

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



Deliverable D3.2.7 Task 8

ELRC Workshop Report for Sweden

Author(s):	Simon Dahlberg and Rickard Domeij, Swedish Language Council
Dissemination Level:	Public
Version No.:	<V1.0>
Date:	2018-07-17



Contents

1	Executive Summary	3
2	Workshop Agenda	4
3	Summary of Content of Sessions	5
3.1	Welcome! 10:00-10:10	5
3.2	Introduction 10:10-10:20	5
3.3	What is the ELRC? 10:20-10:40	5
3.4	Usage of eTranslation at the Swedish National Food Agency 10:40-11:00	5
3.5	Need for automatic translation at the Swedish Agency for Economic and Regional Growth 11:20-11:40	5
3.6	Working with collecting translation data in Norway 11:40-12:00	6
3.7	Panel discussion: eTranslation in Sweden – possibilities and challenges 12:00-12:30	6
3.8	Our work with ELRC and open data 13:30-14:00	7
3.9	What the Swedish state has been doing to enhance open data 14:00-14:20	7
3.10	The role of translation companies in disseminating language data 14:20-14:40	7
3.11	How to collaborate in collecting language data 15:00-15:20	8
3.12	Questions & Answers	8
4	Synthesis of Workshop Discussions	9
4.1	ELRC and Open language Data in Sweden	9
4.2	Success stories and lessons learnt	9

1 Executive Summary

The second ELRC workshop in Sweden was held on May 23rd 2018, at the European Commission's venue in Stockholm, *Europahuset*. In general, the participants appreciated the workshop very much, finding a well-organised arrangement, a pleasant atmosphere and interesting presentations.

The workshop started with two introductions on the ELRC project and its objectives. The first one was given by Rickard Domeij from the Swedish Language Council who gave a broad view of the context, and the second one, given by Andrejs Vasiljevs from Tilde, went deeper into the objectives and the participating partners of the project.

After them, we heard representatives from two Swedish public agencies - Kristina Lagestrand Sjölin, from the National Food Agency, and Anna Hallenbom, from the Agency for Economic and Regional Growth - give their respective views on why automatic translation is useful and necessary. During the break between these two presentations, Kristina Lagestrand Sjölin took time to show a demo of the eTranslation web interface to interested participants.

Following this, the Norwegian National anchor point, Jon Arild Olsen, gave the audience a comparison case by presenting the situation for collecting language data in Norway. Having some similarities but many differences, as expected, the Norwegian experience proved interesting.

Before lunch, a panel then discussed the possibilities, advantages, challenges and problems with automatic translation and eTranslation, as well as what can be the clearest reasons to why our public agencies should find it useful.

After lunch, Simon Dahlberg and Rickard Domeij (the main organisers of this workshop and the authors of this report) from the Language Council explained what they have done so far for the ELRC and for open language data in Sweden.

Following them, Monika Jansson from the National Archives presented what they have been doing to enhance open data in Sweden. She explained what their governmental mission consists in and what they have been doing to fulfil it.

After her, a representative from the translation business trade association of Språkföretagen, Anne-Marie Colliander Lind, held a presentation on how they and translation companies can relate to automatic translation and language data collection. This perspective was very important and gave the workshop a good mixture of opinions and perspectives.

2 Workshop Agenda

09:30 – 10:00	Registration
10:00 – 10:10	Welcome! <i>Rickard Domeij and Simon Dahlberg (Swedish Language Council)</i>
10:10 – 10:20	Introduction <i>Rickard Domeij (Swedish Language Council)</i>
Session 1. Connecting a multilingual Europe: European context and local needs	
10:20 – 10:40	What is the ELRC, European Language Resource Coordination? <i>Andrejs Vasiljevs (Tilde, ELRC representative)</i>
10:40 – 11:00	Usage of eTranslation at the Swedish National Food Agency <i>Kristina Lagestrand Sjölin (Swedish National Food Agency)</i>
11:00 – 11:20	Coffee Break
11:20 – 11:40	Need for automatic translation from the perspective of the Swedish Agency for Economical and Regional Growth <i>Anna Hallenbom (Swedish Agency for Economical and Regional Growth)</i>
11.40 – 12:00	Working with collecting translation data in Norway <i>Jon Arild Olsen (Norwegian National Library)</i>
12:00 – 12:30	Panel discussion eTranslation in Sweden – possibilities and challenges Recent speakers
12:30 – 13:30	Lunch
Session 2. Engage: hands-on data	
13:30 – 14:00	Our work with ELRC and open data <i>Rickard Domeij and Simon Dahlberg (Swedish Language Council)</i>
14:00 – 14:20	What the Swedish State has been doing to enhance open data <i>Monika Jansson (Swedish National Archives)</i>
14:20 – 14:40	The role of translation companies in disseminating language data <i>Anne-Marie Colliander Lind (Språkföretagen Trade association)</i>
14:40 – 15:00	Coffee Break
15:00 – 15:20	How to collaborate in collecting language data <i>Rickard Domeij (Swedish Language Council)</i>
15:20 – 15:55	Questions & Answers
15:55 – 16:00	Conclusions

