European Language Resource Coordination+LOT2 (ELRC+LOT2) is a service contract operating under the EU's Connecting Europe Facility SMART 2015/1091 programme.

European Language Resource Coordination
Connecting Europe Facility

Deliverable D3.2.x
Task 8

ELRC Workshop Report for Denmark

Author(s): Sabine Kirchmeier

Dissemination Level: Public

Version No.: <V1.0>

Date: 2019-01-09

Contents

1 Executive Summary

The second ELRC-workshop in Denmark took place on October 8th 2018 at the Danish Language Council in Copenhagen. There were 30 participants from Danish public and private institutions, from trade unions and from ELRC.

The main points made at the workshop were the following:

- It is necessary for public institutions to become more aware of the value that their language data such as texts and terminologies represent.
- Public institutions should bring into place procedures for identifying data that can be shared and actually share them with the ELRC and if possible also with other developers of language technology.
- The EU is making a huge effort in order to ensure that all EU-language can make use of language technology.
- But Denmark also has to contribute – first and foremost by providing more data.
- Public institutions can become the most important driving force in this process.
- ELRC and the technical national anchor points are ready to help to clean and package data and to solve legal and GDPR-related issues.
-

It was concluded that there is still lack of awareness about how valuable texts and terminologies are for creating better and more consistent content and for research and development of Danish language technology. Many institutions do not see the benefits of maintaining text and terminology databases or they cannot find the necessary resources for these tasks. Some are working systematically with terminology but not with text data.

Many public institutions have outsourced their translation tasks to private vendors without taking measures to keep, organize and reuse the translated data for other purposes.

In order for CEF.at and the continuous collection and contribution of Danish text data to become an integrated part of the workflow in Danish public institutions, it should be possible for private vendors to use CEF.at if they are translating EU-related texts or carrying out other types of translations for public institutions.

A public-private partnership model regarding translation and the use of public language data could probably improve the situation in Denmark considerably.

During the last 2 years, policy makers and state agencies such as the Agency for Digitization have realized that there has to be more focus on language technology for the Danish language. In January 2018, the minister of culture entrusted a committee for language technology lead by the Danish Language Council with the task of proposing a national strategy for language technology for Danish.

2 Workshop Agenda

When:	8. October 2018
Where:	Dansk Sprognævn, Worsaaesvej 19, 4., 1972 Frederiksberg C
Organization:	• The European Language Resource Coordination (ELRC) consortium • Dansk Sprognævn

9.30 - 10.00	Registration and coffee
10.00 - 10.20	Welcome **Sabine Kirchmeier**, Director, Danish Language Council, & **Jens Johansen**, Representation of the European Commission in Denmark
10.20 - 10.40	Digital public services across borders - results and ambitions Video presentation: Claire Bury, Deputy Director General at DG Connect, and Rytis Martikonis, Director General of DG-Translation
10.40 - 11.10	Digital public services in Denmark and the need for open data **Jens Krieger Røyen**, Head of section, Agency for Digitization, Denmark
11.10 - 11.40	Digital public EU-services in Denmark, presentation round **Christel Høst**, European Consumer Consumption Centre Denmark: Experiences with the ODR-platform
11.40 - 12.10	How does CEF eTranslation work? **Claus Larsen**, Language Technology Coordinator, European Commission Directorate-General for Translation
12.10 - 13.10	Lunch
13.10 - 13.40	Overview of the European Language Resource Coordination action **Andrejs Vasiļjevs**, ELRC representative for the Nordic and Baltic region, Tilde

13.40 - 14.00	Development of language technology in Denmark - the role of public institutions **Sabine Kirchmeier**, Danish Language Council
14.00 - 14.30	Collection of language resources in Denmark for ELRC - difficulties and possibilities **Bolette Sandford Petersen**, Professor, Center for Language Technology, University of Copenhagen **Lina Henriksen**, Senior Consultant, Center for Language Technology, University of Copenhagen
14.30 - 15.00	Discussion and conclusion **Sabine Kirchmeier**, Danish Language Council

3 Summary of Content of Sessions

3.1 Welcome and introduction

Jens Johansen, head of the Representation of the European Commission in Denmark, gave a warm welcome to everyone. He stressed that overcoming the language barriers is extremely important if the citizens of Europe are to benefit from the digital single market. The European Commission has a clear focus on ensuring that all European languages are given equal opportunities to be used in all digital contexts first and foremost by providing important tools such as terminology databases such as IATE and online machine translation systems such as CEF.at - the EU machine translation system platform developed by the European Commission's Directorate-General for Translation.

