European Language Resource Coordination (ELRC) is a service contract operating under the EU's Connecting Europe Facility SMART 2019/1083 programme.

Deliverable D3.2.27
Task 3

ELRC Workshop Report for Denmark

Author(s): Carl Frederik Bach Kirchmeier (Agency for Digital Government), Kristīne Vīgante (Tilde), Aivars Bērziņš (Tilde)

Dissemination Level: Public

Version No.: V3.0

Date: 2023-01-19

Contents

1 Executive Summary 3

2 Workshop Agenda 4

2.1.1 ELRC WORKSHOP 5

2.1.2 ELG WORKSHOP 5

3 Summary of Content of Sessions 6

3.1 Session 1: "Welcome and introduction'' 6

3.2 Session 2: "On Danish chat bots and a Danish open source initiative" 6

3.3 Session 3: "Natural language processing: Recent advances and challenges" 7

3.4 Session 4: "What eye movements can teach us about language processing" 7

3.5 Session 5: "Speech Recognition for Digital Inclusion and Support" 7

3.6 Session 6: "Sprogteknologi.dk – A Common Danish Language Resource" 8

3.7 Session 7: "COR – Central Word Register for Danish for Language Technology purposes" 8

3.8 Session 8: "eTranslation" 8

3.9 Session 9: "European Language Resource Coordination (ELRC) and beyond" 9

4 Country Profile: Language data creation, management and sharing 10

5 Workshop Participants 11

6 Workshop Evaluation – Participant Feedback Analysis Error! Bookmark not defined.

7 Appendix 12

7.1 Localised agenda 12

7.2 Localised press release 12

7.3 Workshop Presentations 13

1 Executive Summary

The third ELRC workshop in Denmark took place on 16 November 2021 at 13.20-15.00. The ELRC workshop was part of the broader programme of the day. The Centre for Language Technology at the Department of Nordic Studies and Linguistics, UCPH, and the Danish Agency for Digitisation, hosted the conference "Language Technology conference 2021" with a focus on the language technology developments and status in Denmark and in the European Union. It was decided by the organisers to schedule the workshop within a larger conference, since both organisations had planned a few language technology conferences/seminars etc. soon. Both parties, therefore, agreed that it would be better to hold one larger event, since the community for language technology in Denmark isn't big enough to host multiple events in a short span of time.

The morning session's focus was on research and development in Denmark; the afternoon part was dedicated to two workshops on the European Language Resource Coordination (ELRC) with a focus on EU machine translation (eTranslation), and the European Language Grid (ELG), which is developing a pan-European language technology portal.

2 Workshop Agenda

The final workshop programme in English

09:00 – 09:10	**Welcome and introduction** Bolette S. Pedersen, Centre for Language Technology, UCPH, and Christian Plaschke, The Danish Agency for Digitisation
09:10 – 09:40	**On Danish chat bots and a Danish open source initiative** Henrik Fabrin, Certainly Presentation
09:40 – 10:10	**Natural language processing: Recent advances and challenges** Barbara Planck, IT University of Copenhagen Presentation
10:00 – 10:20	**What eye movements can teach us about language processing** Nora Hollenstein, Centre for Language Technology, University of Copenhagen Presentation
10:20 – 10:30	**Discussion/questions: Speakers from the first morning session in the panel**
10:30 – 11:00	***Coffee Break***
11:00 – 11:30	**Speech Recognition for Digital Inclusion and Support** Klaus Bjørn Larsen, Roskilde Kommune Presentation
11:30 – 11:50	**Sprogteknologi.dk – A Common Danish Language Resource** Danish Agency for Digitisation Presentation
11:50 – 12:10	**COR – Central Word Register for Danish for Language Technology purposes** Peter Juel-Henrichsen, The Danish Language Council Sanni Nimb, Det Danske Sprog- og Litteraturselskab Bolette Sandford Pedersen, Centre for Language Technology University of Copenhagen Presentation pt. 1 Presentation pt. 2
12:10 – 12:20	**Discussion/questions. Speakers from the second morning session in the panel**
12:20 – 13:20	***Lunch Break***

2.1.1 ELRC WORKSHOP

13:20 – 13:45	**eTranslation** Claus Thornby Larsen, the European Commission Presentation
13.45 – 14:10	**Update from the ELRC and new tools on the CEF platform** Thierry Declerck, European Language Resource Coordination
14:10 – 14:30	Panel on ELRC and eTranslation seen in the Danish Context
14:30 – 15:00	**Break**

2.1.2 ELG WORKSHOP

15:00 – 15:30	**The European Language Grid and European Language Equality Projects** Georg Rehm, German Research Center for Artificial Intelligence, Germany
15:30 – 15:50	**Panel on ELG and Synergies with the Danish LT community as presented in the morning sessions**
15:50 – 16:00	**Conclusions**

3 Summary of Content of Sessions

The presentations at the conference were given by 9 different speakers, representing researchers, public and private organisations. The first 6 presentations were focused on development, outlook and use-cases of Danish language technology. It was discussed how Danish language technology could help businesses, what the major obstacles for further developments are, and how to improve the development and use of Danish language technology.

