise awareness of the value and importance of the data which public institutions otherwise just regard as documents. They can be very useful in the work to develop automated translation, so that public institutions have access to systems of better quality that can translate between one's own language and the other European languages.
- to invite Danish public institutions to share data and thereby make an active contribution to improving the development of the EU machine translation service to support the building of Europe — and in fact also support the Danish language.
- to help participants understand and resolve practical and legal problems, through the sharing of data with the European Commission's Directorate-General for Translation with the aim to improve MT results.

Picture 3 Aims and Goal of ELRC, Andrejs Vasiļjevs

4.1.3 Europe and multilingualism

Derrick Kinck Olesen stressed in his presentation that machine translation is already being used, to a certain extent, by translators in the EU institutions, namely the system MT@EC. But the quality of the translations does not yet ensure high productivity gains today (max one page per translator per day in the Danish division). On an annual basis 75 and 85,000 documents are translated from Danish, representing about 4% of all the Commission's translations.

Uffe Sonne Svendsen reported on the experience with machine translation. How Euramis helps to maintain the consistency of all translated documents (Euramis is the data motor of MT@EC). The experience is that the best results are seen for translation between analytic languages (i.e. languages that do not have a strong morphology). Translation between analytic languages and synthetic languages (languages with strong morphology) is still fraught with difficulty since machine translation cannot yet radically change entire sentence structures.

Picture 4 Europe and multilingualism

Picture 5 Europe and multilingualism, MT@EC

4.1.4 Languages and language technology in Denmark

Sabine Kirchmeier, Danish Language Council, expressed concern that Danish language technology is lagging behind and that the expertise on Danish language

technology is disappearing. Although political parties from time to time propose measures to improve the situation, there has been no political agreement on a comprehensive effort to date. There have been many scattered projects, that have been completed and are not followed up. There is a lack of money/grants for research and development. There have been some initiatives to establish private language technology companies, but it is difficult for language technology to develop in Denmark on a private basis, since the market is too small to bear the necessary development investment at the outset. There are few resources spent in the public sector, and there is a lack of coherence between initiatives.

Sabine Kirchmeier pointed out that in contrast, the Netherlands 2006-2009 had a national strategy which was targeted at developing the necessary basic technologies (Basic Language Resource Kit — BLARK). This initiative has made the Netherlands significantly more advanced than Denmark in this area. Also in Norway and Sweden more public investments are made in language technology and in basic technologies for Norwegian and Swedish.

Language technology requires a continuing effort since language is constantly changing and new words are created. For this reason we have to make sure that data for language technology are collected on a regular basis. This is possible in particular with the PSI Directive. Public institutions have an important role pushing the development, demanding better language technology and contributing with useful data.

4.1.5 Multilingualism in the public sector.

Lessons learned from **International House Copenhagen** show that there are big differences in the approaches to language in the municipality of Copenhagen. The major languages are: Urdu, Turkish, Arabic and Somali. The City of Copenhagen has chosen to run a mono-model where the municipality's staff speak English to foreign citizens (except for refugees because they are entitled to an interpreter). **Trine Engelberg** explained that vocabularies are kept on a decentralised basis, and that the only automated translation tool used is Google Translate. There is no routine use of machine translation and linguistic knowledge is not coordinated. The municipality therefore typically places employees with a broad knowledge of languages in their front office.

Anne Mette Olsen from Region Sønderjylland Schleswig explained that significant resources are devoted in the region on providing bilingual material, and that much time is used on translating texts in border commuter advice. Sometimes, there is also a need to edit a translation since a translation may be of such a quality that it cannot be understood because of the frame of reference. Translation in other words is not equal to understanding. There is a strong focus on writing a short and clear and precise language that is easier to translate. The region is considering investing in a translation tool in order to prevent increasing translation costs and to ensure more consistency. The region would like to see public authorities translate more — in particular into German as this may improve the automatic systems.

