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

# Deliverable D3.2.1
# Task 3

# ELRC Workshop Report for Sweden

Author(s): The Institute for Language and Folklore

Dissemination Level: Public

Version No.: <V1.0>

Date: 2021-01-18

# Contents

# 1 Executive Summary

The 12th of November 2020, the Institute for Language and Folklore organised the third Swedish European Language Resource Coordination (ELRC) workshop.

The theme of the workshop was machine translation. In particular, machine translation within the public sector, and the importance of language data for building machine translation systems. Representatives from four different government agencies shared their experiences of using machine translation systems (Swedish Food Agency and Swedish Agency for Economic and Regional Growth) or of building their own in-house machine translation systems (Swedish Migration Agency and the Swedish Tax Agency). There was also a session on guidelines for data sharing within the public sector, as well as a session on the CEF eTranslation Platform.

The workshop also included an extra session on terminology in the public sector. The topics of this session were Sweden's National Term Bank (Rikstermbanken), the Federated eTranslation TermBank Network project (2019-EU-IA-0049), standardised terminology development methods and how language technology can support terminology development.

Due to the Covid-19 pandemic, the workshop was conducted as a virtual event. Close to hundred participants had registered, and we estimate that around **70** persons participated. The main language of the workshop was Swedish, with one of the presentations carried out in English. While a large majority of the participants were based in Sweden and also master Swedish, a live interpretation into English and the virtual format made it possible for participants throughout Europe to attend.

## 2 Workshop Agenda

| | |
|---|---|
| 09:30 – 09:45 | **Welcome, practical info** |
| 09:45 – 10:00 | **Introduction. About ELRC and about resource collection and language technology at the Language Council of Sweden.**<br><br>**Rickard Domeij,** Isof, Språkrådet (Presentation) |
| 10:00 – 10:20 | **The CEF AT Platform – Presentation/Demo Presentation**<br><br>**Khalid Choukri**, ELRC (Presentation) |
| 10:20 – 10:30 | Coffee Break |
| 10:30-11:00 | **eTranslation for the public sector in Sweden**<br><br>**Kristina Lagestrand Sjölin**, Livsmedelsverket (Presentation)<br><br>**Veronica Wiman Nilsson**, språkkonsult och webbredaktör, Tillväxtverket (Presentation) |
| 11:00-11:30 | **Language technology at the public sector in Sweden**<br><br>**Torbjörn Ekholm** och **Anna Sågvall Hein** (Presentation)<br><br>**Andreas Voxberg** och **Camilla Lindholm** (Presentation) |
| 11:30-11:40 | Short Break |
| 11:40-12:20 | **Language data sharing in the public sector: guidelines, practices and challenges**<br><br>**Catharina Ekdahl**, Patent- och registreringsverket (PRV) och **Björn Hagström**, Myndigheten för digital förvaltning (Digg) (Presentation) |
| 12:20-12:30 | **Conclusions** |
| 12:30-13:30 | Lunch |
| 13:30-14:30 | **Extra session on terminology**<br><br>**Karin Webjörn**, Isof, Språkrådet (Presentation)<br><br>**Marie van Dorrestein**, Svenska institutet för standarder (SIS) (Presentation)<br><br>**Magnus Merkel**, Fodina Language Technology (Presentation) |

# 3 Summary of Content of Sessions

## 3.1 Welcome, practical info

Presentation of the local organisers from the Institute for Language and Folklore (Rickard Domeij, Sara Steinholtz Sparby and Maria Skeppstedt) and of the national anchor points for ELRC in Sweden (Arne Jönsson from Linköping University and Rickard Domeij).

## 3.2 Introduction. About ELRC and about resource collection and language technology at the Language Council of Sweden

Rickard Domeij, from the Institute for Language and Folklore introduced ELRC, as well some of the resources provided by ELRC, e.g., the eTranslation services, with practical applications, and the ELRC-SHARE repository.

An overview of the work carried out at the Language Council of Sweden at the Institute for Language and Folklore on collecting language resources for ELRC and on informing public agencies on the importance of sharing language resources for machine translation and for other types of language technology-enabled applications.

## 3.3 The CEF AT Platform – Presentation/Demo Presentation

Khalid Choukri from ELDA presented eTranslation in the CEF AT Platform. A demo of how to use the eTranslation web interface was provided, e.g., how to translate raw text, how to refine the translation by specifying text genre and how to upload pdf files for achieving translated text with the original formatting retained. The eTranslation API was also introduced, and examples of how it can be integrated in web pages for providing multi-lingual content were given.

