Texts and Terms from Swedish Public Agencies in the SB Sam Language Bank

Maria Skeppstedt, Simon Dahlberg, Gunnar Eriksson, Rickard Domeij
The Language Council of Sweden, the Institute for Language and Folklore, Sweden
firstname.lastname@isof.se

Abstract

We here describe data from the *SB Sam Language Bank*, one of three divisions within the Swedish language technology and research infrastructure *The National Language Bank of Sweden*. The SB Sam Language Bank aims at making data gathered by the Institute for Language and Folklore more available, i.e. folklore and dialect archives, terms and dictionaries, as well as language data produced at other Swedish public agencies. The data gathered from public agencies in SB Sam consists of three main repositories, (i) translation memories, i.e. sentence-aligned texts in different languages that have either been extracted from translation tools or from automatically sentence-aligned parallel texts, (ii) terms gathered from public agencies, and (iii) parallel texts in several languages, that either have been crawled from public agency web sites or that were received directly from the agency.

1 Introduction

The *SB Sam Language Bank* is one of three divisions within the Swedish language technology and research infrastructure *The National Language Bank of Sweden*. The National Language Bank offers a Swedish national e-infrastructure that supports research in different areas for which the use of term banks and larger corpora form central components. For instance, the research areas of language technology, linguistics, social science, translation research and digital humanities.

Apart from the (i) SB Sam division, the National Language Bank of Sweden also consists of (ii) *SB Text*, which focuses on Swedish text corpora and on the development of tools for processing these corpora, (iii) *SB Tal* (SB Speech), which focuses on speech corpora and on the development of tools for speech recognition and speech synthesis.

SB Sam – short for *Language bank, society* in Swedish – is administrated by the Institute for Language and Folklore. SB Sam aims at making language data from the Institute for Language and Folklore, as well as data from other Swedish public agencies, more available. That is, (i) data from the institute's folklore and dialect archives, (ii) terms and dictionaries produced or managed by the institute, and (iii) terms and text corpora collected from other Swedish public agencies.

The folklore and dialect archives of the SB Sam infrastructure contain the results of more than 100 years of documentation of folk culture and dialects in Sweden. The collection encompasses many different genres, for instance life stories, folk poetry, legends and food recipes. Within the SB Sam infrastructure, we develop methods for collecting and digitalising folklore and dialect data, e.g. through the use of web-based questionnaires and crowdsourcing. We also develop methods for making the data accessible to researchers and to the general public, for instance visualisation methods based on maps and on automatic text extraction, e.g. topic modelling (Dagsson and Skott, 2019; Skeppstedt et al., 2020).

Another data type within SB Sam consists of terms and dictionaries, for instance official dictionaries with recommended terminology for translations within specific domains, e.g. health care and social insurance, and the Lexin dictionary series, which is targeted towards learners of Swedish.

The two types of data included in the SB Sam infrastructure that will be presented here are (i) multilingual texts published by, or made available by, Swedish public agencies, and (ii) term lists constructed by Swedish public agencies.

2 Texts from Swedish Public Agencies

SB Sam makes collections of language data from different Swedish organisations available, mainly data from Swedish public agencies. The data is divided into (i) translation memories,[1] (ii) terms,[2] and (iii) parallel texts.[3] Content stemming from SB Sam has also been made available through the European Language Resource Coordination Share (ELRC-SHARE)[4] and through the Eurotermbank.[5]

2.1 Translation memories

There is a Swedish public agency framework agreement for procuring translation services. The agreement dictates that translations produced by external translators belong to the agency that contracted the translation. Translation memories must therefore be transferred to the public agency upon request, without extra charge (ELRC, 2019). Thereby, also translation memories stemming from CAT tools (computer aided translation tools) used by external translation agencies can be claimed, and – if they belong to a data type that does not handle data related to persons or to another data type that might contain sensitive data – be made publicly available.

The translation memory repository of SB Sam consists of translation memories from different Swedish public agencies. The repository consists of (i) translation memories that have been directly exported from CAT tools, with the original translation unit alignment from the tool retained, and (ii) translation memories for which the translation units have been automatically aligned from parallel texts. A translation unit typically corresponds to a sentence. The automatic alignments have been performed by an alignment system constructed by the language technology company Tilde.[6] The metadata information of the repository specifies the source of the translation unit alignment.