Picture 6 Anne Mette Olsen from Region Sønderjylland Schleswig

Sigurd Slot Jacobsen from the Danish Competition and Consumer Authority was presented as one of the two Danish users of MT@EC (the other being Sabine Kirchmeier), but did not have much experience with the system yet. English is the obvious second language in the organisation, but there may also be a need for translation from e.g. the procedural languages German and French in a working environment where EU law and the Danish legislation go hand in hand, and where the solution to a problem often lies in

the formulation of a sentence or even in the choice of individual words. There is room for improvement, since the institution has no system for maintenance of multilingual data. It is also a problem that the lack of translation impedes knowledge sharing with the rest of the EU.

Picture 7 Sigurd Slot Jacobsen from the Danish Competition and Consumer Authority

4.1.6 Automated translation: How does it work?

Anders Søgaard, Professor, Centre for Language Technology, explained that there are many ways to translate texts automatically, but that the most widely used method currently is statistical machine translation. A statistical machine translation system performs translation in two steps: First by establishing the likelihood of a given word or word sequence in a language is to be translated by a given word of another language. This entails the use of large amounts of translated texts. Next, by calculating how the translated words or sequences should be combined into sentences to give a correct result in the target language. For this, there is a need for large amounts of original texts in the target language. Other types of linguistic data such as dictionaries, conceptual systems, thesauri, ontologies, etc. can also be incorporated into the systems and increase their quality.

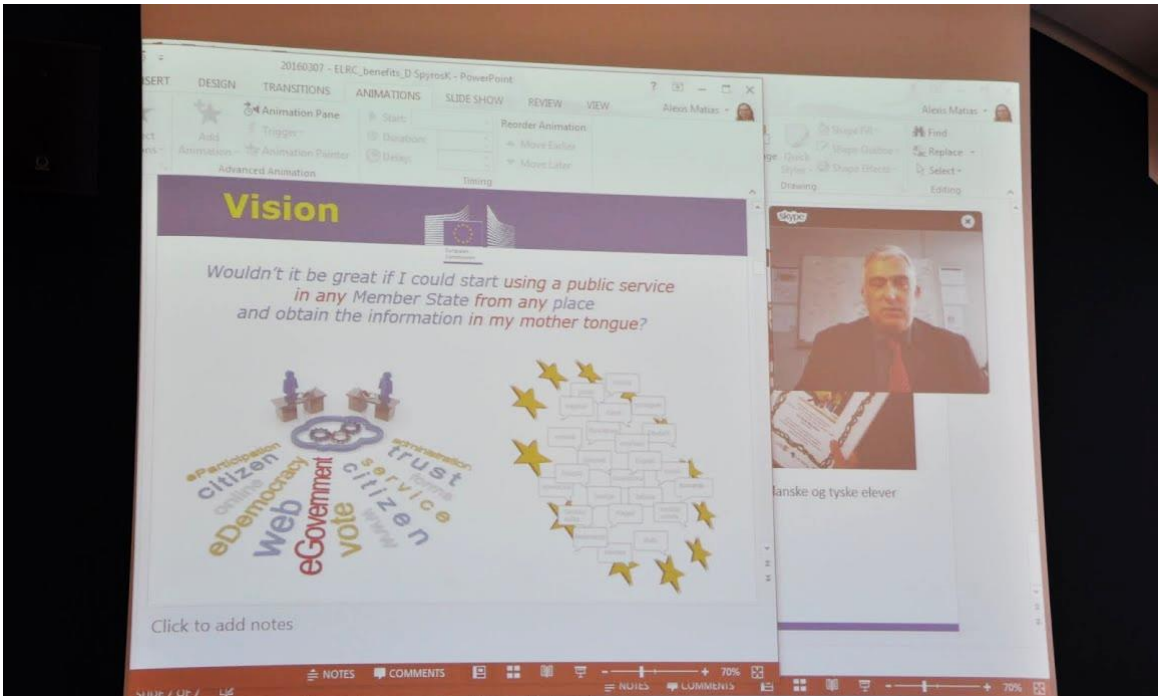
Languages are alive and constantly changing. There will, for example, always be occurring new words and new ways of expression. Therefore, it is fundamental to ensure that data is collected continuously. It is also important to collect texts from many different subject areas, and that different text types are represented in the systems because the quality of the translation will improve the more the systems are trained on these texts. Public institutions have enormous potential to influence the quality of the translation.

