AI FOR A MULTILINGUAL EUROPE

Why Language Data Matters

ELRC White Paper

European Language Resource Coordination

Connecting Europe Facility

Imprint ELRC White Paper

German Research Center for Artificial Intelligence (DFKI)
Multilinguality and Language Technology
Stuhlsatzenhausweg 3
Saarland Informatics Campus D 3 2
66123 Saarbrücken
Germany

Contact:
Phone: +49 681-85775 5285
E-mail: info@lr-coordination.eu
Web: www.lr-coordination.eu

Publisher:
ELRC Consortium:
DFKI
ELDA
ILSP
Tilde

Design: Stefania Racioppa

Print:
OVD.eu | Verlag & Eventagentur
OVD.de | Druck- & Werbeservice
Johanna-Wendel-Str. 13 | 66119 Saarbrücken

Images: Unsplash

Third print edition 2022

First print edition published November 2019
First online edition published November 2019
Second online edition published December 2019
Third print edition published November 2022
Third online edition published November 2022

ISBN: 978-3-943853-07-0

Work underlying the White Paper has been created under the ELRC contract with the European Union (SMART 2019/1083). The opinions expressed are those of the ELRC consortium and Language Resource Board and do not represent the contracting authority's official position.

Authors

The ELRC Consortium:

Eileen Marra, *DFKI*
Andrea Lösch, *DFKI*
Stefania Racioppa, *DFKI*
Hélène Mazo, *ELDA*
Maria Giagkou, *ILSP*

Contributions from the following authors in alphabetical order:

Dimitra Anastasiou, *Luxembourg Institute of Science and Technology*
Natassa Avraamides, *Press and Information Office, Ministry of Interior, Republic of Cyprus*
Carl Frederik Bach Kirchmeier, *Agency for Digital Government, Ministry of Finance, Denmark*
Yngvil Beyer, *National Library of Norway*
António Branco, *University of Lisbon*
Virginijus Dadurkevičius, *Vilnius University*
Hristina Dobreva, *Ministry of Transport, Bulgaria*
Rickard Domeij, *The Language Council of Sweden*
Jane Dunne, *Dublin City University*
Kristine Eide, *The Language Council of Norway*
Maria Gavriilidou, *Institute for Language and Speech Processing*
Stanislava Graf, *Charles University Prague*
Dagmar Gromann, *University of Vienna*
Thibault Grouas, *DGLFLF, Ministère de la Culture, France*
Normunds Grūzītis, *Institute of Mathematics and Computer Science, University of Latvia*
Jan Hajič, *Charles University Prague*
Barbara Heinisch, *University of Vienna*
Veronique Hoste, *Ghent University*
Simon Krek, *Jožef Stefan Institute*
Gauti Kristmannsson, *University of Iceland*
Svetla Koeva, *Bulgarian Academy of Sciences*
Anna Kotarska, *Polish Society for Health Programs*
Kaisamari Kuhmonen, *Prime Minister's Office, Finland*
Krister Lindén, *University of Helsinki*
Teresa Lynn, *Dublin City University*
Kinga Matyus, *Hungarian Academy of Sciences*
Maite Melero, *Barcelona Supercomputing Center*
Laura Mihăilescu, *European Institute in Romania*
Željka Motika, *Central State Office for the Development of Digital Society of the Republic of Croatia*
Tríona Ní Mhathuna, *The Department of Tourism, Culture, Arts, Gaeltacht, Sport and Media, Ireland*
Micheál Ó Conaire, *The Department of Tourism, Culture, Arts, Gaeltacht, Sport and Media, Ireland*
Maciej Ogrodniczuk, *Institute of Computer Science, Polish Academy of Sciences*
Jon Arild Olsen, *National Library of Norway*
Michael Rosner, *University of Malta*
Elisa Schnell, *Austrian National Defense Academy*
Maria Skeppstedt, *The Language Council of Sweden*

Alexandra Soska, *Federal Ministry of the Interior and Community*
Donatienne Spiteri, *Office of the State Advocate, Malta*
Marko Tadić, *University of Zagreb, Faculty of Humanities and Social Sciences*
Carole Tiberius, *The Dutch Language Institute*
Dan Tufiş, *Research Institutte for Artificial Intelligence of the Romanian Academy*
Andrius Utka, *State Commission of the Lithuanian Language*
Tamás Váradi, *Hungarian Academy of Sciences*
Kadri Vare, *Ministry of Education and Research, Estonia*
Andreas Witt, *Leibniz Institute for the German Language*
Josephine Worm Andersson, *Danish Agency for Digitisation*
François Yvon, *French National Centre for Scientific Research (CNRS)*
Jānis Ziediņš, *Culture Information Systems Centre of Latvia*
Bódi Zoltán, *Institute of Hungarian Research*
Miroslav Zumrík, *Slovak Academy of Sciences*

Table of Contents

Foreword

June Lowery-Kingston
Head of Unit CNECT.G3 – Accessibility, Multilingualism and Safer Internet

Welcome to this second revision of the White Paper by the European Language Resource Coordination (ELRC). As the European Commission Head of the Unit charged with supporting ELRC, I am proud to be able to share this revised edition with you and share the progress made since the first report was published in December 2019.

While the European Commission has long encouraged the collection of multilingual data, since 2014 via the ELRC, the European Parliament's 2018 own-initiative report on language equality in the digital age gave the impetus for Member States in the European Union to widen the discussions, which resulted in the first ELRC White Paper. Thanks to continued efforts by the ELRC through their country workshops, initiatives already started have kept and perhaps even increased their momentum in the various Member States. Recognition of the importance of language technologies, resources and tools has become more and more evident and acknowledged. This has culminated in the Commission proposal of the Declaration on European Digital Rights and Principles, which enshrines the right to a trustworthy, diverse and multilingual online environment.

We are convinced that language technologies offer unprecedented opportunities to overcome language barriers in the Union – with both economic and social impact. Our vision of language technologies that automatically produce easy-to-read versions of official communications for those with reading difficulties, cognitive impairment, or no knowledge of a foreign language is within grasp. Advances in artificial intelligence and natural language technologies offer us tools that can be used in more and more situations and support intercultural communication, including between the least widely used European languages. Efforts by the Member States, in conjunction with European funding, are enabling us to keep the cost of developing language technologies down and support (digital) equality between languages with better quality and lower cost. The inclusion of the Commission's eTranslation tool in the online platform for the Conference on the Future of Europe was one example of how such tools can allow multilingual exchanges in a democratic setting. The EU funded tool enabling European SMEs to translate their websites into any of the official European Union languages is another.

An important crossing point has been reached in this language resource journey not only at Member State level. Thanks to the European Data Strategy, as the population at large becomes more aware of the value of their personal data, all sectors of the economy are starting to appreciate the data assets they hold. Whether it is call centre recordings, film, or digitised paper archives on health, tourism or other domains, we see a change in perception of the importance of these data as well as of the language technologies required to manage these riches.

With the launch of the Digital Europe programme and following our European strategy for data, we want to create an active ecosystem around the data associated with languages. The launch of the Language Data Space is a major step forward, increasing the visibility of the actions as well as enlarging the family of the stakeholders, as the private sector will be more involved in the collection and sharing of language resources. The deployment of this European Language Data Space will not only promote the creation, collection, sharing and re-use of language-related data, but will also help to create, share, and re-use those famous computer language models, applying the full potential of artificial intelligence for automatic language processing in many different ways and settings.

It is our intention that the EU should become a lighthouse for language technologies, data, and language equality across the world. We hope that this White Paper is an important steppingstone to reach that goal and will encourage you in your efforts and inspire new stakeholders to join the journey.

Executive summary

Modern Language Technologies (LT) such as Machine Translation (MT), but also Fake News Detection or Text Anonymisation, are based on Machine Learning (ML) – a process where machines improve by learning from sufficient amounts of high-quality training data.

The European Language Resource Coordination (ELRC) was initiated in 2015 to collect such training data – so-called language resources – in all official European languages, as well as Norwegian Bokmål, Norwegian Nynorsk and Icelandic, with special focus on bi- and multilingual language data from various domains. The initial purpose was to collect language resources to train CEF eTranslation, the Machine Translation service of the European Commission that can be used free of charge by all public administrations and public services in the EU Member States, Norway and Iceland, academia, NGOs as well as SMEs.

The usefulness of language data, however, goes far beyond training eTranslation: Language data is the driving force behind all data-based Language Technologies. And in fact, the eTranslation MT application was complemented by a growing number of language tools in the last years, reaching from named entities recognition to translation quality estimation, which are freely available not only for academia and public administrations, but also for SMEs and NGOs.

That is why the data collected by ELRC is still made available to the wider public both for research and commercial applications: approximately 80% of the language resources hosted in the ELRC-SHARE repository are freely re-usable outside ELRC.

In order to further support the sharing of language data in Europe, ELRC conducted a first investigation among public services in 2019 (ELRC, 2019) in order to identify the key stakeholders and mechanisms for the efficient sharing of language data in EU Member States, Norway and Iceland.

The current ELRC White Paper "AI for Multilingual Europe" follows up on the first White Paper version published in 2019: It compares the results of the 2019 analysis, which described European practices for sharing language data as well as corresponding challenges and recommendations on how to address these challenges with the status quo of 2022, illustrating latest developments, recent changes and achievements.

Given the increasing importance of AI and LT across all European countries and sectors, the ELRC White Paper at hand focuses on the role of LT and language resources, both within public administrations and SMEs, while taking into account recent developments in this respect as well as national regulations related to AI.

In the course of this investigation, ELRC gained important new insights into the value and status quo of language-centric AI which actually changed since 2019 (see Section 3). For instance, MT increasingly finds its way into the daily work life of public administrations in 2022 – only 6% of the participating organisations didn't use MT at all. At the same time, we found a massive increase in the use of Computer-Assisted Translation (CAT) Tools. Also, we could observe significant changes on policy level and with regard to actual translation and data sharing practices in the participating organisations in comparison to 2019.

Moreover, the ELRC White Paper illustrates latest developments and approaches to sustainable language data sharing in SMEs and public services (see Section 4). The circumstances that were found to negatively impact or limit the sharing of language data in Europe remained the same as in 2019. However, in addition to the actions that would be most

relevant to overcome these issues and facilitate data sharing, several additional approaches were mentioned in 2022. The survey participants identified six major challenges that organisations involved in the preparation and sharing of language data face in 2022 and beyond, e.g. the development of LT for European minority languages and lack of expertise, e.g. concerning legal provisions and regulations

Last but not least, the Annex contains an updated country profile for each participating CEF country, which provides latest insights into:

- the translation practices and needs in public administrations
- the country's digital and language policy
- the role of LT and language data in public administrations and national regulations
- stakeholders relevant to the sharing of language data
- data collection efforts for LT/AI
- major networks, projects and key players related to LT
- the challenges of sharing language data
- a corresponding action plan to address and overcome these challenges.

Each country profile is a self-contained document supplemented by the main body of the White Paper and the other country profiles. The level of detail may vary from one profile to another. Unless otherwise stated, all information refers to the situation at the national level of the particular country.

With regard to the organisational level, the most important recommendations address:

- **Translation and Data Management**, in particular the designation of Open Data officers in all public administrations and services, the introduction of general rights management in the data management process, the adoption of translation data management plans, the centralisation of translation workflows, and the adaptation of translation procurement contracts.
- **Human Capital,** including in particular the provision of technical and legal training for translators and translation managers.
- **IT Infrastructures, Equipment and Tools**, including in particular the provision of CAT tools, MT, data anonymisation methods and tools etc.
- **Translation process / workflow**, including the appropriate licencing of translation data, the identification (and, where necessary, exclusion) of confidential and personal data and the maximal automatisation of the process of translation/language data creation, curation and collection.

Some of the recently released or updated national AI strategies already pay due attention to the importance of LT and language data. However, according to the results of our investigation, these can only be considered as first steps towards a truly sustainable creation, management and sharing of language resources in Europe. In order to enable the successful implementation of the recommended actions across European countries, future funding schemes should support the proposed activities provided in this document.

Disclaimer:

Please note that the information is based on the experiences of the ELRC consortium and Language Resource Board[1] including individual investigations and expertise as well as information derived from public reports, national strategies and other types of publications. Thus the solutions and actions suggested in this report reflect the expertise of the ELRC consortium and the Language Resource Board and are not national initiatives unless clearly indicated. The information provided cannot be considered complete.

[1] https://www.lr-coordination.eu/anchor-points

Abbreviations

AI	Artificial Intelligence
API	Application Programming Interface
ASR	Automatic Speech Recognition
CAT	Computer-Assisted Translation
CEF	Connecting Europe Facility
ELE	European Language Equality
ELG	European Language Grid
ELRC	European Language Resource Coordination
GDPR	General Data Protection Regulation
IPR	Intellectual Property Right
IT	Information Technology
LR	Language Resource
LSP	Language Service Provider
LT	Language Technology
ML	Machine Learning
MT	Machine Translation
NAP	National Anchor Point
NER	Named Entity Recognition
NGO	Non-governmental organisation
NLP	Natural Language Processing
NLU	Natural Language Understanding
NMT	Neural Machine Translation
QA	Question Answering
SME	Small and medium-sized enterprise
STT	Speech to Text
TM	Translation Memory
TTS	Text to Speech
TU	Translation Unit

Glossary and definitions

Artificial Intelligence:
The simulation of human intelligence processes by machines, especially computer systems. Specific applications of AI include, among others, Natural Language Processing, Speech Recognition and Machine Translation.

Automatic Speech Recognition:
Technology enabling the recognition of spoken language and its conversion into a text. Synonym: Speech to Text (STT).

Chatbot:
System using conversational AI technology to simulate and process human conversation through voice commands and/or text chats, allowing to interact with digital devices as if communicating with a real person.

Computer-Assisted Translation:
Translation performed by human translators with the help of computerised tools. Synonym: computer-aided translation (Azzano, 2011).

CEF Countries:
Countries participating in the Connecting Europe Facility (CEF) programme, a key EU funding instrument to promote growth, jobs and competitiveness through targeted infrastructure investment at European level.

Language Technology:
Language technology (LT), often also referred to as human LT, comprises computational methods, computer programmes and electronic devices that are specialised for analysing, producing, modifying and translating text and speech (Uszkoreit, 2010).

Language-centric AI:
The branch of Artificial Intelligence dedicated to the processing of languages. The term is most often used interchangeably with the term Language Technology.

Language data:
Refers to any textual, audio or audiovisual data produced using human language or data about a human language (such as grammars, language models etc.).

Language data creator:
The person(s) or organisation(s) that generate text or speech in digital form. In the context of translation, the author of the source text and the author of the target text (the translator) are the language data creators.

Language Resource:
Sets of language data and descriptions in machine-readable form, including written and spoken corpora, grammars, and terminology databases. Language resources can be used to build, improve, or evaluate natural language systems such as Machine Translation engines.

Large Language Models:
Statistical and probabilistic tools combining the latest deep learning technology with heavy computing infrastructure to build language models from large amounts of text or speech data. Such models incorporate information that is useful for understanding a language, such as its vocabulary and how it expresses meaning.

Information Extraction:
The task of automatically extracting structured information from unstructured and/or semi-structured machine-readable documents. This mostly concerns processing human language texts by means of Natural Language Processing (NLP). Some of its most common subtasks include Named Entity Recognition, Question Answering, and Relation Extraction.

Intellectual Property Right holder:
The person or organisation that holds the right to benefit from the protection of moral and material interests resulting from authorship of scientific, literary or artistic productions. In the context of translation, the term refers to the authors of the source and target

text, unless otherwise stipulated by specific agreements or contracts.

Less resourced or low-resource language:
A language can be considered a less or low-resource language when it is less studied, a minority language, a less privileged language or a language for which few linguistic resources such as training data are available (Palmer, 2011).

Machine Translation:
The process of automatically translating textual or audio content from one language to another. Synonym: automated translation.

Metadata:
Data about the data, i.e. structured description of a data set with its properties (e.g. title, author/publisher, description of the content, size, topic, IPR holder etc.)

Named Entity Recognition:
Models or systems enabling the extraction of information from an unstructured text and its classification into pre-defined categories, such as person names, organisations, locations, values, etc.

Open Data:
Refers to data which is open in terms of: access, redistribution, reuse, absence of technological restriction, attribution, integrity, no discrimination (cf. European Data Portal, 2017, p. 7f).

Public Sector Information:
Is information generated, created, collected, processed, preserved, maintained, disseminated, or funded by or for the Government or public institution.

Question Answering System:
System built to retrieve the answer to a question from a knowledge base, such as a structured database, but also an unstructured collection of natural language documents.

Relation Extraction:
Models or systems predicting semantic relationships between the entities in a sentence. The extracted relationships usually occur between two or more entities of a certain type (e.g. person, organisation, location) and fall into a number of semantic categories (e.g. married to, employed by, lives in).

Small and medium-sized enterprise:
Business whose personnel numbers fall below certain limits. Its delimitation is based on the definition of the EU recommendation 2003/361 (EU, 2003).

Speech Synthesis:
Technology converting a machine-readable text into a sound imitating the human voice. It is becoming increasingly popular in assistive systems. Synonym: Text to Speech (TTS),

Textual data:
This term refers to systematically collected material consisting of written, printed, or electronically published words, typically either purposefully written or transcribed from speech or from other modalities, e.g. sign languages (Benoit, 2011).

Translation Memory:
A database of previously translated text segments (i.e. sentences, paragraphs, headings etc.). A Translation Memory stores the source segment and its corresponding translation, the target segment, in pairs. These pairs are called “translation units” (TUs).

1. Introduction

The first edition of the ELRC White Paper was published in 2019 and titled "Sustainable Language Data Sharing to Support Language Equality in Multilingual Europe – Why language data matters". Together with the ELRC National Anchor Points from all EU Member States, Iceland and Norway, European practices for sharing language data as well as the related challenges were investigated and recommendations on how to address these challenges in the future were prepared.

In addition, the first White Paper edition provided a country profile for each of the CEF-affiliated countries, which focuses on the following topics:

- National translation practices and information exchange in ministries and public administrations
- Translation needs of the country
- Language data creation and sharing infrastructure
- National open data policies
- Key stakeholders
- Main challenges for sustainable data sharing
- Required actions to overcome the identified challenges

While the initial scope of the White Paper was to report on the practices, challenges and recommendations for sustainable language data sharing within public services, this new publication provides an extended analysis, which also addresses European SMEs and which gives insights into the use of additional AI-based language tools, such as anonymisation or named entity recognition. Besides that, the White Paper at hand includes an analysis of the role of LT and language data in all EU member states, Iceland and Norway and critically discusses if the value of LT and language data has been recognised or if further awareness-raising actions are required.

The updated country profiles – in addition to their original contents – also provide some insights into:

- The role of LT and language data in each country's AI policies
- Major AI networks, projects and players in the particular country
- Data collection efforts and repositories in the country

In consequence, the new title of the white paper is shorter but broader at the same time:

"AI for Multilingual Europe –
Why Language Data Matters".

ELRC's vision has always been to contribute to a true digital single market where all EU citizens can access information independently of the language they speak.

That is also one of the main reasons for collecting language data: language data is the fuel for the development of all the LTs which are increasingly used in our daily lives, thus helping overcome language borders. Such LTs go far beyond automated translation solutions such as eTranslation as evident from the ongoing extension of the EC's Language Tools[2].

Thanks to recent advances, AI can help us address societal challenges related to the environment, health or crisis response, for example. In addition, it allows us to communicate across borders with the help of Machine Translation, to dictate text messages on our mobile phones, using speech recognition and to verify information sources through fake news detectors – to only name a few. It is safe to say that the possibilities are endless and that society has become more open towards exploring them.

[2] https://language-tools.ec.europa.eu/

2. Methodology

The results presented in this white paper have been obtained through numerous actions, namely:

- ELRC Country Workshops
- In-depth analysis based on AI Watch
- Dedicated ELRC White Paper Survey

ELRC Country Workshops

In collaboration with the National Anchor Points, ELRC organises one local workshop in each of the CEF countries, which targets national representatives from the public sector, LT industry, academia, research as well as SMEs with multilingual needs. During the event, participants exchange their experiences and discuss possibilities and requirements for transforming digital interaction in multilingual Europe with the help of LT. In addition, country workshops provide insights into the status and prospects of LT for their official language and discuss how language data can fuel development in AI.

The outcomes of the related discussion rounds, panels and feedback forms were used to shape the contents of the second white paper and to update the country profiles provided in the annex.

In-depth Analysis based on AI Watch

Taking the *AI Watch – National strategies on Artificial Intelligence: A European perspective* (Van Roy et al., 2021) as a starting point, the National Anchor Points and the ELRC consortium analysed how LT and language resources are currently represented in the national AI strategies and which initiatives and activities might be missing to boost the development of language-centric AI in Europe.

The reasons for that were two-fold: An in-depth analysis of LT aspects in the AI strategies could be a useful resource for Member States' policy makers to help them compare their strategies to those of other countries. In addition, it aimed to support the identification of potential areas for collaboration as well as good practices and common strengths in LT on which the EU can reinforce its position for developing AI-based LT.

More precisely, the following information was collected for each CEF country and discussed at the 11th LRB Meeting[3]:

- AI-related LT projects and initiatives
- Available AI funding for LT
- Major LT players in AI
- LT policies
- Data collection efforts/repositories for LT/AI

This was complemented by a detailed analysis of the national AI strategies to be able to assess the visibility and value of LT and language data in the national policies. At the time of the analysis, 24 out of the 29 analysed countries had already published their national AI regulation, while Belgium, Croatia, Greece, Iceland and Romania were still work in progress.

ELRC White Paper Survey

The investigations were round off with the ELRC White Paper Survey, which aims to find out more about:

- The current use and importance of LT
- Common European practices with respect to translation, data management and sharing in public administrations and SMEs
- The contents of national policies and regulations related to LT and AI
- Ideas and priorities to facilitate data sharing and LT development for Europe's multilingual future

The survey was completed by the NAPs, whose feedback set the basis for an in-depth

[3] https://lr-coordination.eu/11thLRB

comparison between the status quo in 2019 and in 2022 (see sections 3.3, 3.4, and 3.5). In addition, external contributors from various sectors, including European SMEs and LT Industry, were invited to participate in the survey, so it was possible to get a broader picture of the overall situation in the EU Member States, Iceland and Norway. In total, 73 people participated in the survey; the distribution is illustrated in Figure 1 below. The complete questionnaire can be found in the Annex section.

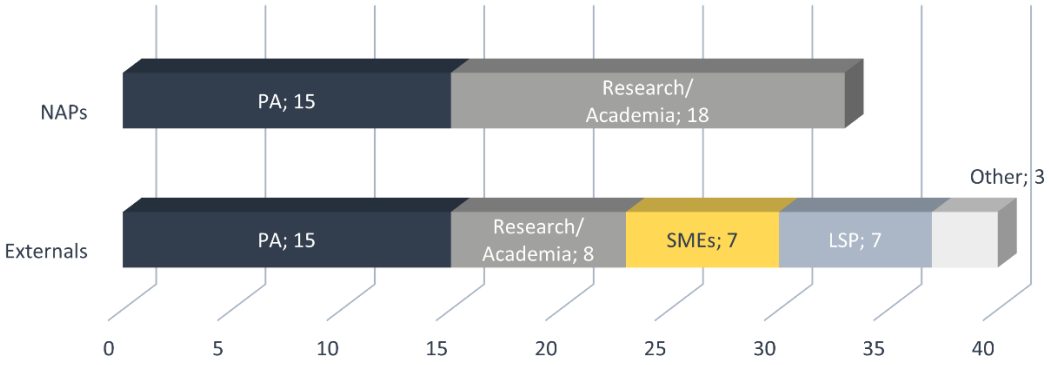


Figure 1: Distribution of survey participants by sector

3. Language-centric AI: Value and Status quo

What is AI? In simple words, Artificial Intelligence can be described as a collection of technologies that combine data, algorithms and computing power[4]. According to the European Commission's high-level expert group on AI[5], the term refers to

> *"systems that display intelligent behaviour by analysing their environment and taking actions – with some degree of autonomy – to achieve specific goals".*

Such systems can either be entirely based on software and exist only in our digital world – such as voice assistants or search engines – or be embedded in hardware devices, like e.g. drones or autonomous cars. While the concept of AI has been existing for decades already, it has become one of today's top trends only recently. Thanks to the growing attention and intensified efforts invested, AI is now advancing rapidly and applies to a variety of fields, including health care, manufacturing, administration – or cross-border communication.

3.1 The value of AI

The ability to communicate and share information across languages and borders has also greatly benefited from the ever-growing popularity of AI. There were major breakthroughs in language processing technology, leading to network architectures that can learn from complex and context-sensitive data. Unlike traditional Machine Translation models that focus on word-by-word or phrase-by-phrase translation, neural Machine Translation (NMT) is now capable of translating entire sentences at a time and of predicting the likelihood of a sequence of words, using deep learning techniques.

Naturally, Machine Translation is only one example of how deep learning can work. Deep learning however applies to various Language Technologies, such as:

- Automatic Speech Recognition (ASR),
- Text to Speech (TTS) systems,
- Dialog systems/Chatbots, Question-Answering (QA) systems,
- Named Entity Recognition (NER),
- Relation Extraction,
- Text Anonymisation,
- Sentiment Analysis, etc.

> *"AI is not a technology of the future, it is a technology of the present"*

This statement from the Irish AI Strategy[6] clearly reflects the growing awareness on today's value of AI. And it is not only true for Ireland, but also for each and every EU country. In addition, the topic has become increasingly important on EU level – and an essential part of our daily lives:

According to the AI Watch Report[7] and latest investigations by ELRC, 24 of the 29 EU Member States, Iceland and Norway have already published their national AI regulation. The remaining AI strategies are already work in progress. This clearly reflects the growing awareness on the usefulness of AI on a national level.

On EU level, the increasing value of AI is reflected by numerous new initiatives and

[4] https://ec.europa.eu/info/sites/default/files/commission-white-paper-artificial-intelligence-feb2020_en.pdf
[5] https://ec.europa.eu/digital-single-market/en/high-level-expert-group-artificial-intelligence
[6] https://www.gov.ie/en/publication/91f74-national-ai-strategy/
[7] https://publications.jrc.ec.europa.eu/repository/handle/JRC122684

projects[8], such as AI4EU, the Face2Face Virtual Agora on EU Artificial Intelligence Centres, etc. Following the OECD.AI Observatory, there are currently 59 AI initiatives in the EU[9], that underline the importance of Artificial Intelligence all across Europe.

There are also very prominent examples of AI in our daily lives, including digital personal assistants (e.g. Siri or Alexa), intelligent cars, chatbots (e.g. in banking or customer support), finance, health care or agriculture.

These examples show that with the advent of AI and NMT and the increasing awareness on EU and national level, a new paradigm and new possibilities emerged. They lay the foundation for safe and powerful technological solutions for the future of a connected and open Europe, where everyone can make a difference, speak up and be heard.

3.2 Today's use and value of LT

Language is a central element of our daily lives, it is part of our identity and culture. While the coexistence of languages is celebrated as one of the core values of the European Union, it can also create barriers for communication and hinder the free flow of information. Language technologies such as Machine Translation systems have become a key enabler for building bridges not only between citizens but also between governmental institutions or industry.

This is proven by the fact that in 2021, more than 204 million pages were translated with eTranslation – which is more than the double of the previous record of almost 95 million pages translated in 2019.

Even more, recent advances led to additional LT tools offered by Europeans for Europeans, such as those offered by the European Commission[10]. Over the last three years, not only the range of languages and engines supported by eTranslation has been continually expanding (including now, among others, Arabic and Ukrainian), but also a number of tools have been made available, such as speech to text, named entities recognition, classification, and automatic text anonymisation.

Since the COVID-19 pandemic, the importance of language-centric AI has significantly increased: Computer-mediated communication became the new and in many cases only modus operandi during the crisis and corresponding LT provided valuable tools and services to facilitate the virtual information exchange. Moreover, LT was also found to be vital in facilitating communication in times of crisis[11]. In consequence, the significant changes in the way we work combined with recent advances in LT thanks to AI contributed to new trends and a greater availability and uptake of language-centric AI in general. Potential use cases of language-centric AI range from solutions to detect disinformation, automated live interpretation of news to chatbots that provide citizens with information about COVID-19 and answer their questions, just to name a few.

3.3 The use of LT in public administrations and SMEs

It is safe to say that Machine Translation has found its way into the daily work life of public administrations. This already became evident in the analysis of 2019, where 38% of the contributors indicated that part of the translation agencies in their countries are making use of MT APIs. The remaining 56% answered that they do not use MT APIs but freely available MT web services during their work, while only 6% didn't use MT at all.