Jens Johansen furthermore pointed out that the success of this initiative highly depends on the commitment by especially public institutions to share their translation data with the Commission in return in order to improve the quality of CEF.at and he encouraged everyone to contribute as much as possible.

Sabine Kirchmeier, Director of the Danish Language Council, welcomed everybody and explained that the workshop was the second workshop held in Denmark. The first workshop took place in 2016. Much has happened since, and there is more focus on language technology than ever before. In Denmark, the minister of culture has entrusted a committee for language technology lead by the Danish Language Council with the task of proposing a national strategy for language technology for Danish. In a number of workshops, the committee is uncovering the needs for language resources and technological resources for Danish. The ELRC-initiative fits perfectly into the work of the committee and the findings of the ELRC-workshop will also become part of the final report to the Danish government.

3.2 Welcome By DG-connect

Claire Bury, Deputy Director General at DG-Connect, and Rytis Martikonis, Director General of DG-Translation, in their video presentation described the ambitions and challenges for a truly multilingual digital Europe. They stated that 90 % of European consumers prefer to browse websites in their own language, and that 42 % never purchase in a language other than their own. They also stressed that according to the European Treaty all languages are considered equal. This is reflected by the fact that European legislation is translated into all 24 official European languages, and that European citizens have the right to address public institutions all over Europe in their own language. The EU-translation services have translated European laws into the languages of all member states and the translation system that they use called, mt@ec, contains 1 billion translated sentences of high professional quality.

The system works best for text with a legal flavor. If it is to be used for other purposes, it needs to be supplied with high-quality data, for instance, professional translations and other language data, from other contexts such as medicine descriptions or banking association terminology. Mr. Martikonis explained the development of the technology behind the system. It has been mainly statistically based, but from 2018 it will make use of neural processing techniques that further improve the quality of the translation.

The CEF.at translation system will be used to support for instance the EU Open Data Portal which enables users to search for national or EU law and certificates in all national languages.
The speakers finally encouraged everyone to express their needs and to bring forward ideas as to how the EU can set up mechanisms to help to manage data efficiently at the national level.

3.3 Digital public services in Denmark and the need for open data

Jens Krieger Røyen, Head of section, Agency for Digitization, Denmark, gave an overview of how digital public services have been developed in Denmark. Through a series of targeted strategies since 2001, Denmark has become one of the most efficiently digitized societies in Europe. In 2011 it became mandatory for citizens and businesses to receive digital mail and to use online self-services provided by public institutions.

In 2016, there was a focus on user-friendliness and coherence of services and on developing an ICT architecture framework for interoperability and an ICT management strategy. In 2018 more than 4 million citizens (90, 8 %) are registered for digital post. The common architecture supports cross-organizational processes and data sharing in government and between the public and private sector.

The ambition is a more efficient, coherent public sector delivering on the needs of the citizens. It is based on the European Interoperability Framework (EIF).

Coherence is ensured for instance by the basic data Agreement of 2012 providing open access to public sector basic data for everyone, including enterprises and individuals through a common single distribution solution: The Data Distributor. There are initiatives for real estate, addresses, geographic data and data for individuals and businesses. Jens Krieger Røyen stressed that the aim for the future is to ensure that digitization is taken into consideration when new legislation is made.

Artificial intelligence is a new aspect that has to be taken into account and put to use in the public sector. Denmark will not be able to compete with the English speaking world with regard to research in this field but has strong advantages when putting new technology into play in the public sector. There is a general agreement that Denmark has to advance with regard to language technology for Danish, and that it is neither sufficient nor efficient to rely on systems that are developed for other languages.

In the following discussion, a number of questions were raised about GDPR – how personal information can be protected in the exchange of data. Furthermore, it was stressed by several people in the audience that as much data as possible must be made freely available.

