Collecting Language Resources from Public Administrations in the Nordic and Baltic Countries

Andrejs Vasiļjevs, Rihards Kalniņš, Roberts Rozis, Aivars Bērziņš
Tilde
Vienibas gatve 75a, LV1004, Riga, Latvia
{andrejs, rihards.kalnins, roberts.rozis, aivars.berzins}@tilde.com

Abstract

This paper presents Tilde's work on collecting language resources from government institutions and other public administrations in the Nordic and Baltic countries. We introduce the activities and results of the European Language Resources Coordination (ELRC) action in this region, provide a synopsis of ELRC workshops held in all countries of the region, identify potential holders and donors of language data suitable for improving machine translation systems, and describe the language resources collected so far. We also describe several national projects and initiatives on sharing of language data accumulated in the public sector and creation of new language resources from this data. Opportunities and challenges in consolidating language data from the public sector are discussed, and related actions and regulatory initiatives are proposed.

Keywords: language resources, national/international projects, parallel corpora, ELRC, machine translation, CEF eTranslation

1. Introduction

Several large-scale projects and initiatives have been undertaken in this century to collect language resources (LR) and create LR repositories and infrastructures on a pan-European scale.

A family of closely related projects co-funded by EC FP7 and ICT-PSP programmes (T4ME, CESAR, META-NORD, META-NET4U) succeeded by establishing the META-NET network of excellence, comprised of 60 research centres in 34 countries (Rehm&Uszkoreit, 2011), and creating META-SHARE[1], an online repository of language resources distributed and automatically synchronized across 28 hosting nodes across Europe (Piperidis, 2012). The META-NORD project contributed to these developments by coordinating language resource collection and setting up META-SHARE repositories in all 8 Nordic and Baltic countries (Vasiļjevs et al., 2011).

Nordic and Baltic countries are part of the CLARIN infrastructure[2] (Wittenburg et al., 2010), which supports the sharing, use, and sustainability of language data and tools for research in the humanities and social sciences (Varadi et al., 2008). National consortiums are established and funded by the Nordic governments, such as CLARINO[3] in Norway, FIN-CLARIN[4] in Finland and CLARIN-DK[5] in Denmark.

The breath and usability of web data is demonstrated by the OPUS corpora[6] (Tiedemann, 2012), in which parallel data is extracted and aligned from numerous web sources, covering from formal (e.g., legal texts) and highly technical (e.g., user interface strings, medicine descriptions) to informal (e.g., movie subtitles, conversational phrases) language.

Consolidation activities have also targeted specific types of language resources. For instance, the EuroTermBank project supported by the EC eContent programme has consolidated heterogeneous multilingual terminology resources in a distributed online termbank (Vasiļjevs et al., 2008).

Still, most of the LRs in these repositories are created by research institutions, or the translation departments of large international organisations (such as the European Commission or United Nations), and have resulted from research projects. At the same time, thousands of translation units are created every day by companies and public sector institutions. Acquisition of commercial data can be very costly (e.g., monetary of reciprocal data exchange at TAUS Data Repository[7]).

Data in the public sector, at least in theory, should be much more available thanks to the Public Sector Information Directive adopted in EU and EEA countries (European Parliament, 2003). The directive basically stipulates that data created by taxpayer's money should be made freely available for any use, including commercial, with only limited exceptions for privacy or confidentiality protection.

In this paper we present our work on collecting language resources from government institutions and other public administrations in the Nordic and Baltic countries.

We introduce the activities and results of the European Language Resources Coordination (ELRC) action in this region, provide a synopsis of ELRC workshops held in all countries of the region, identify potential holders and donors of language data suitable for improving machine translation (MT) systems, and describe the language resources collected so far.

We also describe several national projects and initiatives on sharing of language data accumulated in the public sector and creation of new language resources from this data. Opportunities and challenges in consolidating language data from the public sector are discussed, and related actions and regulatory initiatives are proposed.