Finally, the participants – if belonging to a group eligible to freely use the eTranslation web interface – where encouraged to try the service. Information on how to register for the service, as well as the URL to the eTranslation web page, was provided.

The discussion following the presentation centred on the possibility to use functionality from the CEF AT Platform internally within an organisation for texts too sensitive to leave the organisation. The platform is not meant for this usage scenario. However, the CEF AT platform is built on a number of open-source language technology tools, that could be used within organisations for creating in-house machine translation systems. For such cases, the parallel corpora provided by the ELRC-SHARE repository could form useful resources for training in-house machine translation models.

The question as to whether the source code for the CEF AT platform is open source was also brought up by one of the participants. It is, however, not an open-source platform.

## 3.4 eTranslation for the public sector in Sweden

Kristina Lagestrand Sjölin – who is working within the fields of export and EU-coordination at the Swedish Food Agency – followed up the previous session by presenting eTranslation from a user perspective. She gave examples of how she uses eTranslation in her everyday work, for instance to make content produced in Swedish available to an international project group, and stressed its time saving ability. Compared to Google translate, the eTranslation service performs better for domain-specific vocabulary, and she uses it to translate individual words, paragraphs and entire documents. She typically performs a post-editing of the text produced by the eTranslation service, to correct translation errors.

Opinions on the eTranslation tool from co-workers at the Swedish Food Agency were also presented. They also stressed the usefulness of the tool, for instance for translating content relating to the EU Rapid Alert System for Food and Feed from different EU languages and for translating letters relating to export issues.

As a frequent user of the eTranslation web page, Kristina Lagestrand Sjölin also presented suggestions for improvements. These included (i) a keyboard short-cut for the button that triggers a text being translated, (ii) a button that reverses the source and target languages selected, and (iii) a possibility to provide suggestions for improvements, e.g., regarding translation errors and regarding incorrectly translated (or outdated use of) terms.

Veronica Wiman Nilsson from Swedish Agency for Economic and Regional Growth then presented an evaluation of whether it would be possible to rely on machine translation for translating the Swedish text on the website "verksamt.se" into English (i.e., a web site that collects information on government services to companies). Two different systems for machine translation were evaluated, eTranslation and Google translate.

A similar evaluation was performed three years ago, and the tools were deemed to have improved greatly since then. However, none of the two systems was considered to produce translations of a quality high enough to be used as-is on a public website, and human translators will therefore also in the future be employed for the task of translating the website into English. Apart from general translation errors, the following two more specific problems were found in the machine-translated texts: (i) the English terms chosen often differed from the terms recommended by the government agencies that produce the website, (ii) the Swedish texts, which had been written in plain language to optimise for intelligibility, were translated into an English text by eTranslation that used a more formal and difficult vocabulary than the source text (and that was used in the text produced by Google translate).

## 3.5 Language technology at the public sector in Sweden

Anna Sågvall Hein, professor from Uppsala Universitet and also representing the language technology company Convertus, and Torbjörn Ekholm from the Swedish Migration Agency presented their proof-of-concept systems for machine translation. Their two different proof-of-concept systems targeted different two types of texts, (i) texts for public webpages and (ii) appeals on decisions made by the agency. The first system used output from Google's machine translation (from Swedish to English and Arabic) as a base, and improved that output by using agency specific terminology lists and translation memories. Appeals on the decisions, in contrast, contain sensitive content and the texts are therefore not allowed to leave the agency. A fully in-house system was therefore built for translating these texts from Arabic to Swedish. Due to lack of enough data for training and machine learning model, the translation was built on a rule-based system using transformation rules, a lexicon and agency specific terminology. The proof-of-concepts also included a user interface for post-editing of the machine translated texts. The presenters stressed the importance of sharing and collecting language data to enable machine translation models, as well as the importance of controlled terminologies for achieving high-quality machine translation.

Camilla Lindholm and Andreas Voxberg from the Swedish Tax Agency then presented the language technology work that is carried out at their agency. This includes, for instance, a chat bot, automatic classification of incoming e-mails and machine translation. The machine translation system is built as a fully in-house system, to make sure that no data (including sensitive data) leaves the Swedish Tax Agency when the system is used. It is available through a user interface and through an API. The system is built on a neural networks model that has been trained on, for example, translations of EU documents. Tax-specific texts will also be used in the future for training the system. The machine

translation system is used for translating texts from Swedish to English for collaboration on tax control within EU, and for translating incoming cases, for which the administrator needs to quickly understand what is relevant. It is not meant for texts that are to be used externally, at least not without a post-editing of the automatic translation. The system currently manages around 15 languages, with a focus on official EU languages. However, more languages will need to be added to the service, to be able to handle texts from incoming cases written in other languages.