3.4 Digital public EU-services in Denmark, presentation round

Christel Høst from the European Consumer Consumption Centre Denmark presented her view of automatic translation in the Open Dispute Resolution Portal (ODR) in Denmark. She made three major points:

1) The language of the consumer is often filled with spelling mistakes, typing errors and wrong use of legal terminology
2) It is very difficult to ensure the correct translation of legal terms

3) The system sometimes makes unexpected errors that are hard to explain and sometimes hard to detect.

Mr. Høst gave a number of examples of wrong translations between Danish and English that were made by the system, and discussed various reasons for translation errors. In a number of cases the consumer's spelling errors could cause misleading translations, in other cases the system was robust enough to ignore spelling errors such as *flight ticker* instead of *flight ticket* or *cancelkation* instead of *cancellation,* and render a correct translation. In other cases the translations were quite misleading, such as *Dear Sir* being translated into *Hr. Premierminister* (Mr. Prime Minister) or *internet broad band connection* into *internettilslutningsanordning med et bredt bånd* (internet connection device with a broad ribbon). Finally, linguistic elements that have an important impact on the semantic interpretation of the sentence, such as e.g. negations, are often placed wrongly which in the worst cases renders the opposite meaning.

Mrs. Høst concluded that she would not feel safe about making a legal decision based on the translated texts, and that more effort has to be put into improving the quality of the system before it can be useful. One of the challenges of the ODR portal is the fact that complaints can be about anything, and that it is difficult to fine tune the system to a specific domain.

There has been very little experience with Danish in other public EU-services. The organizers had been in contact with government officials working with ESPD and ESSI, but no experience with the use of CEF.at could be reported. Very few public institutions have in house translators and most of the translation work is outsourced to private vendors.

During the discussion the following points were made:

- The Danish parliament could use a translation system between Danish, Greelandic and Faroese, but these languages are not included in ELRC.
- The Ministry of Foreign Affairs has recently outsourced its in house translation unit which also delivered translation services to other public institution, to a private vendor.
- The Danish tax authorities are working closely with a private vendor but with no systematic approach to collecting and archiving translation memories and translated texts.
- The Danish Agency for Labour Market and Recruitment has a strong need for translation from all EU-languages into Danish in ESSI where electronic documents are exchanged across sectors, and where there is a lot of information given in unrestricted text fields, but little experience with MT has been made so far.

In general, translation tasks are mainly solved by private companies. In most organizations there is very little awareness about issues related to ownership to translation memories and no systematic approach to archiving and reusing translated texts.

3.5 How does CEF eTranslation work?

Claus Larsen, Language Technology Coordinator, European Commission, Directorate-General for Translation, gave an introduction to CEF.at and a demonstration of the system.

During the discussion the following questions were asked:

- Can the system be made more user-friendly? Claus Larsen stated that there is work in progress towards a more intuitive user interface.

- How can the system be accessed? This was demonstrated. In general, accessing the system was considered rather time consuming and was seen as an obstacle for using the CEF.at platform.
- What is the difference between statistical and neural translation? Claus Larsen explained the difference and explained that CEF.at mainly is trained on legal texts and debates in the EU-Parliament which explains some of the strange translation errors.
- How are translations validated? Claus Larsen explained that there are different approaches: BLEU-score which measures the statistical similarity between source and target text, and edit effort, which measures the number of corrections that have to be made in order to render a correct translation.
- Several participants raised concerns about GDPR-issues. How can CEF.at guarantee that personal data are safe at all times?
- Can CEF.at be used to translate the web pages of local municipalities? Some municipalities have implemented Google Translate on their web sites, but might achieve better quality with CEF.at for some languages.
- Can CEF.at be integrated with cat tools?
- Can CEF.at be used by private vendors?

Since many public institutions have outsourced their translation tasks, there were many questions about how a fruitful interaction between CEF.at and private vendors could be established. Claus Larsen pointed towards a number of EU-translation memories that were publicly available for download. Some of the participants suggested that it should be possible for private vendors to use CEF.at if they are translating EU-related texts or carrying out other types of translations for public institutions.