[1] http://www.meta-share.org/
[2] http://www.clarin.eu/
[3] http://clarin.w.uib.no/
[4] http://kitwiki.csc.fi/twiki/bin/view/FinCLARIN
[5] http://info.clarin.dk/en/
[6] http://opus.nlpl.eu/
[7] https://www.tausdata.org/

Work-shop	Location	Organizers	Partici-pants
Denmark workshop, Mar-7-2016	European Environment Agency in Copenhagen	Danish Language Committee, Tilde	71
Estonia workshop, Feb-11-2016	EC Representation Office in Tallinn	Tilde, Estonian Language Resource Centre Office of the EC Representation in Estonia	66
Finland workshop, Feb-19-2016	University of Helsinki	University of Helsinki, Tilde	34
Iceland workshop, Nov-11-2016	Safnahúsið (the Culture House) in Reykjavik	Vigdís Finnbogadóttir Institute Tilde	36
Latvia workshop, Oct-6-2015	EC Representation Office in Riga	Tilde, Culture Information Systems Centre, Office of the EC Representation in Latvia	59
Lithuania workshop, Feb-24-2016	Lithuanian Government Building in Vilnius	State Commission of the Lithuanian Language, Office of the EC Representation in Lithuania, Office of the Government of the Republic of Lithuania, Tilde IT	141
Sweden workshop, Mar-10-2016	Europahuset in Stockholm	Language Council of Sweden, Språkbanken, EC Representation in Sweden, Tilde	47
Norway workshop, Jun-8-2016	Difi's course and conference venue in Oslo	Difi Agency, University of Bergen Tilde	53

Table 1 : ELRC Workshops in Nordic and Baltic countries

2. ELRC in the Nordic and Baltic countries

The aim of the European Language Resource Coordination action is to identify and gather language and translation data relevant to national public services, administrations, and governmental institutions across all 30 European countries participating in the Connecting Europe Facility (CEF) programme[8], i.e. EU Member States, Norway and Iceland (Lösch et al., 2018). This data is used to improve the quality of automated translation systems provided by the European Commission's CEF eTranslation service[9]. In return, CEF eTranslation makes MT services available to public service providers and administrations to support them in their interactions with citizens across language barriers.

The ELRC consortium includes DFKI[10], ELDA[11], ILSP/Athena RC[12] and Tilde[13]. It operates under service contracts SMART 2014/1074 and SMART 2015/1091 with the European Commission. Tilde is responsible for coordination of ELRC activities in Nordic and Baltic countries presented in this paper. ELRC activities in CEF countries are supported by one technological representative (Technology National Anchor Point) and one representative from the public services administration (Public Services National Anchor Point). ELRC National Anchor Points in the Nordic and Baltic region are highly respected representatives from academic sector (University of Helsinki, Vigdís Finnbogadóttir Institute, University of Iceland, IMCS at the University of Latvia), national language policy institutions (Danish Language Committee, State Commission of the Lithuanian Language, Swedish Language Council), government agencies (Estonian Ministry of Education and Research, Finnish Prime Minister's Office, Latvian Culture Information Systems Centre, Norwegian Agency Difi), and language resource centres (Estonian Language Resource Centre, National Library of Norway, Språkbanken at the University of Gothenburg). A full list of National Anchor Points is available on the ELRC online platform[14].

3. Findings from the ELRC Workshops

The first task of ELRC was to inform and engage the public administrations. For this task we organized a series of local workshops in all Nordic and Baltic countries, with the support of national experts (see Table 1). The goals of the workshops were to raise awareness about the importance of language data, understand the needs of national public sector administrations with regard to automated translation, jointly identify relevant sources of multilingual language resources, and discuss technical and legal issues involved in the use of data for automated translations.

Although having the same objectives and similar structure, the workshops reflected differences and particularities in various areas – national policies in the field of language resources and language technologies, openness of the public sector to share linguistic data, awareness of applicability of machine translation and other language technology tools, etc.

The Latvia workshop included a presentation of government activities in creating the national machine translation platform Hugo.lv, as part of Latvia's e-Government service infrastructure (see Section 6.1).

The Latvian workshop also showed that the largest translation volumes are accumulated at the State Language Center (formerly the Translation and Terminology Center), which has a well-established translation process and large translation memories accumulated through the usage of the SDL Trados computer-assisted translation (CAT) tool. This agency also intends to consolidate Latvian terminology data into a national terminology database. We see this agency as a major collaboration partner and provider of valuable parallel data.