The last 3 presentations were given during the ELG and ELRC workshops.

Claus Thornby Larsen from the European Commission presented eTranslation, with a focus on the possibilities and current limitations of the tool.

This was followed by a presentation from Thierry Declerck (representing the ELRC) with the title "Update on ELRC and new tools on the CEF platform". It focused on the importance of gathering language resources, so that tools, such as eTranslation, could be further improved.

In the Panel discussion, the participants Thierry Declerck, Thornby Larsen and Sabine Kirchmeier from EFNIL discussed how European politics could benefit from Danish language technology and how Denmark could further support the initiatives taken by the EU.

Lastly, Georg Rehm held a presentation on the ELG (European Language Grid) platform. It was discussed how the ELG has contributed to the European LT thus far and its effects on the upcoming Language Data Space and the future of AI-related solutions in LT.

3.1 Session 1: "Welcome and introduction"

Bolette S. Pedersen, (Centre for Language Technology) and Christian Plaschke (The Danish Agency for Digitisation), the local ELRC representatives, opened the event by welcoming the audience and introducing the key persons in conceiving and organising the event, namely the ELRC consortium and the EC/DGT representatives.

3.2 Session 2: "On Danish chat bots and a Danish open source initiative"

This session was presented by Henrik Fabrin, CEO and Co-founder of Certainly. The presentation was launched with an introduction to the company. He briefly described Certainly, explaining that it is a Conversational AI Platform made for e-commerce and customer service and was first developed 10 years ago using open-source technologies. A few sentences were also devoted to the history of the company, goals, vision and mission, and main export markets as well aspects of e-commerce and future developments. Henrik Fabrin reported on the growth in conversations within the platform and overall usage of chatbots from January 2019 to October 2021, as well as e-commerce all together.
Henrik Fabrin explained that their choice was to use open source, mainly because it is a perfect alignment between commercial strategy and social responsibility. He mentioned that the company contributed to the open-source community by opening four language models (Danish,

Swedish, Norwegian and Finnish), datasets of questions for Coronavirus in English and Danish as well as Unity test libraries. All listed technologies are part of their chatbots. He emphasised the benefits for companies, and that the use of the platform increases the average order value by 20 percent. The presentation was rounded up with conclusions about the changes that e-commerce has experienced over recent years.

3.3 Session 3: "Natural language processing: Recent advances and challenges"

Professor Barbara Plank (ITU, Copenhagen) began her presentation by raising what she sees as the big goal of Natural Language Processing, that is for AI to understand and produce language, just as humans do. She briefly described the Sources of Bias and how they are not only in the data. She further presented a short historical insight and future growth prospects for natural language processing (NLP). The market size is expected to triple from USD 11.6 billion in 2020 to USD 35.1 billion by 2026. She also briefly touched on the development of language models and the importance of data to sustain the development of large language models.
Barbara Plank introduced various NLP methods for Danish: Dependency Parsing, Coreference Resolution, Nested Named Entities, Lexical Normalisation and provided some examples for these methods. She also presented results for Nested NER: MTL versus STL. The main conclusion of the presentation is that NLP has grown tremendously, biases are everywhere, and the direction should be towards more inclusive & robust NLP.

3.4 Session 4: "What eye movements can teach us about language processing"

Nora Hollenstein (Centre for Language Technology, University of Copenhagen) briefly presented the basics of eye-tracking technology, introducing features such as: First fixation duration, Total reading time, Landing position and Mean fixation duration. She continued with a brief overview of Psycholinguistics.
The speaker then elaborated on the potential of eye-tracking in NLP, and how NLP can be improved on various levels with eye-tracking data. She emphasised that eye-tracking data reflects human text processing and forms Multi-modal NLP models; the compiled data provides modest but consistent improvements on various levels of text processing.
She closed her presentation with two statements:

- Eye-tracking provides online metrics of language comprehension on multiple levels of analysis.
- With eye-tracking technology, a better understanding of the similarities and differences between human and machine language processing is acquired.

3.5 Session 5: "Speech Recognition for Digital Inclusion and Support"

Klaus Bjørn Larsen gave a very informative presentation about a project on intelligent speech recognition and the benefits for the public and businesses. Opening his presentation with the project purpose and benefits, he introduced the project's focus groups: 18-29 year-old, Elderly 70+, Disabled. He also briefly touched on the usage scenarios and explained why the public

services area is good for data collecting and testing. Considering the fact that one of the most important aspects is data protection, he emphasised that the use of the chat bot is non-binding, no personal data is retained, and contact is anonymous.
Klaus Bjørn Larsen concluded his presentation with the statement that business benefits from voice-bots in service availability and compliance (distinctly standardised responses), and that full automation of speech recognition is the new goal.