The repository currently contains translation memories from the Swedish Crime Victim Compensation and Support Authority (Brottsoffermyndigheten), the Swedish Competition Authority (Konkurrensverket), the Swedish Consumer Agency (Konsumentverket), the Swedish Migration Agency (Migrationsverket), the Swedish National Audit Office (Riksrevisionen), the Swedish Agency for Economic and Regional Growth (Tillväxtverket) and Swedish Council for Higher Education (Universitets- och högskolerådet). For most agencies, the translation memories currently consist of alignments between Swedish and English translations. However, for two of the agencies, we have obtained aligned resources for several languages.

2.2 Terms

The terminological resources included in SB Sam currently consist of terms collected from three Swedish public agencies; the Swedish Agency for Economic and Regional Growth, the Swedish Migration Agency and the Social Insurance Agency (Försäkringskassan). Two of them contain parallel terminological resources for Swedish and English and for the third, the Social Insurance Agency, there is also parallel resources for five other large European languages. There are previous studies of what term translation strategies have been used for translating Swedish public agency terms (Dahlberg and Domeij, 2017; Dahlberg, 2017).

2.3 Parallel texts

The parallel texts repository contains parallel texts, i.e. texts that were originally written in Swedish and that have then been translated into other languages. These texts are not sentence-aligned, in contrast to the texts in the translation memory repository. The texts stem from Swedish public agencies (currently thirteen different agencies), as well as from the Swedish nonprofit organisation the Immigrant Institute (Immigrantinstitutet). Most texts have been downloaded from the website of each respective organisation.

Downloaded pdf files, together with text files containing texts that were automatically extracted from the pdf, are published in the repository for some of the organisations. For others, only automatically extracted text files are published, partly due to the photos not being licensed for redistribution. In some cases, documents in text format were delivered directly from the agency.[7]

The repository currently contains 1,734 texts in 40 languages. The original 286 Swedish texts contain a total of 1,398,642 tokens. Most of these texts have been translated into English, and different subsets of the texts have been translated into subsets of the other 38 languages.

[1] sprakresurser.isof.se/myndighetsdata/Oeversaettningsminnen/

[2] sprakresurser.isof.se/myndighetsdata/termer/

[3] sprakresurser.isof.se/myndighetsdata/texter/

[4] elrc-share.eu

[5] www.eurotermbank.com

[6] www.tilde.com

[7] The texts were either extracted using the Poppler pdftotext poppler.freedesktop.org or download using the text download functionality in the w3m text-based web browser w3m.sourceforge.net.

2.4 ELRC-SHARE, Eurotermbank and the Federated eTranslation TermBank Network Action

Some of the documents in the parallel texts repository have been automatically sentence-aligned – again by the language technology company Tilde – and thereafter published on ELRC-SHARE. Among these documents, there are currently 45,786 translation units that contain texts that are parallel for several large languages spoken in the EU (from six to seventeen languages, depending on resource). There are also 59,395 translation units that contain parallel Swedish-English texts.

SB Sam terms from Swedish public agencies, as well as the Lexin dictionaries, have also been published at the Eurotermbank with the eTranslation TermBank terminology collections (Gornostaja et al., 2018). As a follow-up to this collaboration, we are now taking part in the *Federated eTranslation TermBank Network Action* for public organisations and institutions in EU Member States.[8] The project aims at developing an infrastructure that makes it possible to automatically deliver terms from local repositories to the eTranslation TermBank and to ELRC-SHARE.

The technical infrastructure will consist of two main parts, (i) an Open Terminology Management Toolkit which can be deployed locally to function as a national node, and (ii) a central eTranslation TermBank node with which the national nodes can share their data.

The Open Terminology Management Toolkit will include functionality for terminology management and for terminology search. There will also be functionality for regularly synchronising terminology changes with the central eTranslation TermBank node.

There are countries with an existing national terminology database. E.g. in Sweden, Rikstermbanken[9] (Sweden's National Term Bank) has existed as a national terminology database and terminology portal since 2009. For these countries, the technical infrastructure of the Federated eTranslation TermBank will also offer functionality for instead synchronising between this existing national terminology portal and the central eTranslation TermBank node.

The central eTranslation TermBank node, which gathers locally created terminology through the local nodes, will, in turn, be linked to the ELRC-SHARE repository. The eTranslation TermBank will regularly update terminological resources in the repository, and thereby ensure a currentness of the data that is stored in ELRC-SHARE and that is made available to CEF eTranslation.

3 Concluding remarks and future work

To the best of our knowledge, the SB Sam infrastructure is the first public repository for continuously gathering parallel text data stemming from several Swedish public agencies. We plan to extend the SB Sam Language Bank with more public agency data by continuing our ongoing work on informing agencies on the importance of sharing language data.