The Estonia workshop raised a discussion on the applicability of state-of-the-art machine translation systems for such complex, agglutinative languages as Estonian and Finnish. Several participants expressed concerns that current systems like Google Translate produce translations that are not suitable for practical applications or post-editing. This emphasized the need for much larger volume of parallel data to train MT systems of

[8] https://ec.europa.eu/inea/en/connecting-europe-facility
[9] https://ec.europa.eu/cefdigital/wiki/display/CEFDIGITAL/eTranslation
[10] http://dfki.de/en
[11] http://elda.org/en
[12] http:// www.ilsp.gr/en
[13] http://tilde.com
[14] http://lr-coordination.eu/anchor-points

significantly better quality. Public sector participants were interested in contributing to the development of Estonian MT with their language data, making it available not only for the EC, but also for Estonian researchers and developers through the Estonian Language Resource Center.

The Finland workshop explored the complexities of translating into the Finnish language. Several activities such as FIN-CLARIN have successfully mobilized Finland's language technology community to meet this challenge. The workshop demonstrated the acute needs of the public sector in higher quality MT. The Finnish public administration has a pressing need to translate information into both Swedish and Finnish, as well as into major European languages in order to reach the multilingual population. Therefore, the public sector (e.g., Office of the Prime Minister, Municipality of Helsinki, Ministry of Justice, Tax Administration) is actively following the developments in CEF Automated Translation and would be potentially very interested in integrating the platform into public services.

The Lithuania workshop had the largest number of participants (141) and prominent keynote speeches from government representatives, which demonstrated the efficiency of national level coordination by the Lithuanian State Language Commission. A particularly valuable source of translation data was identified in Seimas (the Lithuanian Parliament). Its Legislation Editing and Translation Unit is tasked with the translation, editing, signing, and publication of legislative acts submitted to the Seimas, legislative acts adopted by the Seimas, documents of the Board of the Seimas, and decrees of the President of the Seimas. The 12-member department has collected Translation Memories primarily in Lithuanian, English, and Russian.

The Denmark workshop was a very well-attended event, with nearly 70 participants representing a wide cross-section of public institutions, research organisations, and businesses in Denmark.

The program featured presentations from the ELRC consortium, governmental, translation, and language technology sectors. Both the audience and the speakers had a positive attitude towards the CEF platform and showed interest in how the Danish public administration can provide language resources to the European Commission in order to develop CEF AT. This interest was also compounded by the presence of members of the private sector, particularly representatives from IBM Denmark and translation companies.

The workshop also stimulated discussion on the need to set up national level coordination of public administrations on LRs and application of MT.

The Sweden workshop had a strong showing from actual translators and editors at the Swedish public administration, particularly English language translators at the Ministry of Foreign Affairs. Thanks to the participation of these translators, many of whom are foreign nationals living and working in Sweden, the workshop was able to generate discussion around an actual use case of MT in the public sector, as well as identify multiple sources of translation memories from the Swedish public administration.

The Norway workshop was one of the most successful seminars organized during the ELRC action to date. Norway has a diverse linguistic landscape, due to the co-existence of several official languages: Norwegian variations Bokmål and Nynorsk, Sami, and Kven. Therefore, language resources exist in each of these distinct languages – though, of course, in varying degrees of volume and coverage. As Norwegians are keenly aware of this disparity, they are also aware of the challenges faced by developing machine translation technologies for the public sector. For example, should engines be developed for all official languages, or just one?

To answer these and other questions, the public administration in Norway has already commissioned a large-scale study of the possible costs of collecting sufficient language resources in each of the official languages, as well as the foreseen benefits of utilizing machine translation in the everyday work of the public sector (Oslo Economics, 2016).

The conclusions reached were that national programmes in Norway should work hand-in-hand with the ELRC action to generate the largest possible volumes of language resources for raising the quality of Norwegian MT provided by CEF eTranslation. Adjusting machine translation for the CEF Online Dispute Resolution (ODR) service[15] was identified as a particular priority that requires data that represents more informal language used by users in customer complaints. The workshop stimulated national activities to complement work of ELRC on collecting data related to ODR. The workshop also identified several sources of language resources from the Ministry of Foreign Affairs, which has been collecting Translation Memories generated by translators working in the SDL Trados Studio CAT tool for many years.

The Iceland workshop generated a good deal of interest among a wide range of participants from both the public and private sectors. Interest in language technology in Iceland is certainly driven by the relatively small size of the language community (with just 330 000 native speakers) and the high degree of national pride in keeping the language vital.