[8] https://digital-strategy.ec.europa.eu/en/news/eu-funded-projects-use-artificial-intelligence-technology
[9] https://oecd.ai/dashboards/countries/EuropeanUnion
[10] https://language-tools.ec.europa.eu/
[11] https://lr-coordination.eu/node/453

The White Paper Survey 2022 confirmed this point, as the majority of the contributors keep using freely available MT services (47% in 14 countries: Austria, Bulgaria, Croatia, Czech Republic, Denmark, Germany, Greece, Italy, Luxembourg, Malta, The Netherlands, Poland, Portugal, Romania). However, the use of MT APIs has slightly changed, as 41% of the representatives indicated that at least some of the translation services in their countries have an MT API integrated into the translation process (11 countries: Belgium, Cyprus, Estonia, Finland, Hungary, Latvia, Lithuania, Norway, Slovakia, Slovenia, and Spain). Only two representatives (6%) indicated that MT APIs are used by most translation services (in France and Sweden), while two contributors (Ireland, Iceland) indicated that MT is not used at all (see Figure 2 below).

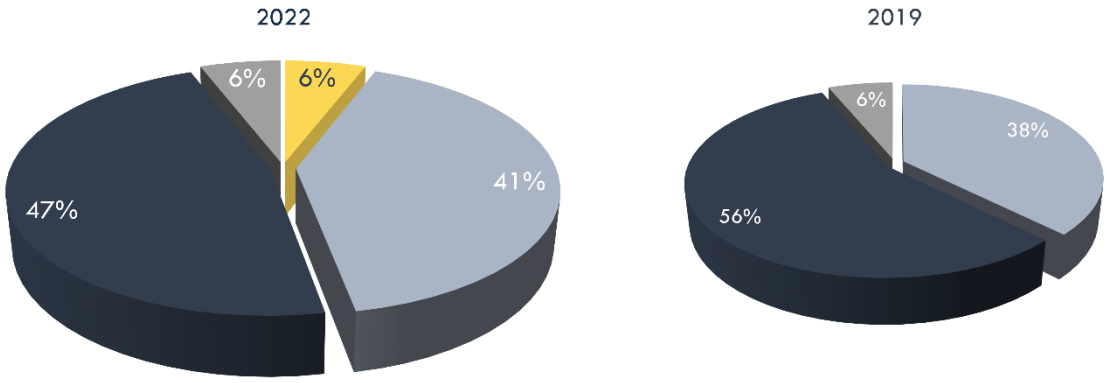


Figure 2: Use of MT APIs in Public Administrations

The answers of the external survey contributors show the same trend: also in this case, the contributors using freely available MT web services are the majority (40%), but the percentage of contributors indicating that the translation services in their countries have MT APIs integrated into the translation process is also noticeable (45% – quite equally divided between most and some translation services). However, at the same time 15% of the contributors indicated that they do not use MT at all (in 4 countries: Germany, Greece, Romania, and Spain – see Figure 3 below).

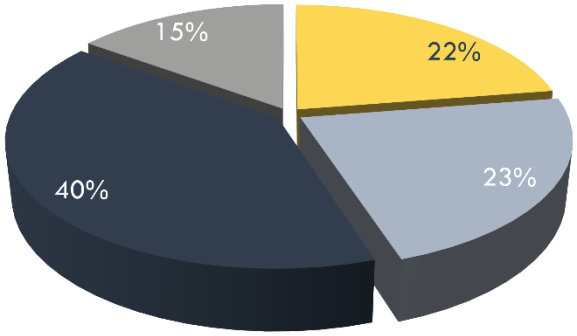


Figure 3: Use of MT APIs in SMEs

These high numbers with regard to the use of freely available MT web services confirm the urgent demand for easily accessible and easily integratable translation solutions that facilitate secure work in a multilingual environment while offering a satisfying quality.

While there were no major changes concerning the use of Machine Translation over the last three years, many other tools gained popularity and some of them are now even used on a regular basis.

According to the findings of the White Paper Survey, this applies e.g. to Classification (24%), Anonymisation (17%), Speech Recognition and Text to Speech (each 15%, see Figure 4 below). These technologies are very often used by 6% of the survey participants (from Bulgaria, Cyprus, Denmark, Finland, Germany, Greece, Hungary, Portugal, Spain, and Norway), but on the other hand, there are still 8% of the participants (from France, Ireland, Lithuania, Luxembourg, Romania, Slovakia, and Slovenia) who indicated that LT are not used at all in their countries.

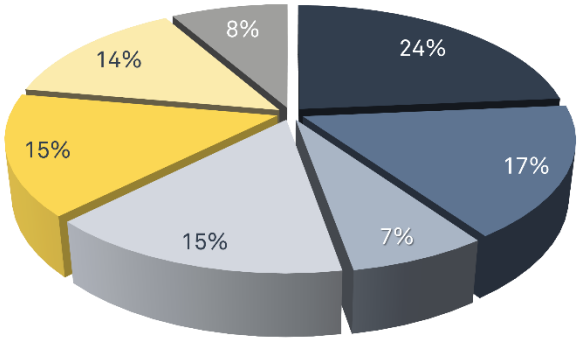


Figure 4: Use of language tools in PA and SMEs

All of these tools have two things in common: they can facilitate our daily multilingual operations – and they are trained with language data!

In line with that, we could see a massive increase in the use of Computer-Assisted Translation (CAT) tools. While in 2019 it was not a standard practice to use them in public administrations, this seems to have changed: the percentage of the representatives indicating the use of CAT tools as common practice for LSPs jumped from 24% to 41%. This can be seen as a great success, because CAT tools are critical for the creation of high-quality multilingual data and therefore a huge asset to the LT community.

As for the external survey contributors, this trend seems to be even more noticeable: 65% of the participants indicated that all LSPs and freelance translators make use of CAT tools, while only 15% stated that CAT tools are not used at all.

Overall, this leads to the conclusion that the use of LT is increasing and is no longer limited to Machine Translation, as more and more organisations have recognised the usefulness of additional Language Technology tools to facilitate their daily operations.

However, despite the increasing popularity of LT and the fast progress in its development and use, there is still room for improvement on policy level, as the value of LT is not reflected in all national AI regulations as we will show in the following section.

3.4 LT in national regulations

The topic of Language Technology is included in 21 out of the 24 national AI strategies published until now. However, the visibility of and emphasis on the topic varies greatly. While some regulations dedicate complete chapters or action pillars to Language Technology (e.g. Denmark, Malta, or Norway), others only mention it in a side note about useful AI application areas (e.g. France, Latvia, and Portugal). Countries where LT is not mentioned explicitly are Sweden, Estonia and the Netherlands. However, it is also important to find out whether there are any complementing policies and regulations, as it is the case in Estonia, for example, where LT is mentioned in the draft strategy for the Estonian language.

Examples of strategic documents and/or regulations covering the topic of LT are provided below:

- In **Spain**, the Plan for the Advancement of Language Technology underlines the importance of collecting and sharing language data as a means to "foster the natural language processing and Machine Translation sectors". Above this, a "New Language Economy" plan was announced In March 2022, aiming at mobilising public and private investments in order to maximise the value of Spanish and the co-official languages of the country [...] towards a global level.[12]
- In **Hungary**, the importance of the development of LT for the Hungarian language is highlighted, with the aim of integrating it in all customer service processes ("zero level support") to facilitate administrative processes.
- In **Bulgaria**, the use of LT is foreseen to support foreign language learning.

[12] https://european-language-equality.eu/2022/03/04/spanish-government-invests-into-new-language-economy/

In our analysis, the contributors were asked to rate the role of Language Technology in their country's language plan or AI policy.

Compared to 2019, the number of countries where LT is not mentioned at all has dropped dramatically from 36% to 9% in 2022 only in 3 countries: Austria, Portugal, and Romania. Unfortunately, this doesn't correspond to a similar increase in countries where LT is explicitly mentioned in the national AI strategies. However, LT is clearly no longer a topic that can be disregarded: the number of countries in which it is mentioned at least as a side note has exploded in 2022 from 7% to 37% (in Austria, Croatia, Cyprus, France, Finland, Germany, Ireland, Italy, Luxembourg, and Slovakia).

However, the goal should be to raise awareness on the usefulness of LT and to include the topic explicitly in all national AI regulations, ideally along with a detailed strategic plan, covering financial and structural matters.

3.5 The value of language data

For all LT applications, language data plays a crucial role. This is even more true when we consider the exponential growth of the digital communication platforms, which in turn increase the need for more efficient and reliable LT.

Organisations, however, can only collect the necessary amount of language data required for the development of competitive language-centric AI if they invest considerable efforts – both in terms of time and resources. For this reason, data sharing is increasingly considered as the best way towards a truly sustainable language data management.

Nonetheless, in many countries of the EU, the sharing of language data is still not common practice, even though tons of data are produced in public administrations, research and industry on a daily basis.

Against this background, ELRC started investigating the translation practices and common data management and storing procedures in public administrations and SMEs of the EU member states, Iceland and Norway. More precisely, the analysis focused on the following questions:

- Are translations produced in-house or outsourced?
- Are the translation memories (TMs) or other by-products of outsourced translations requested back by default?
- To what extent are European organisations storing language data like tmx files, translations, audio files, video recordings, etc.?

The key findings are summarised below.

Translation Practices

In 2019 and 2022, the National Anchor Points were asked how multilingual needs are being addressed in the public sector. Possible answers were:

1 More than 50% of translations carried out by language services and translation professionals in-house
2 More than 50% of translations carried out in-house by professional translators or bilingual/multilingual staff members
3 Only single ministries with in-house translation services, mostly outsourcing of translations through central purchasing body
4 Only single ministries with in-house translation services, mostly independent outsourcing of translations
5 All translations are outsourced via central purchasing body
6 Independent outsourcing of all translations

According to our latest results, most translations are still being outsourced, but the overall share has decreased (from 79% to 64%), reflecting a trend towards in-house translation (from 17% to 27%).

In particular, only the Slovenian representative chose answer 1. Answer 2 was indicated as predominant practice in 6 countries (Germany, Iceland, Latvia, Luxembourg, Malta, and Norway). Answer 3 holds true for Cyprus, the Czech Republic, Estonia, Hungary, Lithuania, and the Netherlands. Bulgaria and

Finland chose answer 5, Ireland answer 6. The Belgian representative didn't specify any preferred practice. All other country representatives indicated that translations are mostly independently outsourced (see Figure 5).

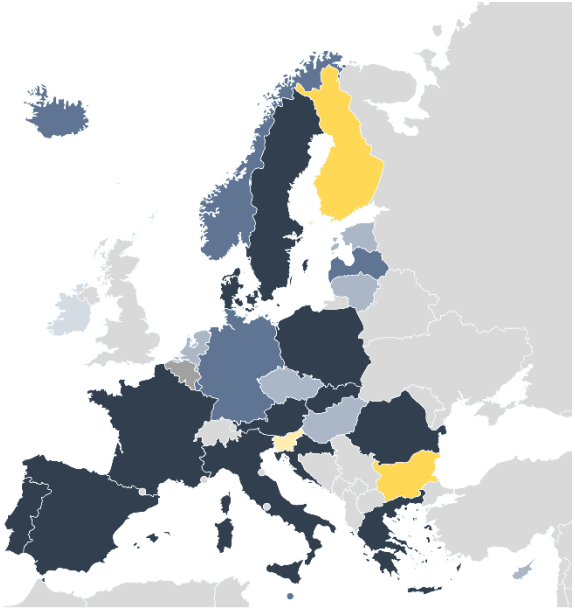


Figure 5: Translation practices in Public Administrations

Given that most of the translations are still being outsourced, it is also important to find out what happens to the translation memories or other by-products if the translation was produced outside the organisation. Translation memories (TM) are the desired language data input for MT systems, as they require little or even no preprocessing before they are fed into the MT system. Requesting them back is therefore an important step towards sustainable language data management.

As Figure 6 below shows, there is a clear decrease in cases where TMs or by-products are not requested back (from 50% to 32%). At the same time, the cases where TMs are "sometimes" requested back increased significantly (from 33% to 48%). Nonetheless, the overall percentage is still high, and it is still not common practice in Europe to request them back by default. In fact, this is the case only in 2 countries: Finland and Bulgaria (7%). 13% indicated that they request back TMs for "most" outsourced translations (2019: 17%).

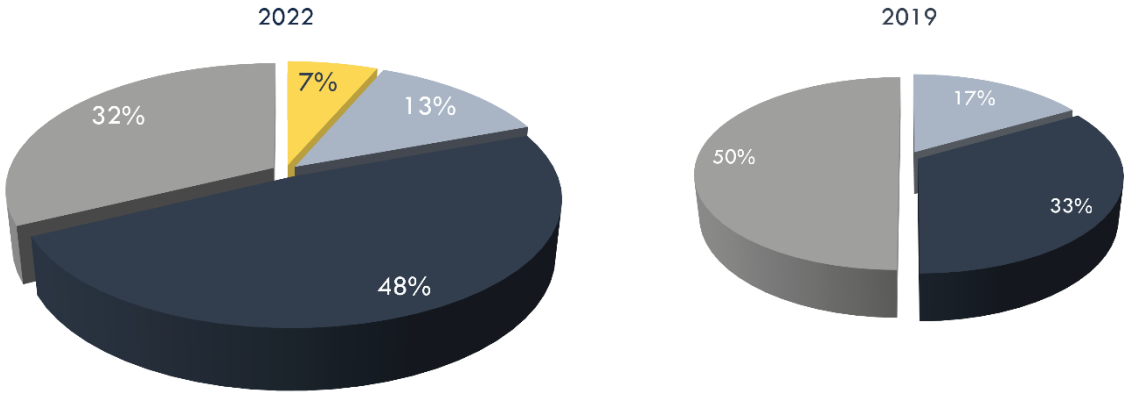


Figure 6: TM files and MT by-products in Public Administrations

Surprisingly, the external survey contributors have drawn a quite different picture. 30% of the participants (in 7 countries: Belgium, Bulgaria, Greece, Poland, Portugal, Romania, and Spain) indicated that they request back the TM files and other MT by-products by default. However, also in this case the percentage of contributors not requesting back these data remains fairly high (37%, see Figure 7 below).

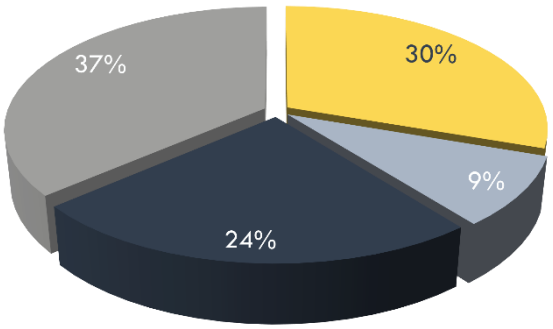


Figure 7: TM files and MT by-products in SMEs

Language Data Management and Sharing

According to the latest results, 17 country representatives answered that their organisations are storing language data whenever possible. This holds true for:

Austria	Belgium
Cyprus	Czech Republic
Finland	Germany
Greece	Hungary
Iceland	Latvia
The Netherlands	Norway
Poland	Romania
Slovenia	Spain
Sweden	

Only 4 country representatives indicated that language data is hardly or never stored in their organisation. This applies to Croatia, Estonia, Malta and Portugal.

The remaining representatives stated that in their countries language data is sometimes stored (see Figure 8 below).

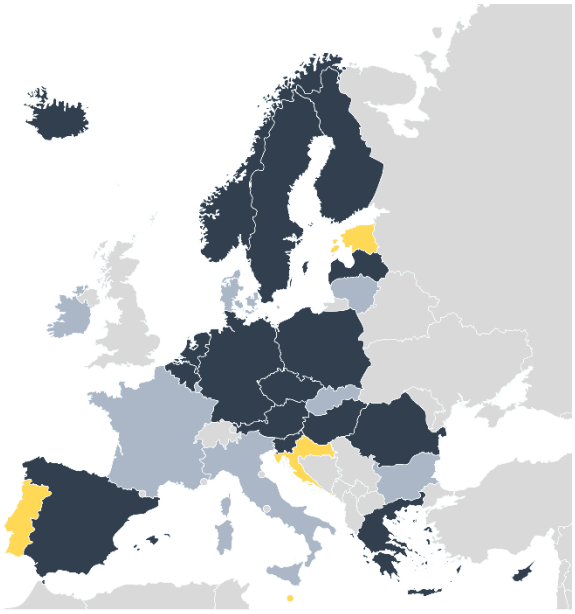


Figure 8: Storage of language data in Public Administrations

Similarly, the large majority of the external survey contributors indicated that language data are stored whenever possible in their organisation (59% in 9 countries: Bulgaria, Denmark, Finland, Germany, Italy, Poland, Portugal, Romania, and Spain), but the percentage of those who indicated that they hardly or never store such data is not minimal (19%).

On the one hand, this confirms that the value of language data is being increasingly recognised all over Europe, but on the other hand, it also demonstrates that there is still a need for more awareness-raising efforts – also on the part of the governments.

It is no coincidence that as we asked our representatives at the 6th ELRC Conference[13] whether they think that the value of language data has been recognised in their country, we could not establish a clear trend. While 36% of the participants answered that this is definitely the case, 32% think that the value of language data has still not been recognised, and further 32% could not give a clear answer.

This is also reflected in the results of the 2022 White Paper Survey: although 17 representatives know that language data are explicitly mentioned in the AI regulations of their countries, only 4 are aware of a corresponding strategic plan (Iceland, Lithuania, Norway, and Slovenia). Moreover, 6 representatives stated that language data are mentioned only as a side note – e.g. as useful example of AI (Austria, Bulgaria, Czech Republic, Ireland, Luxembourg, and Malta). Finally, 5 indicated that language data are not mentioned at all in the AI regulations of their countries (Belgium, Cyprus, Germany, Romania, and Portugal), while the Italian representative chose the answer Other (see Figure 9 below).

As for the external contributors, 32% indicated that language data are explicitly mentioned in the AI regulations of their countries, even if only 10% know about a corresponding strategic plan. 19% saw language data mentioned as side note, while 15% stated that language data are not mentioned at all.

[13] 30th March 2022 via Zoom: https://lr-coordination.eu/6thELRC

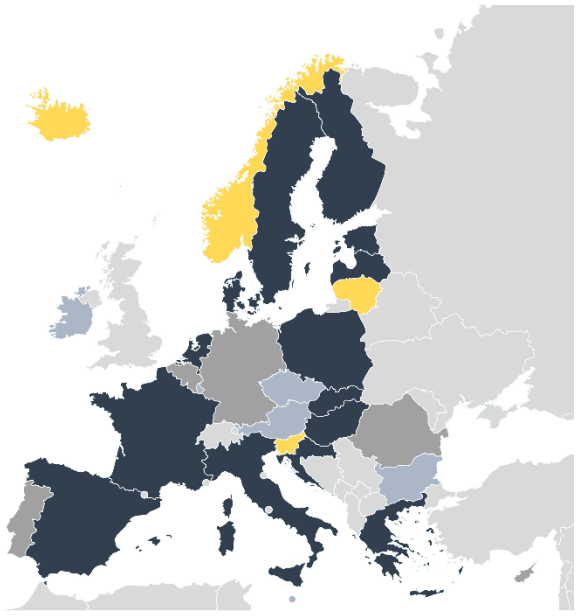


Figure 9: Role of language data in Europe's AI regulations

However, the external contributors from Spain gave partly different answers, so that it is probably true that there is a lack of communication/information when it comes to the country's official language plans. This also explains why the majority (34%) openly stated that they don't know whether language data are mentioned in the national AI regulations or not.

And in fact, following the AI Watch, only in 11 of the 24 national AI Strategies published until now[14] language resources are explicitly mentioned. This applies to Denmark, Finland, France, Hungary, Ireland, Latvia, Malta, Norway, Portugal, Slovenia, and Spain.

However, while this accounts for less than 42%, there are already numerous best practice examples that could be found in Europe, for instance:

- The Norwegian strategy includes a full chapter about LT and language data, which highlights the crucial importance of language resources, especially for the NLP systems targeting less-resourced languages like the Sami languages.
- The Spanish AI Strategy mentions boosting the National LT Plan and the creation of resources in the Spanish Language as one of their action items.
- In Ireland, the value of language data is publicised, because one of their action items is to move away from US-based language data and use sources that include everyday language used by Irish citizens. In addition to that, the development of language resources for Irish is mentioned as one of the key enablers to provide digital services in Irish.

Such developments reiterate that the value of language data has significantly increased and will continue to increase – both within organisations and in national regulations.

[14] See chapter 2, Methodology.

4. Latest developments and approaches to sustainable Language Data Sharing in EU public services and SMEs

4.1 Limitations and supporting activities

Several **circumstances** were found to **negatively impact or limit the sharing of language data** in Europe. These include above all several characteristics associated with each organisation, namely:

- The lack of recognition of the value of textual data and language data in general
- Lack of digital skills
- Lack of adequate language data management practices/plans
- Limited access to translation memories of outsourced translations

Last but not least, as in the 2019 White Paper, legal concerns (such as GDPR, copyright) were found to particularly complicate and limit the sharing of language data.

When looking at the analysis of the **actions that are considered most relevant to facilitate data sharing**, it is not surprising that in 2022 they have not changed much compared to 2019, with the difference between the first and second placing being minimal. Respondents from PA and SMEs considered the following activities as most important for fostering the sharing of language data in the future:

O2 Increasing interest in MT/LT as part of the national digital policy
O1 Raising awareness of language data as open data and a valuable asset
O5 Establishing good data management practices in organisations
O3 Tackling legal concerns
O4 Gaining access to outsourced translations

Figure 10 below illustrates how often these action items were mentioned in the top three positions. The overall ranking is given by the sum of these positions.

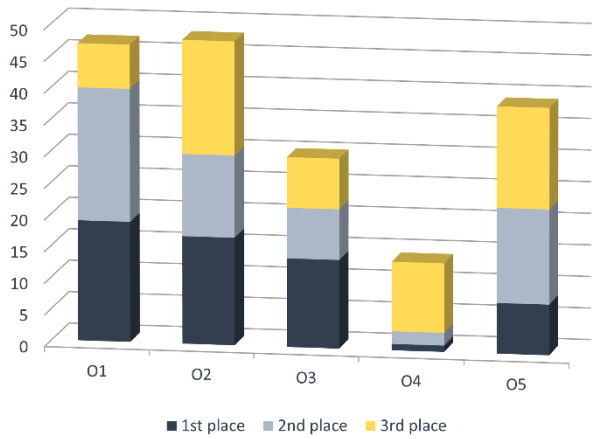


Figure 10: Objectives to facilitate data sharing: top-three ranking

However, 19 respondents also suggested additional approaches relevant for facilitating the sharing of language data. One major obstacle mentioned by the survey participants was that it is almost impossible for organisations to justify the resources and investment needed for collecting, cleaning, structuring and releasing language data (e.g. text corpora) especially with regard to the return of investment for these organisations. As such, other approaches mentioned to overcome this problem were centred around (i) making language data collection, preparation and sharing less costly and (ii) providing monetary support for activities targeted to language data sharing and collecting.

The provision and availability of **corresponding funding** for the sharing of language data was one of the most frequently made proposals by respondents, ranging from national public funding in the form of a language programme (as for instance in Spain), national funding programmes for LT, as well as LT-specific funding for the public sector, industry and research.

Another important suggestion that is in some way linked to establishing good data management practices in organisations is to

develop **central infrastructures** that facilitate data collection. Suggestions here ranged from "a centralised institution with a good overview on how language data is managed in the different ministries" to establishing "national or language nodes" that manage language data in a centralised way. Even a top-down approach and corresponding legislative obligations with regard to the sharing of language data in the public sector were mentioned. Such infrastructures would also need to "safeguard [...] that languages other than those official in the EU (e.g. Catalan) will have full rights and a level playing ground to remain an active part along these avenues of technological development/evolution."

Last but not least, the **provision of relevant tools and technologies** to support the sharing of language data was considered important for minimising the obstacles. For public services, the availability and accessibility of easy-to-use tools for generating and assembling large corpora from resources is of utmost importance. Like this, high-quality bi-/multilingual corpora could be built directly e.g. by translators or out of existing available texts. There was also the suggestion to create custom text corpora directly from a text cloud.

Another important tool / technology mentioned concerned the support needed for easily and quickly anonymising language data to make them sharable. For instance, one of the respondents requested to provide"de-identification tools at user's end, so that with one click of a button, the in-house data are cleaned, with one browse of a 'list of suspected issues', the de-identified documents can be reviewed, and the user is then not afraid of sharing [the language data]."

All these suggestions could indeed help overcome current limitations in organisations that want to share their language data.

4.2 Recent advances

Survey participants were also asked where they saw "recent advances related to LT development, digitalisation and data collection" and whether they noticed "any big changes compared to the situation three years ago". The responses confirmed that since 2019, LT as a whole, digitalisation and data collection have progressed significantly. Among the 74 responses received to this question, the majority mentioned three central developments:

1. Greater availability of language data;
2. Better LT – covering not only MT, but also Natural Language Processing (NLP) and Natural Language Understanding (NLU) in general;
3. Increased uptake of LT.

One respondent rightly summarises: "The single biggest difference is the use of large language models, and the finished transfer of practically all MT development to Deep Learning. In addition, speech technology (ASR/TTS) is used [...] as the ASR and TTS quality improves. [...] [Much] **more data is available** in general (for example, in the UD collection, the number of languages grew from 70 to 130+ and the number of treebanks to more than 200."

In Croatia, for instance, significant **improvements in LT** were also explicitly noted: "In the last three years we have built a high-quality NMT-based MT system that outperforms Google Translate in both (hr->en, en->hr) directions for several BLEU points. This has been achieved through the support from the CEF project "EU Presidency Translator".

The improvements in MT and LT lead to a **greater uptake of LT** not only in the participating public administrations but also in general.

Also, several activities mentioned by participants underline the **great advances** that have been made **with regard to the sharing of language data** in the past three years in Europe. For instance, in Lithuania, there is a "hope that the accepted Lithuania's Recovery and

Resilience Plan (2022-2026)[15] will stimulate the creation of AI-oriented language resources, which will stimulate the creation of quality language services." In Finland, the "Donate Your Speech Campaign" was launched in 2020 "to help researchers and application developers to create better working Finnish language AI by gathering speech donations from Finnish speaking people". At the same time, the National Library of Finland created Annif – an open source toolkit for automated subject indexing and classification which is based on a combination of existing natural language processing and Machine Learning tools.

In Estonia, a new regulation of the Minister of Education and Research – "The list of language data and the conditions and rules of publishing and reusing language data" is being established on the basis of the Public Information Act (PIA). As a matter of fact, not only Estonia but also Slovenia were found to excel in a recent investigation by the European Commission on "Open Data Best Practices in Europe"[16]. In 2020, Slovenia, for instance, funded a large initiative called "Slovenscina"[17] that supports the development of Slovene in a digital environment. In Malta, a "National Language Technology Platform"[18] is currently under development.

In France, similar developments took place, with respondents observing "better data collection and sharing as well as enforcement through national regulation and guidelines. [...] There was also a major national plan for the development of AI which included some AI language parts, especially on the data collection side (official public collection of French question-answer type data sets for training AI for example)." In 2021, the French Prime Minister Jean Castex announced a renewed open and shared data strategy following a 2020 Report by MP Eric Bothorel which includes data and data sets for AI.

Another success story mentioned by respondents is the "Smart Industry" approach in the Netherlands "which started with a handful of field labs and grew into a nationwide network of 46 field labs and five regional hubs. At the end of 2019, the government launched its "Strategic Action Plan for AI (SAPAI)"[19]. The associated AI Coalition has grown, the government has made extra funding available (a budget is allocated to the NL AI Coalition in the context of the so-called Groeifonds funding), and the Netherlands are working at national, European and global levels to strengthen the Dutch AI ecosystem. The government recognises that a flourishing data economy is essential in order to achieve this. Based on the data sharing vision, the government is facilitating voluntary data sharing between sectors through the so-called "Data Sharing Coalition"[20] in order to utilise data capabilities for the benefit of the economy and society in a responsible way."

Based on the National Strategy for Artificial Intelligence from 2019 in Denmark, the so-called "Digitisation pact" set out to develop "A Common Danish Language Resource"[21]. The purpose is to support Danish Language Technology companies in developing Danish-language solutions within AI. As one respondent explains: While before, "Danish language data was scattered across various organisations and hence locating and esti-

[15] https://ec.europa.eu/info/business-economy-euro/recovery-coronavirus/recovery-and-resilience-facility/lithuanias-recovery-and-resilience-plan_en
[16] https://data.europa.eu/sites/default/files/report/Open_Data_Best_Practices_in_Europe_Estonia_Slovenia_and_Ukraine.pdf
[17] https://www.slovenscina.eu/en
[18] https://aclanthology.org/2021.mmtlrl-1.3/
[19] https://www.rijksoverheid.nl/ministeries/ministerie-van-economische-zaken-en-klimaat/documenten/beleidsnotas/2019/10/08/strategisch-actieplan-voor-artificiele-intelligentie
[20] https://nlaic.com/en/partner/data-sharing-coalition/
[21] https://en.digst.dk/policy/new-technologies/a-common-danish-language-resource/

mating the value of language data was a time-consuming process, especially for SMEs, [...] this problem has largely been resolved through the establishment of "A Common Danish Language Resource".