The discussion following the two presentations focused on resources needed to train in-house systems built on machine learning, i.e., language data to train models on and pre-trained models that can be refined within the organization. The two agencies agreed on that texts that are specific to the agency and its domain will always be required for training their models, but that common language resources are also very useful. For instance, as exemplified by the approach of the tax agency, which currently are using translations of EU documents for training their models. It might also be useful with such common resources written within other text genres than formal EU-documents.

## 3.6 Language data sharing in the public sector: guidelines, practices and challenges

Björn Hagström from Agency for Digital Government and Catharina Ekdahl from the Swedish Intellectual Property Office (PRV) presented their work on creating formal recommendations for open licenses and intellectual properties for public agencies, with the aim of simplifying the agencies work in making data accessible for innovation and artificial intelligence.

The Agency for Digital Government supports public agencies in their work in making data accessible for re-use, in the form of developing principles and guidelines for licensing, dissemination, and sharing data. The agency bases their work on “Seven Principles of Accessibility”: Open data by default, Conduct a systematic work on information security, Make sure the shared data is up-to-date and user-centred, Make the data easy to process, Use a licensing that promotes wide usage, Document and describe the data, Encourage usage and dialogue.

The formal recommendations that they develop should be possible to apply on all types of data and information that are developed within the agency, and the recommendations are based on the Creative Commons licenses. The following is recommended: (i) Data created at the agency that is not protected by copyright should have PDM or CC0 licenses, (ii) Databases created at the agency should have a CC0 license, and (iii) Data created at the agency that is protected by copyright should have a CC-BY 4.0 license.

The discussion following the presentation was about how these formal recommendations can be applied on texts and terminologies. Decisions, protocols, official texts and webpages from public agencies are normally not protected by copyright, but there might be exceptions, such as photos included in these publications. These texts should thereby typically have a PDM or CC0 license. No definite answer was given on copyright for terminologies, but it might be the case that they should be looked upon as databases, with a possible exception for longer, descriptive text segments. There was also a number of requests formulated to the authors of the recommendations, for instance that more concrete examples of how to license different types of texts could be added, and that a translation of the recommendations into English would be helpful for similar efforts within EU.

## 3.7 Conclusion

Rickard Domeij summarised the day and presented the main findings of the workshop (see section 4 below).

## 3.8 Extra session on terminology

After the lunch break, an extra session on terminology within the public sector was offered. Most of the participants from the morning stayed on also for this session, and there were also some new participants. Rickard Domeij started the session by giving an introduction to the topic and the other three presenters. He also mentioned Rikstermbanken (Sweden's national term bank) and introduced its connection to the Federated eTranslation TermBank Network Action for public organisations and institutions in EU Member States. The Federated eTranslation TermBank aims at developing an infrastructure that makes it possible to automatically deliver terms from local repositories – e.g., from Rikstermbanken – to the eTranslation TermBank and to the ELRC-SHARE repository.

Karin Webjörn from the Institute for Language and Folklore then presented Rikstermbanken and the responsibilities of the Institute for Language and Folklore when it comes to supporting other public agencies in the work of terminology development. Paragraph 12 of the Language Act (2009:600) in the Swedish Code of Statutes is concerned with Swedish terminology. The paragraph states that: "Government agencies have a special responsibility for ensuring that Swedish terminology in their various areas of expertise is accessible, and that it is used and developed." The Institute for Language and Folklore is commissioned by the Swedish Government to provide other public agencies with support for implementing the Language Act. This includes the task of maintaining Rikstermbanken, which is a national term bank with 130,000 term entries from both the public and private sectors. Rikstermbanken was originally developed by "Terminologicentrum TNC (The Swedish Centre for Terminology, TNC)", which in 2006 was commissioned by the Swedish government to develop a national term bank. The responsibility of maintaining the terminological content as well as the technical product has now been handed over to the Institute for Language and Folklore. Rikstermbanken is an important tool for language and terminology work, for instance for (i) reusing existing terminological content, (ii) developing new terminology within different subject areas, (iii) making results of terminology work available.

