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

European Language Resource Coordination
Connecting Europe Facility

Deliverable D3.2.5
Task 3

ELRC Workshop Report for Norway

Author(s): Kristine Eide, Language Council of Norway

Dissemination Level: Public

Version No.: 2

Date: 2022-02-01

Contents

1 Executive Summary 3

2 Workshop Agenda 4

3 Summary of Content of Sessions 5

3.1 Welcome and introduction 5

3.2 Implementing chatbots in the municipality of Fredrikstad 5

3.3 Using machine translation in the Ministry of Foreign affairs 6

3.4 Using machine learning to classify criminal cases. 6

3.5 Language Technology: Latest trends and pitfalls 6

3.6 What is language data and how do we collect them? 7

3.7 Norwegian language technology and international collaborations 8

4 Synthesis of Workshop Discussions 10

5 Country Profile: Language data creation, management and sharing 11

1 Executive Summary

The workshop was organised by the Norwegian Language Council and The National Library of Norway, as a virtual event on March 3rd 2021. The title of the workshop was "Kan roboten egentlig snakke?", best translated as "does the chatbot actually speak?" – with the original pun getting somewhat lost in the translation.

International projects and initiatives, like ELRC, ELG and ELE are especially important for smaller countries like Norway, particularly since Norway is not in the EU. The theme of the workshop was a response to the action plan from the ELRC white paper and to the demand from the public sector for knowledge about AI, language technology and its implementation and the demand for language data from the side of the developers. The workshop focused on:

1) implementation of LT solutions, including machine translation, in the public sector.

2) Language technology and AI, latest trends and common pitfalls.

3) Language data creation, management and sharing.

A report from 2020 on the use of language technology in the Norwegian public sector showed, not surprisingly, that knowledge of language technology within an organisation correlated with the organisation's use of language technology. Also, knowledge of language technology among decision makers was the most important factor for deciding whether or not to implement new language technology.

Previous presentations of the importance of language data on national arenas have revealed that the public sector, including decision makers, computer scientists and system administrators are not really aware of what language data are nor of their importance. Although large organisations with in-house translation units, such as the Department of Foreign Affairs, are very aware of the value of their data, smaller organisation do not share the same awareness of data value and the possibilities of machine translation.

The two main goals of the workshop were therefore to fill the knowledge gap on language data in the public sector and to provide examples of the usage of language technology and artificial intelligence.

Another goal is to provide organisations with a simple solution as to where they should send their data. One of the most common complaints we hear, is that 1) there are several data repositories, and it is difficult to know which one to choose and 2) that different types of language data should be sent to different places.

In order to solve this problem, the Norwegian Digitalisation Agency, the National Library and the Norwegian Language Council have joined forces and produced guidelines for how to handle language data. The Norwegian solution is to have everything sent to the language bank at the National Library, who is then responsible for passing the data on to for instance, the ELRC.

The Language Council is also currently working on guidelines for the implementation of chatbots in the public sector. It was therefore important to present a case of implementation of chatbots in the workshop programme.

The conference was streamed live on youtube: https://www.youtube.com/watch?v=bVJHNeeAVvI

An edited version can be found here:
https://mediasite.nb.no/Mediasite/Showcase/arrangementsarkiv/Presentation/ea3753c1c6574186bc4b84f395491b9d1d

2 Workshop Agenda

10:30	Welcome by Andrea Lösch, ELRC
10:40	Implementing and running chatbots in the Municipality of Fredrikstad
	Ketil Johansen, Municipality of Fredrikstad
11:10	Using Machine translation in the Department of Foreign Affairs
	Stein Gabrielsen, Department of Foreign Affairs (PRINCIPLE Early Adopter)
11:40	Using artificial intelligence in the classification of criminal cases
	Jan Roar Beckstrøm, Office of the Auditor General
12:10	Break
13:00	Language technology: latest trends and common pitfalls
	Pierre Lison, Norwegian Computing Centre
13:45	What is language data and how do we collect it?
	Magnus Breder Birkenes, National Library/The Language Bank
14:15	Norwegian language technology and international collaborations:
	European Language Resource Coordination (ELRC), eTranslation and the European Language Grid (ELG)
	Kristine Eide, The Language Council of Norway
14:45	End of the workshop

3 Summary of Content of Sessions

3.1 Welcome and introduction

Andrea Lösch presented ELRC at a glance, focusing on the types of language data that the ELRC collects, and why they are collected. The eTranslation service was also presented, along with the additional, expected upcoming services on the CEF AT platform and the language repository ELRC share.

3.2 Implementing chatbots in the Municipality of Fredrikstad

Chatbots are one of the most commonly used LT solutions in the public sector. They are used for external communication as well as internal communication.

The Municipality of Fredrikstad has implemented a municipal chatbot, presented by Ketil Johansen, chief IT manager who has also been on the advisory board for the Language Council's planned "Guidelines for implementing chatbots in the public sector".