The emergence of Open Data Portals supporting the collection of data is also seen as a positive development: "The best example of data collection in Bulgaria still is the Open Data Portal. There are 10,563 data sets and 536 organisations which are sharing their data in 14 main thematic areas." Moreover, following the proposal of the Bulgarian Public Services National Anchor Point, the process for outsourcing translations was adapted: It is now a technical requirement within the public procurement process that the selected translation agencies provide their translation memories stored in CAT tools for the outsourced translations.

Similar developments happened in Germany: "At federal level, language services lead a project [...] to explore whether and how Machine Translation could be introduced in the public administration, including an extensive survey on the current and potential use of MT in Germany's federal administration and translation needs in general". This would also support the language data creation and sharing process among the different organisations involved.

Important changes, however, have not only happened with regard to new policies and new initiatives as the investigation proves. A good development also seems to be the fact that "the major publication venues now request that both data and source code are made available upon publication of research results, which is a major improvement not only with respect to the replicability of results, but also to facilitate[ing] follow-up research and avoid[ing] work being done over and over again".

Respondents also acknowledge the contribution that ELRC made in this respect: "The pan-European data collection campaigns (e.g. ELRC Workshops) represent a major breakthrough that will change the general attitude towards the preservation of digital textual data, mono- and multilingual."

Most interestingly, several participants also assume that the pandemic has "substantially accelerated a number of developments" in this respect.

With regard to the near future, it is expected that "tools for fake news detection and anonymisation will be a game-changer". Explainable AI is expected to be a major challenge in the near future.

4.3 New challenges

Survey participants were asked what the currently biggest challenges related to LT/LR were in their country. 44 responses were received to this open ended question, identifying six major challenges in Europe:

- The availability of high-quality language data
- The development of LT for European minority languages
- Limiting legal provisions and regulations
- Data management / continuous process for sharing language data
- Continuous funding and support
- Human capital

Availability of high-quality language data

Several respondents claim that "[the] availability of data is still the biggest concern." This particularly refers to the lack of parallel corpora to improve bi- or multilingual MT. However, it was also noted that the "existence, availability and collection of massive language resources (e.g. 100+ GB of quality plain-text) for training large, pre-trained language models (like GTP-3)" is still something Europe lacks as well as the "collection of domain specific LR for the adaption/finetuning of [multilingual] pre-trained models." Similarly, it was noted that "convenient and well-regulated access to public data" is still an unresolved challenge. One key issue certainly is the persistent problem of "convincing data holders to make their data available".

Development of LT for European minority languages

More than half of the respondents also point out that especially less spoken languages in Europe require substantial support in the future. Some of the EU official languages like Finnish, Czech, Croatian, and Swedish are less spoken languages, as well as some regional languages like Valencian, Basque, Galician and Catalan in Spain. According to the survey responses, "[t]he actual technology works pretty good with English, French and Spanish but it is not yet prepared to deal with minor official languages." Another respondent even claims that the "development of LT for national minority languages is neglected."

As a different respondent explains: "Finnish and the indigenous Sámi languages are small languages and not a priority for big AI developers. The availability of high-quality and open language data" for these languages hence is not to be taken for granted. In other cases, the situation is also considered as a result of a biased language policy, as another survey participant explains: "In Spain, for instance, there is one official language (the Spanish), and also there are three co-official languages (Euskera, Galician, Catalan). The national plans and efforts [however] tend to be in Spanish and that fact puts in disadvantage the support and development of LT for the co-official languages." Nonetheless, the lack of language data also applies to EU official languages as the comments of several other survey participants show. For instance, "[t]he size of Czech resources, even if covered quite well overall, is still well below the major languages." For Croatian, speech processing is "seriously lagging behind other languages" because of the lack of corresponding language resources. For Bulgarian too it is noted that "[m]any commonly used and necessary technologies are still not available".

The lack of language support also becomes apparent for different variants of a language as the following comment underlines: "For application and development of LT tools at national level, the distinction between Dutch as used in Belgium and Dutch as used in The Netherlands is important. This distinction is needed as both countries have their own terms for specific concepts. This distinction may not be important at European level, but it is important at the national level. It would be good if data repositories could include this information in the metadata to increase reusability of the data." The situation applies also to German or, for instance, Spanish as the survey illustrates: "Another important problem is the lack of diversity in the LT Spanish data set. Spanish is a very spread language with many local variants and accents. However, there are few data sets that collect that diversity."

As such, "[c]reating high-quality resources for all languages, not only [for] the mostly spoken ones, but also for low-resource languages – this is currently the next challenge for overcoming the digital divide for the speakers of these languages and assuring equity in the access to digital services", as one respondent summarises. And it is apparently also the place, where LT can make the greatest difference: "[I]t's for the minority languages where LT could make the most impact [because] there simply are not enough translators or minority language speakers" to support "manual" translation or direct interpretation in situations needed (e.g. also in daily settings in public services and the healthcare sector).

Last but not least, it was pointed out that "we lose so much content and meaning" if we try to always resort to English – apart from the fact that non-native speakers are never perfect at speaking English.

Limiting legal provisions and regulations

15 respondents also mentioned persistent legal issues as a key challenge for the upcoming years. This includes in many cases problems of anonymising language data and/or being GDPR compliant, as well as adequately respecting copyright of potential language data. One respondent summarises: "Data sharing is further complicated by

unclarity and uncertainty on how to comply with the GDPR framework. There is also a certain contradiction: Working on open science can be at odds with sharing data with commercial bodies."

Another respondent also acknowledges the prevalence of legal issues that hinder the sharing of language data while at the same time proposing one way out of the dilemma at least with regard to copyright: "Legal concerns are of the greatest importance. Data should be treated as material, not as copyrighted work so it can be gathered and processed to improve language models without any legal risk."

As regards the issue of personal data, the following suggestion was made by another participant: "Data sharing can be improved by providing tools at national or European level for e.g. automated anonymisation / pseudonymisation, such that everyone uses the same tools with the same quality instead of different tools with different quality."

While issues of copyrighted and/or personal data can be overcome in the near future, the issue of confidential data remains: "LRs often contain confidential data which cannot be released even if copyright and personal data issues can be overcome." However, since confidential data represent only a very small fraction of the language data produced, this does not constitute a significant problem.

Apart from the legal issues, respondents also point out that the "lack of legislative obligation to make publicly funded data available" significantly hinders the sharing of language data. So a general requirement to clean, anonymise and share language data could greatly improve the amount of available language data in the future.

Continuous funding and support

Several respondents also mention that unstable funding for enabling the provision and sharing of language data is a major problem. This does not only apply to the research sector, but also the public sector and companies. Securing stable and consistent funding could lead to a significant increase in available language data. As one respondent claims, "Without initiatives such as the ELRC and ELG, there would be no events or funding for LTs/LRs". On the other hand, the National Language Technology Programme in Iceland illustrates how national funding and support can make a difference: "Since 2019, Iceland has a National Language Technology Programme. Within this programme, a great number of LT resources have been developed, including resources for speech technology and MT."

Data management / continuous process for sharing language data

In some way linked to the challenge of having funding for the sharing of language data is the appropriate data management and/or set-up of a continuous process for sharing language data. 14 survey participants address this challenge in their responses.

One frequently mentioned problem is the lack of coordination between different entities in the public sector: This can be for instance the "[l]ack of coordination between translation departments from different public and local authorities", but also "the lack of collaboration between the ministries" or even within an organisation, as the following respondent explains: "The biggest challenge is – still – to set up a continuous process of preparing and sharing language data. While awareness of the importance of language data (in particular among translation services) has been established, finding the willingness and resources to share the data is still difficult. One reason is that as a "service provider", translation professionals are not involved in AI/LT projects planned in other parts of the public administration." As such, it is not surprising to find that respondents "do not perceive any structured approaches towards creating a multilingual environment." Another participant illustrates: "A major obstacle to data sharing is that there is no organised or centralised exchange of

language data at national level. There are no clear roles and responsibilities at the different levels. Maybe a separate ministry for Digital Affairs is needed." In Denmark, where "data was scattered around society and were in the hands of different organisations", the problem could be solved through the creation of "A Common Danish Language Resource" (see above, 4.2: Recent advances).

Similar improvements were reported from The Netherlands, where "[p]ublications from public administrations can be found on officielebekendmakingen.nl, overheid.nl and data.overheid.nl, coordinated by KOOP (the Dutch publications office)." As such, it can be concluded that still, "[a] stable data governance and shared practices are required in the public sector" for the successful sharing of language data. Directly linked to this is the need for "further digitisation of public services which requires the use of language and Language Technologies."

However, the problem is not only limited to the public sector as the survey reveals, but also to other areas, including research: "Further challenges are posed in research contexts where there are not always clear guidelines as for where to deposit data to comply with rules for the management of research data."

As such, "[c]ollaboration between different platforms and infrastructures is key to future service provision." Similarly, standards and in particular "standards for metadata are essential for sharing data" in order to avoid problems with the retrievability of language data and translations. Also, internal processes need to be adapted and persistently implemented as the following feedback illustrates: "What I mean is that people and processes constantly change. An optimisation [...] discovered by a worker is not preserved and reused if the agenda is handled by a new worker. Data collected on the go are not made available to subsequent tasks of similar kind. There is no "life net" in organisations that would make sure the LT knowledge and resources are preserved." So there is still a lot of room for improvement and the necessity to "increase the capacity for innovation in public services [and elsewhere], which has yet a long way to go".

Human capital

5 respondents also mentioned issues related to human resources as a major challenge. Specifically, they mentioned the "lack of competent specialists" and the "difficulty to find the right persons" for the task. The problem is also realised in academia where "there is limited institutional capacity for supporting researchers".

Other challenges

In addition to the aforementioned six major challenges frequently identified by survey respondents, a number of individual challenges were additionally reported:

- One respondent states that "a major challenge for future AI and LT is being able to truly understand the meaning of texts." This is similar to the experience of another survey participant who explains that "Machine Translation quality still varies a lot depending on the text; sometimes it contains critical mistakes in terms of content." As such, survey respondents conclude that "[t]he overall objective is to have methods, algorithms and ready-made system(s) for full Natural Language Understanding. Whether it is done by Deep Learning alone or in combination with symbolic methods and/or databases is not that important, but data is certainly important. Identifying gaps in technology and data is the next important goal. It is still not clear which applications are or are not possible now and in eight years time with current technology, or which improvements are possible with incremental development and which will need breakthroughs. Availability of high quality, clean data is next."
- Until today, the "lack of information on the benefits of LT/LR" seems to persist as the survey shows. This is supported by another comment acknowledging that "[t]here was

considerable progress in the field, but awareness of the opportunities available [thanks to LT] are not widespread." Moreover, "[m]any commonly used and necessary technologies are still not available (human-computer interaction, multimodal processing, etc.) and for others, if some advance in technologies is recorded, there are no available applications (summarisation, question answering, etc.)." It was also pointed out that "LT/LR are to some degree still seen as a side note to other branches of AI."

- One respondent is also troubled by the fact that with regard to LT, there "is a focus on in-house development which is not always aligned with emerging standards rooted in developments elsewhere."
- Another emerging issue identified by respondents is linked to the availability of "[c]omputational resources for pretraining large language models" which are still insufficient and/or inaccessible. This is especially true for smaller European languages that need to have larger than current LMs available (BERT, GPT-like).
- Last but not least, there are concerns about the lack of European LT solutions. As one respondent explains: "People's distrust of these technologies is causing companies not to be interested in developing these areas. Both the EU and investors find companies in this area uninteresting because the market in which they operate is not global. Companies then go bankrupt or are bought by American multinational players." This often leads to the so-called "locked in" problem for many European languages, i.e. the fact that the widely used commercial products are not open to adaptations and fine-tuning for specific languages. As one of the respondents illustrates: "Basically, Apple / Microsoft / Google / Amazon / Baiduu / Samsung / Alibaba etc. generally do not provide solutions where the best e.g. Finnish LT can be plugged in. And they will never prioritise Finnish [...], so how can we facilitate state-of-the-art performance in all kinds of applications for the "less important" European languages?" The main way out from the respondents' perspective is to ensure the following three basic ingredients in Europe:

  - "Massive quality data corpora available for all EU languages (i.e. restricted access, but available for research and development).
  - Large pretrained language models available for all EU languages.
  - Computational resources comparable to Facebook, Open AI etc. available to EU research centres ("CERN for LT")."

5. Conclusions and Outlook

In 2020, the President of the European Commission Ursula von der Leyen presented her vision of how to shape Europe's digital future:

> *"I am a tech optimist. My belief in technology as a force for good comes from my experience as a medical student. I learnt and saw first-hand its ability to change fates, save lives and make mundane what once would have been a miracle.*
>
> *Thanks to technology, these miracles are becoming more breathtaking and more regular by the day. They are helping to better detect cancer, support high-precision surgery or tailor treatment for the needs of each patient.*
>
> *This is all happening right now, right here in Europe. But I want this to be only the start. And I want it to become the norm right across our society: from farming to finance, from culture to construction, from fighting climate change to combatting terrorism."*

With regard to Language Technologies, the future is already here: the digital revolution has penetrated virtually all areas of our lives and as with any other revolution, it has significantly changed the professional and personal lives of people.

As part of this White Paper, several changes in the value and status quo of language-centric AI could be identified (see Section 3). For instance, in comparison to 2019, Machine Translation has found its way into the daily procedures and practices of public administrations in 2022. Only 6% of the participating organisations didn't use MT at all. At the same time, we found a massive increase in the use of Computer-Assisted Translation (CAT) Tools. The increasing importance of Language Technologies is also mirrored on policy level: In 21 out of the 24 national AI strategies published until now, the topic of Language Technology was included. Similarly, translation practices changed significantly. According to our latest results, most translations are still being outsourced, but the overall share has decreased (from 79% to 61%), reflecting a trend towards in-house translation (from 17% to 26%). At the same time, we could identify an increase with regard to the storing of language data: The majority (59%) of the survey contributors indicated that language data are now stored whenever possible in their organisations.

Also, the ELRC White Paper reveals latest developments and approaches to sustainable language data sharing in SMEs and public services (see Section 4). The circumstances that were found to negatively impact or limit the sharing of language data in Europe remained the same as in 2019. However, in addition to the actions that would be most relevant to overcome these issues and facilitate data sharing, several additional approaches were mentioned in 2022. They were all centred around (i) making language data collection, preparation and sharing less costly and (ii) providing financial support for activities targeted to language data sharing and collecting. The investigation also confirmed that since 2019, significant advances could be made related to LT development, digitalisation, and data collection both in the public sector and in SMEs, leading to greater availability of language data, better LT, and increased uptake of LT. Last but not least, survey responses identified six major challenges that organisations involved in the preparation and sharing of language data face in 2022 and beyond, e.g. the development of LT for European minority languages and human capital.

Coming back to Europe's digital future as expressed by Ursula von der Leyen, at least in the area of LT, the start into the digital decade was a very successful one. It is now in the hands of the European Union, each Member State and each organisation, to follow-up and tackle the remaining challenges and make Language Technologies "the norm right across our society".

References

[Azzano 2011] Azzano, Dino: *Placeable and localisable elements in translation memory systems, A comparative study*, 2011, https://edoc.ub.uni-muenchen.de/13841/2/Azzano_Dino.pdf.

[Benoit, 2011] Benoit, Kenneth: *Data, Textual*. In: International Encyclopedia of Political Science, 2011.

[EC, 2003] Document 32003H0361: *Commission Recommendation of 6 May 2003 concerning the definition of micro, small and medium-sized enterprises*.

[ELRC, 2019] European Language Resource Coordination: *ELRC White Paper – Sustainable Language Data Sharing to Support Language Equality in Multilingual Europe: why Language Data Matters*, 2019, ISBN: 978-3-943853-05-6.

[European Data Portal, 2017] European Data Portal: *Analytical Report 9: The Economic Benefits of Open Data*, 2017, https://data.europa.eu/sites/default/files/analytical_report_n9_economic_benefits_of_open_data.pdf.

[Palmer, 2011] Palmer, Alexis: *Computational Linguistics for Low-Resource Languages*, 2011, http://www.coli.uni-saarland.de/courses/CL4LRL/slides/cl4lrl_intro.pdf.

[Uszkoreit, 2010] Uszkoreit, Hans: *What is LT?*, 2010.

[Van Roy et al., 2021] Van Roy, V., Rossetti, F., Perset, K. and Galindo-Romero, L., *AI Watch – National strategies on Artificial Intelligence: A European perspective*, 2021 edition, EUR 30745 EN, Publications Office of the European Union, Luxembourg, 2021, ISBN 978-92-76-39081-7, doi:10.2760/069178, JRC122684.

Annex

ELRC White Paper survey

Which country are you representing?

- Austria
- Belgium
- Bulgaria
- Croatia
- Cyprus
- Czech Republic
- Denmark
- Estonia
- Finland
- France
- Germany
- Greece
- Iceland
- Ireland
- Italy
- Latvia
- Lithuania
- Luxembourg
- Malta
- The Netherlands
- Norway
- Poland
- Portugal
- Romania
- Slovak Republic
- Slovenia
- Spain
- Sweden

Which sector do you represent?

- Public Administration
- Research / Academia
- Language Services Provider
- Small and Medium-Sized Enterprises
- Other

Please define

.....

Are you part of the ELRC Language Resource Board (LRB)?

- Yes, Public Services NAP
- Yes, Technology NAP
- Yes, Consortium Member
- No

Translation practices

Are translations mostly produced in-house or are they outsourced? What is more common in your organisation/department?

- More than 50% of translations carried out by language services and translation professionals in-house
- More than 50% of translations carried out in-house by professional translators or bilingual/multilingual staff members
- Only single ministries with in-house translation services, mostly outsourcing of translations through central purchasing body
- Only single ministries with in-house translation services, mostly independent outsourcing of translations
- All translations are outsourced via central purchasing body
- Independent outsourcing of all translations

If the translations are being outsourced, are the translation memories (short TMs) or other by-products requested back?

- TMs and any other by-product of translation are requested back by default

- TMs and other by-product of translation are requested back for most outsourced translations
- Some request back TMs and/or other by-product of translations
- TMs or any other by-product of translation are not requested back

The importance of language data

Is your organisation storing language data like tmx files, translations, audio files, video recordings, etc.?

- Language data is stored whenever possible (high importance)
- Language data is sometimes stored (medium importance)
- Language data is hardly/never stored (low importance)

Use of Language Technology (LT)

To what extent does your organisation/department make use of Machine Translation (MT)?

- Most translation services have an MT API integrated into the translation process
- Some translation services have an MT API
- No MT API but use of freely available MT web services
- No use of MT at all

Which of the following language technologies are commonly used in your organisation/department?

- ☐ Classification
- ☐ Anonymisation
- ☐ Summarisation
- ☐ Speech Recognition
- ☐ Text to Speech
- ☐ None of the above

Please indicate to what extent this technology is being used (rarely, occasionally or regularly)

.....

To what extent do Language Service Providers (LSPs) use CAT Tools?

- All LSPs and freelance translators use CAT
- It is common practice for LSPs but not freelance translators to use CAT
- Only single LSPs (and/or freelance translators) use CAT
- No use of CAT tools

To what extent does your organisation/department use CAT Tools?

- All translations are carried out with the aid of CAT tools (language services, translation professionals and other translating staff members)
- It is common practice that translation services/translation professionals use CAT tools
- Only single translation services or translators use CAT tools
- No use of CAT tools

National Regulations related to LT/AI

Is there a language policy in your country?

- Yes
- No
- I don’t know

How would you rate the role of LT in your country's language plan or AI regulation?

- ○ Special mention of language technology including a financial plan
- ○ Special mention of language technology but no information about financial plan
- ○ Language technology is mentioned as a side note (e.g. as useful example of AI)
- ○ No mention of LT
- ○ Other
- ○ I don't know

How would you rate the role of language data in your country's language plan or AI regulation?

- ○ Special mention of language data including a strategic plan on its collection and/or provision
- ○ Special mention of language data but no detailed plan
- ○ Language data is mentioned as a side note
- ○ No mention of language data at all
- ○ Other
- ○ I don't know

Latest developments & approaches to data sharing

Which of the following action items would be most relevant to facilitate data sharing? Please list them in descending priority (first-mentioned approach = most relevant)

- ☐ Raising awareness of language data as open data and a valuable asset
- ☐ Increasing interest in MT/LT in public services as part of the national digital policy
- ☐ Tackle legal concerns
- ☐ Identify and gain access to outsourced translations
- ☐ Establish good data management practices in public services
- ☐ Other

Please define which other approaches would be relevant (if any)

.....

Where do you see recent advances related to LT development, digitalisation and data collection? Were there any big changes compared to the situation three years ago?

.....

What are currently the biggest challenges related to LT/LR in your country? Which ones have been addressed or even solved in the last three years?

.....

What are the key objectives (new or old) that should be reached by 2030 when it comes to LT development, language data management and sharing? Please also indicate how you would prioritise them (from high to low)

.....

Any additional suggestions, ideas, remarks you would like to share with us?

.....

Annex

Country Profile Austria

State of Play:

Translation practices and information exchange in ministries and public administrations:

Within Austrian federal government organisations, integrated language services are the exception. Most public administrations either outsource translations or translate in-house but not with staffed translators. Overall, the translation process on the federal level is decentralised, which means that every public administration meets their own translation needs, there is no central management tool for translation requests and no formalised exchange of translation memories or expertise. However, an informal working group ARG GUT (Arbeitsgruppe Gouvernementaler Uebersetzungs- und Terminologiedienste) was initiated by the Language Institute of the Austrian Armed Forces in the Federal Ministry of Defence. The working group consisting of translators and terminologists for Austrian German <> English (and partly also for French), not only exchanges information and expertise but also creates resources such as the administrative glossary that is freely available on the Austrian Language Resource Portal (Sprachressourcenportal Österreichs). This portal was created as an aid for the Austrian EU Council Presidency in 2018. However, it is continuously developed further.

In addition to the Austrian Armed Forces Language Institute (SIB) subordinated to the Federal Ministry of Defence, the Federal Ministry of Interior and the Federal Ministry of Agriculture, Forestry, Regions and Water Management, other public administrations have integrated language services, such as the Austrian Financial Market Authority, the National Bank of Austria as well as the Vienna City Administration. The Austrian Armed Forces Language Institute provides translations, terminology work, and language teaching in English, French, Italian, Czech, Slovak, Hungarian, Slovenian, Russian, Ukrainian and Balkan languages. The Federal Ministry of the Interior has several translation cells in various agencies of the ministry. There is the Language Service of the Criminal Intelligence Service with about 12 translators, interpreters and terminologists and small translation cells both in the Federal Bureau of Anti-Corruption and in the Directorate State Protection and Intelligence Service. There is little coordination and exchange between those three language services, partly due to secrecy and security reasons, partly because there is no language governance on the superordinate ministerial level. However, translators and terminologists from all three services work together within the informal terminology working group ARG GUT.

The other ministry that has an in-house translation service is in the Federal Ministry of Agriculture, Forestry, Regions and Water Management (formerly known as the Ministry for Agriculture) with three translators/interpreters. Although it is common practice that the in-house translation services use computer-assisted translation (CAT) tools, including translation memories (TMs), the TMs are not managed in a way that allows for easy language data sharing with e.g. the Austrian Open Data Portal. In addition to the translation services within the ministries, two government owned agencies who also offer translation services, the Justizbetreuungsagentur (JBA – judicial support agency) and the Bundesagentur für Betreuungs- und Unterstützungsleistungen (BBU – Federal Agency for Reception and Support Services) were founded.

All other administrations meet their translation needs by either outsourcing to freelancers or language service providers or by their own employees who are not professional translators but are making use of their language skills or commercial machine translation systems. In these cases, no CAT tools are applied in the translation process and TMs are not requested back from LSP to whom the translations are outsourced.

The Bundesbeschaffung GmbH (BBG – Federal Procurement Agency) stipulates framework contracts for translation services, that can be used by the public administration if the outsourcing of translation services is needed.

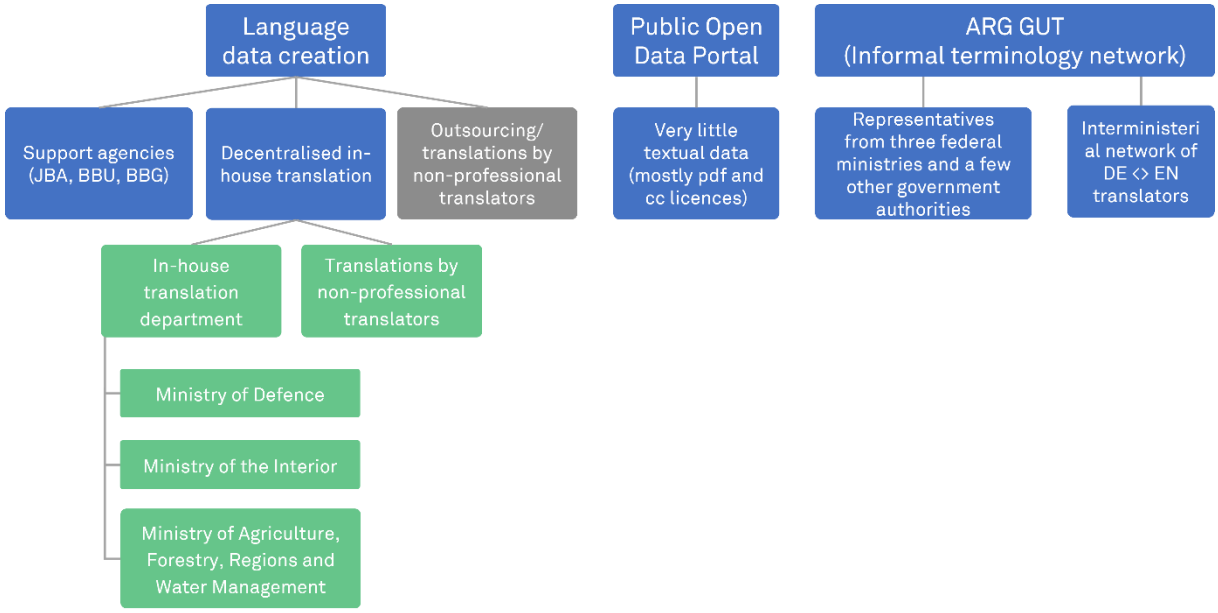
Interesting fact:

The Austrian Armed Forces Language Institute (SIB) developed apps based on language resources to enable basic intercultural communication, primarily for soldiers in international operations but also for the general public. The apps can be downloaded for free.

Translation needs:

The translation needs in Austrian public administration are threefold. There is a need for international communication and translation, a need for adaptation between different varieties of German and the need for translation between the language register used by public officials and everyday language citizens use, or also plain language. The Austrian Language Resource Portal mainly addresses the need for appropriate international communication (into and from English) and translation specifically tailored to the Austrian German variety.

The current language data creation and sharing infrastructure in Austrian public bodies looks as follows:



Open Data in Austria:

While a lot has been done already in the direction of creating language data, language resources, and data infrastructures, the discussion of the third Austrian ELRC Workshop in 2021 has clearly shown that there is room for improvement. Open Data as a concept and practice has been initiated in Austria with pioneer projects, such as the Open Data Project, but little interconnection between institutions and data portals can yet be perceived. In fact, with the existing resources there is a lack of overview of which resources have been considered and which are yet to be considered.

The Austrian Data Portal data.gv.at has won the United Nations Public Service Award for Open Government Data in 2014 and has over 37,000 data sets, documents and applications available for reuse. It centralises metadata of decentral data catalogues and is the single point of contact to the European Data Portal. However, there is a strong focus on numerical data, exemplified by the fact that there is no explicit category for language data and textual data currently falls in the category of “documents” available predominantly in PDF format. After the second ELRC Workshop in Austria, however, some of the textual data that were PDF files were converted into TMX files and are now available in machine-readable format on ELRC-SHARE. The Austrian Data Portal has a filter for the file format, but TMX files are still not available. Currently, a metadata core, optional attributes and a vocabulary for the metadata catalogue for open government data (OGD metadata) are available in German and English.

Digital policy and language policy in Austria:

In 2021, Austria published its first Digitalisation Report, which describes the main initiatives and projects that are boosting digital transformation in administration services for society and businesses.

The Austrian Digital Action Plan puts forward initiatives targeted at improving digital services for citizens, increasing the use of data for an economical growth, data security and resilience.

The Open Government strategy is coordinated by the "Cooperation Open Government Data Austria", whose main objective is to create an environment that encourages the sharing of government data as well reaping its benefits. As mentioned above, the focus is currently on numerical data and language data are not yet acknowledged as valuable Open Data resources on the policy level.

The e-government programmes are coordinated by the Federal Ministry of Finance. Digitalisation is seen as a cross-sectional topic which is coordinated by Chief Digital Officers (CDO) appointed for every area of responsibility across all ministries. Together these officers compose the "CDO Taskforce" that is tasked with optimising the coordination of digitalisation activities. The former Ministry of Digital and Economic Affairs is also responsible for the implementation of the Digital Single Market in Austria[22].

Part of the eGovernment initiative is Austria's central platform for digital public services oesterreich.gv.at. The platform offers citizens services such as an electronic signature, electronic payments, changing residency and others. However, the website is mainly available in German. Some general information is also available in English.