Picture 8 Anders Søgaard, Professor, Centre for Language Technology

Sabine Kirchmeier stressed that not only the translated texts are of interest — the system can use all possible kinds of texts, also monolingual ones. Public institutions should therefore make an effort to find both translated and non-translated texts and other linguistic material which may be used for language technology. A very important resource are the translations which institutions send to private translators. They are typically translated with the so-called translation memory systems. Public institutions should be aware of this, and make agreements with the translation bureaus in order to get these translation memories back to the institution together with the translated texts, since these memories can be directly integrated in a machine translation system. In principle, it is possible to send the translation memories directly to CEF.AT/MT@EC if they do not contain personal data etc.

4.1.7 How can public institution benefit from CEF.AT?

On Skype from Brussels **Spyridon Pilos**, **Head of Sector "Language Applications" Directorate General for Translation, European Commission,** presented a very clear vision: Wouldn't it be great to be able to obtain information in your mother tongue, no matter where you are in the world — to translate in real time? Citizens should have the possibility to use their own language. Citizens must have the opportunity to be addressed in their own language. According to Spyridon Pilos this is the only solution to the linguistic challenges of the EU: to promote information transfer across borders. If you do not understand the target language, you can use machine translation. This way you can quickly get an overview of whether a text is relevant or not. If you want to work with the text, you may choose to send it to a human translator. A machine-translated text can never stand alone if the text is to be used in an official context. It must always be validated by human translators.



Picture 9 On Skype from Brussels Spyridon Pilos, Head of Sector "Language Applications" Directorate General for Translation, European Commission

4.1.8 Tilde — MT for e-government

Mr Rihards Kalniņš, Tilde, Latvia, reported on translation use for e-commerce and e-Government on the example of e-Gov machine translation service in Latvia. All websites would like to have more customers buying more, but the majority of the EU population does not purchase anything on websites which are not in their own language. Machine translation is therefore used in large e-commerce companies such as Amazon, eBay, Etsy and Airbnb, to mention just a few. Automatic translation is also used for consumer ratings, because it is important for most consumers to read product reviews in their own language. When you look at machine translation in these contexts, you also see that consumers ignore grammatical errors. They look at the whole and this is sufficient to make a decision.

In e-government the focus is on servicing the citizen. In Latvia, Tilde developed Hugo.lv, which can translate websites and various texts from the public sector. Machine translation is fully integrated into various e-services and public websites. The user interface is easy to use, and it is intended both for government employees and citizens. It is also integrated on latvia.lv – the government e-portal. It is possible to add input to the translation, and users can therefore help to improve the machine translation system.

Picture 10 Rihards Kalniņš talks about the role of MT for the Digital Single Market

4.1.9 The legal framework for the provision of data

There are a number of legal and copyright challenges associated with submitting data for the improvement of machine translation, and **Cathrine Lippert from the Danish Agency for Digitisation (Digitaliseringsstyrelsen)** could contribute with basic information about both the PSI-legislation (access to Public Sector Information) and how public organizations can use these data as raw material to develop other products and services.

The national legislation interacts with PSI in the way that PSI is secondary to national legislation. This means that national legislation has to be respected, as in Denmark, for example, personal data must not be made public (see the Personal Data Act). This means that attention must be paid to the texts public institutions make available for machine translation. There are methods to identify and encrypt personal data and it would be smart to develop a general procedure for the public sector to handle such cases. Institutions must be aware that once the data are set free, they can no longer control them. Everyone, including private companies now have access.

Picture 11 Cathrine Lippert from the Danish Agency for Digitisation (Digitaliseringsstyrelsen)

4.1.10 Data and language resources in Denmark

Peter Juel Henrichsen, Copenhagen Business School, described his research and the emergence of open source speech recognition for the Danish language. He said that there is a great recycling potential in data for voice recognition systems, and that municipalities and other public authorities now are working together to explore this. There is a discussion about who owns the data when developing language technology systems, such as speech recognition. Private providers tend to keep data for themselves, and this makes it expensive for the municipalities because you are locked in to a specific vendor and may not be able to take advantage of using competing vendors. Therefore, the municipalities have agreed to go in a joint procurement where they retain the availability of data and thus can benefit from lower prices and higher quality.