Johansen described the implementation of the two chatbots in the Municipality of Fredrikstad: *Ivar*, the internal chatbot, designed to answer questions from employees, and *Kari*, the external chatbot who answers questions from the citizens.

The chatbots have increased efficiency in the municipality and the results have been overwhelming.

Johansen stressed that a strong organisation is necessary for the chatbot implementation. In his organisation, each section has its own, designated chatbot-monitor, who registers new questions and supervises the answers. New intents are registered by monitoring questions the robot cannot answer. New answers are accepted by an editorial committee before being fed to the chatbot.

Without such a strong foothold in the organisation, the implementation would not have worked. Similar projects in other municipalities who have not invested as much in implementing the structures in the organisation have been less successful.

Fredrikstad has also cooperated with other municipalities. The questions from the public are likely to be similar, and they can re-use each other's intents and answers.

Several questions were raised after the presentation, such as whether or not the robot can have integrated speech recognition and machine translation, if it supports both varieties of Norwegian (Nynorsk and Bokmål) and what the greatest challenge has been during the implementation process

In Johansens view, speech recognition can be implemented in the future, although they are not quite there yet, the same is true for machine translation. As for supporting the two varieties of Norwegian, this would require quite a bit of manual work, since the answers have to be written by a person. As of yet, Fredrikstad only has Bokmål. (Other municipalities have Nynorsk, though)

The greatest challenge has been the implementation of the process in the line organisation, giving people time to work on the project. The technology works, but learning the complexity of the software is time consuming. Both the implementation of the chatbots as well as subsequent maintenance and monitoring has to be part of people's job descriptions.

All in all, the development has been substantial: a few years back, Fredrikstad Municipality did not even know which questions people were going to ask. Now we know, and we provide answers. The application can be developed further and needs to integrate functions such as speech recognition and machine translation.

3.3 Using machine translation in the Ministry of Foreign affairs

Implementing eTranslation through the Principle (early adopter) project in the Ministry of foreign affairs.

Stein Gabrielsen, Ministry of Foreign Affairs gave us a short history of translation in the Ministry of foreign affairs. This talk was performed without slides, making it more like a podcast, and made for a very welcome break from the usual slideshows.

The Ministry of Foreign affairs started using SMT when it became commercially available, integrated in the CAT Tool, using only internal input data. The machine translation worked to a certain point, but required heavy supervision. A solution with neural machine translation was tested, and there was general agreement that the quality was better, even though the efficiency was not scientifically measured. However, the price was too high and the agreement was not renewed.

A new opportunity for machine translation appeared with the PRINCIPLE project, using data from the Ministry.

Comparing PRINCIPLE to Google translate, in many cases there was not a big difference, probably because Google has used data from the ELRC and the Norwegian Language bank. Even though the Iconic system has some particularities, and it is a little too early to draw conclusions, results from the first phase are promising. Overall, machine translation has been very useful, but that is probably due to a large amount of internal input data.

3.4 Using machine learning to classify criminal cases.

Jan Roar Beckstrøm Division Director/Chief Data Scientist at the Innovation Lab, Office of the Auditor General of Norway.

Use of Machine learning in the classification of criminal cases in the ICT area. With the purpose of finding out how many cases were ICT cases. – a case study

The Office of the Auditor General wanted an answer to the question: Out of all criminal cases, how many cases involve ICT-crime?. The presentation was a detailed description of the process of making training data, cleaning, classification and training of the model, followed by lessons learned during the process.

Lessons learned:

- Test alternative algorithms – find out which serves you the best. Overfitted models do not predict well. Four models were tested: Naive Bayes (bad result), Random forest (over-fitted), XGBoost (over-fitted), and finally Support Vector Machine which was chosen.
- You need training data? – Make training data!

3.5 Language Technology: Latest trends and pitfalls

Pierre Lison, Senior Research Scientist at the Norwegian Computing Center

Lison outlined the latest trends in Language technology, focusing on three topics: Neural language models, ethics and explainability and small datasets.

Data driven models reproduce stereotypes and prejudices that are expressed in the training data. The lack of representation of minority groups in the language data can lead to consequences such as some groups being less understood in automated speed recognition. Predictions from these neural models are hard to explain and more research is needed.

As for small/inadequate data, this can be partially remedied by

a) reuse from related domains/languages
b) data increase: produce new training material from modified version of older training sets
c) weak supervision

These are some of the pitfalls that many stumble into when they want to make use of the latest technologies:

a) Expectations are too high:
    - the model's "understanding" is still very superficial
    - do not fall for the hype: If someone tries to sell a "revolutionary" system with a human "understanding" of language, run in the opposite direction!
b) The task is vaguely defined:
    - Inputs and outputs must be clearly defined
    - "good" and "bad" outputs must be clearly defined
    - Do not underestimate the importance of preprocessing
c) Bad or insufficient data:
    - Data is essential for LT solutions, including rule based systems.
    - Data collection and annotations require plenty of time and resources
    - High quality data is important
d) Using the newest, most complicated models is often a bad idea
    - Desire to use the latest, most advanced language models is often a bad idea: they are more difficult to use, and it is harder to understand the mistakes that are produced.
    - If money is limited, it often pays off to invest in data resources instead.
e) Not enough focus on evaluation.
    - Many public enterprises buy packets form a developer, and the evaluation of the system may be anecdotal.
    - Systematic evaluations, both quantitative and qualitative pay off! The evaluations need to be repeated over time to reflect new language usage, new topics etc.