As Austria's sole official language, all government communication is exclusively in German and only partially translated into English, underlining the strong dominance and the status of Austrian German as the only official language. Recognised minority languages are Croatian, Romani, Slovak, Slovene, Czech and Hungarian. As some regions have a large number of native speakers of Slovene, Croatian and Hungarian, these languages have minority status and school education is offered bilingually in these regions. Additionally, some schools offer mother-tongue teaching in a total of over 26 languages. Article 19 of the Basic Law of 21 December 1867 on the General Rights of Nationals in the Kingdoms and Länder represented in the Council of the Realm specifies that all ethnic entities have a right to the preservation and fostering of their language, including schools, administration and public life.

The role of LT and language data in Austria's AI regulations:

Although the Artificial Intelligence Mission Austria 2030[23] mentions language technologies, such as speech recognition and voice control as potential fields of application of AI, language data are not mentioned in the strategy.

Stakeholders and major networks:

The Federal Ministry for Climate Action, Environment, Energy, Mobility, Innovation and Technology is responsible for the CEF agenda including matters related to eTranslation and therefore an important stakeholder. Still, since translations are carried out in all federal ministries and all areas of responsibility have CDOs, all ministries are important stakeholders. So far, about 20 institutions and public administrations, including the former Ministry for Digital and Economic Affairs have participated in ELRC events. Language resources in various formats such as plain text, TMX and XML were contributed by the Austrian Armed Forces Language Institute, the Federal Ministry of the Interior, the National Bank of Austria, the City of Vienna and the Centre for Translation Studies at the University of Vienna, among others.

A valuable overview of major AI players including LT-related companies was presented in 2020 by enliteAI in the document "AI Landscape Austria"[24]. In the area of research, Austria is involved in ELG, ELE, COST NexusLinguarum, CLARIN and DARIAH, among others.

[22] N.B.: The responsibilities might change with the new government.
[23] https://www.bmk.gv.at/themen/innovation/publikationen/ikt/ai/strategie-bundesregierung.html
[24] https://www.enlite.ai/insights/ai-landscape-austria

Main challenges for sustainable data sharing:

Austria faces a number of challenges when it comes to sharing language data continuously and sustainably. The main challenges are:

- Clear lack and need of expertise in the field of LT and data science: Such experts would be needed on the microlevel of individual organisations but also on a macro-level to detect blind spots in the communication and interaction between organisations to provide a more coherent language technology and language data network within Austria.
- Lack of awareness of the benefits that can be offered by national LT and linguistic data science expertise
- General lack of funding for language technology and linguistic data science in industry, the public sector, and research: While there is a general AI policy in Austria, its terms are underspecified in terms of dedicated funding and no programmes for the globally extremely economically successful field of language technologies have been defined.
- Little awareness of the importance of language technologies and language resources
- Legal and security-related concerns and inadequate practices for data management due to a lack of a centralised initiative or plan. Furthermore, a lack of reliability of data and a lack of interoperability between data without realistic measures to address these issues was indicated
- Internal translation workflows are not taking into consideration that the produced translations have linguistic value apart from their inherent purpose, which leads to a number of issues such as:
  - Translation memories are not requested back when translations are outsourced, hence the translations are not available in their most useful file format, i.e. TMX
  - Language data is not managed/filed in a way that allows for easy data sharing
  - Privacy, copyright and IPR are not clearly indicated or transferred
  - It is unclear who can authorise sharing of language resources
  - Superiors do not acknowledge the value of sharing data and therefore do not initiate necessary changes in the workflow or processes
- Language policy or multilingualism is not in the portfolio of a specific ministry (only the Ministry of Education, Science and Research covers language and multilingualism at schools) making it difficult to identify decision makers in relation to multilingualism, language data as Open Data or the translation and procurement processes

Action plan:

In order to address the identified challenges in Austria, the following actions are suggested:

- **The main objective is to improve and establish data management practices that allow for reaping the maximum benefit from language data. Specific actions include:**
  - The identification of data managers
  - Further investigation of data management practices
  - Guidelines for the identification of confidential and personal data
  - Clear indication of confidential and personal data as well as copyright in the translation process to make data sharing in the future easier

As these actions need to be addressed top-down, support and guidelines from the European Commission and ELRC would be very helpful.

- **The second objective is to include machine translation and language technology in the national digital policy and increase the interest in these topics in public services. Specific actions include:**
  - Secure the support of decision makers to include language technology in the national policy
  - Establish synergies with national actions and initiatives related to language technology, machine translation and language data
  - Inform about amounts of language data that are needed to develop language technologies as well as processes that are available to make data sharing safe and secure
  - Stress the importance of the Austrian variety of the German language (in order to receive high-quality machine translation output for Austrian German)

- **The third objective is to gain access to outsourced translation:**
  - Gaining access to outsource translations could be a valuable asset, since this data has enormous potential and value.
- **Another objective is to generally raise awareness about the value of language data:**
  - This includes its potential when shared and used for machine translation but also more generally in many different areas of artificial intelligence.

References and links:

Austrian Language Resource Portal: https://sprachressourcen.at.

Metadata Catalogue for Open Government Data (OGD Metadata):
https://www.data.gv.at/katalog/dataset/metadaten-von-ogd-osterreich

Open Data Österreich: https://www.data.gv.at.

Open Science Network Austria: https://oana.at/.

[BMWF] Bundesministerium Bildung, Wissenschaft und Forschung: *Sprachliche Bildung*,
https://bildung.bmbwf.gv.at/schulen/unterricht/ba/sprachenpolitik.html.

[OGD] Cooperation OGD Österreich: *Infos Cooperation OGD Österreich*,
https://www.data.gv.at/infos/cooperation-ogd-austria/.

[Roadmap] Federal Ministry Republic of Austria Digital and Economic Affairs: *Digital Roadmap Austria*,
https://www.digitalroadmap.gv.at/fileadmin/downloads/digital_road_map_broschuere.pdf.

[eGovernment, 2017] Federal Chancellery Republic of Austria: *Behörden im Netz, Das österreichische E-Government ABC*, 2017, https://www.digitales.oesterreich.gv.at/documents/22124/30428/E-Government-ABC.pdf/b552f453-7ae9-4d12-9608-30da166d710b,
English version: https://www.bmdw.gv.at/dam/jcr:8fc815bb-1dc7-4e45-9610-78d63560944a/E-Government-ABC_2019_EN.pdf.

[ELRC, 2018] Heinisch, Kotzian: *ELRC Workshop Report for Austria*, 2018,
http://lr-coordination.eu/sites/default/files/Austria/2018/ELRC_Workshop_Austria_Report_public_v1_FINAL.PDF.

[StGG] English translation of the Basic Law of 21 December 1867 on the General Rights of Nationals in the Kingdoms and Länder represented in the Council of the Realm (in German: *Staatsgrundgesetz vom 21. December 1867, über die allgemeinen Rechte der Staatsbürger für die im Reichsrathe vertretenen Königreiche und Länder* – StGG):
https://www.ris.bka.gv.at/Dokumente/Erv/ERV_1867_142/ERV_1867_142.pdf.

[Digital Agenda, 2019] Stadt Wien: *Digitale Agenda Wien 2025*, 2019,
https://digitales.wien.gv.at/wp-content/uploads/sites/47/2019/09/20190830_DigitaleAgendaWien_2025.pdf.

Annex

Country Profile Belgium

State of Play:

Translation practices and information exchange in ministries and public administrations:

In Belgium, each institution is responsible for the translation of their data. Translation needs are often handled on demand and the applied translation practices are diverse. Consequently, there are public administrations, which outsource all their translations, whereas other institutions are solely building on in-house translation. In addition, there are administrations applying a combination of both approaches.

Data exchange and translation are currently not coordinated in Belgium. In public administrations, all outsourced translations are part of call for tenders. Since there is not one call for all administrations, each department has its own tender. Further information is available in the country reports produced by NEC TM.

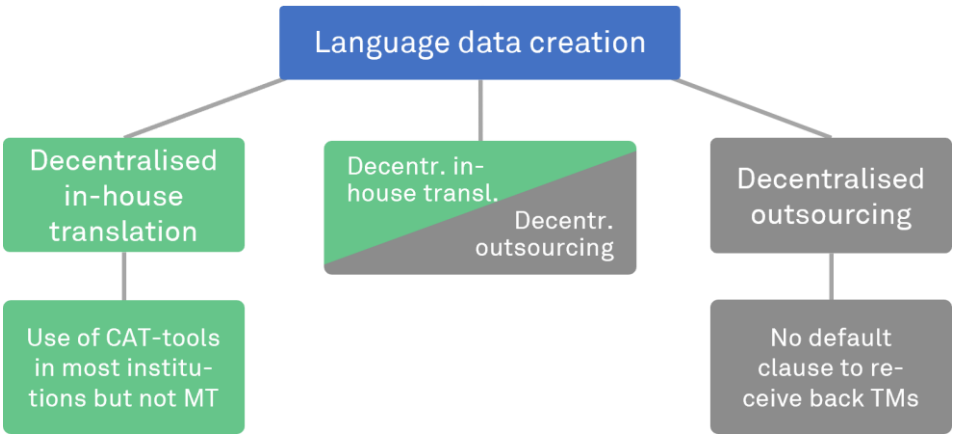
CAT Tools are used by the vast majority of Belgian language service providers (LSPs) and freelance translators. This is also common practice in Belgian institutions, which use e.g. computer-assisted translation software suites or translation management systems. Although there is a growing awareness that machine translation (MT) can be a valuable asset and facilitate the translation process, it is only rarely used. If the translations were outsourced, the corresponding translation memories (TMs) or any other by-products are usually not transferred back.

> **Interesting fact:**
>
> **In a few public administrations, data management plans have been developed and integrated. Public administrations are provided with a set of guidelines and handle their data management individually.**

In the academic area, Belgian funding agencies such as the research foundation FWO are now required to submit a data management plan together with their project proposal. Apart from that, in 2017, the DMPbelgium Consortium was founded by a number of Belgian universities, including Ghent University, Hasselt University and University of Antwerp, among others. They developed a shared data management planning tool, offering common data management plan templates, institutional templates and guidance. Further information about data management at Belgian universities is available at https://dmponline.be.

Although there are currently no specific data sharing infrastructures, the focus on data management can be seen as a preparatory step. There is no obligation to share data, but it is increasingly encouraged and corresponding platforms are becoming more and more popular among researchers.

The current language data creation and sharing infrastructure in Belgian public bodies looks as follows:



Open Data and data collection in Belgium:

In compliance with the PSI Directive, the Belgian Federal Council of Ministers agreed to adopt a federal Open Data strategy with an ambitious roadmap for 2015-2020 (Ferri et al., 2015), introducing the principle of open public data by default. According to this strategy, all data collected by the Belgian public administrations has to be freely available and reusable. Exceptions are only acceptable if the data contains private information or content with the potential to harm public security. The federal Open Data strategy includes a set of fifteen practical guidelines to facilitate the reuse of data. The primary goals include (DLA, 2015).

- The free use of PSI without any reference to the source to facilitate the combination of data sets.
- The provision of data in machine-readable formats to facilitate reuse, identification and extraction (e.g. Excel instead of PDF, CSV instead of Excel, etc.) whenever possible.
- The provision of public sector information by the federal government not only upon request, but proactively by 2020.
- The development of an Open Data strategy in each federal public service and the appointment of an "Open Data champion", acting as the Open Data contact point within the organisation.
- The set-up of a web portal providing continuous access to open data sets (https://data.gov.be/en).

The above-mentioned web portal already includes more than 10,000 data sets covering a variety of fields, e.g. public sector, science and technology or environment. It provides a search function and filtering by topic, licence, data format or organisation. According to the Belgian federation for the technology industry Agoria, making public sector information available to non-public entities could also lead to a considerable economic benefit. In more concrete terms, Agoria expects a net gain of around 900 million EUR (van Tilborg, 2018).

An important aspect, which may prevent public administrations from sharing their translated data is the fact that the translator holds the ownership of the produced text by default (Deene, 2018). In order to avoid any copyright or GDPR-related issues, it is thus important that public administrations are granted the rights to the translated text when receiving an outsourced translation.

Digital policy and language policy in Belgium:

Multilingualism plays an important role in Belgium. Belgium can be divided into the three regions Flanders, Brussels and Wallonia and has three official languages, i.e. Dutch (approx. 5.2 million speakers), French (4.6 million speakers) and German (72,000 speakers). Language policy in Belgium is based on two principles, the constitutional principle of linguistic freedom and the territoriality principle as described in Article 4 of the Belgian Constitution. According to the latter, Belgium can be divided into four linguistic areas, i.e. the Dutch-speaking region, the French-speaking region, the bilingual region of Brussels-Capital and the German-speaking region (Hüning et al., 2010).

> **Interesting fact:**
>
> **Belgium has three official languages and can be divided into four language areas. Official language use is determined by the territoriality principle.**

Each municipality in Belgium is part of only one of the four language areas. The official language use therefore depends on the territorial boundaries and varies from one linguistic area to another. Whereas in Brussels, there are two official languages (Dutch and French), all other regions are monolingual.

However, the linguistic boundary does not quash the constitutional principle of linguistic freedom, stating that "the use of languages spoken in Belgium is optional; only the law can rule on this matter, and only for acts of the public authorities and for judicial affairs."(Article 30). The territoriality principle is restricted to a limited number of domains, including public authority and administration, court, education, social relations (between employer and employees) and official documents. In addition to the three official languages, English also plays an important role in Belgium, especially regarding increased visibility outside the country borders and in multilingual contexts (ELRC, 2018).

Multilingualism is also of utmost importance for Belgian public digital services. At federal level, all public digital services are at least bi-lingual, since their provision in Dutch and French is obligatory. This also applies to Brussels, while in Flanders, all information must be provided in Dutch. In Wallonia,

public digital services must be available in French, whereas in the Eastern part of Wallonia, the additional provision of public digital services in German is obligatory, too.

The role of LT and language data in Belgium's AI regulations

Since 2019, the Flanders AI research programme aims to promote research and education in AI. The main goal of the programme is the successful adoption of AI in Flanders, in order to contribute to a sustainable and prosperous human-centred digital future in Flanders, as well to economic growth and innovation in the region[25]. The research activities are also aligned with the European ambition and strategies for AI and the digital agenda for 2030. In the research challenge "Human-like AI", language technologies and language resources play a major role.

Stakeholders and major networks:

The ELRC National Anchor Points represent two relevant stakeholders, namely the Chancellery of the Prime Minister and Ghent University. The collection of language data is also strongly supported by federal and regional public services, including e.g. the National Bank of Belgium and RIZIV, the National Institute for Health and Disability Insurance. Since 2016, the National Bank of Belgium donated more than 140 term bank entries in all three languages plus English and RIZIV made a translation memory available that consists of more than 30,000 translation units in French and Dutch. Local ELRC events and workshops were attended by representatives of more than 40 institutions, including e.g. FPS Chancellery of the Prime Minister or Nederlandse Taalunie. This clearly demonstrates that many Belgian institutions are already aware of the importance of collecting, managing and sharing language data to facilitate information exchange not only across the four linguistic regions of Belgium, but also across the European Union.

Main challenges for sustainable data sharing:

- Anonymisation is often an obstacle to sharing translations and translation memories. Although automatic processes for anonymising data may be helpful to overcome this issue, the output is not 100% reliable. Especially when dealing with unstructured data, this problem can hardly be fixed.
- Legal issues remain an important obstacle, and the people responsible for big data sets tend to err on the side of caution and not release any data, rather than risk e.g., GDPR issues. Being able to refer to ELRC for legal assistance in this regard is important, because GDPR and privacy are too often used as an excuse to not share anything.
- Data sharing is often hindered by the hierarchical organisation of institutions where the decision-making process of sharing does not include the language professionals who are aware of the importance of data contribution.
- Belgium's dense bureaucratic system, with a complicated government system leading to many different institutions and little contact between institutions, can make it difficult to find and address the right people. This problem was exacerbated by the withdrawal of the Belgian NAP a few months before the ELRC Workshop. Related to this issue, it is not always easy for people from the Flemish side to have the necessary contacts in the Walloon region.

Action plan:

For Belgium, the following objectives could be defined to address the identified challenges. In the order of their priority, they are:

- **To raise awareness of the value of language data:**
  As language data is currently not included in the Belgian digital policy, it is important to promote the benefits of sharing language data. This could be achieved with the help of concrete examples of how data contributions had a positive impact on machine translation systems. In addition, it is important to establish practical guidelines for LR as Open Data and to broaden the definition of textual resources by adding speech data, data for AI and other types of language resources.

[25] https://www.flandersairesearch.be/en

- **To establish good data management practices in public services:**
  Although there is already a focus on data management plans in the academic field, the above-mentioned developments will need to be continued and extended to all public services.
- **To increase interest in MT/LT in public services as part of the national digital policy:**
  Concrete examples of how public administrations can benefit from language technologies in their daily operations would raise the institutions' awareness and increase their interest in MT/LT. In addition, the use of MT/LT services could be promoted by identifying and establishing synergies with national projects and initiatives, wherever possible.
- **To tackle legal concerns:**
  Since legal concerns are one of the key challenges when it comes to sharing data in Belgium, it is important to develop and share easy-to-apply guidelines for IPR and privacy issues. Apart from that, the possibility to implement rights management along with data management needs to be investigated.

References and links:

Research Foundation Flanders (Fonds voor Wetenschappelijk Onderzoek – Vlaanderen, FWO): https://www.fwo.be/en/.

[Digital, 2017] Belgian Government: *Digital Belgium*, 2017, http://digitalbelgium.be/wp-content/uploads/2017/07/compressed_Brochure_DB_FINAL.pdf.

[Deene, 2018] Deene, Joris: *Can language data be shared and how?*, 2018, http://www.lr-coordination.eu/sites/default/files/Belgium/2018/S2.5_Language%20Data%20Sharing.pdf.

[DLA, 2015] DLA Piper: *Belgian Government gives green light to a new Open Data strategy*, 2015, https://www.dlapiper.com/en/uk/insights/publications/2015/12/spotlight-on-belgium-issue-8/belgian-government-approves-open-data-strategy/.

[ELRC, 2018] De Smeytere, Hoste, Terryn: *ELRC Workshop Report for Belgium*, 2018, https://lr-coordination.eu/sites/default/files/Belgium/2018/ELRC%2B2%20Workshop_Public_Belgium-.pdf.

[Ferri et al., 2015] Ferri, Springael: *Federale open data-strategie*, 2015, https://www.presscenter.org/nl/pressrelease/20150724/federale-open-data-strategie?lang=fr.

[Hoste, 2018] Hoste, Véronique: *ELRC in Belgium*, 2018, http://lr-coordination.eu/sites/default/files/Belgium/2018/S2.2_ELRC%20in%20Belgium.pdf.

[Hüning et al., 2010] Hüning, Vogl: *One Nation, One Language? The case of Belgium*, 2010, https://www.academia.edu/1056036/One_nation_one_language_The_case_of_Belgium.

[van Tiborg, 2018] van Tilborg, Luc: *National Initiatives for Digital Public Services and (Open) Data*, 2018, http://www.lr-coordination.eu/sites/default/files/Belgium/2018/S1.2_National%20Initiatives%20for%20Digital%20Public%20Services.pdf.

Annex

Country Profile Bulgaria

State of Play:

Translation practices and information exchange in ministries and public administrations:

In Bulgaria, translation services are subject to procurement through the Central Purchasing Body (Ministry of Finance).The centralised public procurements (CPP) are for the provision of interpretation (simultaneous and consecutive) and second for written translation from Bulgarian to foreign language and from foreign language to Bulgarian for the needs of the administrations. The Central Purchasing Body (CPB) which by definition is a contracting authority, setting up a framework agreement with several contractors for each of the two CPP. Public administrations are able to make their choice between the agencies, part of the framework agreement for the relevant period. When the conclusion of the Framework Agreement by the Central purchasing body is postponed or cancelled, the contracting entities shall apply the general rules and award the procurement individually.

However, the application of the public procurement procedure depends on the value of the contract. Public procurement data is available through the Public Procurement Agency (http://www.aop.bg/index.php?ln=1) and the Public Procurement Portal of Bulgaria. The proposal of the National Anchor Point of the Administration for the translation memories (TMs) stored on CAT instruments for the realised translations to be available by request was accepted and up to now they are part of the technical requirements of the PP for the selected language service providers (LSPs). Similarly, there is no coordinated exchange of translations or language data among ministries (and/or other public bodies) in place at the moment.

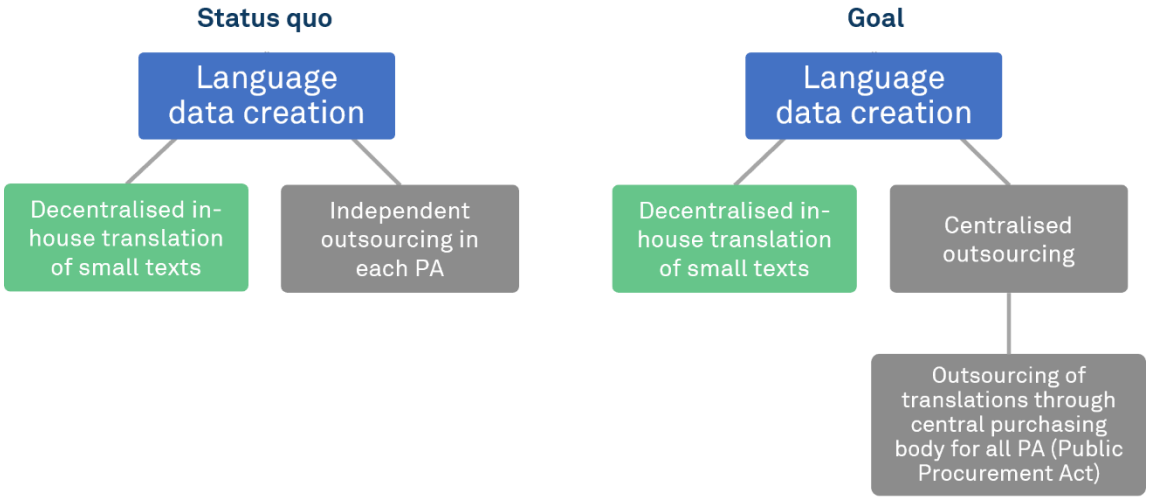
The use of computer-assisted translation (CAT) tools or machine translation (MT) is currently not a common practice in administrations and ministries: They mainly use the spelling tool in their Office package and some of them use the European Commission's eTranslation MT system as well as the translator of the Bulgarian EU Council Presidency.

Bulgaria keeps working for raising the awareness of the importance of language resources. One of the main purposes is to make LR part of the data sets of the National Open Data Portal. The opportunities offered by the automatic translation platform e-Translation and possibilities for free access by the representatives of the administration, SMEs and NGOs continue to be promoted. Bulgaria is planning to reach out to the local authorities encouraging them to share their language data with the ELRC platform. Efforts to raise the awareness for the importance of the reusing of the available documents will be made.

Interesting fact:

The Ministry of Transport and Communications has made general registration of all its employees so they can use the CEF Automated Translation (AT) platform without the need of individual requests or registrations.

The current language data creation and sharing infrastructure in Bulgarian public bodies looks as follows:



Open Data and data collection in Bulgaria:

The best example of data collection in the country is the Open Data Portal "data.egov.bg", maintained by the Ministry of eGovernment. Up to August 2022 there are 10,805 data sets and 537 registered organisations which are sharing their data in 14 main thematic areas. The Bulgarian administrations and the public authorities all have an obligation to prioritise and publish data for re-use on the portal.

Overall, the portal is built in accordance with the requirements of the Access to Public Information Act, and the information it publishes is set out in the regulation on standard terms and conditions for reuse of public sector information. There is a list of data sets per priority area, which has to be published in an open format and is updated on an annual basis, as well as a synthesis report on the availability of public sector information every three years. The following key players have a direct influence on the Open Data strategy in Bulgaria:

- Ministry of eGovernment
- The Administration of the Council of Ministers The Institute for Public Administration – regularly organises Open Data Training.

Digital policy and language policy in Bulgaria:

The Ministry of eGovernment is responsible for the e-Government strategy and the faster implementation of e-governance in Bulgaria. Core issues covered in the State e-Government Agency's domain are semantic, legal, technological and organisational interoperability. Data is considered the new most valuable resource and there is an explicit intent to make as much data available as possible for re-use through the public services Bulgaria offers to people and businesses.

Bulgaria has also implemented the secure electronic delivery system "eDelivery", which is also part of the single application model for payment and provision of electronic public services at national level.

Since June 2020 there is a fully functional Bulgarian Portal for Open Science (BPOS) and a national repository for open access to scientific information, maintained by the National Centre for Information and Documentation.

With regard to the legal framework for sharing data, special attention needs to be paid to the Directive on the reuse of public sector information, the legal framework on the reuse of information and in particular the Access to Public Information Act, which introduced the Directive in Bulgaria through a special regime substituting the older access to information regime.

Stakeholders and major networks:

The national AI strategy of Bulgaria emphasises the need to strengthen research and innovation capacities and the uptake of AI technologies by means of active collaborations between research institutions and industry at national and international level. This is achieved through:

- Promoting cooperation spaces between researchers and AI professionals and encouraging the creation of collaborative networks in AI between universities, vocational schools and companies;
- Participating in European testing and experimentation centres related to health, robotics and agriculture.

The national procedure for selection of project proposals for the establishment of European digital innovation hubs (EDIH) in Bulgaria has already started. The role of future European digital innovation hubs (between 3 and 6 in Bulgaria) is to provide businesses and local administrations with innovative digital solutions and their integration into their day-to-day operations. Digital innovation hubs should offer the opportunity to experiment and test new technologies according to the specific needs and activities of each company or institution in the public sector.

The digital hubs will be funded under the programme "DIGITAL EUROPE" 2021-2027 and will be the core for the development and implementation of digital innovations. The purpose is to integrate language technologies into systems for supporting the learning of foreign languages.

Above this, a consortium of five Bulgarian academic institutions is working to create and develop an integrated national academic infrastructure for language resources.

The 13 Bulgarian research and innovation centres include four "centres of excellence" i.e. fundamental research institutions and nine "centres of competence", focused on applied research activities with potential for industrial uptake. It concerns sectors such as mechatronics, digital technologies, creative and gaming industries and biotechnology and other areas in line with the priorities of Bulgaria's smart specialisation strategy, its industrial and innovation strategy based on local competitive strengths. Sofia Tech Park is part of the Bulgarian entrepreneurship and innovation ecosystem that could contribute towards boosting early stage entrepreneurship activity, strengthening cooperation between research and industry, as well as commercialisation of R&D and development of competitive innovative products and processes.

Within ELRC, more than 150 potential stakeholders that are involved in the creation or sharing of language resources, related activities and/or policy setting were identified. They also participated in the latest ELRC Workshop. Most importantly, the stakeholder base includes all the ministries, different state and executive agencies, administrations of the Presidency and the National Assembly which are potential providers of language resources.

Digital services and public organisations with multilingual needs that could benefit from the eTranslation platform include in particular:

- All institutions part of the public administration;
- Local authorities;
- The NRA (National Revenue Agency), which is the public administration in Bulgaria with the highest number of electronic administrative services and could benefit from the instrument.
- Taxation and health insurance services;
- SOLVIT;
- National Institute of Immovable Cultural Heritage;
- All organisations providing public services.

Main challenges for sustainable data sharing:

- Difficulty to identify and convince high-level officials to authorise data sharing;
- No established procedures for the translation of documents on administrative level, which leads to:
  - Potential legal issues, e.g. concerning the ownership of the data, personal data issues, copyright issues
  - Technical difficulties relating to data processing
- Resistance to new technologies;

- Lack of resources, which are required for supporting the technical and legal preparation and sharing of language data;
- Concerns about the quality of translations, which could be shared (feeling that their quality may not be high enough for sharing).

Action Plan:

Key actions for improving the sharing of language resources in Bulgaria are mainly targeted at (i) raising awareness of language data as Open Data and valuable asset, (ii) increasing interest in MT/LT in public services as part of the national digital policy and (iii) improving access to outsourced translations.

Regarding the awareness raising of language data as Open Data and valuable asset, several activities are planned and/or on the way:

- **Integrating language data in the national Open Data policy, digital agenda, etc.:**
  The National Anchor Points (NAPs) sent letters to all Ministries and their second level spending units on the benefits of sharing their translations through the ELRC-SHARE. A corresponding high-level meeting to stimulate the process of sharing language resources was organised involving the different administrations. The Commissioner for Digital Economy and Society and the Prime Minister in charge in this field are aware of the ongoing efforts.
- **Increasing collaboration with the Open Data officer on national level & establishing practical guidelines for LR as Open Data.**
- **Promoting the value and benefits of sharing language data for language activities:**
  Language data sharing should be embraced by a wider audience, while the data contribution processes should be fine-tuned. The Bulgarian NAPs are planning to reach out to the municipal authorities in Bulgaria, which also have materials that could be useful for the eTranslation platform.
- **To increase interest in MT/LT in public services as part of the national digital policy, several important efforts are also planned/have already been started:**
  - **Secure support of decision makers to change/adapt national policy:**
    As indicated above, a high-level meeting to stimulate the process of sharing language resources was organised involving the different administrations. The Commissioner for Digital Economy and Society and the Prime Minister are aware of the ongoing efforts.
  - **Ensuring central accessibility to eTranslation:**
    The Ministry of Transport and Communications has made general registration of all its employees, so they can use the CEF AT platform without the need of individual requests or registrations.
- **Regarding the improvement of access to outsourced translations, several efforts are in planning/implementation:**
  - **Centralising procurement of translations:**
    There are major efforts, in particular through the Public Procurement Agency, to centralise the procurement process and set common standards. In this respect, procurement of translations should also be considered and coordinated between the different ministries and public bodies.
  - **Establishing practice of receiving any by-product of outsourced translations:**
    The proposal of the National Anchor Point of the Administration for the translation memories (TMs) stored on CAT instruments for the realised translations to be available by request was accepted and up to now they are part of the technical requirements of the PP for the selected language service providers (LSPs).