3 Summary of Content of Sessions

3.1 Welcome! 10:00-10:10

A general presentation was given by Simon Dahlberg and Rickard Domeij, welcoming all the participants and giving basic information on the agenda and the venue.

3.2 Introduction 10:10-10:20

An introduction to ELRC and eTranslation. Rickard Domeij explained shortly what eTranslation is, how it can be used as an integrated tool in a digital service, giving the example of Tenders Electronic Daily and comparing with a Dutch initiative that saw automated translation reach high quality for the Dutch language. Following this, he also talked of what is needed from Sweden to accompany this effort and to enable eTranslation to reach high quality for the Swedish language in different domains, affirming that in the future automated translation will actually be necessary for digital services. He also presented a few things that have happened since the first workshop in March 2016, most importantly the introduction of neural engines into machine translation and how that has increased its quality but also made it more difficult to discover errors.

3.3 What is the ELRC? 10:20-10:40

Andrejs Vasiljevs went deeper into the ELRC and eTranslation, coming from the picture of Connecting Europe Facility. His presentation explained why these initiatives are needed (for example, why we cannot just rely on Google Translate instead for something so important as our languages), what the ELRC goals are and why fighting the language barriers is important for a European Single Digital market. It showed that eTranslation is one of the building blocks of the Connecting Europe Facility program, and thus in the European Commission's intention of creating cross-border digital services that automate and streamline many aspects of public administration activities. Andrejs Vasiljevs also informed of how the ELRC works, what organisations are involved, what makes data so valuable and how one can get involved and be assisted in the data contribution process.

3.4 Usage of eTranslation at the Swedish National Food Agency 10:40-11:00

Kristina Lagestrand Sjölin from the Swedish National Food Agency gave a very good presentation on how she and her organisation uses and benefits from eTranslation. She explained what their mission is, how they are connected to the rest of the Union and what their multilingual needs are. She described a few use cases from her own work to show why eTranslation is useful for her. She highlighted the importance of using your judgement and knowing how to use it, as one cannot simply suppose that whatever results is of sufficient quality.

This presentation made the workshop take an excellent step from the big visionary picture into the context of how eTranslation can benefit Swedish public agencies in practice and how it can directly help workshop participants in their daily work.

3.5 Need for automatic translation at the Swedish Agency for Economic and Regional Growth 11:20-11:40

Anna Hallenbom spoke of the need for automatic translation at her agency, the Swedish Agency for Economical and Regional Growth. She explained that it comes to them mainly for two reasons. The first reason is a number of directives and projects from the European Commission, such as SDG, (Single Digital Gateway), TOOP (the once only principle) and EIDAS, which require their information to be available in other EU languages, to adapt it to citizens from the rest of union. The second reason is a

web portal called *Verksam* that is the contact point for entrepreneurs who need to find information on how to start and run a business in Sweden (in order to avoid frustration in having to contact several public agencies to find the information they need). As this information is also supposed to be adapted to other EU citizens, the webpage has a big need for translation services, and automated translation can cater for a part of this need.

3.6 Working with collecting translation data in Norway 11:40-12:00

Jon Arild Olsen, from the Norwegian National Library, who is one of Norway's National Anchor Points, described the Norwegian situation concerning collection of language data. Of course, it differs a lot from the Swedish situation, above all because of their national language policy, Norway having two official Norwegian written languages, *Bokmål* and *Nynorsk* and not being a part of the EU. His presentation provided very good input to give perspective on the Swedish context. The one piece of information that was perhaps the most interesting and directly useful in the work with ERLC, was the fact that one of the bigger translation companies in the Nordics, Amesto – which is now part of Semantix – had been asked to share all the translations memories that had been created from deliveries to Norwegian public administrations. To everyone's surprise, they responded positively with a large delivery. We are hoping this support will see a replication in Sweden.

He also described what has been done for the position of the Norwegian language in eTranslation so far. Naturally, English is the one language that most translations are made to from *Bokmål*, meaning that, currently, translation between *Bokmål* and any other language than English needs to go through English, giving it a pivot language position. Unfortunately, there is very little translation data for *Nynorsk*, leaving it in need of *Bokmål* as a pivot language to reach English.