Picture 12 Peter Juel Henrichsen, Copenhagen Business School

Bolette Sandford Petersen, Copenhagen University, spoke further on the data submitted. One problem is that there are only few translated texts available for Danish, and therefore research is being pursued in order to extract bilingual information from comparable texts in several languages and to explore how monolingual documents can be used as data.

Picture 13 Bolette Sandford Petersen, Copenhagen University

Bodil Nistrup Madsen spoke about how public institutions cooperate on the clarification of concepts and on the need for structured information. Public authorities do not always speak the same language as the citizens, e.g. in public self-service solutions, and it is therefore essential that public institutions are aware of their terminology and use them consistently with precise definitions.

Many public institutions maintain vocabularies which not only may be of great value to MT systems and other language technology, but also for citizens. It is a good idea to be aware of their importance and to maintain them and to make them available. In addition, she pointed out that there is a huge demand for translating the subpages of public institutions into English, in particular self-service solutions.

Picture 14 Bodil Nistrup Madsen, Professor, CBS –Department of International Business Communication/DANTERMcentret

4.1.11 Data — and language resources: Technical and practical aspects

Andrejs Vasiljevs, Tilde/ELRC, spoke about the data needed**.** The point is that the more information public authorities make available, the better quality output of MT results they get. Public authorities across Europe need to provide data — vocabularies, reports, speeches, documents, brochures, websites etc. Only 4% of the texts produced by public institutions can be found on the internet in the public domain. It is the remaining 96% ("the deep web"), that public authorities are invited to provide for CEF.AT. There is a need for the public sector to point out visible and invisible data and to grant CEF-AT access to them.

4.1.12 Interactive session: How can we engage?

Andrejs Vasiljevs concluded by describing the technical and practical aspects of gathering data and language resources. One should start by identifying the open sources (i.e. the 4% of the iceberg) and to decide how to collect these data. Next authorities should be encouraged to provide the remaining 96% of data from their internal data repositories. It is up to the authorities to analyse the data which may be submitted, but the EU can assist public authorities in:

- assessing the usefulness of data for improvement of machine translation
- assessing the documentation of data
- applying different tools — public authorites provide raw data and ELRC extracts data from the raw data (e.g. deletion/anonymization of personal data, separation of phrases in another languages within the same document etc.)
- handling the legal aspects.

Public authorities may submit data as digitised documents. Scanned and printed documents cannot be used. PDF can be used, but the format the PDF file is based on, e.g. the original Word-file, is preferred. If the original files are not available, it may be possible to extract data from the PDF.

Many public institutions choose to outsource the translation task to private translation agencies. ELRC proposes that public institutions make agreements with the agencies in order to get these translation memories back to the institution together with the translated texts, since these memories can be directly integrated in a machine translation system. This will also make it easier to send translation services out to tender, and thus save the authority money.

On http://lr-coordination.eu you can find the events, workshops, technical and practical support.

Here you find web forms, telephone numbers and email addresses to get in touch with a legal and/or a technical expert in relation to data. You can also get individual replies. The same website also provides a practical guide on how to upload data: "How to submit your data".

Public authorities may:

- send a URL and the system will pick up the data automatically
- submit documents
- upload data directly
- submit data via a tangible medium if they do not wish to upload or if there are too many or too specific data.

Raw data will not be shared —only language resources created of them. The sharable data will be made available on the European Data portal.

4.1.13 A language technology network in Denmark

Bolette Sandford Petersen (Professor, University of Copenhagen, president of the department council for terminology and language technology, Danish Language Council) stressed that there is a need to bring together language technology stakeholders in a language technology network.

During the meeting the participants therefore were invited to approach Sabine or Bolette to join the network. The network will meet 1-2 times per year and discuss major relevant issues.

4.1.14 Closing remarks by Sabine Kirchmeier

The conclusion of a day with many presenters who travelled far in order to explore the many facets of automatic translation, is that there are many exciting opportunities in machine translation, and that public authorities can save both time and money by being more aware of how they handle their translation activities.

Public institutions have a major influence on how the future of machine translation develops. The ability of EU's MT systems to handle Danish can be significantly improved

if public institutions decide to take an active part by contributing their data. The technology is still in its infancy, and there are a number of issues with, for example personal data protection, which the individual authorities need to cater for. But the prospects are huge, and the likelihood that we will have better tools to manage our multilingual reality is far greater if we engage and provides data, than if we do nothing.