References and links:

Access to Public Information Act: https://www.me.government.bg/en/library/access-to-public-information-act-448-c25-m258-2.html

Bulgarian Open Data Portal: https://data.egov.bg/

Public Procurement Portal of Bulgaria: https://www.aop.bg/

SOLVIT: https://ec.europa.eu/solvit/index_en.htm.

Ministry of eGovernment: https://egov.government.bg.

Annex

Country Profile Croatia

State of Play:

Translation practices and information exchange in ministries and public administrations:

Translation practices in Croatian public administration are fully decentralised and organised independently. Most public administrations outsource translations to Language Service Providers (LSPs), only the Ministry of Foreign and European Affairs and the Ministry of Economy, Entrepreneurship and Crafts have in-house translation services. There is no regulation that would enforce the usage of computer-assisted translation (CAT) tools or translation memories (TMs) in public administration and its usage is left to an individual initiative and non-standardised licencing. Also, there is no central service that is responsible for translations at the level of government, ministries or state offices or agencies. Consequently, when translations are outsourced, TMs are not requested back by any of the outsourcing public administration bodies. Currently, there is no infrastructure in place to exchange translations or glossaries between ministries or to share translations with the national Open Data Portal.

Interesting fact:

The Central State Office for the Development of the Digital Society and the Faculty of Humanities and Social Sciences of the University of Zagreb are partners in the CEF funded action National Language Technology Platform (NLTP) that by 2023 aims to develop the CAT environment accompanied by NMT services and TM management, that would become part of eGov infrastructure, thus serving all public administration bodies.[26]

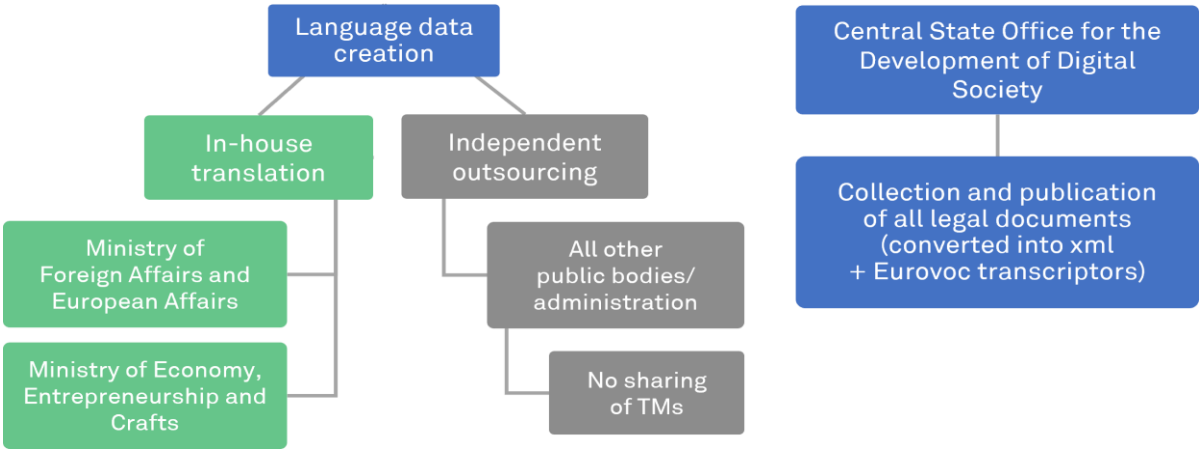
As part of the NLTP project activities, in March 2022 a survey on the importance of LT and their use in public administration was conducted. The survey was sent to more than 500 public authorities, with a total of 308 responses received. According to the responses, most state institutions do not use LT in their daily work – 45%, against 39% of those who claim to use them. CAT tools, terminology solutions and MT are the most used LTs. The most sought-after LTs that would greatly facilitate the respondents everyday work are CAT tools, ASR, Transcription (e.g. for meeting minutes), Terminology solutions and MT.

Respondents believe that users of public services could benefit the most from CAT tools, TTS/ASR and MT solutions, which would greatly improve access to public services.

The conclusion derived from the conducted survey is that there is great interest in LT and their introduction into the work of state administration bodies. The benefits of language solutions for digital public services users have also been recognised. However, it turned out that despite their interest, civil servants still know too little about LT and the importance of managing the language resources they produce.

[26] cf. https://www.nltp-info.eu

The current language data creation and sharing infrastructure in Croatian public bodies look as follows:



Open data and data collection in Croatia:

The Croatian Open Data Portal "data.gov.hr" is administered and governed by the Ministry of Public Administration and the Central State Office for the Development of the Digital Society. The Central State Office for the Development of the Digital Society is also responsible for the collection and publication of all legal documents that are converted into XML format and accompanied by Eurovoc transcriptors as metadata. The implementation of the Public Sector Information (PSI) directive is under the jurisdiction of the Information Commissioner's Officer and is fully transposed in Croatian national legislation (cf. Act 403/13, 2013 and extracts from Act 1065/09, 2009). All bodies of the public administration are obliged to make public sector information available and accessible in a digital and open format with appropriate metadata. This data is published on the national Open Data Portal, which is the main access point for the re-use of public sector information. However, there is very little language data available on the Open Data Portal as it currently mostly holds geolocation data, transportation data, meteorological data, environmental data, and other types of statistical data (ELRC, 2019, p.10).

Interesting fact:

The Central State Office for the Development of the Digital Society collects and publishes all legal documents of the Republic of Croatia in machine-readable format accompanied with relevant EUROVOC descriptors as metadata. They are openly accessible using the language-sensitive search engine CADIAL[27].

Digital policy and language policy in Croatia:

The usage of the Croatian language and Latin script as the official language and script is regulated by Article 12 of the Croatian Constitution. Other languages and Cyrillic or other scripts may be used together with the Croatian language and Latin script in individual local units and "under conditions specified by law" (Constitution, 2010, Article 12). Apart from that, there is no explicit language policy and language technologies are not mentioned in the National Strategy of Education, Science and Technology from 2014. The strategy was developed with the help of more than 130 experts in 19 different working and thematic groups and is under the responsibility of the Croatian State (National Strategy, 2014).

The role of LT and language data in Croatia's AI regulations:

The draft of the national strategy for AI was proposed in late 2019, but it was withdrawn after the announcement of changing EU regulations and the passing of the new directive on AI in 2021. In 2022 as part of Croatia's Recovery and Resilience Plan, the Government paid special attention to reforms and investments related to digital transition and transformation. This should contribute to the finalisation of both the National Plan for the Digital Transformation of the Croatian Economy and the National Strategy for the Development of Artificial Intelligence, as well as define the guidelines on the financing

[27] https://sredisnjikatalogrh.gov.hr/

needed to achieve the goals stated in the documents. Now a new strategy is expected to be drafted taking into account changed EU regulations as well as Croatia's recovery and resilience plan. Whether LT will be part of this new strategy remains to be seen.

Stakeholders and major networks:

The Central State Office for the Development of the Digital Society is an important administrative stakeholder as the State Office is the driving force behind the digitalisation process and is responsible for the national Open Data Portal together with the Ministry of Public Administration. CEF Telecom is coordinated by the Ministry of Economy, Entrepreneurship and Crafts and is therefore also an important stakeholder.

More than 30 organisations have participated in past ELRC events and several public administrations have already shared language data with ELRC. Among the data, donors are the Ministry of Regional Development and EU Funds, the Ministry of Agriculture and the Ministry of Physical Planning, Construction and State Assets.

Regarding the academic stakeholders, in Croatia there used to be a nationally funded programme for the development of LT for the Croatian language from 2007 to 2012 (Bašić et al., 2007), which set the foundations for the widening of the seminal research from the Faculty of Humanities and Social Sciences, the University of Zagreb to a number of other public institutions in Croatia that became relevant in LT, such as the Faculty of Electrical Engineering and Computing, University of Zagreb, Institute of Croatian Language and Linguistics, University of Rijeka, University of Split, University of Zagreb Computing Centre (SRCE), but it also involved private companies, such as Ciklopea or Integra. The Croatian Language Technologies Society established in 2004, has a mission to loosely coordinate LT activities in Croatia (Tadić, 2022).

Main challenges for sustainable data sharing:

- The public sector is much slower in adapting digital infrastructures/innovations than the private sector
- General lack of interest and awareness of the importance of sharing language data among the higher-level officials
- Concerns with respect to:
  - the control of the quality of the language data used in training the systems for machine translation (in terms of both the relevance of the documents and the quality of translations)
  - the impact of the type of the text on the quality of machine translation
  - the accessibility of the translation system to a wider audience (universities, translation agencies) (ELRC, 2019)

Action plan:

To address the above-mentioned challenges, the following recommended actions are considered vital:

- Raising awareness among decision-makers is regarded as the most important future step of the ELRC action in Croatia.
- Raising awareness about the importance of sharing language data that originated from public funding.
- Starting the initiative to establish a central translation office that would serve the Government, ministries, state offices and agencies with translation to and from Croatian when needed. Such a translation office could use the latest state-of-the-art resources and CAT tools in the translation process (centralised TMs, domain-dependent MT, general domain MT, etc.). The participation of two major stakeholders (The Central State Office for the Development of the Digital Society and Faculty of Humanities and Social Sciences of the University of Zagreb) as partners in the CEF-funded NLTP project, will provide technological fundaments for this initiative by 2023.
- More research in several LT domains:
  - Speech processing for Croatian in both directions: ASR and TTS.
  - More domain-sensitive Language Models that would boost the results in "traditional" tasks (sentence splitting, tokenisation, lemmatisation, POS/MSD-tagging, NERC, dependency parsing, semantic role labelling).
  - Very large corpora (mono- and multilingual) annotated with the tools mentioned above.

- Natural Language Understanding research: WSD, anaphora resolution, semantic parsing, semantic web technologies, automatic RDF-triples population from texts in Croatian, links between running text and conceptual spaces (e.g. WordNet, Wikipedia/DBpedia/Babelnet, etc.).

References and links:

CADIAL search engine for Croatian legal documents: http://www.digured.hr.

Central State Office for the development of Digital Society: https://rdd.gov.hr.

[Act 403/13, 2013] Right of Access to Information, *Zakon o pravu na pristup informacijama*, Act Nr. 403/13 of 8 March 2013: http://digarhiv.gov.hr/arhiva/263/100541/narodne-novine.nn.hr/clanci/sluzbeni/2013_02_25_403.html.

[Act 1065/09, 2009] General Administrative Procedure Act, *Zakon o općem upravnom postupku*, Act Nr. 1065/2009 of 1 April 2009, http://digarhiv.gov.hr/arhiva/263/44262/narodne-novine.nn.hr/clanci/sluzbeni/2009_04_47_1065.html.

[Bašić et al., 2007] Dalbelo Bašić, Bojana; Dovedan, Zdravko; Raffaelli, Ida; Seljan, Sanja; Tadić, Marko: *Computational linguistic models and language technologies for Croatian*. In: Lužar-Stiffler, Vesna; Hljuz Dobrić, Vesna (eds.), Proceedings of the 29th International Conference on Information Technology Interfaces (ITI 2007), pages 521-528, Cavtat, Croatia, Srce, Zagreb, 2007.

[Constitution, 2010] Committee on the Constitution: Standing Orders and Political System of the Croatian Parliament: *Constitution of the Republic of Croatia*, Consolidated Text, 2010, https://www.wipo.int/edocs/lexdocs/laws/en/hr/hr060en.pdf.

[ELRC, 2019] Tadić, Marko: *ELRC Workshop Report for Croatia*, 2019, http://www.lr-coordination.eu/sites/default/files/Croatia/2019/ELRC%2B%20Workshop%20Public%20Report%20Croatia_FINAL.PDF.

[National Strategy, 2014] Government of the Republic of Croatia: Strategy of Education, Science and Technology, *Nove Boje Znanja*, https://vlada.gov.hr/highlights-15141/archives/strategy-of-education-science-and-technology-nove-boje-znanja/17784.

[Tadić, 2022] Tadić, Marko: *Report on the Croatian Language (D1.7)*, In Giagkou, Maria; Piperidis, Stelios; Rehm, Georg; Dunne, Jane (eds.) Language Technology Support of Europe's Languages in 2020/2021, European Language Equality Project deliverable, Dublin, 2022.

Annex

Country Profile Cyprus

State of Play:

Translation practices and information exchange in ministries and public administrations:

The Press and Information Office (PIO) is the official communication service of the government, subordinated to the Ministry of Interior. Between 1990 and 2019, the PIO was the national authority for certified translations, recruiting associate translators from the private sector. After the passing of the Law on Sworn Translators in March 2019[28], the service for certified translations was transposed to the Sworn Translators.

As regards translation practices, there are no specialised in-house translation services in Cypriot public administrations. If the documents do not require certification (e.g. press releases), these are translated by bilingual or multilingual employees as part of their work within the administration. In the case of certified translations, it is obligatory for public administrations to outsource them to the Sworn Translators, as provided by the relative Law.

Currently, there is no central translation/terminology database yet and the use of machine translation (MT) or computer-assisted translation (CAT) tools is not standard (no full digitisation of translations).

With regard to the use of language technology, the European Commission's machine translation system, eTranslation,, has been available for all European small and medium-sized enterprises, including SMEs based in Cyprus, since 23 March 2020, through the NAP for Public Services.

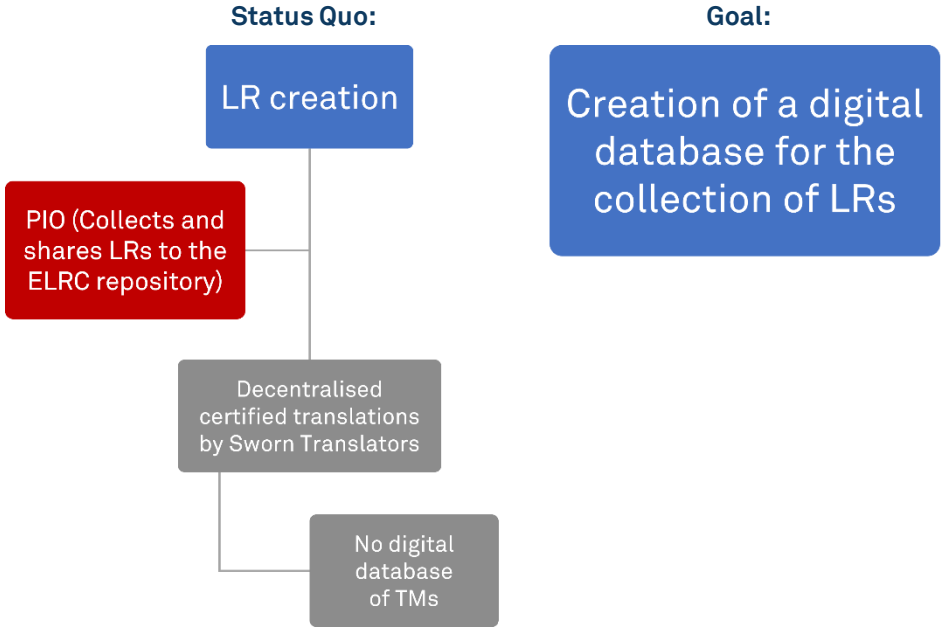
During the third Cypriot ELRC Workshop, the Public Services NAP pointed out that Language Technologies are behind almost every digital product we use. Language technologies such as machine translation are a means of overcoming language barriers and supporting linguistic diversity. It is important for the Greek language to survive in the new digital era and thus for the Greek culture along with the Cypriot culture to travel to the new digital world and this can be achieved by collecting as much language data as possible in order to improve the digitisation of the Greek language and with it also the Cypriot Greek dialect.

Interesting fact:

The PIO supported a system of translations both for the private and the public sector in 26 languages from 1990 until 2019.

[28] https://bit.ly/3PrjZ8f

The current language data creation and sharing infrastructure in public bodies of Cyprus looks as follows:



Open Data and data collection in Cyprus:

The PIO has been the major contributor of Language Resources to the ELRC depository since 2017. The PIO coordinated and informed the SMEs in Cyprus about accessing the eTranslation tool within the framework of the related pilot ELRC programme. Since 2017, the PIO has collected LRs both from public administrations, but mainly from its own data resources and shared them to the ELRC depository for the purposes of developing eTranslation.

Digital policy and language policy in Cyprus:

The National Digital Strategy of the Deputy Ministry of Innovation, Research and Digital Policy is based on four pillars: Digital Government, Digital Infrastructure, Digital Economy and Digital Society[29].

Stakeholders and major networks:

So far, several bi- and multilingual language resources were contributed by different public agents and the PIO, which has been the major provider of language resources for ELRC in Cyprus. Also in 2021-2022, the PIO has remained the main contributor of language data to the ELRC-SHARE Repository.

Main challenges for sustainable data sharing:

While the lack of digitalisation was mentioned as the key challenge back in 2019, during the 3rd Cypriot ELRC Workshop in December 2021, it was stressed that Cyprus has made tremendous progress in digitalisation of the public sector and it still continues to improve digitalisation in different areas. Also, the government had shown willingness to look into the integration of language technologies as part of the public services portfolio. With regard to the collection of language data it was stressed that the need to enhance the network of contacts in the different ministries remains a challenge. Furthermore, it was made clear that it is important that smaller languages like Greek and the Greek Cypriot dialect survive in the new era of digitisation and not Have the same fate as many other small languages/dialects in the past.

[29] https://bit.ly/3z2XuQi

Action plan:

In order to overcome the central challenge mentioned above and to enable sustainable data sharing, the following objectives were defined:

- **To develop good data management practices:**
  The establishment of a corresponding infrastructure for sharing language resources (through the creation of a central database) is an important step towards language data sharing. As a direct consequence, it is of utmost importance to develop and establish good data management practices: Existing data management practices need to be thoroughly investigated and updated, e.g. with regard to establishing responsible data managers, introducing a clear separation between confidential/personal data from public sector information, and establishing a practice of receiving any by-product of outsourced translations. Most importantly, the basis for the collection of linguistic data (corresponding system and processes on a national level) needs to be established.
- **To increase interest in MT and Language Technology (LT):**
  On policy level, several efforts are needed or already in progress to increase the interest in MT and LT as part of the national digital policy, including in particular:
  - Securing support of decision makers to change/adapt national policy with regard to language data: The Press and Information Office communicated the developments in the area of Language Technologies in the EU to the Deputy Ministry for Innovation, Research and Digital Policy, responsible for digital policy.
  - Creating synergies with related national agents.
- **To tackle legal issues:**
  In order to pave the way for the sharing of language resources, corresponding legal issues need to be tackled. On the one hand, this includes the development and sharing of easy-to-use guidelines for IPR and privacy issues. On the other hand, further support (in particular training) is needed with regard to the anonymisation of textual data.
- **To raise awareness of language data as Open Data and valuable asset:**
  In this context, several important activities, which require support are foreseen or already in progress.
- **Integrating language data in the national Open Data policy and digital agenda.**
- **Broadening the definition of textual resources by adding speech data, data for AI and other types of language resources:**
  First steps need to be taken in this direction.

References and links:

Cyprus Open Data Portal: https://www.data.gov.cy/?language=en.

Government Portal, Digital Services of Cyprus Government: https://www.gov.cy/en/.

Law on Sworn Translators in March 2019: https://bit.ly/3PrjZ8f.

The Registration and Regulation of Certified Translator Services in the Republic of Cyprus Law 2019: http://bit.ly/2lABxGc.

[Hadjioannou et al.] X. Hadjioannou et al.: Language Policy and language planning in Cyprus, 2011, https://www.researchgate.net/publication/232995910_Language_policy_and_language_planning_in_Cyprus.

Annex

Country Profile Czech Republic

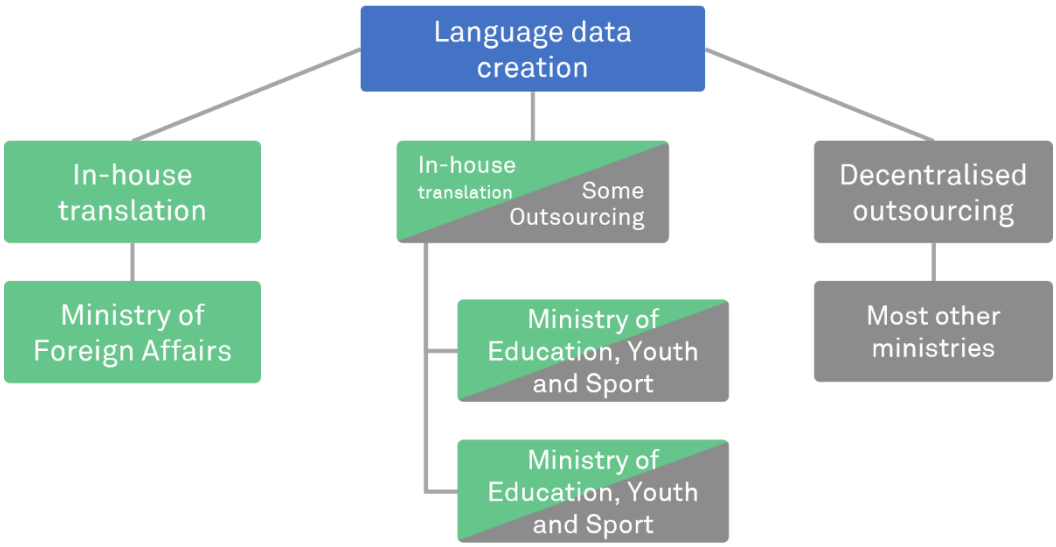
State of Play:

Translation practices and information exchange in ministries and public administrations:

The translation process in Czech public administrations is completely decentralised on the national level. Most public administrations meet their translation needs by outsourcing translations to language service providers (LSPs). The threshold for public procurement is 200,000 Kč (~8,000 EUR). Only the Ministry of Foreign Affairs, the Ministry of Education, Youth and Sports and the Czech Statistical Office have in-house translation services, whereas the latter two also outsource part of their translations.

As regards the applied translation practices, all LSPs and freelance translators are using computer-assisted translation (CAT) tools in their daily operations. The corresponding translation memories (TMs) or any other by-products of the outsourced translations are usually not requested nor automatically referred back, although the use of CAT tools is also common practice in public administrations and ministries. Some public services are also using a machine translation (MT) API or translate their documents with the help of freely available MT web services.

The current language data creation and sharing infrastructure in the public bodies of the Czech Republic looks as follows:



The situation related to LT development, digitalisation and data collection has changed over the last three years. The biggest difference is the use of large language models, and the transfer of practically all MT development to Deep Learning. In addition, speech technology (ASR/TTS) is used more and more in research – not only at the Institute of Formal and Applied Linguistics but also country-wide – as the ASR and TTS quality improves. More data is available in general; for example, in the Universal Dependencies collection the number of languages grew from 70 to 130+ and the number of treebanks to more than 200; for Czech three more treebanks have been added in the past three years. With respect to language resources and tools, the major improvement is a much larger morphological dictionary of Czech (Morfflex CZ), now fully compatible with the PDT-C 1.0 treebank, fully manually morphologically annotated with almost 4M tokens. Some new services have also been implemented and made freely available in the LINDAT/CLARIAH-CZ research infrastructure services list. In the same infrastructure, many Digital Humanities and Arts (DHA) data sets have been made available (such as digitised news lips and speeches released by the National Film Archive). A nationwide catalogue of language and DHA metadata and data is being created and will be available soon. It includes catalogued data from all

national libraries in the Czech Republic. However, only some of them are available for LT development because of copyright issues.

Open Data in the Czech Republic:

There is no central repository for translation or language data from public services in the Czech Republic and no infrastructure for continuous language data sharing yet. However, the National Open Data Portal was launched in 2018 encouraging organisations and citizens to share data under open licences although, currently, very little textual or language data is available on the portal.

Digital policy and language policy in the Czech Republic:

The Czech language is spoken natively by more than 9 million people, which corresponds to almost 90% of the population in the Czech Republic according to the 2011 census. In descending order, the following languages are spoken natively by minorities in the Czech Republic: Slovak, Moravian, Ukrainian, Polish, Vietnamese and German. Despite the dominant use of the Czech language, Czech is not declared the official or state language in the constitution from 1992 or in a language act (cf. Srpova, 2018, p. 296 ff.). It is however mentioned implicitly as the state language in other laws (cf. Zwilling, 2004, p. 3). Although there is no official language act that regulates the public use of Czech or protects minority languages, the Czech Republic accepted the European Charter for Regional or Minority Languages in 2006, granting national minorities rights such as "the facilitation and/or encouragement of the use of regional or minority languages, in speech and writing, in public and private life" (Council, 1992, p. 3) including for example the availability of "pre-school education in the relevant regional or minority languages" (cf. Srpova, 2018, p. 300; Council, 2018, p. 4).

Although not institutionalised, the Czech Language Institute is widely accepted as the regulatory body for the standard Czech language (Srpova, 2018, p. 293). However, since there is no "language act" or similar legislation, and several varieties of the Czech Language exist, there is no officially binding rule for the use of a standard variety except for the "codified orthographic and grammatical standard" (Srpova, 2018, p. 296 ff.) in Czech language lessons in the education system (only).

The undefined use of the Czech Language permeates in the digital policy, i.e. there is no explicit agenda for language policy or language technology in the digital agenda.

The government Digital agenda governed by the Ministry of the Interior is compiled of four pillars:

- The "Digital Czech" programme
- The "Digital Economy and Society" programme (long-term broad impact and sustainability, "which covers all aspects of government involvement, from legal aspects to direct support to research, development and innovation in the economy" in the digitisation process
- The "Information strategy of the Czech Republic"
- The "Czech Republic in the Digital Europe" agenda (cf. ELRC, 2018, p. 5)

The Digital Economy and Society programme has the widest scope as it "covers all aspects of government involvement, from legal aspects to direct support to research, development and innovation in the economy" (ELRC, 2018, p. 5) in the digitisation process. Overall, the Czech Republic offers more than 700 public services although only some of them are available digitally or multilingually so far.

The role of LT and language data in the Czech Republic's AI regulations

The National Strategy for Artificial Intelligence published in 2019 mentions speech and language processing in a separate chapter and includes language and speech technology among the most relevant ones within AI-classified Industry and research fields, with the third largest workforce among nine priority AI themes. Nevertheless, there is no commitment for a language programme (or similar) yet (cf. Digital Strategy, 2019).

Stakeholders and major networks:

The Ministry of the Interior is not only mainly responsible for the digital agenda, they also operate the Czech Open Data Portal and are therefore an important stakeholder. More than 30 institutions have participated in past ELRC events, among them the Ministry of the Interior, Social Security Office and Supreme Audit Office who are three of the Czech data donors who already contributed data to ELRC.

Main challenges for sustainable data sharing:

- The availability of data is still the biggest concern in the Czech Republic, also in connection with the fact that the 2019 CD has not been transformed into a national legal system yet. Also, the size of Czech resources, even if covered quite well overall, is still clearly below the major languages. However, the quality of MT (ENCS specifically) has improved dramatically in the past three years.
- The overall objective is to have methods, algorithms and ready-made system(s) for full Natural Language Understanding, be it performed by Deep Learning alone or in combination with symbolic methods and/or databases. In any case, data is certainly important.
- Identifying gaps in technology and data is the next important goal. It is still not clear which applications are possible now and in the next decade with current technologies, which improvements are possible with incremental development, and which will need a breakthrough.
- Another big challenge is to ensure the availability of high quality, clean data. Czech needs to have larger Language Models than those currently available (BERT, GPT-like, TMs).
- In order to secure sustainable development of language technologies on the national level, funding should be available long-term with longer perspective, which is currently not the case in the Czech Republic, since all research, including infrastructural support, is project-based only. Nationwide public funding in the form of a language programme, like the one in Spain, would be a relevant approach.

Action plan:

The following action items would be most relevant to facilitate data sharing:

- To tackle legal problems – current legislation mentions language data marginally without any detailed plan (finance-wise including)
- Raising awareness of language data as open data and valuable asset
- Increasing interest in MT/ LT in public services and SMEs as part of the national digital policy
- Establish good data management practices in public services and SMEs
- Identify and gain access to outsourced translations

References and links:

Centre for Language Research Infrastructure in the Czech Republic: https://lindat.cz.