The Icelandic workshop was also candid about the limits of developing high-quality machine translation services for a language with such a small amount of input data. Therefore, the participants at the workshop were strongly supportive of mobilizing the public administration to gather as much language resources as possible for developing CEF eTranslation for Icelandic.

Detailed reports on the workshops as well as presentations and video recordings are available on the ELRC online platform[16].

4. Language Resource Collection

To facilitate LR collection, the ELRC consortium created the online LR repository ELRC-SHARE[17], based on the META-SHARE open source distributive. The ELRC-SHARE data model is an extension of the META-SHARE schema (Gavrilidou, 2012) with added fields to support LR management.

ELRC dissemination activities have already spurred the interest of numerous public administrations to assess and

[15] http://ec.europa.eu/consumers/odr

[16] http://lr-coordination.eu/events

[17] https://www.elrc-share.eu

share their language data. Some of them have directly uploaded their data on the ELRC-SHARE repository. Others consulted with ELRC experts or National Anchor Points (NAPs) and provided their data through them.
By the end of the first phase of the ELRC action, 65 language resources were provided from all countries of the region (see Table 2). Most LRs (39) are parallel corpora, and only 6 are monolingual corpora. This is because ELRC has prioritized parallel data as the scarcest and most necessary resource to improve the quality of MT.
Instead of providing language data, several institutions pointed to their websites with parallel multilingual content. ELRC partners then crawled these sites to extract parallel data from it, process, align and build a parallel corpus from the data collected.
The collected data is first provided to the EC for use in training their MT systems. After an assessment by the EC, these language resources will also be distributed on the META-SHARE platform and the European Open Data Portal.

Type of Text	Corpus	Lexical Resource
Danish	2	2
Estonian	7	1
Finnish	11	4
Icelandic	1	1
Latvian	9	1
Lithuanian	4	1
Norwegian	2	1
Swedish	9	9
Total	**45**	**20**

Table 2 : LRs collected and processed in the first phase of ELRC action by language and type

Among all the data collected and donated in the project there were some outstanding resources in terms of their volume. The main difference with them is that the content of these resources was created in a managed way: as part of a centralized translation workflow or as part of a document or terminology content management system. This clearly illustrates the importance and yield from managed workflows in the area of language resource processing. Also, most of those resources were donated to the ELRC action as part of direct communication on the part of NAPs with the resource holder, a relationship built during ELRC events and follow-up communication. The following summarize some of the most outstanding resources:

- Translations of Lithuanian legislation from Seimas of the Republic of Lithuania – the entire translation memory of Lithuanian-English translations exported from the translation server of Lithuanian Seimas – a total of over 130 000 translation units donated to the ELRC action by the Legal Acts Editing and Translation Unit of the Document Department at the Seimas of the Republic of Lithuania.
- Corpus of the Translations of Estonian legislation with 47 255 translation units, donated for the ELRC by the Ministry of Justice of the Republic of Estonia.
- Translation memories from the Ministry of Foreign Affairs containing Norwegian Bokmål and Nynorsk translations of Acquis Communautaire with 733 081 translation units licensed under CC-BY license. This resource is hosted by the Norwegian Language Bank (Språkbanken).
- Icelandic Termbank by The Árni Magnússon Institute for Icelandic Studies containing 103 753 term entries in 41 term collections in various domains, licensed under CC-BY-SA license.

5. Intellectual Property Rights

The ELRC consortium takes particular care about the clearance of Intellectual Property Rights (IPR) for all resources collected or received from donors. To do so, the Consortium follows a set of guidelines established by ELRC.
The basic principles behind this workflow consist of (1) checking whether the data under consideration are available under an Open License; otherwise (2) see whether they are under PSI scope, or (3) if they need to be negotiated. If neither step 1 nor step 2 suffice, the Consortium contacted the data owners to negotiate and agree upon usage conditions.

6. National Projects

Besides the ELRC action, which was initiated and funded by the EC, there are several other activities on a national level that contribute to the collection of language resources from public administrations.