4.2 Summary in Danish

4.2.1 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.

4.2.2 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.

4.2.3 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.

4.2.4 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.

4.2.5 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.

4.2.6 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.

4.2.7 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 https://ec.europa.eu/cefdigital/wiki/display/CEFDIGITAL/eTranslation

4.2.8 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.

4.2.9 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.

4.2.10 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.

4.2.11 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.

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.

4.2.12 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.

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.

4.2.13 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.

5 Synthesis of workshop discussions

5.1.1 In English

QUESTIONS: Why does the EU does not grant access to private companies to MT@EC and, in the long term, CEF.AT?

ANSWER: V. Andrejs Vasiljevs: The system is created mainly with a view to help public institutions to act in several languages. If you open MT@EC, to the private sector there will be the risk to interfere in commercial interests, as private companies have been specialising in developing and therefore also selling machine translation systems. The initiative does not wish to distort competition, and it is still being considered what role the European Commission should to play in the future in relation to the private and public actors in the areas where market fails to address the needs of multlingual Europeans.

Derrick Kinck Olesen added that 25 % of the translations that are made in the Danish Linguistic Unit of the EU, are made by freelance translators. This corresponds to approximately 16-18,000 cases per year. In this way, the EU already shares information with its external translators and both parties benefit from it.

QUESTIONS: Could literature or speech recognition be used as data?

ANSWER: Sabine Kirchmeier — fiction is one of the most difficult text types to acquire. Publishers and authors are extremely reluctant to provide texts because of copyright issues, which is understandable because it is their earning base. But language technology is not in competition with the book market. We are not interested in the work of art, but in the words and phrases. Right now there is a dialogue between the Danish Language Council and the Ministry of Culture on language technology and copyright, as well as a dialogue with the Danish Academy to provide literary texts for linguistic research. For machine translation, fiction poses another challenge as it represents completely different text genres. Authors use things like dialogue and a more emotional language that is not found in texts from the public sector, and they coin very imaginative images and phrases. In short: a different kind of language. If you use fiction as data for machine translation, for example in the EU, you risk getting some pretty alternative translations because the domain becomes too wide.

QUESTIONS: Can universities have access to MT@AC and, in the long term, CEF.AT?

ANSWER: Derrick Kinck Olesen — universities offering European Master of Translation have access to use the system. In Denmark, there is one translator training program at Aarhus University — and they have access. The system will eventually be extended so that other universities may have access, too.

QUESTIONS: Are there any plans to include non-EU languages, for instance Urdu, Arabic, etc. and is there a willingness to serve citizens with these?

ANSWER: V. Derrick Kinck Olsen — in our system the focus is exclusively on EU languages, but we offer translation from some of the major world languages (Chinese, Russian, Urdu) internally. We use the freelance market and assure the quality afterwards. The resulting data (e.g. translation memories) are included in our database, but we are at an embryonic stage as the demand is not large right now.

QUESTIONS: As employees at the Nordic Council of Ministers, can we use the system? We spend a lot of money on translating Finnish and Icelandic.

ANSWER: Derrick Kinck Olesen — it should be possible**.** The problem is that the system does not perform particularly well for Nordic languages such as Finnish at this stage, due to the nature of the language as it is highly inflectional. Norwegian gives a better result. **Sabine Kirchmeier** suggested that the Nordic Council of Ministers could try to feed their translated texts to the database in order to improve the quality for the Nordic languages, for instance as a smaller project within the organisation. Most translations and translation memories could be put to use here.

QUESTIONS: Can private companies download translation memory bases from the European Data Portal?

ANSWER: It is completely open — and does not even require that you register yourself. There are no restrictions relating to the European Entry Point Services. Only the automatic translation system is protected with a password.

QUESTIONS: It is possible to strike the right tone with machine translation?

ANSWER: Andrejs Vasiljevs — machine translation systems learn from data. And if there are sufficient data to model style and tone, then the system can learn it. But the need for data is large in order to achieve this. Machine translation does not distinguish between formal and informal language.

5.1.2 Questions and answers In Danish

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.