Czech Open Data Portal: https://data.gov.cz.

[Council, 1992] Council of Europe: *European Charter for Regional or Minority Languages*, 1992, https://rm.coe.int/1680695175.

[ELRC, 2018] Hajic, Jan; Pecina, Pavel: *ELRC Workshop Report for the Czech Republic*, 2018, http://lr-coordination.eu/sites/default/files/Czech%20Republic/2018/ELRC%2BWorkshop Report_Public_CZ-FINAL.pdf.

[National Strategy, 2019] Ministry of Industry and Trade: *National Artificial Intelligence Strategy of the Czech Republic*, 2019, http://www.vlada.cz/assets/evropske-zalezitosti/umela-inteligence/NAIS_kveten_2019.pdf.

[Srpova, 2018] Srpova, Hana: *Forms of Language Planning and Policy in the Czech Republic*. In: Andrews E. (eds) Language Planning in the Post-Communist Era, 2018, https://link.springer.com/chapter/10.1007/978-3-319-70926-0_12.

[Zwilling, 2004] Zwilling, Carolin: *Minority Protection and Language Policy in the Czech Republic*, 2004, http://www.gencat.cat/llengua/noves/noves/hm04tardor/docs/zwilling.pdf.

Annex

Country Profile Denmark

State of Play:

Translation practices and information exchange in ministries and public administrations:

Only a few public institutions have in-house translation services such as the Region South Jutland-Schleswig, the Danish Tax Authorities and the Nordic Council of Ministers. Until the end of 2017, the Ministry of Foreign Affairs had an in-house translation service, which also delivered translation services to other public institutions. In 2018, the translation unit was dissolved and all translations are now outsourced to private vendors. Latest number indicates that about 80-90% of all translations in the public sector are outsourced to private vendors with a growth tendency.

Translation memories and other by-products of translations are not systematically requested from the private vendors together with the translation. The lack of a systematic approach to archiving and re-using translated text indicates that the importance and value of language data collection is not a recognised priority. In addition, there is currently only very few small Danish providers for translation memory systems.

Translation needs:

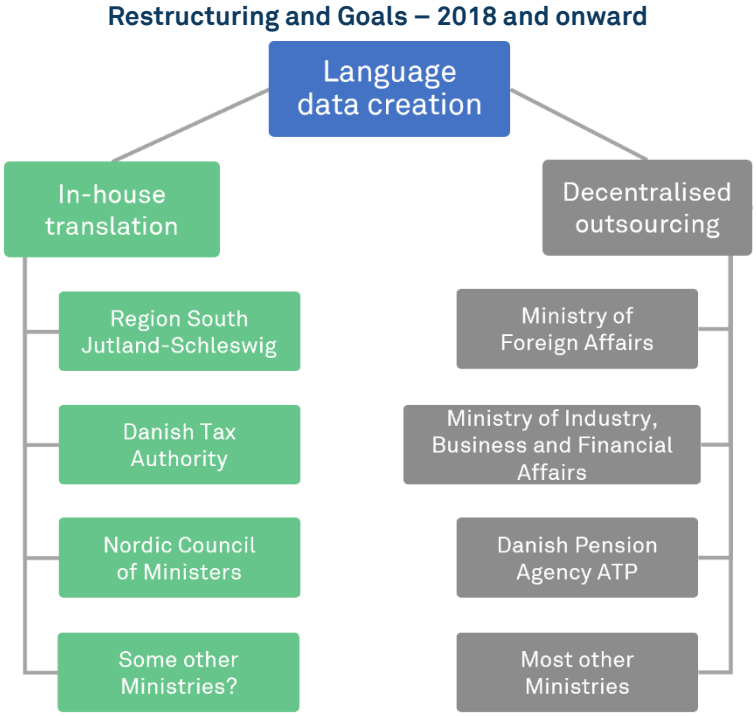
Denmark has a fairly small population of about 5.8 million residents constituting a small linguistic area for Danish – its sole official language. Still Denmark is a multilingual country with Faroese, Greenlandic and German as recognised regional languages, and Swedish also commonly spoken in the area around Copenhagen. Danish Sign Language is not officially recognised, but Danish Sign Language users are supported by the state, and the Danish Sign Language Council is responsible for providing documentation of and information about Danish Sign Language. Among the main immigrant languages are Arabic, Turkish, Polish and Romanian/Romani. Overall, English is the dominant second language, and English language competence among Danish citizens is very elevated. Translation demands, however, are still high for official EU-languages, non-official EU-languages and immigrant languages.

On the national level, for example, the Danish Agency for Labour Market and Recruitment has a strong need for translation from all EU-languages into Danish in the Electronic Exchange of Social Security Information system (EESSI), where electronic documents are exchanged across sectors. These electronic documents are often filled out by citizens through a combination of standardised text and open-entry text fields. This problem became even more evident as the number for Ukrainian refugees started to rise. Here Danish municipalities expected their translation resources to be exhausted within weeks.

Few examples are being brought up when it comes to automatic text translation. For example, the webpage www.lifeindenmark.com was translated into Ukrainian from English within a work day, with the help of eTranslation.

Also, the Danish parliament regularly needs translations into and from Greenlandic and Faroese, both not official EU-languages. One of the main needs, especially at the municipal level, is the translation of websites. Very few official websites are available in other languages. Currently, some municipalities have integrated a popular freely available machine translation service into their websites to meet the translation demands from foreigners living in Denmark.

The current language data creation and sharing infrastructure in Danish public bodies look as follows:



Digital policy and language policy in Denmark:

Although Denmark is a highly digitised country, not many (financial and human) resources have been allocated for the development of language technologies in the past. Recent developments, however, show a significant shift in the perception of the importance of language technologies for the preservation of the Danish language.

To this end, a language technology committee was established in 2018 by the Minister of Culture tasked with the development of a proposition for a national strategy for language technology in Denmark. This committee, headed by the Danish Language Council, produced a report in April 2019 including a list of recommendations and an overview of available resources. The committee simultaneously advised the Ministry of Finance who announced in October 2018 a new strategy for providing world class digital services including world class language technology – recognising that the two fields cannot be separated from each other. The report pointed at the following main issues as a barrier for the further development of Danish language technology.

- Danish is a small language community, which is why there is generally less language data available and the market for LT for Danish is small
- General lack of coordination of research programmes and strategic research projects complicate LR collection
- Little recognition of textual data being valuable in the public sector, and therefore no systematic approach to the curation and sharing of public data resources
- Only a few local developers of language technology products
- Strong tendency to outsource translation projects.

The Danish language technology council gave the following recommendations, for the further development of Danish language technology:

- A central organisation should be established to plan and manage a national Danish language resource bank.
- The language resource bank should contain high quality resources of the following kind:
  - A time annotated Danish speech technology corpus
  - Danish text corpora and annotated gold standards for machine learning

  - A comprehensive lexical database
  - A Danish terminology bank
  - A language technology toolkit
  - A language portal for the distribution of resources
- More education of training of experts in language technology for Danish
- More research on language technology for Danish (cf. Danish LT Report, 2019, p. 63 and 72).

Besides the governmental strategies Danish universities has also taken initiative on AI and LT areas. In 2021, a large group of Danish universities announced the start of a new collaborative centre, the AI Pioneer Centre. The centre fosters coordination and collaboration between researchers on a variety of different AI related topics. Seven collaboratories has already been established, with one collaboratory specifically focused on speech and language technology.

Open Data and data collection in Denmark:

When it comes to open government data in general, access is provided through The Data Distributor (datafordeler.dk), the Danish Open Data Portal[30], The Data Catalogue[31] and several other portals. Based on the Data Agreement from 2012, public sector data must be made available to public bodies, citizens and businesses.

In the 2022 government budget plan, €7.3 million were allocated to the development of a new Data Portal. The new data portal should provide users with an overview of the many data sets available for re-use by companies, researchers and citizens. In order to solve the challenge users currently have in figuring out which of the many data platforms to go to in order to find a particular data set. The new data portal was released in September 2022 by the Agency for Digital Government and the Danish Business Authority[32].

There has been a surge in activities related to the collection and sharing of Danish language resources as of late. The Centre for Language Technology at the University of Copenhagen is still effectively the Danish node for DK-CLARIN and is hosting one of the infrastructures/portals in Meta-SHARE. There are also similar initiatives in place at other universities. In fact, all university centres involved in language technology and language data, tend to have their own GitHub repository where Danish and other language data are shared. The Alexandra Institute (an advanced non-profit technology group) are creating, gathering and sharing Danish language resources and is hosting a Danish NLP network and GitHub repository, DaNLP[33]. The Danish Agency for Digital Government is hosting the catalogue sprogteknologi.dk, that collects and shares metadata on already available language resources for Danish. Further, The Danish Agency for Digital Government is also working to create new resources and to publicise existing resources not yet available.

A handful of public organisations, for example The Danish Research Council and the Ministry of Higher Education and Science, have from time to time allocated funds for the collection of language resources. These resources consists of scrapes of their Danish webpage and the English translation hereof that are transformed into a multilingual corpus, which is now available on the ELRC platform. These initiatives and processes have not been coordinated and standardised so far. In addition to this, debates in the Danish parliament are being broadcasted through TV and radio. These recordings along with extensive summaries of the debates are made available for the public.

The role of LT and language data in Denmark's AI regulations

In the "National Strategy for Artificial Intelligence" published by the Ministry of Finance and the Ministry of Industry, Business and Financial Affairs on behalf of the Danish Government in March 2019, the Danish Government emphasised the use of new technologies, such as AI in a variety of sectors, in order to improve public services. One of four key elements are the creation of language resources that can be widely used to develop language technologies for Danish. In the strategy, the Danish Government

[30] https://portal.opendata.dk
[31] https://data.virk.dk
[32] https://www.datavejviser.dk
[33] https://github.com/alexandrainst/danlp

acknowledged the importance of data collection by naming more and better data as one of four focus areas next to: a responsible foundation for AI, strong competences and new insights, and increased investment (cf. National Strategy, 2019).

In the National AI strategy one initiative were explicitly related to the development of Danish LT: the establishment of "the Common Danish language resource". The aim of the initiative is to enhance development of Danish language technology solutions by providing access and overview of existing Danish language resources. In the following "Joint Government Digital Strategy", which is a collaboration between central, regional and local governments the "Common Danish Language Resource" received €2.6 million running 2019 – 2026. The Danish Agency for Digital Government is responsible for the initiative.

In 2020, the agency launched the platform sprogteknologi.dk, which is the digital platform of the Common Danish Language Resource. The main purpose of the portal is to make Danish language resources accessible and thus make LT approachable for all types of organisations. However, the initiative is currently running multiple different activities to support the development of Danish LT and will continue to do so in the future, i.e. contributing to the organisation of the national ELRC Workshops, the development of the "Central Word Register" for AI purposes, collection and curation of language resources, and development of new "high-value" speech data set containing conversations and read-aloud.

The development of Danish language technology has also been included into the latest Joint Government Digital Strategy 2022 – 2025, which is a collaboration between central, regional and local governments in Denmark, the initiative ("A Common Danish Language Resource") was reemphasised once again.

Within the 2022 annual state budget, a new digitalisation fund received 67€ million in funding. These funds will be distributed to a wide array of initiatives that supports further digitisation of Denmark through new technologies and AI. Further, 7.3€ million has been allocated to the establishment of the "Datavejviser" to enhance public data sharing and re-usage.

Stakeholders and major networks:

- Danish Agency for Digitisation (Ministry of Finance)
- Danish Business Authority
- Centre for Language Technology/University of Copenhagen (Ministry of Education and Science)
- The Danish Language and Literature Society (https://dsl.dk/)
- The Danish Language Council (Ministry of Culture)
- IT University Denmark
- Technical University Denmark
- Aarhus University
- DaNLP (https://medium.com/danlp).
- StrømbergNLP (https://stromberg.ai/)
- Center for Humanities Computing at Aarhus University (https://chcaa.io/#/)
- AI Pioneer Centre, speech and language (https://www.aicentre.dk/)

Next steps:

The action plan for the initiative on Danish language technology is managed by The Danish Agency for Digital Government, who is responsible for the continuous development of the earlier mentioned platform sprogteknologi.dk. The action plan is developed in an agile method, which means that the action plan undergoes yearly reviews. This allows for a more responsive development that takes into account recent developments within Danish language technology and the needs of different organisations.

- The Central Word Register for Danish, which consist of incorporating and upgrading existing Danish digital dictionaries, terms and lexical resources is being developed
- The development and implementation of a time-encoded Danish speech corpus
- Further identifying and developing new language resources in cooperation with stakeholders
- Expand knowledge and promote new language technology development and uptake in the public sector

References and links:

The Common Danish Language Resource: https://sprogteknologi.dk.

[AI Strategy, 2019] Danish Government: *National Strategy for Artificial Intelligence*, 2019, https://en.digst.dk/media/19337/305755_gb_version_final-a.pdf.

[Danish LT Report, 2019] *Danish Language Technology Report*, 2019, https://dsn.dk/wp-content/uploads/2021/01/sprogarbejdet-i-danske-kommuner-2016.pdf.

[Digital Strategy, 2019] Agency for Digitisation, Ministry of Finance: *New national strategy: Artificial intelligence should benefit individuals, businesses and society as a whole*, 2019, https://en.digst.dk/news/news-archive/2019/march/new-national-strategy-artificial-intelligence-should-benefit-individuals-businesses-and-society-as-a-whole/.

[Digital Strategy, 2022] Agency for Digital Government: *New Joint Government Digital Strategy aims to overcome societal challenges*, 2022, https://en.digst.dk/news/news-archive/2022/june/new-joint-government-digital-strategy-aims-to-overcome-societal-challenges/.

[ELRC, 2018] Kirchmeier Sabine: *ELRC Workshop Report for Denmark*, 2018, http://lr-coordination.eu/sites/default/files/Denmark/2018/ELRC%2B%20Workshop%20Report_Denmark.pdf.

[Ingemansson et al., 2017] Ingemansson, Meyer, Kjærgaard & Kirchmeier: *Sprogarbejdet i danske kommuner. Rapport. Dansk Sprognævn*, 2017, https://dsn.dk/wp-content/uploads/2021/01/sprogarbejdet-i-danske-kommuner-2016.pdf.

[Kirchmeier et al., 2019] Kirchmeier, Diderichsen, Hansen & Henrichsen: *Dansk Sprogteknologi i Verdensklasse, Rapport fra sprogteknologiudvalget under Dansk Sprognævn nedsat af Kulturministeriet. Dansk Sprognævn*, 2019, https://dsn.dk/udgivelser/sprognaevnets-udgivelser/sprognaevnets-rapporter/sprogteknologi-i-verdensklasse.

[National Strategy, 2022] Danish Government: *National Strategy for Digitalisation*, 2022, https://en.digst.dk/news/news-archive/2022/may/the-government-launches-the-new-national-strategy-for-digitalisation/.

[Rigsrevisionen, 2019] Folketinget Statsrevisorerne: Extract from *Rigsrevisionen's report* submitted to the Public Accounts Committee Open data, 2019, https://uk.rigsrevisionen.dk/Media/6/5/12-2018.pdf.

Annex

Country Profile Estonia

State of Play:

Translation practices and information exchange in ministries and public administrations:

In Estonia, each public sector institution is responsible for its own translation services. Translation needs or procurement services are not centralised although all public bodies outsource at least part of their translations, either independently or through public procurement for order amounts above the threshold of 10,000 EUR. Only the Ministry of Foreign Affairs and the Ministry of Justice have in-house translation services but they also outsource part of their translations. So far, translation memories (TMs) are not requested back, as there does not seem to be a need for it, especially without an in-house translation service that could make use of it and/or maintain the TMs. Still, awareness raising in the past couple of years triggered a shift in public procurement processes as the benefits of re-using TMs have been recognised by state authorities (cf. ELRC, 2018, p. 15).

Together with the Ministry of Education and Research, the Ministry of Economic Affairs and Communications conducted a survey about the usage of TMs and translation arrangements during 2018 in the public sector in September 2019. One of the survey's objectives was to raise awareness of re-using TMs. 58 public sector organisations participated in the survey. Only a few institutions have in-house translators, who use different TMs and only one institution is using eTranslation officially. Other machine translation (MT) systems are used irregularly and for personal use only. Most translations are to and from English, but also Russian, Finnish and other Baltic languages are being translated. Exact numbers of translated pages cannot be determined, but based on the survey, around 1 million pages were reached in 2018 and translation costs sum up to more than 1 million euros per year. According to the survey, public administrations and ministries would be interested in a centralised MT of TM-based domain-dependent systems, which considers their specific terminology.

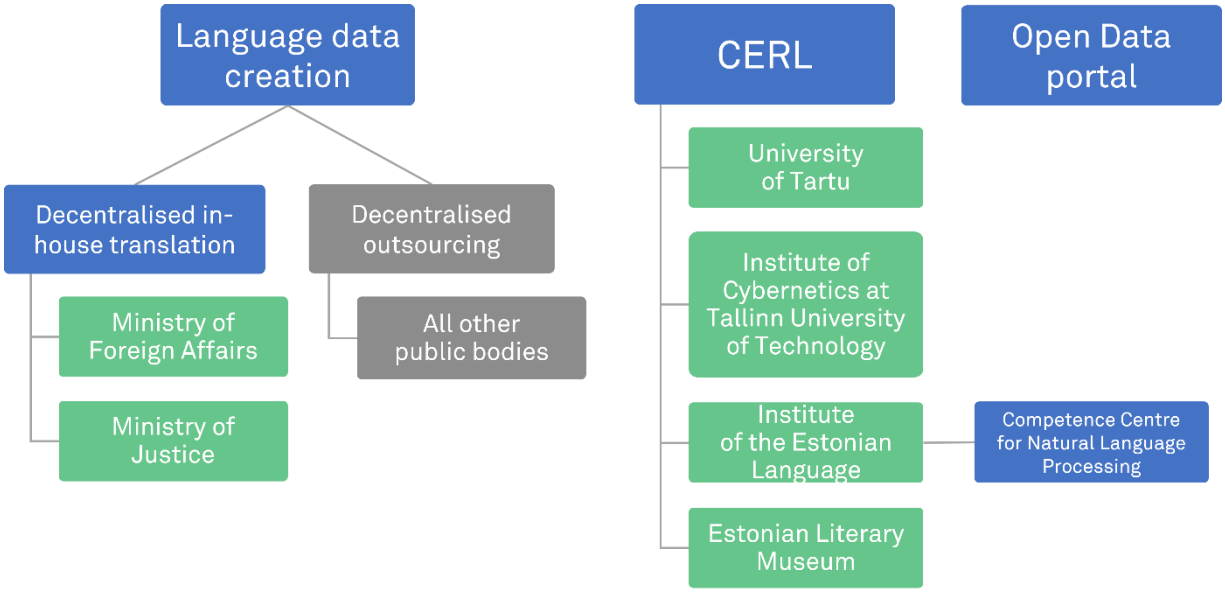
Data sharing practices:

There is no central language-related data exchange infrastructure for the public sector on the national level in Estonia. However, language technology resources are collected and shared mainly through the repository in the Center of Estonian Language Resources – CERL. The CERL also serves as the national node of the CLARIN infrastructure (cf. Center of Estonian Language Resources).

During the third Estonian ELRC Workshop, the following conclusions were drawn:

- Language resources for training domain-specific engines are scarce and require much effort to collect or create manually or synthetically
- There is a legal framework soon to be established to collect public sector translations and TMs which is expected to help in MT training tasks.

The current language data creation and sharing infrastructure in Estonian public bodies looks as follows:



Open Data in Estonia:

The Estonian Open Data Portal (https://opendata.riik.ee/en) provides a single point of access for the general public to unrestricted public sector data with the permission to re-use and redistribute such data for both commercial and non-commercial purposes.

Digital policy and language policy in Estonia:

Estonian is constitutionally the state language in Estonia and also one of the official EU languages. The current Estonian language development plan (2021-2035) is used as a basis for the sustainable development of the Estonian language. The strategy is used as a blueprint for planning and financing all four areas covered with a special focus on Estonian as a first language (L1).

Strategic planning for the development of the Estonian language started in 1998. The current strategy covers four areas: Estonian as first language; Estonian as second language; Estonian abroad and multilingualism, including foreign languages. In Estonia, the Ministry of Education and Research is responsible for the development of language policy.

The digital future of the Estonian language highly depends on the state of Estonian language technology (LT). By building resources and investing into technologies required for machine translation, speech recognition and speech synthesis, the position of Estonian in the digital sphere will be strengthened.

The role of LT and language data in Estonia's AI regulations

During the third Estonian ELRC Workshop, the Estonian Language Technology Programme was presented (cf. LT Programme, 2018). The position of Estonian language technology in the languages with the same number of speakers is stable. As a result of consistent work, important basic technologies have been developed, and applications that are actually used by the end-user for speech recognition, speech synthesis, and machine translation have been created. Applications are based on extensive language resources and text analysis tools. Through the programme, the state supports a field where it is not always profitable for the private sector to take on the risks associated with the development of technology for a language with a small number of speakers – as a small number of speakers also means a small market. The activities of the research and development programme "Estonian Language Technology 2018-2027" supporting the development of language technology will implement the objectives of two sectoral strategies: the research and development and innovation strategy "Estonian Research and Development, Innovation and Entrepreneurship Strategy 2021-2035" and "Estonian Language Strategy 2018-2027". The programme focuses on LT-based technologies: speech, machine translation, text analytics tools, and corpuses. The outcomes of the programme are freely available.

The recent important projects were presented: subtitling broadcasts, the Machine Translation project Mtee, Grammar Checkers and the public sector virtual assistant #bürokratt. A LT competence centre has been established to speed up the work in the area.

Stakeholders and main networks:

Key stakeholders include the Ministry of Education and Research, the Ministry of Economic Affairs, the Ministry of Justice, Competence Centre for Natural Language Processing in The Institute of the Estonian Language, and the Center of Estonian Language Resources. The second Estonian ELRC Workshop received 73 registrations, spanning a wide range of ministries and public organisations, but also language service providers (LSPs) and academia. Before the workshop, targeted communication activities were organised to ensure that key relevant public administrations were represented. With almost 50 participants, the workshop was well-attended. Over 40% of the participants were representatives from public services and public administrations, 35% were participants from the technology eco-system and 14% were language service providers. The remaining participants were part of the organising committee and the European Commission.

Main challenges for sustainable data sharing:

- Incoherent data sharing practices
- The low value of language data
- Little awareness of the potential of language data
- Lack of available language data to train domain-specific engines
- Legal issues relating to the creation and use of language data and language models

Action plan:

- Based on the identified challenges, the following six objectives were defined:
- Raising awareness of language data as Open Data and making language data as open as possible.
- Establishing good data management practices in public services.
- Identifying and gaining access to outsourced translations.
- Increasing interest in MT in public services.
- To build a Translation Hub platform http://tolkevarav.eki.ee/ which provides text and speech translation services for public sector and for the general public. The platform will combine human and machine translation, facilitate TM maintenance and sharing, and take advantage of termbases. Customised translation workflows allow each user organisation to meet its specific needs. A roadmap foresees to develop the platform's code in cooperation with other countries interested.
- The Ministry of Education and Research is currently preparing a regulation "List of Language Data Sets conditions and procedure for their publication and re-use" based on Public Information Act, where certain types of language data created within the public sector will be considered as high-value data sets.

References and links:

Artificial Intelligence for Estonia: https://www.kratid.ee/in-english.

Center of Estonian Language Resources: https://www.etag.ee/en/funding/infrastructure-funding/core-infrastructures/center-of-estonian-language-resources/.

[ELRC, 2018] Luts, Martin: *ELRC Workshop report for Estonia*, 2018, https://lr-coordination.eu/sites/default/files/Estonia/2018/ELRC%2B%20Workshop%20Report%20Tallinn%202018_v1.0.pdf.

[ELRC, 2021] Tilde Eesti: *ELRC Workshop report for Estonia*, 2021, https://lr-coordination.eu/sites/default/files/Estonia/2021/ELRC%20Tallinn%202021%20Report%20v3_final%20-%20public.pdf.

[Estonian LT, 2018] Ministry of Education and Research: *Estonian Language Technology 2018-2027*, https://www.hm.ee/sites/default/files/estonian_language_technology_2018-2027.pdf.

[Estonian Language, 2019] Ministry of Education and Research: *Estonian Language and Culture in the Digital Age 2019-2027*: https://www.hm.ee/en/activities/research-and-development/research-programmes.

[Estonian Strategy, 2021] Estonian Research and Development, Innovation and Entrepreneurship: *Strategy 2021-2035,* https://www.hm.ee/sites/default/files/taie_arengukava_kinnitatud_15.07.2021_211109a_en_final.pdf.

[Language Strategy, 2020] Ministry of Education and Research: *Estonian Language Strategy 2021-2035*, https://www.hm.ee/sites/default/files/htm_eesti_keele_arengukava_2020_a4_web_en.pdf.

[LT Programme, 2018] Ministry of Education and Research: *Estonian Language Technology 2018-2027*, https://www.keeletehnoloogia.ee/en.

Annex

Country Profile Finland

State of Play:

Translation practices and information exchange in ministries and public administrations:

Finland is one of the few countries that has a fully centralised translation service on the ministerial level. In 2015, the translation services from all the 12 government ministries were regrouped in a centrally organised Translation and Language Services Division (TLD) located at the Prime Minister's Office. Its 77 language specialists now provide translation, language and terminology services to all the ministries covering about 50% of the translation needs.

Interesting fact:

Finland has a centralised translation service for all 12 government ministries.

The incoming requests are handled with an internal service management tool called "Virkku", which is also used for the management of other services. TLD uses computer-assisted translation (CAT) tools including Government Machine Translation Service Aura for all translations and has an internal term bank and translation memories (TMs) as well as the government external term bank "Valter" (www.valter.fi). New term bases published in Valter will also be published in Open Data format on the Finnish Open Data Portal (www.avoindata.fi/en). However, the Government hesitates to share TMs as they also contain all non-public pre-final versions.

50% of the translations are centrally outsourced to language service providers. It is stipulated in the translation contracts that language service providers (LSPs) have to transfer translations, copyrights of the translations and all TMs to TLD as a minimum criterion. This criterion is not negotiable. TLD provides exports of their TMs to LSPs if they are relevant to the requested translation. However, the returned TMs are archived separately. Occasionally, TLD also exchanges TMs with the Finnish Parliament or other public bodies for specific translations.

The TLD coordinates the government level translations and language resource exchange and in addition, supervises and develops language usage in the ministries. The TLD introduced two customised public neural machine translators (EU Council Presidency Translator) 2019 and Government Machine Translation Service Aura 2021, which is tailored for ministries. It would have been impossible to develop these without the open sharing of language data.

Other government agencies take care of their translations independently with either in-house translation or outsourcing translations to language service providers. The same process holds true for translation needs on the regional and local level. Because of the bilinguality of the country, the demand for translations to and from Swedish is considerable.

Finnish Language Technology

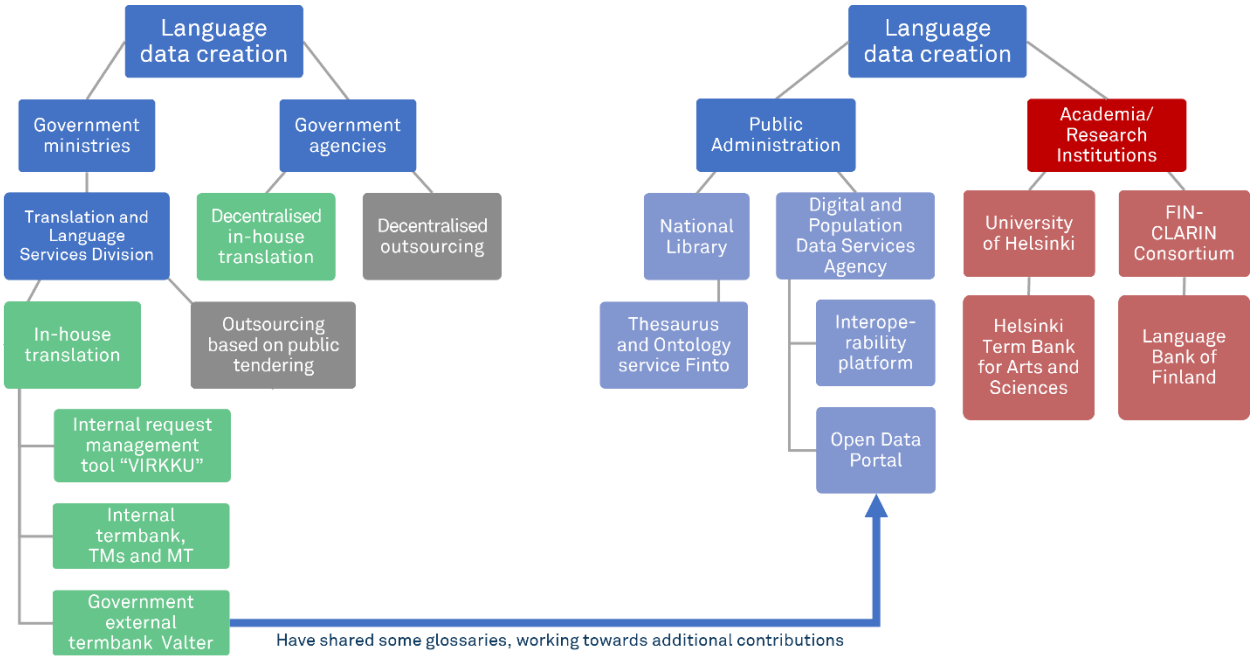
Looking back at its traditions in computational linguistics and software engineering, Finland is strong in Language Technologies. The attitudes towards MT have become more positive in Finland during the past few years. MT systems are developed actively and used more and more widely. Especially in the public sector, the EU Council Presidency Translator has been a success story, and has encouraged many public sector organisations to introduce their own tailored machine translation engines.