3.6 Overview of the European Language Resource Coordination action

Andrejs Vasiļjevs, ELRC representative for Nordic and Baltic region, Tilde, gave an overview of ELRC. He stressed the importance of accessing translated texts or translation memories between Danish and other languages from public bodies. He also mentioned that other resources such as bilingual or multilingual terminology could be extremely useful.

Most of the participants agreed that this would be a good idea, but also mentioned various obstacles such as the fact that most translation jobs were outsourced and that there were no resources to keep track of translations and translation memories.

During the discussion a number of open resources were mentioned such as

- The open data portal of the Danish Parliament https://oda.ft.dk/
- The DGT's translation memories
- DK-Clarin
- The Danish Spelling Dictionary
- Meta-Share

3.7 Development of language technology in Denmark - the role of public institutions

Sabine Kirchmeier, Danish Language Council, described the role of public institutions with regard to language technology and gave a status on the work of the language technology committee established by the minister of culture.

She stressed the difference between language data and other types of data, mainly that language data are complex, ambiguous and seemingly unstructured. Although languages have structure, the structure is highly complex and for many languages not sufficiently well researched and described. Language is ruled by conventions that may differ from text to text and in different situations. Language is closely connected to human identity – not two people express themselves in exactly the same way. Language is infinite both in terms of number of elements and their combinatory potential. Language is perpetually changing and language technology needs to be constantly updated in order to keep up with these changes.

Furthermore, she explained the importance - especially for less resourced languages like Danish - of sharing language data in order to support the development of language technology. The role of public institutions is extremely important, since public institutions produce a lot of texts that are highly relevant for language technology providers. If the collection and sharing of public texts and other language data can be organized and systematized this would benefit the development of language technology immensely. One of the major obstacles for this to happen is the fact that public institutions are not aware of the importance of their language data and do not have routines in place for identifying shareable data, archiving them and sharing them with others.

Danish is not the first priority for language technology providers, which means that Danish companies and public institutions usually have to wait several years before they can use new technology and profit from the increase in productivity it offers. During the last 2 years, there seems to be a growing awareness among policy makers and state agencies such as the Agency for Digitization that there has to be more focus on language technology for the Danish language. Thus, in the beginning of 2018 the Danish Minister of Culture entrusted a new committee on language technology with a survey of the perspectives and challenges for language technology in a Danish context. The committee was also asked to present recommendations as to how Denmark can ensure that Danish and other languages can be used in digital services in the future. Furthermore, the committee was to investigate the need and perspectives for the establishment of a national term bank.

The committee consists of representatives of private and public institutions, researchers and large and small language technology providers, and is led by the Danish Language Council.
The committee is currently organizing workshops and questionnaires in order to provide a roadmap of the situation for language technology for Danish involving users, distributers, developers, researcher and teachers within the field of language technology. Furthermore, it will draw on experiences from the EU (MetaNet, ELRC, Human Language Project etc.) and from similar initiatives in the Nordic countries. The experiences from the past and present ELRC-workshops are also considered important input for the committee.
The recommendations of the committee will be presented in March 2019.

3.8 Collection of language resources in Denmark for ELRC - difficulties and possibilities

Bolette Sandford Petersen, Professor, Center for Language Technology, University of Copenhagen, described the role of the national anchor points (NAPs)and the data collection work that is carried out for ELRC at the University of Copenhagen. She stressed that the only way automatic translation between Danish and other languages can be improved is that Denmark can provide high quality language data such as translated texts and terminologies.

A number of public institutions have already provided some data, but much more is needed. Bolette Sandford Petersen stressed that there is too little awareness about how valuable texts and terminologies are for research and development of Danish language technology. She explained that one of the main reasons is the fact that there is a lack of interest in language policy and lack of focus on language issues in the public sector, which for instance results in the fact that some public institutions do not use terminological resources or other tools that can ensure better text quality, and that public institution do not share language data. This lack of awareness and the skepticism towards data sharing has severe consequences for the development of language technology in Denmark.

Furthermore, the new rules for GDPR have created a sense of responsibility but also anxiety which has made public employees even more reluctant to share their data.
A more proactive and politically supported approach is necessary to make it easier and more acceptable to collect and share language data for instance in a common public infrastructure.
Bolette Sandford Petersen encouraged everybody to track down resources and to send links and relevant data sets to her and her colleagues at the technical NAP. She also offered help and support for everyone in the process.