6 Appendix

6.1 Workshop presentations

Målsætninger

Andrejs Vasiļjevs, ELRC/Tilde **(præsentation)**

Europa og flersprogethed

Derrick Kinck Olesen & **Uffe Sonne Svendsen,** EU-Kommissionen, Generaldirektoratet for Oversættelse **(præsentation)**

Sprog og sprogteknologi i Danmark

Sabine Kirchmeier, Dansk Sprognævn (præsentation)

Diskussion: Flersprogethed i den offentlige sektor – hvordan imødegås udfordringen?

Deltagere:

Sigurd Slot Jacobsen, Specialkonsulent, Konkurrence- og Forbrugerstyrelsen **(præsentation)**

Anne-Mette Olsen, Konsulent, Region Sønderjylland – Schleswig **(præsentation)**

Automatisk oversættelse: Hvordan fungerer det?

Anders Søgaard, Professor Center for Sprogteknologi, Københavns Universitet **(præsentation)**

Hvordan får de offentlige institutioner gavn af CEF.AT-platformen?

Spyridon Pilos, Head of sector "Language Applications", Directorate General for Translation, European Commission **(præsentation)**

Tilde – Maskinoversættelse til den digitale offentlige sektor - et eksempel

Rihards Kalniņš, ELRC/Tilde **(præsentation)**

Hvilke data er der behov for? Hvorfor?

Andrejs Vasiļjevs, ELRC/Tilde **(præsentation)**

De juridiske rammer for levering af data, European Data Portal

Cathrine Lippert, Specialkonsulent, Digitaliseringsstyrelsen **(præsentation)**

Diskussion: Data og sprogressourcer i Danmark

Peter Juel Henrichsen, CBS **(præsentation)**

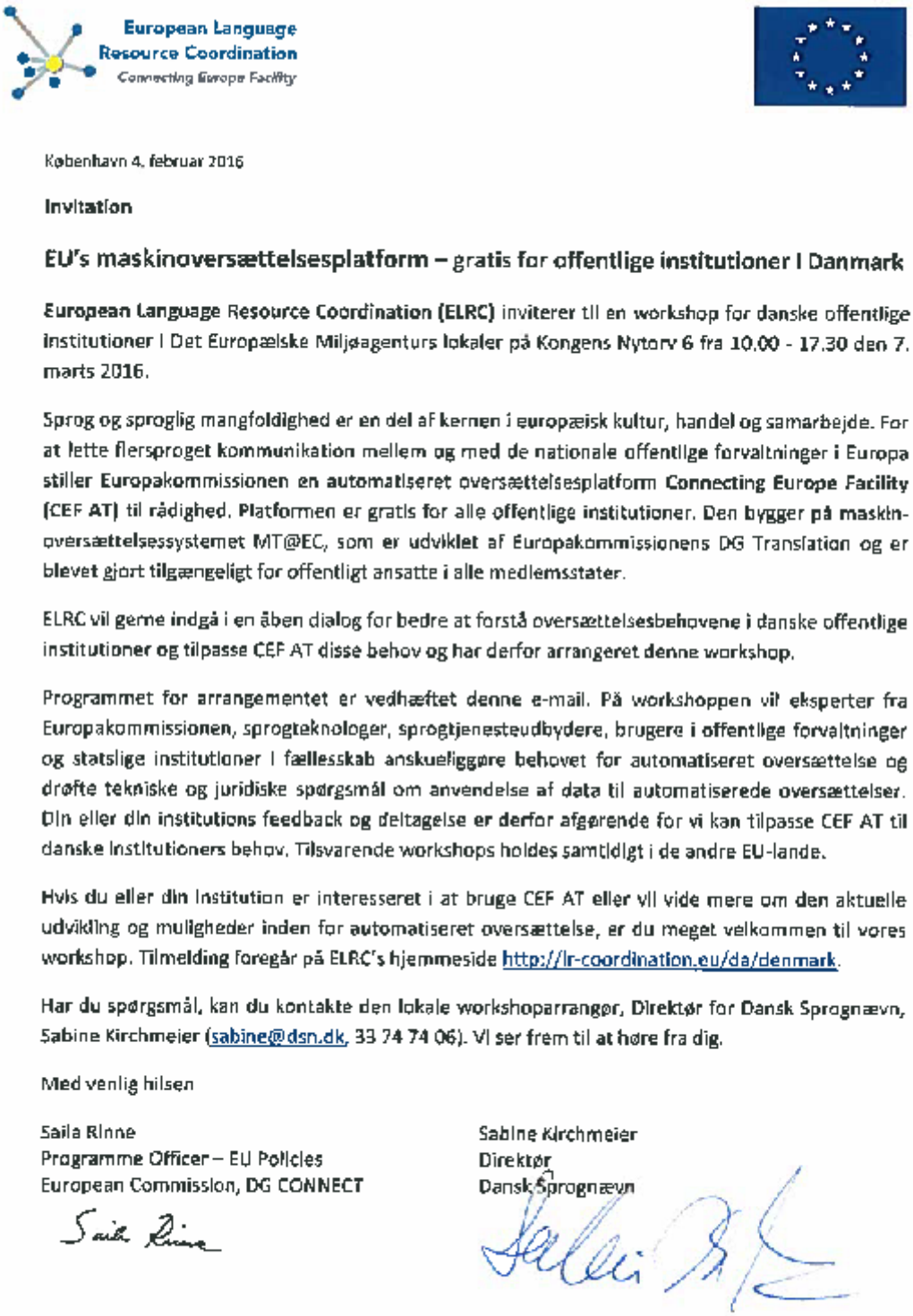
Bolette Sandford Petersen, Professor Københavns Universitet (præsentation)