For developing LT models and tools with good coverage for various domains, it is essential to have access to large quantities of training data. However, there are still legal, technical and practical issues to solve in order to put more incentive on sharing data and to make data sharing and reuse more convenient for various stakeholders.

The current trends in LT tools include multimodality and interactivity. User interfaces must be able to process dialogue and conversation. Speech data is recorded in audio and video format instead of text,

and sign languages will also need attention. Many stakeholders are looking forward to speech interfaces that could support Finnish sufficiently well. One of the topics highlighted during the third Finnish ELRC Workshop in 2021 was the Donate Speech campaign that aims to collect Finnish speech data that can be legally used for academic research as well as for AI development in the private sector. The general concept, the speech recording interface and the experiences obtained during the campaign will be useful in other countries where similar data are required.

The current language data creation and sharing infrastructure in Finnish public administrations looks as follows:



During the third Finnish ELRC Workshop in 2021, it was noted that data management skills are becoming increasingly important for researchers and companies. Legal issues, such as copyrights and the evolving country-specific practices with regard to personal data processing still have a negative impact on data sharing.

Open Data and data collection in Finland:

Apart from the access to terminological data through Valter, the interoperability platform maintained by the Digital and Population Data Services Agency gives public bodies the tools and the method for specifying and managing interoperable data and information content. The platform consists of the terminologies, code sets, and data models needed for flows of data and other forms of information management. The Agency also provides the Opendata.fi service for publishing and utilising open data.

The TEPA term bank maintained by the Finnish Terminology Centre (Sanastokeskus ry) contains special language terms and definitions in approximately 365,000 terminological entries.

Another term bank is the Helsinki Term Bank for Arts and Sciences.

Additionally, the Finnish thesaurus and ontology service Finto is available online. The Annif is an open source tool for automated subject indexing and classification. It uses a combination of natural language processing and machine learning tools. The National Library of Finland provides both of these services.

Digital policy and language policy in Finland:

The two official languages in Finland are Finnish and Swedish. The Ministry of Justice is responsible for promoting the realisation of linguistic rights by monitoring the implementation and application of the Language Act and by issuing recommendations. The Ministry of Justice has also developed an indicator tool for following-up the implementation of the linguistic rights.

The revised Strategy for the National Languages of Finland was published in 2021. The aim of the strategy is to ensure that Finland continues to have two viable national languages. The strategy also examines the national languages of Finland from the perspective of digitalisation and artificial intelligence.

Finland is one of the few European countries that has an agreed artificial intelligence strategy, which includes aspects of language technology to make public services available to citizens in their own languages.

The role of LT and language data in Finland's AI regulations

In 2017, the Finnish Ministry of Economic Affairs and Employment published its first national AI strategy entitled Finland's age of artificial intelligence. This report fits under the umbrella of a broader Artificial Intelligence Programme in Finland (also labelled as AI Finland) with a view to establishing AI and robotics as the cornerstones of success for Finnish companies.

In 2019, VAKE (currently the Climate Fund) published a report on language-centric artificial intelligence development in Finland (Jauhiainen et al., 2019) pointing to neural networks suitable for deep learning as well as more traditional methods for machine learning. The report specified the next phase of the language-centric artificial intelligence development programme and collected topics in need of interventions.

In 2020, the Ministry of Finance launched the AuroraAI programme. The goal of AuroraAI is to develop an operating model for arranging public administration activities to support people in different life situations and events so that services provided by organisations function seamlessly between service providers in different sectors. Continuing until the end of 2022, the programme lays the foundation for using artificial intelligence to bring services and people together in a better way.

In 2020, Finland launched an updated national AI strategy. The Artificial Intelligence 4.0 Programme promotes the development and introduction of AI and other digital technologies in companies, with a special focus on SMEs. Finland's AI 4.0 Programme includes the following aims:

- strengthen digitalisation and economic growth in Finland
- encourage cooperation between different sectors, increase investments in digitalisation and improve digital skills in SMEs
- contribute to the recovery of companies and the economy from the coronavirus pandemic.

Stakeholders and major networks:

The key stakeholders related to language policy and language data sharing are:

- The Ministry of Justice who is responsible for monitoring the implementation of the language policy.
- The Translation and Language Services Division at the Prime Minister's Office who has the largest public administration translation service and is therefore the main language resource creator and holder in the public sector in Finland.
- The Digital and Population Data Services Agency (known as *Population Register Centre* until the end of 2019) that is responsible for the Interoperability Platform and the Open Data Portal, among other things.

The key stakeholders and decision makers for digitalisation and technology are:

- The Ministry of Finance provides preconditions for the digitalisation of the public sector and sets a strong example. This is done for instance, by promoting interoperability, AI and robotisations across administration and enabling the security of authorities' activities.
- The Ministry of Economic Affairs and Employment that is responsible for Finland's innovation and technology policy, among other things. The Ministry also promotes business digitalisation.
- The Ministry of Transport and Communications whose key duties include improving access to data, providing opportunities for data-based businesses by means of regulation, drafting legislation concerning data resources and the use of information.

In the past few years, almost 300 Finnish stakeholders that represent ministries and agencies on the government and the local level, public online services, language service providers and research institutions have been identified. Among the stakeholders that already shared language data with ELRC are the Translation and Language Services Division of the Prime Minister's Office, Statistics Finland, the University of Helsinki, the Finnish Terminology Centre TSK and the City of Helsinki.

Main challenges for sustainable data sharing:

Finland is one of the most advanced countries with respect to sharing language data that falls in the scope of public sector information. Still, it faces a few challenges that hinder language data sharing.

- Government hesitates to share TMs as they contain all non-public pre-final versions also. This is also a problem in government agencies and in other public bodies.
- Lacking awareness and knowledge of secure use of MT
- Anonymisation of language data

Action plan:

For Finland, 5 objectives were defined that will help to foster language data sharing infrastructures and awareness of the value of language data. Ranked according to their priorities, these recommended objectives and actions are:

- **Raising awareness of language data as Open Data**
  Some targeted actions are:
  - Including language data in the national Open Data policy and digital agenda
  - Establishing practical guidelines for Language Resources as Open Data
  - Sharing the benefits of sharing data
- **Increasing interest in MT/LT in public services as part of the national digital policy**
  Specific actions include:
  - Establishing synergies with national projects and initiatives
  - Diffusing best practices, where technology proves cost-cutting and increases productivity
  - Securing support of decision makers to adapt the national policy
  - Communicating facts about language data, such as how much data is needed to improve a MT system or details about the anonymisation process of data
- **Tackling legal concerns**
  This objective mainly addresses the development and distribution of easy to apply guidelines for Intellectual Property Rights (IPR) and privacy issues in textual data.
- **Raising awareness of secure training of MT engines and use of MT**
  This objective is targeted at establishing practical guidelines for secure MT training, taking into account information security and the safe use of MT services.
- **Establishing good language data management practices in public services**
  This objective is specifically targeted at the translation process. It includes:
  - A solution for the separation between confidential texts (e.g. working versions of official documents) from public information such as official publications.
  - Defining the best and most efficient process for sharing language data with minimal extra effort for anyone involved in the process.

References and links:

Finnish Interoperability Platform: https://dvv.fi/en/interoperability-platform

Finnish Open Data Portal: https://www.avoindata.fi/en

Finnish Thesaurus and Ontology Service (Finto): http://finto.fi/en/

Helsinki Term Bank for Arts and Sciences: https://tieteentermipankki.fi/wiki/Termipankki:Etusivu/en.

TEPA Termbank: https://termipankki.fi/tepa/en.

[AI Programme] Ministry of Economic Affairs and Employment of Finland: *Artificial Intelligence 4.0 programme*, https://tem.fi/en/artificial-intelligence-4.0-programme.

[AI Report, 2019] Ministry of Economic Affairs and Employment of Finland: *Leading the way into the age of artificial intelligence. Final report of Finland's Artificial Intelligence Programme 2019*, 2019, http://julkaisut.valtioneuvosto.fi/bitstream/handle/10024/161688/41_19_Leading%20the%20way%20into%20the%20age%20of%20artificial%20intelligence.pdf.

[AuroraAI] Ministry of Finance: *National Artificial Intelligence Programme AuroraAI*, https://vm.fi/en/national-artificial-intelligence-programme-auroraai.

[National Strategy, 2021] *Strategy for the National Languages of Finland*. Publications of the Finnish Government 2021:87, 2021, http://urn.fi/URN:ISBN:978-952-383-967-0.

[Recommendations, 2017] Ministry of Economic Affairs and Employment of Finland: *Finland's age of Artificial Intelligence. Turning Finland into a leading country in the application of artificial intelligence. Objective and recommendations for measures*, 2017, https://julkaisut.valtioneuvosto.fi/bitstream/handle/10024/160391/TEMrap_47_2017_verkkojulkaisu.pdf.

Annex

Country Profile France

State of Play:

Translation practices and information exchange in ministries and public administrations:

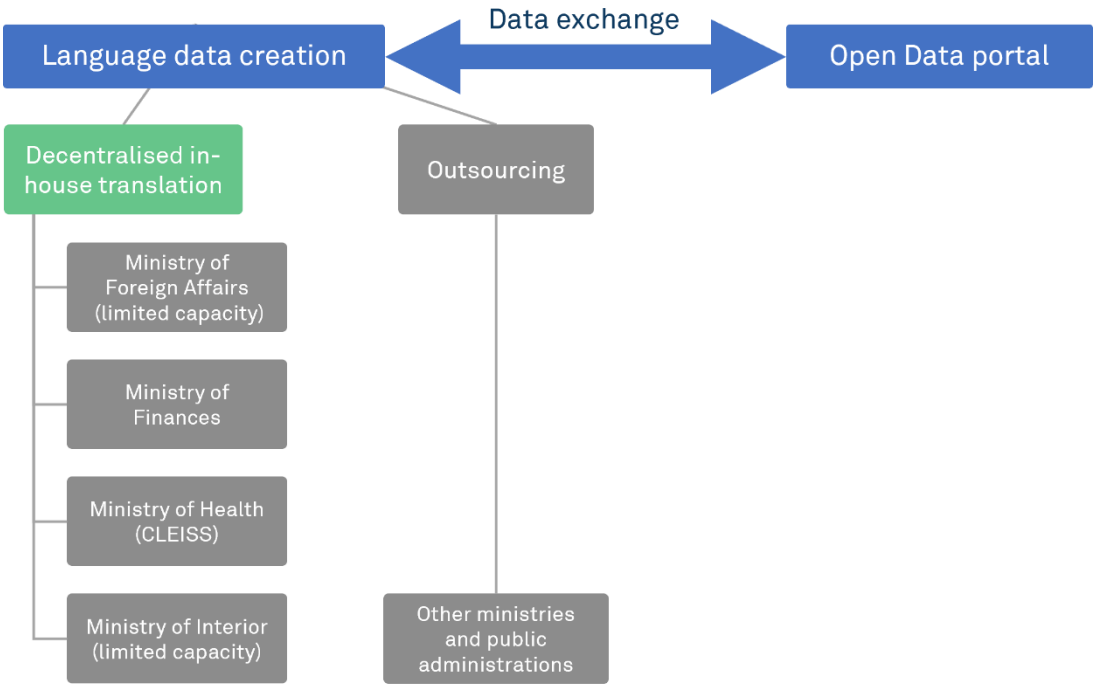
In France, there is no central translation service or procurement contract and no systematic exchange of translations and/or knowledge between public administrations. Translations for public administration bodies are mainly outsourced to Language Service Providers (LSPs) or freelance translators.

At the central level, two ministries have well-structured in-house services employing translators on a permanent basis with a specific civil servant status.

- The Ministry of Finances, which employs a team of 20 translators, produces approximately 30,000 pages each year using computer-aided translation (CAT) tools and maintains the Minéfiterm, a terminological database of 80,000 terms (in 15 languages, covering 40 domains). They mainly work for Tax and Budget services in the Ministry of Finances, but also provide translations for other administrations such as Customs. They outsource to trained freelance translators.
- The Ministry of Foreign Affairs' team is tighter with 13 translators working on translating civil status documents, but also speeches and diplomatic reports and documents for the President and the Prime Minister. They resort to freelance translators for all translations over 5,000 words. They also use CAT tools and maintain a terminological database. Some in-house translation departments can be found in other ministries, but they are usually very small and work only for their own administration. The service in charge of international affairs at the Ministry of Interior Affairs for instance has a very limited team and addresses the internal translation requests, mainly confidential.

The only other public administration with in-house translation services is CLEISS (Centre of European and International Liaisons for Social Security), France's single helpdesk for international mobility and social security which meets the translation needs of French social security institutions. With 50,000 pages being translated on average each year from 40+ languages, it is France's premier public translator.

The current language data creation and sharing infrastructure in French public bodies looks as follows:



Open Data and data collection in France:

The French government has acknowledged that sharing and enhancement of data and algorithms supports innovation and research and contributes to create and boosts the development of new uses, such as artificial intelligence. In April 2021, as part of the implementation of the data policy, data administrators (AMD) have been appointed in each ministry to develop the strategy for data, algorithms, and source code. A roadmap detailing the objectives related to the opening and sharing of data has been produced by each ministry (cf. roadmap of the Ministry of Culture). Following quality and interoperability standards, the open data of each ministry must be referenced on the data.gouv.fr, the portal managed by Etalab.

In addition to the official move to push for data sharing, there are still collaborative initiatives, such as the Inter-ministerial Working Group on Translation (GIT) co-chaired by the Ministry of Finances and the Ministry of Culture (DGLFLF), which takes place every 6 months for almost 15 years, and gathers all the translation professionals working in the public sphere, including translators, terminologists, academics, decision makers to exchange information and best practices. But this remains quite informal. The translation services in French public administrations remain reluctant to share their data, whether translation memories or translated documents, mainly out of confidentiality concerns.

Digital policy and language policy in France:

French is the official language of France and appears as such in the Constitution (article 2) since 1992. However, as the white paper on The French Language in Digital Age (Mariani et al., 2012) reminds us, the European Charter for Regional or Minority Languages the Constitution (article 75-1) acknowledges regional languages spoken in France as part of French cultural heritage, since 2009. Nevertheless, French remains the language mainly used in France, and there are strong constraints for its use in the public sphere (schools, public services, media, etc.)

In France, the Délégation générale à la Langue Française et aux Langues de France (DGLFLF), a body of the Ministry of Culture, oversees coordinating the French Government's language policy. Currently, DGLFLF plays a prominent role in the implementation of President Macron's plan: "An ambition for the French language and multilingualism" presented on 20 March 2018 at the Académie française, which encompasses measures such as the teaching of two European languages in addition to the native language as well as language training in European and international institutions. As part of the workplan

Culture 2018-2022, both translation and digital technology as a tool for multilingualism are the key objectives in the French Government's strategy. Both issues have also been highlighted during the French EU Presidency in 2022.

The role of LT and language data in France's AI regulations:

The objective of French National Strategy for AI is to make France a global leader in AI.

Investments will focus on research and education, data, IT infrastructures, etc. Numerous initiatives, programmes, projects, partnerships, etc, have been or will be set up at the national and European levels. The most recent, announced in November 2021, is the launching of the 'Artificial Intelligence' (AI) PEPR is part of the national acceleration strategy. It is dedicated to research to help breakthrough technologies emerge, and will focus notably on embedded AI, certified AI or the use of AI to accelerate the ecological transition.

Natural Language Processing, as a branch of AI, will also benefit from these mechanisms boosting research programmes, technology developments and partnerships in Language technology start-ups and companies. AI progress in the field of machine learning has had a strong impact on numerous fields including speech processing and NLP. Within the framework of the National AI Research Programme, the supercomputer Jean Zay has been setup and is increasingly being used by NLP projects. The largest of these projects to date is BigScience, led by Hugging Face, whose purpose was to train Bloom, the largest multilingual open-source language model, released in July 2022.

Stakeholders and major networks:

The French National Points represent two key institutions related to language policy (P-NAP from DGLFLF) and language technology (T-NAP from CNRS), which highlights the interest and importance of this topic in France. All the local ELRC events were attended by representatives from many public institutions, including ANR (research agency), Ministry of Finances, CLEISS, Etalab (the Open Data agency), CNRS, Banque de France and the French translators' union (SFT). Some of these institutions have already contributed language data to ELRC, namely the Ministry of Finances and the ANR. Other stakeholders like the Banque de France-ACPR have developed their own NMT solution as the confidentiality of the data is key to their activity.

On the French scene, there are several French and international institutions based in Paris which, having to deal with the growing volume of translation, are looking into the support that can be provided by language technologies, with a specific interest in the eTranslation developments. For instance, the OECD, based in Paris, is currently developing its own NMT engine for both their official languages French and English. There are also French public services such as OFPRA (French Office for the Protection of Refugees and stateless Persons) where daily activities involve interviews in up to 127 different languages or the Inter-ministerial Delegation for refugees' reception and integration (DIAIR), a body of the Ministry of Internal Affairs, which helps refugees make their way into the administrative procedures by providing multilingual information on a dedicated collaborative platform.

Main challenges for sustainable data sharing:

The challenges France is facing in sharing data are as follows:

- IPR Issues
  - Privacy concerns prevent ministries from sharing data (Finances, Interior, Foreign Affairs)
  - By law, translators own the translation and the translation memory, resulting IPR issues can hinder the sharing of language data.
- Translation workflow
  - LSPs and in-house translators know the management of translation data, which is not always the case in the public administrations.
  - CAT tools are not always used, even in institutions with in-house translators, because they deal with documents or formats for which CAT tools are not considered useful/needed.
  - TMs are not always part of the final translation deliverable (with the translated material) and some institutions simply lack technical staff to process the parallel texts into TMs a posteriori.

- Data sharing
  - There is no TM exchange because there is no infrastructure for this at the State level. The Open Data Portal is not used for this purpose in France. However, in some cases and for public documents and reports that do not contain sensitive information, translation memory files were shared by public translation centres with the DGT. It led to substantial progress in eTranslation output for the given domain.
  - In some institutions, the common perception is that since most of their documents are confidential, they cannot share even the reports that are public and could be useful for training the MT engine. The anonymisation feature, that is available in the NLP Tools section from https://language-tools.ec.europa.eu/, could help translation centres overcome the issue of sharing documents containing personal information.
  - In many cases, the sharing could be easily processed if the decision-maker at the administration level could be easily identified/convinced.

Action plan:

- Continue to promote the in-domain data as means to improve the performance of MT engines.
- Promote anonymisation features to convince the translation centres to anonymise their data before submitting them to eTranslation
- Work with the data administrator in the Ministry of Culture to identify data that can be shared on ELRC-SHARE

References and links:

Etalab, the "Chief Data Officer" for France: https://www.etalab.gouv.fr/.

French Open Data Platform: https://www.data.gouv.fr/en/.

[Bothorel et al., 2020] Eric Bothorel, Stéphanie Combes, Renaud Vedel: *Pour une politique publique de la donnée*, 2021, https://www.vie-publique.fr/rapport/277879-pour-une-politique-publique-de-la-donnee.

[Mariani et al., 2012] Mariani et al.: *The French Language in the Digital Age*. In: META-NET White Paper Series, 2012, http://www.meta-net.eu/whitepapers/e-book/french.pdf.

[Roadmap, 2021] The Ministry of Culture: *AMD Roadmap* (in French), 2021, https://www.culture.gouv.fr/Espace-documentation/Documentation-administrative/Feuille-de-route-Donnees-et-contenus-culturels.

Annex

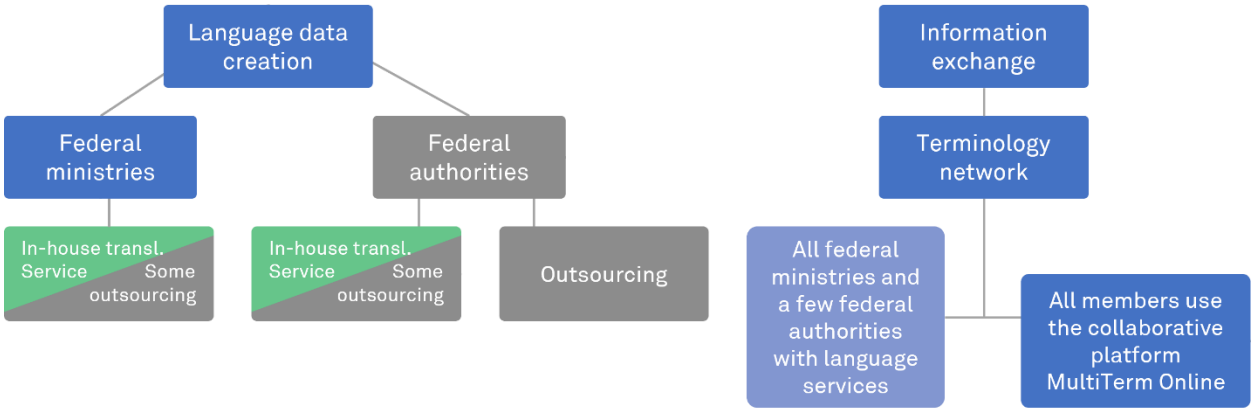
Country Profile Germany

State of Play:

Translation practices and information exchange in federal ministries and public administrations:

The way translations are carried out in Germany varies depending on the administrative level. All federal ministries have an in-house translation service whereas only some federal authorities and very few state ministries have their own translation service. It is the common practice to use computer-assisted translation (CAT) tools, including translation memories (TMs) and terminology systems (termbases), in the translation process both by in-house translation services of the federal ministries as well as by language service providers (LSPs) and freelance translators. To fully meet their translation needs, all public authorities outsource at least some translations to either freelance translators or language service providers. When it comes to outsourcing, three different scenarios exist: public authorities that have in-house translation services have a smaller demand for outsourcing translations and therefore do not usually call for tenders but vendor small contracts to freelance translators that provide high quality translations and are generally willing to share their translation memories with the contracting authority. Public administrations that do not have their own translation service outsource translations to language service providers and usually do not request that the translation memories and new terminological entries are returned to them as they do not see the need for it since the TMs will not be reused in-house. In general, all federal authorities may also order translations under a central call for tenders but are not obliged to do so.

The current language data creation and sharing infrastructure in German public bodies looks as follows:



Open Data and data collection in Germany:

There is no formalised exchange of information or data between public services unless they are part of the terminology network ("Terminologiedatenbankverbund"). The majority of federal translation services are part of the terminology network. Despite its name, terminology is not the main focus of this network. Its main purpose is to exchange information about translation practices and developments in the field. Additionally, all members use the collaborative platform MultiTerm Online granting reading access to terminology databases of other public services. However, there is no active exchange of TMs or textual data.

Interesting fact:

In Germany, the copyright (das Urheberrecht) of translations belongs to the translator by default and is not transferable. For rightful reuse of the texts, respective licences are needed.

In the Federal Government's National Action Plan to implement the G8 Open Data Charter, it is explicitly stated that: "Unstructured information such as notes, files, studies, reports or other texts do not constitute data in this sense" and therefore do not fall in the category of Open Data (cf. Action Plan, 2014). However, in the course of the past years, several contacts were established with the German Open Data community. The German Public Services NAP for example spoke to the Open Data coordinator at the Federal Ministry of the Interior and offered to participate in an Open Data pilot at the ministry to advocate for language data to be included in the pilot project. Some interest was also shown in publishing language resources (LR) on GovData, the German Federal Open Data portal. However, currently there is no function on GovData to search for language data, language resources or textual data on the portal.

According to the E-Goverment Act (see link below), public authorities ought to make data that is of public interest and can be shared according to the "Informationsverarbeitungsgesetz" (cf. Data Use Act, 2021), the German implementation of the Public Sector Information directive (2003/98/EC), available in machine readable format online. However, data is defined as structured information mainly available in tables and lists (cf. eGovernment Act). Although this definition does not explicitly exclude language data, it does show the implicit interest of statistical and numerical data and little interest in language data. It can be said that sharing data (textual or other) is still not an inherent part of the everyday work in public administrations in Germany and is a field with much unused potential: Germany is ranked 24th on the Global Open Data Index, in 10th place on the Open Data Barometer and comes in 8th place on the Open Data Maturity Dashboard on the European Data Portal (EDP)[34] underlining the fact that although data is an important resource in the 21st century, Germany only barely uses its potential (cf. Open Data, 2016). However, comparing the current Maturity Level Rating with those of 2018 (17th place) and 2019 (12th place), it can be stated that the situation in Germany keeps improving.

During the third German ELRC Workshop in April 2021, it could be noted that the situation with regard to language data creation, management and sharing practices in Germany has not significantly changed. Looking at the availability of language data in the organisations of the workshop participants, the vast majority of workshop participants (90%) confirmed to hold language resources. Unfortunately, the polls also revealed that less than one quarter (24%) of the respondents' organisations had a data management plan in place. The practices for the creation of multilingual parallel data (in particular translations) remain fragmented in Germany. There is no formal translation procedure common to all public administrations on the federal level, let alone all public services across the federal states. Some (especially the larger) authorities maintain in-house translation departments, others outsource their translations. In the majority of cases, the outputs are not managed according to a data management plan, translation memories of outsourced translations are not requested back, and, in some cases, translations are only available in Microsoft Word format. Legal issues as well as the absence of data management plans (or even guidelines governing the sharing of language data) in organisations remain the main barriers hindering the sharing of language data in Germany.

Several important developments, however, may help overcome this situation in the future and pave the way for an increased sharing of language data. The first major improvement may be seen in the Data Use Act adopted in July 2021 in Germany (cf. Data Use Act, 2021) which aims to significantly extend the provision of open administrative data on the federal level and hence to facilitate the re-use of such data (also explicitly allowing commercial usage of such data). Based on the new Data Use Act, public sector information shall be freely usable and available in machine-readable format, in all available languages, free of charge, with relevant metadata, via the National Open Data Portal[35], and, where necessary, with an open licence (e.g. CC-BY, DL-DE/zero, DL-DE/namensnennung). Especially the organised collection (including standardised metadata) and licencing are expected to contribute to the improved organisation, sharing and hence availability of language data. The second important development is the increased take-up of MT (see above) – and hence the pressure to prepare and share language data. So more and more organisations will need to internally review their processes for language data sharing and above all, address one key question: How to make relevant language data retrievable? This central question covers different aspects of the organisation, including:

[34] Information as of 25 July 2019
[35] https://www.govdata.de/

- the translation process – How can it be adapted in a way to enable the re-use of language data?
- the classification and metadata descriptions – How can metadata be usefully extended or modified to facilitate the identification/export of shareable data? How can contents be best classified?

With regard to the actual cleaning of data for MT training, organisations can outsource this process using a corresponding confidentiality agreement (in case of a one-time affair) or build corresponding human resources in-house in case this task will frequently arise.

In the private sector, the sharing of language data appears to be slightly more difficult given that such data are often considered as trade secrets and/or as information providing a particular competitive advantage which is why there is a considerable reluctance against the open sharing of this data. Even when sharing language data in a defined circle, the issue of appropriate remuneration is always present.

While it was stressed repeatedly that the availability of sufficient parallel data (e.g. translations) remains key for the development of MT systems in SMEs and public administrations, it could also be shown during the workshop that in cases where sufficient parallel data may not be available, scientific advances may soon provide viable MT solutions in the form of unsupervised MT or even self-supervised MT.

Recent developments:

In January 2021, the Federal Government adopted its Data Strategy[36]. The strategy's aims are: creating effective and sustainable data infrastructures; increasing innovative and responsible data use; improving data skills and establishing a data culture; and making the Federal Government a world leader in data use. Based on this strategy, the position of Chief Data Scientist or Chief Data Office will be created and data laboratories will be established in federal authorities.

In addition, the Federal Ministry of Economic Affairs and the Federal Ministry of the Interior and Community are currently working to establish a Data Institute („Dateninstitut") to facilitate data access and sharing[37].

Whether the labs and Data Institute will address language resources as well remains to be seen.

Digital policy and language policy in Germany:

The fact that language data or textual data is not considered a valuable asset may relate to the German language policy or better the lack thereof. Due to the federal organisation of the country, there is not a single ministry or public authority that is in charge of digital policy or language policy in Germany. In addition, there is no Language Council per se but a supranational Council for German Orthography (Rat für deutsche Rechtschreibung) representing several countries with German as (one) official language observing the developments of the German language and proposing corresponding adjustments that are then implemented into national legislation (cf. Adler et al., 2018, p. 227).