3.6 What is language data and how do we collect them?

Magnus Breder Birkenes, National Library/The Language Bank

Breder Birkenes gave an introduction to language data and their use in language technology and the importance of representation in the language data that is being gathered.

He introduced the Norwegian Language Bank and exemplified how they work there:

Example 1: Creating a parallel corpus from translations from the Norwegian database for public procurement, resulting in 300.000 translation units.

Example 2: Creating *Målfrid*. *Målfrid* was originally made for statistical purposes, to facilitate the language report each national public enterprise has to deliver every year on the use of *Bokmål* and *Nynorsk* in their publications. For these purposes, The National Library harvested all webpages from the national public sector, processed them, and detected the languages of the material.

The harvesting down to level 12 and its subsequent processing, including OCR-scanning of pdf files, deduplication and classification according to topic, resulted in what is to date probably the largest existing free corpus of clean Norwegian texts, with 3.2 bn words in Bokmål and 289 million Nynorsk. (This compared to for instance Common Crawl's 804 millions for Bokmål and 9,4 million for Nynorsk.)

Birkenes then presented a new project, which is the creation of a digital solution for depositing language data for public and private institutions. The solution will provide the opportunity to upload anonymized material that can be published in the Language Bank's catalogue.

3.7 Norwegian language technology and international collaborations

Kristine Eide, Språkrådet

European Language Resource Coordination (ELRC), eTranslation and the European Language Grid (ELG)

The talk focused on the advantages of joining European projects and initiatives such as the ELRC and the ELG even though Norway is not a member state of the EU. We want the same technologies as EU countries, namely:

- Language technology that works just as well in our own language as in larger languages.
- Language technology that is independent of the major players.
- Language technology that makes it possible to communicate across languages.

To achieve this, we need data. A comparison between the Norwegian and the Swedish eTranslation, shows that the Swedish translation from English is much better than the one to Norwegian Bokmål (Nynorsk is not an option). This is probably due to the lack of Norwegian data.

The new guide for public enterprises on how and where to send their language data was presented. The guide is a cooperation between the Language Council, The Norwegian Digitalisation Agency and The National Library. It is a result of a demand for information on identifying language data and what to do with them. Many public enterprises have expressed interest in sharing their data, and language data is mentioned specifically in the yearly circular from the Norwegian Digitalisation Agency. One question that comes up quite frequently, is what happens to the data that has been sent. We think that by explaining where it is sent, how the data is processed and its practical usage, the value of the data becomes more evident, and public enterprises will be more likely to spend the time needed to organise their own data.

The CEF AT platform was presented, focusing on what we can expect of the extensions of domain coverage, language coverage and additional language technologies.


European Language
Resource Coordination
Connecting Europe Facility

4 Synthesis of Workshop Discussions

The main issues that came up in the workshop:

1) The importance of a strong foothold within the organisation when implementing new technologies.
2) Machine translation between Norwegian and English works quite well, depending on the right set of data. Area specific data give better results.
3) Investing in proper training data is just as important as the model.
4) The importance of constant evaluation of new system.
5) The benefits of taking part in international projects and cooperations.

Other discussions among the participant circulated around whether or not a chatbot can have integrated speech recognition and machine translation, if it supports both varieties of Norwegian (Nynorsk and Bokmål) and what the greatest challenge has been during the implementation process.

The Norwegian Language Bank at the National Library has definitely been a success, considering the amount of data they have gathered and made publicly available. In particular, the parallel Bokmål-Nynorsk texts are important for machine translation from English to Nynorsk.

5 Country Profile: Language data creation, management and sharing

The main challenges for Norway described in the Country profile in the ELRC white paper, were the lack of data for Nynorsk, lack of parallel texts for Norwegian in General and for Nynorsk in particular, lack of awareness of the value of language data in the public sector and of anonymization of for instance translation memories.

As this workshop has shown, many of these challenges have already been addressed, through the action points defined in the Country profile: The amount of data has increased substantially for both Bokmål and Nynorsk, there are growing corpora of parallel Bokmål-Nynorsk texts as well as translation to and from English. The National Library, the Language Council and The Norwegian Digitalisation Agency have produced a guide for how to identify translations and other language data, and to make them available for reuse. With this in mind, The National Library has initiated the creation of a digital solution for depositing language data for public and private institutions.

This does not mean that there is enough data. Norway lags behind the EU countries with regard to the amount of translation units that are available for use in machine translation as well as when it comes to area specific data. Raising the awareness of what language data is, its value and its uses in language technology must be continued.