3.9 Conclusions

Sabine Kirchmeier, Danish Language Council, concluded the workshop with a number of summarizing statements:

- It is necessary for public institutions to become more aware of the value that their language data such as texts and terminologies represent.
- Public institutions should on bring into place procedures for identifying data that can be shared and actually share them with the ELRC and if possible also with other developers of language technology.
- The EU is making a huge effort in order to ensure that all EU-language can make use of language technology.
- But Denmark also has to contribute – first and foremost by providing more data.
- Public institutions can become the most important driving force in this process.
- ELRC and the technical national anchor points are ready to help to clean and package data and to solve legal and GDPR-related issues.

4 Synthesis of Workshop Discussions

Danish is not the first priority for language technology providers, which means that Danish companies and public institutions usually have to wait several years before they can use new technology with Danish input or output and profit from the increase in productivity this offers.
Only very few public institutions have experience with the use of CEF.at, and those who have, do not think that the system has the necessary quality and reliability with regard to translation to and from Danish.

Some of the main obstacles apart from translation quality were accessibility, user-friendliness of the system and integration with other CAT-tools especially in cooperation with private vendors. Furthermore, GDPR-issues were not sufficiently clear, e.g. how can it be made sure that names, and addresses and other personal data are safe.

There is still lack of awareness about how valuable texts and terminologies are for creating better and more consistent content and for research and development of Danish language technology. Most institutions do not see the benefits of maintaining text and terminology databases or they cannot find the necessary resources for these tasks.

Many public institutions have outsourced their translation tasks to private vendors without taking measures to keep, organize and reuse the translated data for other purposes.

In order for CEF.at and the continuous collection and contribution of Danish text data to become an integrated part of the workflow in Danish public institutions, it should be possible for private vendors to use CEF.at if they are translating EU-related texts or carrying out other types of translations for public institutions. Private vendors could become a part of the language data supply chain and perhaps they could also offer language data management to public institutions.

A public-private partnership model about translation and the use of public language data could improve the situation in Denmark considerably.

During the last 2 years, there seems to be a growing awareness among policy makers and state agencies such as the Agency for Digitization that there has to be more focus on language technology for the Danish language. In January 2018, the minister of culture has entrusted a committee for language technology lead by the Danish Language Council with the task of proposing a national strategy for language technology for Danish.

4.1 ELRC and Open language Data in Denmark

Through a series of targeted strategies since 2001, Denmark has become one of the most efficiently digitized societies in Europe.
The ambition is a more efficient, coherent public sector delivering on the needs of the citizens. It is based on the European Interoperability Framework (EIF).
Coherence is ensured for instance by the basic data Agreement of 2012 providing open access to public sector basic data for everyone, including enterprises and individuals through a common single distribution solution: The open data portal Datafordeleren (The Data Distributor).

Artificial intelligence is a new aspect that has to be taken into account and put to use in the public sector. There is a general agreement that Denmark has to advance with regard to language technology for Danish, and that it is neither sufficient nor efficient to rely on systems that are developed for other languages.

There is in principle no problem sharing language data if GDPR-regulations and other legal regulations under the PSI Directive are respected, but there are no general policies in place that support sharing of language resources. Consequently, there are hardly any language data in the Danish open data portal (except for place names). Language data for Danish are currently shared through portals such as DK-Clarin, MetaShare and others and there is no structured overview of the data available.

4.2 Success stories and lessons learnt

Highlight:

- There is awareness now on the political level that language data are important for the development of language technology and artificial intelligence and that the state has to provide measures to organize and distribute language data for Danish, and appropriate measures are in preparation. This also includes translated data.

Suggestions:

- A key problem for the general use of CEF.at in Danish public institutions is the relation to private vendors since most translation tasks are outsourced and there is little in house translation in public institutions. A governance model for the collection and distribution of language data needs to cater for this.
- The access facilities and the quality of CEF.at are still a major obstacle. A better user interface, integration into other CAT tools and a facility to translate web sites was considered to be very useful. There was general agreement that the quality could only be improved by making more data available.