Marie van Dorrestein from the Swedish Institute for Standards thereafter gave a presentation on terminology within the public sector and standardized methods for terminology development. The Swedish Institute for Standards is commissioned by the Swedish Government to conduct work on standardisation and to represent Sweden in CEN and ISO. The institute is responsible for a number of publications on certified methods for how to develop and standardise terminology. Benefits of standardised terminology include, (i) it facilitates comparisons within and between areas of expertise, (ii) it makes it easier to maintain and further develop the terminology, (iii) it facilitates communication between employees, between public agencies and citizens, and between public agencies and providers of language services. The Institute for Language and Folklore and The Swedish Institute for Standards are both representatives for Sweden in the Federated eTranslation TermBank Network Action. The role of The Swedish Institute for Standards in the project was mentioned, and the aims of project were further explained.

Magnus Merkel from the language technology company Fodina was the last presenter of this session, presenting the company's language technology products that support terminology development. He also had a list of advantages with standardised terminology, including (i) fewer misunderstandings by readers and translators, (ii) reduced translation costs and increased translation quality, and (iii) improved findability. The company's product for terminology development (i) automatically identifies and extracts term candidates in existing texts, (ii) clusters identified terms into synonyms and spelling variants, (iii) helps the user to make fact-based decisions about what terms to use, and (iv) publishes standardised terms to term bases and writer support systems.

The discussions reiterated one of the key points of the morning with regards to terminology, i.e., the importance of incorporating information from high-quality terminology resources for making the output of machine translation systems usable for applications within the public sector. The need for cooperation between public agencies, and to increase the knowledge of terminological issues within the agencies was also discussed, as well as the need for providing normative and harmonised terminologies that can also be made publicly available, not only through a search portal, but also as open data.

# 4 Synthesis of Workshop Discussions

We noted that a lot has changed since the ELRC workshops started. It used to be the case that representatives from public agencies were informed about usefulness of language technology and the importance of sharing data. Today, representatives from public agencies instead present their language technology solutions and inform on the importance of data and on problems caused by the lack of enough language data. There is also ongoing work on creating formal recommendations for open licenses, with the explicit aim of simplifying the public agencies work in making data accessible for innovation and artificial intelligence. It was also noted that eTranslation has been improved during this time, and examples were given of how it is used as a practical tool in a multi-lingual environment.

The main findings of the workshop are the following:

It was shown how eTranslation can be a practical tool, which saves a lot of time.

The presenters mentioned that the output of eTranslation, as well as the output from in-house solutions for machine translation, was often post-edited/corrected. We therefore would like to suggest that there seems to be a need for machine translation that functions as a support for a human translator, not only a machine translation that produces content without human intervention. There is, for instance, to the best of our knowledge, no interface for performing such a post-editing within the eTranslation web interface. Also, the results of the post-editing/correction of the texts could form valuable training data for improving the machine translation models. This is also the case for other types of user input on the functionality of eTranslation. As there is no interface for performing the post-editing within the eTranslation web interface, the user's post-editing of the texts is not currently collected to be used as training data to the model.

A problem with the eTranslation, shown in an evaluation of the tool, was that Swedish text that had specifically been written in plain language (i.e., language that is meant to be easy to understand) was translated into a more formal text in English, in which difficult words were used. We guess that this might be a result of the model being trained on formal EU documents. We therefore suggest that among the settings for text genre in the eTranslation web interface, an additional setting might be needed. That is, the setting that plain language is to be used.

Two examples of why it is difficult to share language data from public agencies were given. First, the data might contain sensitive information, which is the case for appeals on the decisions. Such texts cannot leave the agency. Second, licensing issues might be difficult. The suggested solution to the second problem is to give the agencies support in this matter, by providing formal recommendations on how to license text data of different types.

eTranslation cannot be used at public agencies that need to translate texts that are too sensitive to leave the agency. This highlights the importance of not only offering eTranslation, but to also offer data – e.g., as it is done through the ELRC-SHARE repository – that can be used for training in-house models. Texts specific to each agency and its domain will also be required for training the in-house models, but the common language resources are also very useful.

We believe that it might also be useful with such common corpus resources written in other text genres that occur frequently within public agencies than formal EU-documents. Appeals on decisions was one example of a frequent text genre within one public agency. Perhaps ELRC could survey which types of genres that occur frequently at several agencies, and for which it might be worth gathering texts. Or, in the case of text genres that are too sensitive to share, perhaps synthesised texts need to be created.

It was also given examples of that machine translation is needed for other languages than the official EU languages, e.g., in this case languages common among immigrants to Sweden.

A point repeated in many of the workshop sessions was the importance of incorporating information from high-quality terminology resources for making the output of machine translation systems usable. A presenter of one of the in-house machine translation systems stressed this point. The evaluation of eTranslation showed that one of the problems with the tool was that vocabulary used in the translation did not adhere to the terminology recommended within the agency. This point was also discussed in the terminology session, and the importance of high-quality terminologies that can function as normative (rather than descriptive) resources—both for humans and for machine translation systems—was mentioned.