3.7 Panel discussion: eTranslation in Sweden – possibilities and challenges 12:00-12:30

A panel discussion followed, including the recent speakers; Rickard Domeij, Andrejs Vasiljevs, Kristina Lagestrand Sjölin, Anna Hallenbom and Jon Arild Olsen. Below follow the questions of the discussions, together with the answers from the panellists.

Question 1: What possibilities and advantages do you see with automated translation and eTranslation?

- It can make publishing information faster, as one will no longer have to select and request for human translation.
- It can improve web pages.
- It gives the possibility of reaching more people, providing information in more languages.
- eTranslation gives every organisation the possibility of affecting the quality of the translation by contributing with their language data.
- There is an opportunity of using the Commission's initiative and funding instrument to keep our local languages, not to be obliged to change to English.

Question 2: What challenges and problems do you see with automatic translation and eTranslation?

- The challenge of adapting to the type of language used by certain organisations. For example, some organisations want a language that is less formal and thus better adapted to "ordinary" citizens.
- The challenge of marketing and spreading the conscience of the existence of eTranslation and its benefits.
- The challenge of filling the data gaps.

- The need for working together, in a coordinated way, to gather and process data

Question 3: What are the clearest reasons for our public agencies to share their language data with ELRC and to make benefit of using eTranslation?

- There are no disadvantages with sharing data.
- Quality of eTranslation increases.
- Terminology correctness becomes up-to-date and improves.
- It is a way of releasing open data, to let the public see what they can make of it.
- It is actually a democratic duty to share your data.
- The result of processing the data that is done for eTranslation is also useful for human translators.

3.8 Our work with ELRC and open data 13:30-14:00

This presentation was held by us, Simon Dahlberg and Rickard Domeij, from the Swedish Language Council. Rickard Domeij explained the progress that we have seen in Sweden in terms of language data collection at a national and broader level. Simon Dahlberg spoke of what we have done in the ELRC project, showing what has been shared with its repository until now and where it comes from. He also mentioned the license issue, the Creative Commons system and the value of terminology in automated translation.

3.9 What the Swedish state has been doing to enhance open data 14:00-14:20

Monika Jansson from the Swedish National Archives gave a presentation on their mission to enhance open data in Sweden and on what they have been doing to fulfil it, explaining what open data is and showing what the national open data website, opnadata.se, looks like. In their mission, they have created guidelines for publishing open data and stimulating other organisations in Sweden to do so. A part of that work has consisted in collecting PSI lists (lists of open data resources) from other public agencies. The Swedish National Archives' mission terminates this year, to be taken over by a new public agency.

3.10 The role of translation companies in disseminating language data 14:20-14:40

Anna-Marie Colliander Lind from the trade association of *Språkföretagen* gave very valuable input in the discussion on how to collaborate in language data collection, presenting what perspective the private translation firms have in this matter.

Her organisation represents only a few translation companies in Sweden but still more than half the market of our country, regarding the amount of translation and interpretation services provided every year.

Her presentation raised many important questions, such as what services we will need from the translation companies if more and more translation is done by machines. We cannot take for granted that the result is perfect or even good enough; we will need a professional quality check and correction. This situation, in turn, will mean that the procurement processes for translation services cannot be designed the same way they are today, as the requested services and competences will not be the same.

This also means that there needs to be a stronger conscience and competence from the side of the public agency employees ordering the translation services. This includes, among other things,

knowledge of what they are actually paying for and an understanding of the value of qualitative terminology. If the translation companies will be expected to share their data to improve automated translation, they have to be included in the discussion about what implications it will have for translators and for the business as a whole, in order to adapt the procurement routines and the translation requests accordingly.

3.11 How to collaborate in collecting language data 15:00-15:20

With the intention of starting a discussion on language data collection and the need for automated translation and open data, Rickard Domeij gave a contextual framework on the possibilities and intentions that we (the Language Council) and the Swedish State have for the development of language technology in Sweden. He talked of what multilingual and linguistic needs Swedish public agencies have, what visions and governmental missions we have, as well as how one can contribute with data to the ELRC, clarifying that one does not necessarily go through us. This way, it came even clearer that ELRC is a part of the effort for making Swedish and the other languages in Sweden follow the language technology development through national data collection and coordination of language services. This introduction gave path to questions and reflections from the audience, engaging them in the topic and the discussion.