We will particularly focus on the task of adding texts written in the Swedish national minority languages to the repository, i.e. texts written in varieties of Meänkieli, Finnish, Sami, Romani Chib and Yiddish (Lag (2009:724) om nationella minoriteter och minoritetsspråk, 2009). As typical for minority languages, the commercial interest in Sweden in creating resources for the Swedish national minority languages has generally been low [p. 114](Domeij et al., 2019). In addition, the publicly funded large national corpora/language technology infrastructures in Sweden have so far mainly focused on the Swedish language. Although there are corpus collection efforts in other Nordic countries for some of the national minority languages (Moshagen et al., 2014; Jauhiainen et al., 2015; Giellatekno, 2020a,b), many of the national minority languages of Sweden are still heavily under-resourced.

Results of the future work outlined here will be continuously reported on the web page of the National Language Bank of Sweden.[10]

Acknowledgements

The National Language Bank of Sweden is funded by the Swedish Research Council (project number 2017-00626). The Federated eTranslation TermBank is co-financed by the Connecting Europe Facility of the European Union.

The contents of this paper are the sole responsibility of the authors and do not necessarily reflect the opinion of the European Union.

[8] www.lr-coordination.eu/node/1022

[9] www.rikstermbanken.se

[10] www.sprakbanken.se

References

Trausti Dagsson and Fredrik Skott. 2019. Digital cultural heritage — a digital folklore archive. https://sweclarin.se/eng/digital-cultural-heritage—–digital-folklore-archive.

Simon Dahlberg. 2017. Tre svenska myndigheters strategier för termöversättning till spanska och franska. Bachelor's thesis, Stockholm University, Department of Linguistics.

Simon Dahlberg and Rickard Domeij. 2017. Översättning av termer i myndighetstexter: En studie om översättning av myndighetstermer i arbetet med nationell språkinfrastruktur på språkrådet. In *Workshop Termplanering och termbruk i svenskan på Svenskans beskrivning 36*.

Rickard Domeij, Ola Karlsson, Sjur Moshagen, and Trond Trosterud. 2019. Enhancing information accessibility and digital literacy for minorities using language technology — the example of Sami and other national minority languages in sweden. In *Perspectives on Indigenous Writing and Literacies*. Brill.

ELRC. 2019. ELRC white paper: Sustainable language data sharing to support language equality in multilingual europe. http://www.lr-coordination.eu/sites/default/files/ELRC_Conference/ELRCWhitePaper.pdf.

Giellatekno. 2020a. Giellatekno, the research group for Saami language technology. http://giellatekno.uit.no/index.eng.html.

Giellatekno. 2020b. Korp (sami). http://gtweb.uit.no/korp/.

Tatjana Gornostaja, Albina Auksoriūtė, Simon Dahlberg, Rickard Domeij, Marie van Dorrestein, Katja Hallberg, Lina Henriksen, Jelena Kallas, Simon Krek, Andis Lagzdiņš, Kelly Lilles, Asta Mitkevičienė, Sussi Olsen, Bolette Sandford Pedersen, Eglė Pesliakaitė, Claus Povlsen, Andraž Repar, Roberts Rozis, Gabriele Sauberer, Ágústa Thorbergsdóttir, Andrejs Vasiļjevs, Artūrs Vasiļevskis, Mari Vaus, and Jolanta Zabarskaitė. 2018. eTranslation TermBank: stimulating the collection of terminological resources for automated translation. In *Proceeding of the XVIII EURALEX International Congress*.

Tommi Jauhiainen, Heidi Jauhiainen, and Krister Linden. 2015. The Finno-Ugric languages and the internet project. In *Proceedings of the First International Workshop on Computational Linguistics for Uralic Languages*.

Lag (2009:724) om nationella minoriteter och minoritetsspråk. 2009. Kulturdepartementet. https://www.riksdagen.se/sv/dokument-lagar/dokument/svensk-forfattningssamling/lag-2009724-om-nationella-minoriteter-och_sfs-2009-724.

Sjur Moshagen, Jack Rueter, Tommi Pirinen, Trond Trosterud, and Francis M. Tyers. 2014. Open-source infrastructures for collaborative work on under-resourced languages. In *CCURL 2014: Collaboration and Computing for Under-Resourced Languages in the Linked Open Data Era*.

Maria Skeppstedt, Rickard Domeij, and Fredrik Skott. 2020. Adapting a topic modelling tool to the task of finding recurring themes in folk legends. In *Proceedings of the Digital Humanities in the Nordic Countries*, pages 388–392. CEUR Workshop Proceedings.