6.1 HUGO.LV in Latvia

The collection of parallel data is a part of the work on the machine translation platform for the e-Government infrastructure of Latvia. The platform, called HUGO.LV and developed by Tilde, includes machine translation systems for Latvian, English, and Russian tailored for the requirements of various e-Government services (e.g., the state e-services portal Latvia.lv). As part of the platform's development efforts, a large corpus of parallel and monolingual data was collected.
The project has raised awareness in the state administration about the need to manage translation service tenders in a way that not only fulfils the direct goal of acquiring translations, but also requires translation memories to be returned in order to build an open corpus of public sector translations, which, in turn, can be further reused on newer builds of the Hugo.lv service. To simplify the data workflow, a LR data upload facility has been created in HUGO.LV. Within this facility, registered state officials can upload parallel content in any format: translation memories (.tmx .sdlxliff, etc.), text files (TXT, DOC, DOCX, etc), PDF, or JPEG. This ensures that language technicians can transform the content to reusable corpora, to be used in the training of HUGO.LV MT systems, as well as to be distributed on open data portals. Figure 1 contains a screenshot of the parallel data upload page.

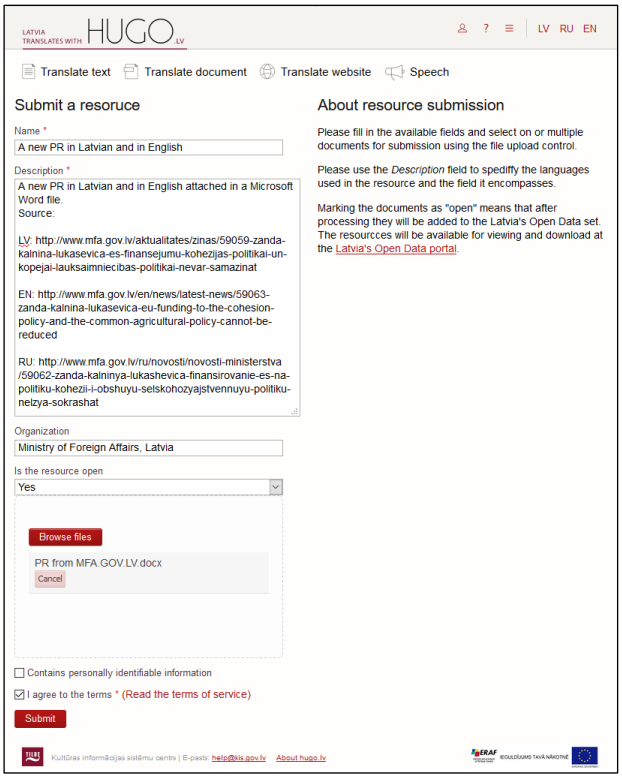

Figure 1: HUGO.LV submission page for parallel textual data

The government institution in charge of HUGO.LV – Culture Information Systems Centre (CISC) – has several cooperation partners. The Latvian State Language Centre – *Valsts valodas centrs* (*VVC)* – performs official translation work of national legislation, international conventions and agreements, EU legal documents between Latvian and English as well as maintains official terminology for the translation of legal texts. VVC is committed to donate TMs with at least 250 000 translation segments and 100 000 terminology entries for the benefit of the second phase of the HUGO.lv project. The Latvian Court Administration, in turn, is anticipated to provide 2000 anonymized judicial decisions and other parallel documents. The Information Society Committee, led by the Prime Minister, has tasked CISC with compiling a full list of Translation Memories available in the Latvian public sector, to be submitted by the end of 2018.

CISC has agreed to cooperate with ELRC and contribute to its data collection activities.

6.2 Data Creation for ODR in Norway

The Norwegian Consumer Centre (Forbruker Europa) contributes to ELRC by creating and providing data that will support Norwegian in the ODR service. The project is focused on the production of translated texts from consumer enquiries. The first batch of 4400 translated segments are already submitted to ELRC and EC to train ODR-specific CEF eTranslation engine.

6.3 Estonian Open Parallel Corpus

The National Programme for Estonian Language Technology supports the development of Estonian-related language resources and technologies. In 2016, the Estonian Open Parallel Corpus project collected, aligned and published into the EOPC corpus as many as 413 000 English-Estonian and 155 000 Estonian-Russian translation units from websites of Estonian public sector and EU institutions. In 2017, the EOPC project collected a parallel corpus of 226 631 translation units from Estonian to English, French, German and Russian languages from public parallel web sites. The resulting corpora is distributed on the META-SHARE[18] repository.