Bodil Nistrup Madsen, Professor, CBS –Department of International Business Communication/DANTERMcentret (**præsentation**)

Data- og sprogressourcer: Tekniske og praktiske aspekter

Andrejs Vasiļjevs, ELRC/Tilde **(præsentation)**

6.2 Invitation



European Language Resource Coordination
Connecting Europe Facility

København 4. februar 2016

Invitation

EU's maskinoversættelsesplatform – gratis for offentlige institutioner i Danmark

European Language Resource Coordination (ELRC) inviterer til en workshop for danske offentlige institutioner i Det Europæiske Miljøagenturs lokaler på Kongens Nytorv 6 fra 10.00 - 17.30 den 7. marts 2016.

Sprog og sproglig mangfoldighed er en del af kernen i europæisk kultur, handel og samarbejde. For at lette flersproget kommunikation mellem og med de nationale offentlige forvaltninger i Europa stiller Europakommissionen en automatiseret oversættelsesplatform **Connecting Europe Facility (CEF AT)** til rådighed. Platformen er gratis for alle offentlige institutioner. Den bygger på maskinoversættelsessystemet MT@EC, som er udviklet af Europakommissionens DG Translation og er blevet gjort tilgængeligt for offentligt ansatte i alle medlemsstater.

ELRC vil gerne indgå i en åben dialog for bedre at forstå oversættelsesbehovene i danske offentlige institutioner og tilpasse CEF AT disse behov og har derfor arrangeret denne workshop.

Programmet for arrangementet er vedhæftet denne e-mail. På workshoppen vil eksperter fra Europakommissionen, sprogteknologer, sprogtjenesteudbydere, brugere i offentlige forvaltninger og statslige institutioner i fællesskab anskueliggøre behovet for automatiseret oversættelse og drøfte tekniske og juridiske spørgsmål om anvendelse af data til automatiserede oversættelser. Din eller din institutions feedback og deltagelse er derfor afgørende for vi kan tilpasse CEF AT til danske institutioners behov. Tilsvarende workshops holdes samtidigt i de andre EU-lande.

Hvis du eller din institution er interesseret i at bruge CEF AT eller vil vide mere om den aktuelle udvikling og muligheder inden for automatiseret oversættelse, er du meget velkommen til vores workshop. Tilmelding foregår på ELRC's hjemmeside http://lr-coordination.eu/da/denmark.

Har du spørgsmål, kan du kontakte den lokale workshoparrangør, Direktør for Dansk Sprognævn, Sabine Kirchmeier (sabine@dsn.dk, 33 74 74 06). Vi ser frem til at høre fra dig.

Med venlig hilsen

Saila Rinne
Programme Officer – EU Policies
European Commission, DG CONNECT

Sabine Kirchmeier
Direktør
Dansk Sprognævn