3.6 Session 6: "Sprogteknologi.dk – A Common Danish Language Resource"

This session was presented by Christian Plaschke (Danish Agency for Digitisation). He began his presentation by introducing what the Sprogteknologi.dk initiative is and its purpose – to strengthen the development of solutions for the Danish language using AI based technologies. He then continued by describing why this initiative is important to language technology companies as well as to end users, whether they be public administrations, researchers, companies or citizens. He concluded his presentation with a recommendation on how to ensure that the Danish language thrives in the digital age: share Danish language resources, Open-source as much as possible and think about the future collection of language resources.

3.7 Session 7: "COR – Central Word Register for Danish for Language Technology purposes"

Bolette Sandford Pedersen presented COR – Central Word Register, and explained how Danish companies now cooperate to a large extent to create Danish-orientated AI. She explained and gave examples of why COR must facilitate language-centred AI systems for Danish and is therefore based on locally rooted knowledge of the Danish language and society rather than simply adapting applications from English. She continued by explaining how COR develops basic guidelines for the amalgamation of meanings from the Danish Dictionary and deletion of obsolete or overly specialised meanings.

3.8 Session 8: "eTranslation"

The presentation began by introducing the Directorate-General for Translation in numbers and facts. Claus Thornby Larsen (Language Technology Coordinator Directorate-General for Translation) compared machine translation from the past to nowadays. He continued his presentation with a deeper introduction into eTranslation, emphasising the main disadvantages of machine translation, such as inconsistent terminology, omissions or additions, providing several examples. He gave some examples of how MT mistakes can be politically sensitive. He emphasised that MT quality depends on the available data and data consistency. He continued by explaining how each language pair is trained and what amount time is necessary to do so. He gave as an example of fields such as legislation, where a legal act is often composed of bits and pieces from other legal acts and defined terms must be used consistently.
He concluded his presentation with the statement that AI is of great help to professionals.

3.9 Session 9: "European Language Resource Coordination (ELRC) and beyond"

Thierry Declerck gave an in-depth overview of the European Language Resource Coordination activities, starting his presentation with a brief history, a description of the consortium members, and ELRC objectives - to provide language resources to overcome language barriers in public services and the European Digital Single Market in general. He characterised various types of language data which are under the focus of the ELRC, as well as described the progress achieved through several phases of ELRC activities. He paid special attention to the importance of achievements in numbers of unique Language Resources (LR) by type, especially large multi-language multilingual corpora. Although much has been achieved since 2015, he emphasised that this is only the tip of the iceberg and there is still much to be done. As he continued his presentation, he highlighted that there are now more than 676 language tools/services offered by CEF eTranslation and third-party services. He explained that besides activities on language data collection and processing, ELRC also offers advice on legal questions and assists in improving data management processes. He encouraged the participants to contact the Helpdesk with any questions on technical or legal aspects related to the use, production, collection, processing and sharing of LR.

To conclude the presentation he highlighted opportunities, challenges, and the future of ELRC and how this has been impacted by the increasing demand for new languages, new domains and modalities.

4 Country Profile: Language data creation, management and sharing

An updated profile for Denmark with respect to language data creation and management has recently been published in the 2022 ELRC White Paper (see: https://lr-coordination.eu/sites/default/files/LRB/LRB-12/ELRC-White-Paper.pdf)

5 Workshop Participants

There were 143 participants who completed the online registration form maintained by the Danish Agency for Digitisation. This reflects the number of participants as well, as some came without registration and some people who had registered didn't attend.

6 Appendix

6.1 Localised agenda

6.2 Localised press release

https://cst.ku.dk/kalender/sprogteknologisk-konference-2021/

https://sprogteknologi.dk/blog/program-for-sprogteknologisk-konference

Blogpost about conference

https://finnaarupnielsen.wordpress.com/2021/11/18/sprogteknologisk-konference-2021/

6.3 Workshop Presentations

The presentations are available on the ELRC website: https://lr-coordination.eu/denmark3rd.

All the presentations are also available on the local website (via Centre for Language Technology, Department of Nordic Studies and Linguistics): https://cst.ku.dk/kalender/sprogteknologisk-konference-2021/.

	(Facilitator Manex Agirrezabal)
	ELRC-workshop Ordstyrer Carl Frederik Bach Kirchmeier
13.20-13.45	Claus Thornby Larsen, EU-kommissionen: ***EU-Kommissionens maskinoversættelsessystem: eTranslation.*** Download præsentation.
13.45-14.10	Thierry Declerck, European Language Resource Coordination (ELRC): *Update from the ELRC and new tools on the CEF platform.* Download præsentation.
14.10-14.30	Panel on ELRC and eTranslation seen in the Danish context
14.30-15.00	**Pause** Kaffe Posters fra forskere, brugere, udviklere og studerende (Facilitator Manex Agirrezabal)