During the terminology session, it was mentioned that existing and available terminologies form important resources also for developing new terminologies and for harmonising terminology between different organisations. This, together with the previous point, shows the importance of developing and sharing high-quality terminology resources.

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

Due to the virtual format of the workshop, it was difficult to get that the participants to fill out the two surveys. Only one person filled out the country survey. We will therefore here mainly discuss what was mentioned during the workshop with regards to data sharing. The one person that filled out the country survey works at an organisation that uses or intends to use eTranslation, and no other language technologies or services. The person states that there are no language resources or translations within their organisation and that there is no data management plan. The main difficulties that may prevent the sharing of language data are legal issues and inadequate practices for the management of language data.

Since the last country profile, there is generally more activity going on within public agencies when it comes to working with and planning for open data in general. There is also more work on in-house language technology solutions, as was exemplified by the Swedish Migration Agency and the Swedish Tax Agency, which leads to an increased understanding for the importance of language data. There are also many organisations that use eTranslation (see list below), which might also increase the understanding for the need for data for improving it, and thereby the need for more data.

Two challenges were discussed, (i) that knowledge with regards to copyright and licensing of language data created at public agencies is often lacking, and (ii) that much text data for which machine translation is needed is sensitive data (appeals on decisions was given as an example), which is difficult to share even within the organisation, and thereby impossible to share outside of the organisation.

The formal recommendations for sharing and licensing of data created at public agencies, which are created by the Agency for Digital Government and the Swedish Intellectual Property Office, might be a resource for addressing the first of these two challenges. These recommendations currently lack (i) concrete examples of common types of text data produced at agencies and their recommended licensing, (ii) information on licensing of terminology resources. This was discussed during the workshop, and we hope that the recommendations will be updated with that information. In the future, we plan to inform about these guidelines with formal recommendations when contacting agencies regarding text data sharing.

For the second challenge related to sensitive data, we currently do not have a clear plan for how to approach it. It might be good to focus on non-sensitive data for a start, since despite sharing non-sensitive texts is a much easier task, there is still large amounts of non-sensitive data that is not being shared. However, in the long run, machine translation systems will also need texts that belong to text genres that typically contain sensitive content, in order to produce high-quality translations for these texts. One possibility is to anonymise the sensitive data before sharing it, but that is likely to be very difficult from a legal point of view. The Swedish Migration Agency, which had problems accessing enough sensitive data within the organisation for training a machine learning model, constructed a small corpus of made-up data that was similar to the real data for evaluating their rule-based machine learning system. We have come across similar solutions within other domains with sensitive data, for instance the domain of health record text. Although each corpus of constructed data is likely to be very small, due to the costs of writing fictional texts that are similar to real ones, the collection or construction of many such small corpora of made-up texts might be a valuable contribution to an ELRC-SHARE repository, for which it is difficult to collect texts belonging to sensitive text genres.

The challenge of separating sensitive text data from other types of text data, which has been mentioned in previous country profiles, was not explicitly discussed during the workshop. However,

the general increased focus on open data and guidelines for how to license it, might be a first step towards creating procedures for how to handle this separation.

Given the difficulty of receiving input from the participants regarding resources – and given that the workshop has shown the importance of high-quality terminology resources for machine translation – we aim to follow up this seminar by conducting a more structured survey focusing on the existence of publicly sharable terminology resources at public agencies.

We received the following information regarding usage of eTranslation in Sweden from Markus Foti, Head of Sector for Machine Translation, European Commission.

CONNECTED USERS: 109,

PAGES: 75 968,

REQUESTS: 4 439

REGISTRED ORGANISATIONS:

- forsakringskassan.se
- slv.se
- socialstyrelsen.se
- additude.se
- aklagare.se
- alvesta.se
- av.se
- avesta.se
- business-sweden.se
- datainspektionen.se
- deciduous.se
- digg.se
- fk.se
- folkhalsomyndigheten.se
- gov.se
- havochvatten.se
- jordbruksverket.se
- knowit.se
- konsumenteuropa.se
- konsumentverket.se
- lakemedelsverket.se
- migrationsverket.se
- mil.se
- msb.se
- norrbotten.se
- riksbank.se
- riksdagen.se
- riksrevisionen.se
- sala.se
- skatteverket.se
- slv.se
- sprakochfolkminnen.se

- tillvaxtverket.se
- tullverket.se
- u-lift.se
- vanerhamn.se
- vinnab.se