6.4 Government Translation and Language Services in Finland

The Finnish government has begun an initiative to improve data management in their country[19]. The Translation and Language Services Division (TLD), established in 2015 as part of the Government Administration Department, provides translation, language, and terminology services to all ministries in the Finnish government. Among the duties of TLD is the management of a Translation Memory system and internal termbases, as well as maintenance of the government's online term bank VALTER[20].

TLD has also created a system for the ordering and management of translation requests, called SHAKE. The system manages language resources by storing all documents sent for translation automatically on a network drive, accessible to the entire TLD (all translations, XLIFF files, and background material are stored in the same place). In addition, SHAKE allows users to search for and retrieve earlier translations via automatic archiving and enables the establishment of common practices for the use of TMs and term banks, including via server-based TMs with set attributes that enable the retrieval of "theme-based" translation memories for the use of external service providers.

In effect, the Finnish government is successfully implementing in the public sector a language resource management process that help to manage TMs, leveraging them for use in translation projects, as well as maintain a high degree of order in the management and storage of data.

7. PSI Directive in Practice

The ELRC action shows that public sector institutions are very interested in using MT and they have ample resources to contribute. However, there are multiple barriers that must be overcome to make that happen.

Often translation data is not organized within an institution. Employees of government institutions refer to their website as the ultimate data source. The drawbacks of this approach are that just a small fraction of the data is readily available; it must yet be cleaned and aligned both on a document- and sentence-level.

The statements of the PSI directive are not yet adapted effectively. Often the data is there, but motivation is missing to go the last mile to actually share the data. It takes pleading, clarifying, and sending official letters from EC or ELRC consortium for someone in the institution to take the responsibility to give an internal order to donate the data. Instead, we anticipate proactive sharing of data online so

[18] http://metashare.tilde.com/repository/search/?q=estonian+open+parallel+corpus

[19] http://www.lr-coordination.eu/sites/default/files/Belgium/2017/Brussels_3rd_ELRC_Conference/Taru%20Virtanen_Case%20Study%20Finland.pdf

[20] http://www.valter.fi

that anyone who has interest may find and download the data that they need.

There is a need to raise awareness that written texts produced in public sector is valuable public data, which may be reused for language technology research and development; therefore it must be organized, saved, and made available publicly whenever possible. This may require a common approach on all levels – starting from submitting tenders for translation services; introducing clauses in contracts with translation service providers for submitting TMs as part of delivered data; and the need for common state-wide infrastructure for managing the textual data – monolingual documents, translated documents, TMs, source files prior to their publishing or conversion to PDF.

There is a limited understanding within institutions about their language data assets. Government officials being addressed do not know what data the institution has, and often see data collection activities as an additional burden. Surprisingly, in some cases it turns out that it takes very little effort to export translation memories and to upload them on ELRC-SHARE.

ELRC is actively promoting best practice examples such as data contribution by the Parliament of the Republic of Lithuania or HUGO.LV project in Latvia to encourage other institutions and national governments to follow.

8. Conclusions

The results of the ELRC action have shown that public administrations have valuable language data in their possession. More than 500 participants in ELRC workshops and numerous data contributions demonstrate that public administrations in the region can be effectively engaged in language resource identification and collection. The workshops also showed that the public administrations in Northern Europe have a pressing need for integrating machine translation into public services, as the region is highly multilingual.

At the same time, the languages of the region are extremely complex, with relatively small volumes of available data. However, by continuing to mobilize stakeholders in the region, the ELRC action shows how to identify and gather valuable language resources for improving the quality of MT services for Northern European languages.

65 new Baltic/Nordic language resources were collected in the first phase of the ELRC action. In the second phase by the end of 2017, 72 additional resources were collected or donated by public administrations and are in processing to produce LRs ready for use in MT training.

Acknowledgments

The research leading to these results has received funding from the research project "Competence Centre of Information and Communication Technologies" of EU Structural funds, contract No. 1.2.1.1/16/A/007 signed between IT Competence Centre and Central Finance and Contracting Agency, Research No. 2.2. “Prototype of a Software and Hardware Platform for Integration of Machine Translation in Corporate Infrastructure".