3.12 Questions & Answers

Among other topics, the discussion brought up:

- How automated translation can help professional human translators in their work
- What should be done with data from private companies - some say that their language resources can be shared for eTranslation but should not be published as open data for the broader public, as it has been built by private companies that are already struggling
- How to ensure the quality of the data
- How to keep track of the terminology

4 Synthesis of Workshop Discussions

The discussions brought up, among other things:

- How automated translation will affect human translators and the translation market as a whole. The answer given was that automated translation is actually increasing the need for translation and thus expanding the business, and that there is no intention whatsoever of completely replacing human translators with automated translation.
- How to ensure data quality. One answer given was that the public agency employees that request translation services need to have competence and understanding of the business of what their organisation is paying for. For instance, it is important to understand that a result is not necessarily the final result, there needs to be a follow-up quality check and feedback to reach it. This way, the market will produce better translation data.
- Whether the ELRC wants to collect corpora that have not been created with the purpose of building a large national database, for example literature and poetry. The answer is that literature and poetry is not convenient for training eTranslation, as the focus is on translating texts from public administrations.
- The value of putting in place a national coordination for language data collection, not to let private companies get hold of the big sets of data before the public institutions do it. By making it national, the data becomes open and public, which is very important for the technological development of our language(s). For this purpose, it is important to create a conscience among public agencies that language data collection has great value for society.
- How to handle the complexity of translation memories. There are many different formats and many different alignment systems. In top of this, it is not something static, as it needs to be updated with consistency. The answer given was that the TMX format generally works for most translation memory files.

4.1 ELRC and Open language Data in Sweden

The Swedish state has the ambition of making Sweden the best country in the world (or among the best) regarding publishing and managing open data. In charge of the open data mission is the Swedish National Archives. In their mission, they have created guidelines for publishing open data and they have stimulated open data publishing among public agencies and other organisations. A part of their work has consisted in collecting PSI lists (lists of open data resources) from other public agencies. There now exists a national open data portal, www.opnadata.se, where public agencies and other organisations are welcome to publish metadata according to DCAT-AP together with links to their open data, meaning that the data is not duplicated, the portal only points the websites where each respective agency stores its data. All public agencies are also recommended to create a webpage ending in “/psidata” (such as <http://www.sprakochfolkminnen.se/psidata>), to make it easier for the web user to find their open data.

This mission started in July 2016 and will terminate in August this year, to be taken over by a new authority, “Myndigheten för digital förvaltning” (“The Digitalisation Authority”).

The idea is to let the Swedish data shared with ELRC to become open data, whenever the data holder has permitted it. For this purpose, usage of Creative Commons is recommended to define how a certain resource can be reused and disseminated. Generally, the *Public Domain Mark* is recommended for Swedish public data.

4.2 Success stories and lessons learnt

There is now awareness among several Swedish public agencies that there is a need for automatic translation and that eTranslation is an option. Since the last workshop in 2016, we (The Swedish

Language Council) have delivered an initial set of language resources, and we are confident that there will be more soon.

Since 2016, we have also established contact with representatives from a small number of Swedish public agencies, trying to create a best practice model for language data collection with them. The idea is for this model to be applied on more and more agencies in the future, once it is better defined. The agencies with whom we have contact today come from different sectors with different needs of automated translation, which gives us a good idea of the context and the different challenges. Two of these agencies were represented by speakers at the workshop, whereas all the others were represented in the audience.

From this cooperation, it seems that a good strategy to make public agencies interested in the ELRC is to stress the benefit of combining ELRC engagement with other efforts that are closer (maybe even internal) to the agency itself, to satisfy needs that the employees see in their daily work. Examples of such efforts can be managing internal terminology, language policy guidelines and routines for requesting translation. In order to accomplish this, face-to-face conversations seem to be necessary to get an understanding of each public agency's situation and to create a stronger confidence for eTranslation and belief in the ELRC project.

From this workshop, we have initiated contacts with several public agencies that were not involved before. The feedback from the engagement form questions (particularly those answered by e-mail) seems to promise good results for the follow-up. The Swedish public agencies that now are involved or likely to be involved shortly in the ELRC engagement are:

- The National Food Agency (Livsmedelsverket)
- The Agency for Regional and Economic Growth (Tillväxtverket)
- The Company Registration Office (Bolagverket)
- The Migration Board (Migrationsverket)
- The Ministry of Foreign Affairs (Utrikesdepartementet)
- The Police Authority (Polismyndigheten)
- The National Archives (Riksarkivet)
- The Public Health Agency (Folkhälsomyndigheten)
- The National Agency for Education (Skolverket)
- The National Board of Health and Welfare (Socialstyrelsen)
- The Swedish Parliament (Riksdagsförvaltningen)
- The Swedish NAO (Riksrevisionen)