Bibliographical References

Gavrilidou, M., Labropoulou, P., Desypri, E., Piperidis, S., Papageorgiou, H., Monachini, M., Frontini, F., Declerck, T., Francopoulo, G., Arranz, V. and Mapelli, V. (2012). The META-SHARE Metadata Schema for the Description of Language Resources. Proceedings of the Eighth International Conference on Language Resources and Evaluation (LREC 2012), Turkey.

European Parliament (2003). Directive 2003/98/EC of the European Parliament and of the Council of 17 November 2003 on the re-use of public sector information, Official Journal of the European Union, L 345, 31.12.2003, p. 90–96

Lösch, A., Mapelli, V., Piperidis, S., Vasiļjevs, A., Smal, L., Declerck, T.; Schnur, E., Choukri, K., and van Genabith, J. (2018). European Language Resource Coordination: Collecting Language Resources for Public Sector Information Management. In Proceedings of the 11th Language Resources and Evaluation Conference (LREC 2018), Miyazaki, Japan, May 2018. European Language Resources Association (ELRA)

Oslo Economics (2016). Kartlegging av behovet for automatisk oversettelse i statlig sector, *Utarbeidet for Kommunal- og moderniseringsdepartementet OE-rapport 2016-15*

Piperidis, S. (2012). The META-SHARE Language Resources Sharing Infrastructure: Principles, Challenges, Solutions. Proceedings of the Eight International Conference on Language Resources and Evaluation (LREC'12), Istanbul, Turkey, pages 36-42.

Rehm G. and Uszkoreit H. (2011). Multilingual Europe: A challenge for language tech. *MultiLingual*, 22(3):5152, April/May 2011.

Wittenburg, P., Bel, N., Borin, L., Budin, G., Calzolari, N., Hajicova, E., Koskenniemi, K., Lemnitzer, L., Maegaard, B., Piasecki, M. and Pierrel, J.M. (2010). Resource and service centres as the backbone for a sustainable service infrastructure. In Proceedings of the Seventh International Conference on Language Resources and Evaluation (LREC 2010) (pp. 5-9).

Skadiņa, I., Vasiljevs, A., Borin, L., De Smedt, K., Linden, K. and Rognvaldsson, E. (2011). META-NORD: Towards Sharing of Language Resources in Nordic and Baltic Countries. Proceedings of Workshop on Language Resources, Technology and Services in the Sharing Paradigm, Chiang Mai, Thailand, pp. 107-114.

Tiedemann J. (2012). Parallel data, tools and interfaces in OPUS. *Proceedings of the Eight International Conference on Language Resources and Evaluation (LREC'12)*, Istanbul, Turkey, may. European Language Resources Association (ELRA).

Váradi, T., Krauwer, S., Wittenburg P., Wynne, M., Koskenniemi, K. (2008). CLARIN: common language resources and technology infrastructure. Proceedings of the Sixth International Language Resources and Evaluation Conference (LREC 2008).

Vasiļjevs, A., Pedersen, B. S., De Smedt, K., Borin, L., Skadiņa, I. (2011). META-NORD: Baltic and Nordic Branch of the European Open Linguistic Infrastructure. Proceedings of the NODALIDA 2011 workshop on Visibility and Availability of LT Resources, Riga, Latvia, pp. 18-22.

Vasiljevs, A., Rirdance, S. and Liedskalnins. A. (2008). EuroTermBank: Towards Greater Interoperability of Dispersed Multilingual Terminology Data. Proceedings

of the First International Conference on Global Interoperability for Language Resources (ICGL 2008), Hong Kong

Language Resource References

The Árni Magnússon Institute for Icelandic Studies. (2016). The Icelandic Term Bank, accessible online at http://www.malfong.is/index.php?lang=en&pg=idord

Nasjonalbiblioteket Språkbanken. (2016). Translation memories from The Ministry of Foreign Affairs, dowloadable at http://www.nb.no/sprakbanken/show?serial=oai%3Anb.no%3Asbr-36&lang=en

Tilde. (2016). Estonian Open Parallel Corpus, https://www.keeletehnoloogia.ee/en/projects-2011-2017/estonian-open-parallel-corpus-1, distributed via META-SHARE

The Prime Minister's Office Finland. (2016). Hallituskausi 2011–2015, distributed via The Prime Minister's Office

The Prime Minister's Office Finland. (2015). Hallituskausi 2007–2011, The Prime Minister's Office