German is the sole official language in Germany and has a strong tendency towards being the single language for public usage. This is exemplified by the fact that the role of the German language for a number of historic and other reasons is not even mentioned in the German constitution (ibid., p 222 ff.). There are however four autochthonous minority languages and one regional language, namely Danish, Frisian, Sorbian, Romani and Low German, that are recognised as minority languages (ibid., p 222 ff.).

In a survey conducted by the Leibniz Institute for the German Language, 90.2% of the respondents indicated German as their first language (L1), but generally there is very little information about the use of language(s) in Germany (ibid., p. 221). A decreasing use of regiolects or dialects can be observed and immigration history shows that allochthonous languages are spoken but no figures are available to indicate their use in the population.

Given Germany's federal structure, its digital policy landscape is complex and multi-layered.[38]

---

[36] https://www.bundesregierung.de/breg-en/service/information-material-issued-by-the-federal-government/data-strategy-of-the-federal-german-government-1950612

[37] https://www.bmi.bund.de/SharedDocs/pressemitteilungen/DE/2022/10/dateninstitut-startschuss.html

[38] This diagram (in German) by the National Regulatory Control Council (Normenkontrollrat) illustrates the complexity of Germany's landscape of digital transformation in the public administration:

The Online Access Act of 14 August 2017 states that public services have to be made available electronically within 5 years on the federal and state level (cf. Access Act). All public services (including services of third parties, cf. Leika-plus, 2015) are in the process of being described in a dedicated catalogue ("Leistungskatalog") many of which are to be described multilingually.

Stakeholders and major networks:

In the context of sharing language data, the Federal Government Commissioner for IT as well as the heads of the translation departments across the ministries are the key decision makers. The Federal Office of Administration, an executive agency of the Federal Ministry of the Interior and Community, has a consultancy function for Open Data and is therefore also an important stakeholder. Overall, more than 30 institutions representing federal and state ministries, language service providers and research institutions attended past ELRC events. Some federal ministries actively contributed some of their language resources.

Among them are the Federal Foreign Office, the Federal Ministry of the Interior and Community, the Federal Ministry of Transport and Digital Infrastructure and the Federal Financial Supervisory Authority.

Language Technologies in Germany

When it comes to the take-up and acceptance of machine translation, the third ELRC Workshop in Germany revealed that MT is already part of German public administrations and SMEs. 97% of the workshop participants indicated that they already used MT, about half of them both eTranslation and other free systems. Most interestingly, almost 50% were either fully satisfied or very satisfied with the quality of MT for German which shows a great improvement in terms of translation quality compared to earlier years and earlier workshops. As such, it can be concluded that in the past 5 years, MT has become a core technology in and for public administrations and SMEs – a finding that is supported by the recent European Language Industry Survey (ELIS 2021). There was great interest among the workshop participants in the evaluation of MT systems for the German federal authorities. Key questions were about the process of the evaluation and how the translation quality was assessed. As could be shown, the involvement of translators from various translation services in the review process is vital for the success of such a large-scale evaluation: There must be a dedicated working group with staff members from key institutions on the one hand in order to gather and prepare relevant materials and frameworks. On the other hand, translators working as independent evaluators for assessing the quality of the MT outputs are needed. Last but not least, the development of a dedicated framework for quality assessment that would fit the needs of the participating institutions is vital.

As became evident from the live polls, information search and retrieval are the second most important LT according to the workshop participants (MT remains the most widely used technology by German public services and SMEs). In the particular context of the Federal Statistical Office, classification, however, was the key technology. Virtual assistants/chatbots, speech recognition and text-to-speech solutions were not yet widely used among the audience. One reason could be the low technological maturity of these technologies, e.g. for German. As shown by the live polls, more than two thirds (69%) of the workshop participants prefer to use LT in their own language (i.e. German). Less than one quarter use them in English. However, more than two thirds were also not really satisfied with the quality and reliability of the LT used in their language. Even more, only 8% of the workshop participants were actually satisfied with the digital readiness of German public services and SMEs. They see the main role of SMEs and public services as data stewards and users/service providers of LT, rather than as a regulator. With regard to the solutions side, the Catalogue of Services shows that more than 130 LT solution providers are actually based in Germany. This fact was also underlined when looking at the start-up scene in Germany: The majority of AI start-ups in 2021 in Germany could be assigned to the domain of LT companies, being either directly from the computer linguistics domain or supporting customer service and/or marketing functions of businesses. As such, LT other than machine translation are on the rise, and it can be expected that within the next 5 years, take-up will also increase in Germany.

---

https://www.normenkontrollrat.bund.de/resource/blob/72494/1957834/6021b69d1c029ea1f2a5bce908f14917/220320-wimmelbild-monitor-digitale-verwaltung-data.jpg

Main challenges for sustainable data sharing:

- Translators do not see benefits for themselves although their data could be useful for various other administrations
- It is noted that freely available online machine translation services are being used fairly frequently although the exploitation of the results is not known
- Texts are by default copyrighted by the author/translator
- Language data is not considered valuable
- No central coordination of translations or language policy in general

Action Plan:

To address the identified challenges, the following objectives and actions are proposed:

- **To tackle legal concerns:**
  - Develop and share easy to apply guidelines for Intellectual Property Rights (IPR) and privacy issues that can be followed by data creators and holders in order to decide whether their data can be shared
  - Investigate the idea to implement rights management along with data management, i.e. legal in-house support for data sharing in general
- **To identify and gain access to outsourced translations:**
  - Establish the practice of receiving any by-product of outsourced translations whenever translations are outsourced, irrespective of whether the contracting authority has an in-house translation service
- **To establish good data management practices in public services:**
  - The identification of data managers is ongoing and is considered important to introduce changes such as:
  - The practice of clear separation between texts that contain confidential and personal data from texts that fall under the public sector information directive (or the Data Use Act)
  - Establish practice that public administrations have the right to share and publish translations although the copyright belongs to the author/translator and is as such not transferable
- **To raise awareness of language data as Open Data and a valuable asset:**
  - To integrate language data in the national Open Data policy and digital agenda is an ongoing process. The language services division of the interior ministry offered to participate in an Open Data pilot to share LR with Germany's Open Data portal Govdata. The federal ministries' Open Data officers have been made aware of the value of LR as Open Data.
  - Establish practical guidelines for LR as Open Data
  - Continue to share benefits of sharing language data is considered crucial to achieve the above-mentioned objectives
- **To increase the interest in machine translation (MT) and language technology (LT) in public services as part of the national digital policy:**
  - Even though the interest in MT and LT has increased since the publication of the first country profile, it would be important to establish synergies with national projects and initiatives such as the evaluation of the need for an MT system for the federal administration.
  - Build networks to be able to exchange knowledge and experiences and to get a better picture of the use of LT and MT in Germany.
  - Secure the support of decision makers to change/adapt the national policy is an ongoing activity. The topic of MT and language data is frequently brought up in internal meetings with the language services divisions of the federal ministries and will be continued
  - Anonymisation is a central topic, important to many language data holders, however, internal expertise is lacking in this area, therefore support is needed

References and links:

Federal Government information page about the Online Access Act: https://www.onlinezugangsgesetz.de/Webs/OZG/EN/home/home-node.html.

Germany in the Global Data Index: https://index.okfn.org/place/de.html.

Germany's Government Cloud Strategy: https://www.it-planungsrat.de/fileadmin/it-planungsrat/foederale-zusammenarbeit/Gremien/AG_Cloud/20210813_DVS_-_Germanys_government_cloud_strategy__-_target_architecture_framework_v1.0_final_EN.pdf.

Federal Office of Administration: *Handbuch für offene Verwaltungsdaten*, https://www.bva.bund.de/DE/Services/Behoerden/Beratung/Beratungszentrum/Methoden/_documents/stda_open_data.html.

NEGZ-Kurzstudie: https://www.researchgate.net/publication/343318794_Kunstliche_Intelligenz_in_der_offentlichen_Verwaltung.

Open Data Barometer: https://opendatabarometer.org/?_year=2017&indicator=ODB.

Open Data Maturity Dashboard: https://data.europa.eu/en/dashboard/2021.

[Access Act] Federal Ministry of the Interior and Community: *Online Access Act*, https://www.gesetze-im-internet.de/ozg/.

[Action Plan, 2014] Federal Ministry of the Interior and Community: *Nationaler Aktionsplan der Bundesregierung zur Umsetzung der Open-Data-Charta der G8*, 2014, https://www.bmi.bund.de/SharedDocs/downloads/DE/publikationen/themen/moderne-verwaltung/aktionsplan-open-data.html.

[Adler et al., 2018] Adler, Astrid; Beyer, Rahel: *Languages and language policies in Germany / Sprachen und Sprachenpolitik in Deutschland*. In: National language institutions and national languages. Contributions to the EFNIL Conference 2017 in Mannheim, 2018, urn:nbn:de:bsz:mh39-78536.

[Data Ethics] Data Ethics Commission: *Gutachten der Datenethikkommission/Opinion of the Data Ethics Commission*, https://www.bmi.bund.de/SharedDocs/downloads/EN/themen/it-digital-policy/datenethikkommission-abschlussgutachten-lang.pdf?__blob=publicationFile&v=5.

[Data Use Act, 2021] Federal Ministry of Economic Affairs and Climate Action: *Data Use Act*, https://www.gesetze-im-internet.de/englisch_dng/index.html.

[eGovernment Act] Federal Ministry of the Interior and Community: *E-Government Act*, article 12 a, http://www.gesetze-im-internet.de/englisch_egovg/index.html.

[ELRC, 2021] Soska, Alexandra; Witt, Andreas: *ELRC Workshop Report 2021*: https://lr-coordination.eu/sites/default/files/Germany/2021/ELRC3_Workshop%20Report%20Germany_PUBLIC_final_clean.pdf.

[Fraunhofer, 2019] Fraunhofer Fokus: *Leitfaden für qualitative hochwertige Daten und Metadaten*, 2019, https://cdn0.scrvt.com/fokus/e472f1bf447f370f/32c99a36d8b3/NQDM_Leitfaden-f-r-qualitativ-hochwertige-Daten-und-Metadaten_2019.pdf.

[IT-Index, 2021] Kompetenzzentrum Öffentliche IT: *Deutschland-Index der Digitalisierung 2021*: https://www.oeffentliche-it.de/publikationen?doc=196440&title=Deutschland-Index%20der%20Digitalisierung%202021.

[LeiKa-plus, 2015] IT-Planungsrat: *National E-Government Strategy Update*, 2015, https://www.it-planungsrat.de/der-it-planungsrat/nationale-e-government-strategie.

[Open Data, 2016] Konrad Adenauer Stiftung: *Open Data. The Benefits, Das volkswirtschaftliche Potential für Deutschland*, 2016, https://www.kas.de/c/document_library/get_file?uuid=3fbb9ec5-096c-076e-1cc4-473cd84784df&groupId=252038.

Annex

Country Profile Greece

State of Play:

Translation practices and information exchange in ministries and public administrations:

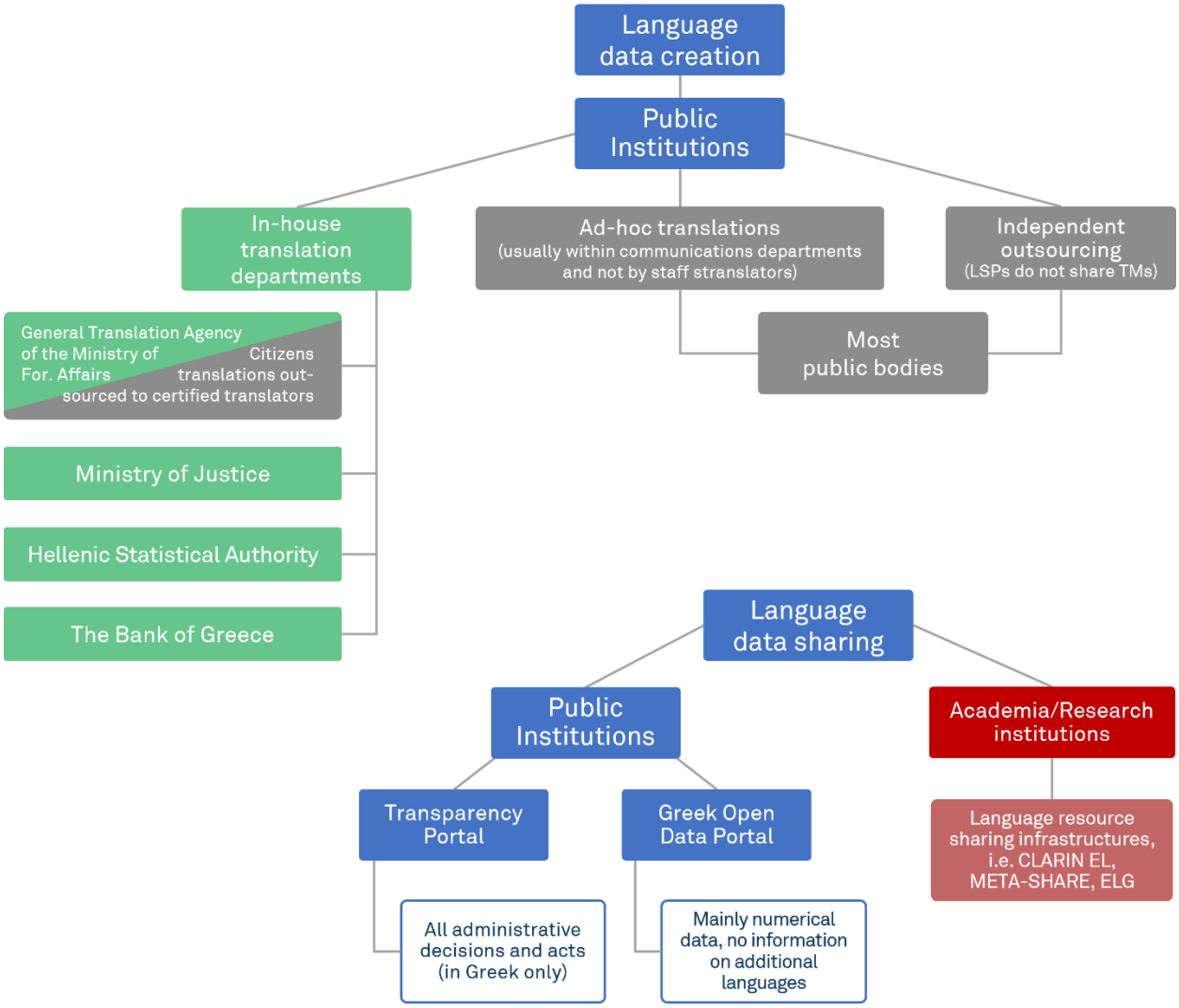
Greek is mostly used in the country by its inhabitants, while outside the country it is used as heritage language by Greek expatriates; therefore, the need for translation of documents from and into Greek is great. There is a continuous need for translation of all EU-documents, mainly from English (and to a lesser extent from French) to Greek, but an increasing need for translation between Greek and the immigrant languages spoken in the country is also attested.

The situation in Greece with regard to language data creation, management and sharing practices has not changed since the publication of the country profile as part of the ELRC White Paper in 2019. The practices for the creation of multilingual data (translations) in the Greek public sector remain fragmented. Very few of the Public Administration institutions meet their translation needs in-house, with dedicated translation departments. These are mainly ministries that have increased translation needs, either due to frequent exchange of documents with EU services or foreign countries, or their mission and objectives entail heavy communication with citizens (Greek nationals or immigrants). Some of the bodies that have in-house translation departments are the Ministry of Justice, the Ministry of Foreign Affairs, the Hellenic Statistical Authority, The Bank of Greece, and the Hellenic Army General Staff. Broadly used practices include (a) outsourcing of translations to LSPs and (b) in-house translation by non-authorised personnel, for internal, unofficial needs, usually undertaken by communication, media and/or publications departments.

However, official certified translations, such as documents that need to be submitted to public authorities, are the responsibility of the Translation Service of the Ministry of Foreign Affairs, whose task is to validly translate public and private documents. Official translation needs can occur either between Public Administration bodies at the national or international level (such as documents from/to foreign governments, embassies, etc., documents exchanged between national services that include text in a foreign language e.g. texts from Europol/Interpol, etc.) or between Greek Public Administration and national or foreign citizens (such as translation of identity papers, University degrees, etc.). Citizens' needs are covered by translators listed in the Official Registry of Certified Translators, available through the Government's official website, the so-called Common Digital Gateway. Certification depends on success in special examinations organised by the Ministry of Foreign Affairs. Lawyers, registered in one of the Greek Lawyers Associations after having succeeded in the relevant exams, also have official translation rights and their translations have full validity, even in courts of Justice.

There is no centralised and uniform procedure for the procurement of translation tasks, apart from the procedure described above concerning the Translation Service of the Ministry of Foreign Affairs. An official quality evaluation process is also lacking. In some cases, the translation output is not even stored in digital form. However, a steadily increasing number of LSPs use CAT tools for their translation services they offer. The Public Administration bodies, however, do not request the Translation Memories (or other by-products, such as term lists) to be submitted together with the translated documents. As a consequence, the degree of usage of CAT tools by the translation companies cannot be documented. What is more important, however, is the fact that Public Administration does not benefit from CAT technology and that similar or exactly identical documents need to be translated anew. The single identified exception is the Bank of Greece, a partially state-owned S.A., which is quite advanced in terms of in-house CAT and Translation Memories production.

The current language data creation and sharing infrastructure in Greek public bodies looks as follows:



Open Data and data collection in Greece:

Access to Public Information: the Transparency Portal

Since October 2010, with the Transparency Programme initiative, all government institutions are obliged to upload their acts and decisions on the Transparency portal (https://diavgeia.gov.gr/en) with special attention to issues of national security and sensitive personal data. Following the latest legislative initiative (Access to Information Law 4210/2013) of the (then) Ministry of Administrative Reform and e-Governance, administrative acts and decisions are not valid unless published online; publication in the Transparency Portal overrides the validity of the Government's Official Gazette itself. All the acts and decisions published on the Transparency portal are exclusively monolingual, i.e. in Greek only.

Open Access in Public Administration

While the Transparency portal hosts Public Administration's acts and decisions, the dedicated Open Government Data Portal hosts the central catalogue of Open Government Data and offers open access to digital resources of the Greek government institutions to citizens, services and information systems for reuse for any purpose. It implements the Open Data policy adopted following the transposition of the EU Directive 2013/37/EE (Law 4305/2014).

Currently, the Greek Open Data portal is being redesigned (Beta version). The new portal presents faceted views on the data, with visualisations of the data sets it contains; however, not the total volume of data sets has been migrated to the current version (approximately 10,600 data sets are hosted at the

archived version. The redesign includes an improved use of metadata, allowing the data sets to be categorised according to Publisher, Topic and Subtopic, thus facilitating searching and filtering the data.

Training public servants on the value of Open Data and, most importantly, raising awareness on the value of language data has been a demanding procedure. The National School of Public Administration and Local Government (ESDDA) has the mission to create a body of specialised officials of the Public Administration with comprehensive professional training; Digital Policy and Digital Governance feature among the subjects taught, while e-ESDDA, the digital repository of the School is responsible for the preservation and dissemination of the School's digital material.

The situation regarding data sharing varies among the Public Administration bodies. Several Ministries have been moving forward as regards the digitisation of their services and workflows and are keen on making their data openly available. Indicatively, but not exclusively, the Ministry of Environment and Energy with its dedicated Open Data portals on various subjects of its responsibility, the Ministry of Justice with its plan for eJustice in place and its various electronic services for the citizens, the Ministry of Finance and the Ministry of Health with their eServices and the Hellenic Statistical Authority, an independent authority which digests data to produce statistics useful for public policy, the economy, and more broadly the lives of citizens. The Governmental website, the Common Digital Gateway, lists all eServices offered to the citizens on 11 domains (Business, Justice, Health, Agriculture, Income and Tax, Work and Insurance, Contacting Public Administration, Education, Family, Culture and Sports.

Dedicated language data sharing infrastructures

There are several language resources repositories and research infrastructures in Greece, stemming from R&D activities and initiatives related to Language Resources and Technology, either national or European. The CLARIN:EL infrastructure with its portal and its inventory of language resources, hosted by ILSP/Athena RC which coordinates a distributed network of 14 nodes, caters for language resources and technologies sharing, as well as for training and raising awareness on the significance of language technology. Language resources and technologies for Greek (but also other languages) are also being shared by META-SHARE, the open and secure network of repositories for sharing and exchanging language data, tools and related web services and also by the more recent European Language Grid Platform, which aims at listing data sets and language technology services as well as relevant stakeholders, from technology development to research centres, from small and medium-sized companies to large enterprises.

Digital policy and language policy in Greece:

Greece is a small country (approximately 10 million inhabitants), whose official language is Greek. The only minority language which is officially recognised is Turkish, which has the status of minority language in Thrace (a region in North-eastern Greece). Due to the immigration flow attested in recent years, there are also numerous immigrant languages spoken in Greece: mainly languages from Balkan, Central and Eastern European countries, but also Chinese, Pashto (Afghanistan), Urdu (Pakistan), Kurdish, Arabic etc.

The language policy of Greece is designed by the Ministry of Education and Religious Affairs, the Ministry of Foreign Affairs (through its General Secretariat for Intercultural Education and Greek Studies Abroad) and the Ministry of Culture, and is implemented by the Centre for the Greek Language (mainly in what concerns teaching Greek as a foreign language) and the Institute of Educational Policy. It defines the educational goals, the principles and the structure of teaching of Greek as first (L1) and second language (L2) in the country and abroad, but does not set priority axes as regards language and language technology research.

Since the publication of the first version of the country profile, there were two important developments on policy level:

- The establishment of the Ministry of Digital Governance: This brings together all critical IT and telecommunications structures related to the provision of electronic services to citizens and the wider digital transformation of the country, previously scattered in different public organisations. As a result, most of the critical ELRC stakeholders, including the units involved in the management and

publication of public open data, are now part of the Ministry of Digital Governance. The design and implementation of the Governmental website is owed to this Ministry.

- The publication of the "Digital Transformation Bible 2020-2025": The Digital Transformation Bible is a record of the necessary interventions in the technological infrastructure of the state, in the education and training of the population for the acquisition of digital skills, as well as in the way Greece utilises digital technology in all sectors of the economy and public administration. Its main role is to describe the vision, philosophy and goals of the national strategy for the digital transformation of the country. It describes the guiding principles, the model of governance and implementation, but also the strategic axes of digital transformation. Furthermore, it describes more than 400 specific projects, classified into short-term and medium-term, horizontal and sectoral, which implement the strategy for Digital Greece. The Bible includes special provisions for the release and exploitation of public data. Among the anticipated provisions and actions is the establishment of Thematic Data Repositories in selected vertical sectors. A number of data that are considered of high-value are mentioned, e.g. geodata, meteorological, environmental and cultural. Unfortunately, language data are not explicitly mentioned nor there seems to be any special provision for their inclusion in a specialised Thematic Data Repository.

Staying on the policy level, the Hellenic National Strategy for AI has been finalised and is planned to be officially published soon.

Stakeholders and major networks:

The key decision makers in Greece for the topics adherent to this white paper are the following:

- For the national digital agenda: Ministry of Digital Governance and Ministry of Development and Funds, General Secretariat for Research and Innovation
- For Open Data and open government: Ministry of Digital Governance and Ministry of the Interior (former Ministry of Administrative Reform)
- For language policy: Ministry of Education and Religious Affairs, and its affiliated bodies, the Centre for the Greek Language and the Institute of Educational Policy.

While 64 unique organisations have attended the second ELRC Workshop in 2017 (of which more than 50% represented public sector bodies), the third ELRC Workshop held in 2021 attracted more than 160 participants from a variety of sectors. This included research and academia (33%), the public sector (30%), but also SMEs (11%) and Industry LT Providers (2%) among others. A considerable number of public bodies have already shared their language data either under permissive or restrictive licences with ELRC. Indicatively but not exhaustively, these include the Central Bank of Greece, the Ministry of Justice, the Ministry of Environment, and the Ministry of Finance.

Language Technologies in Greece

When it comes to LT in Greece and for the Greek language, the third ELRC Workshop showed that the picture is fragmented. Some tools and services exist, but the research and industry providers mainly rely on adapting language-independent systems due to the lack of Greek language data. Three main factors were identified as prerequisites for developing language-centric AI: data, trained human experts and access to powerful computing infrastructure. With respect to the take-up and acceptance of machine translation, the third ELRC Workshop in Greece revealed that MT is already part of Greek public administrations and SMEs. Almost 96% of the workshop participants indicated that they already use MT. Almost 35% of them have already used eTranslation, while 60% made use of other commercial or freely available systems. Most interestingly, no participant rated his/her satisfaction level with the machine translation results for Greek as being "excellent"; most participants indicated a satisfaction level of "fair" (60%) or "good" (21%). As lesser spoken languages like Greek are threatened by digital extinction if their digital presence is not catered for, the development of language-centric AI is key to the digital preservation of the language. The Greek public administration is currently working towards expanding digitisation of all public services and centralising their availability through the Common Digital Gateway (https://www.gov.gr/). This portal, which currently offers 1161 e-services, will be available in a number of languages and the integration of eTranslation is currently being investigated. In addition to machine translation, the need for chatbots arose from the contributions of the public sector

representatives. Chatbots are envisaged for providing information to citizens on administrative procedures and case routing, as part of the Gateway's central Helpdesk. Finally, language technologies for text classification and information extraction, especially for the automatic codification of legislation, are considered to be valuable additions for a constant demand of the Greek public sector and all of stakeholders involved. Such technologies are already additionally used for building some of the public administration core registries, such as the central common registry of administrative procedures and the registry of public organisations.

Main challenges for sustainable data sharing:

Regarding the challenges identified when it comes to sustainable language data management and sharing by and in the public organisations, the discussions at the 3rd ELRC Workshop in Greece have confirmed previous findings. The main obstacles that prevent public administrations from effectively adopting sharing practices and integrating them in their workflows are:

- the lack of openness and sharing culture; the lack of appreciation of the value of this endeavour;
- the lack of explicit endorsement of the task by the political or high-level managerial personnel in public administration and of a subsequent inclusion in the public bodies' organisational charts and structures.
- legal concerns constantly appear to be present in the list of perceived challenges, although the national legal and institutional frame is considered to be in place and it provides the theoretical framework and the guidelines for making public data as open as possible.

Action plan:

An important step towards tackling the administrative obstacles and easing the procedures for collecting language data within a public organisation was the Presidential Decree 40/2020, which assigned specific responsibilities for supporting ELRC to the Department of Information Systems for Open Government of the Min. of Digital Governance. This support is explicitly detailed as support in collecting language data from the Ministries of Finance and Digital Governance. This is a best practice example and should be included in the organisational charts and responsibilities of other ministries as well. Engagement with ELRC was made feasible because of the aforementioned circular of the Ministry of the Interior, which encouraged all directorates to assign one person per directorate as operationally responsible for collecting data sets for ELRC. Of course, not all directorates reacted, depending on whether their activities included the creation and management of language data, especially multilingual data. The fact that this type of involvement is not mandatory, nor monitored with deadlines and that no responsibilities are defined within the organisations' structures, in addition to the fact that the value of the endeavour is not widely recognised, hinder further involvement of the public administrations. It also undermines its sustainability.

A repository of language data produced by the public sector, whether as a subdomain of data.gov.gr or a separate domain, would boost LT development, which in turn would be for the benefit of the public, through relevant services.

In order to address the identified challenges, the following actions are proposed:

- Raise awareness on language data as valuable asset with policy makers, academia and public administration
- Establish good data management policies in public services
- Raise awareness on textual data as valuable Open Data
- Increase interest in MT/LT in public services as part of the national digital policy
- Train public administration officials on the use of digital tools, e.g. of CAT tools
- Need for change of workflows and procedures regarding translation within Ministries
- Tackle legal issues and train public administration personnel on legal issues
- Identify and gain access to outsourced translations

References and links:

Centre for the Greek Language: https://www.greeklanguage.gr/.

CLARIN:EL: https://www.clarin.gr/.

Common Digital Gateway: https://www.gov.gr/

Digital Repository e-ESDDA: https://www.ekdd.gr/en/the-school/digital-repository-e-esdda/.

Digital Transformation Bible 2020-2025: https://digitalstrategy.gov.gr/ https://digitalstrategy.gov.gr/vivlos_pdf.

European Language Grid Platform: https://www.european-language-grid.eu/grid/.

General Secretariat for Research and Innovation, Ministry of Development and Funds: https://gsri.gov.gr/.

Greek Government Open Data portal: http://data.gov.gr/.

Greek Government Open Data Portal Data Sets (archived): http://archive.data.gov.gr/

Institute of Educational Policy: http://iep.edu.gr/en/.

Legal framework for open access and reuse of public domain documents, information and data: http://repository.data.gov.gr/pages/thesmikoplaisio.

Ministry of Digital Governance: https://mindigital.gr/.

Ministry of the Interior: http://www.ypes.gr.

National Digital Strategy 2016-2021: http://www.opengov.gr/digitalandbrief/wp-content/uploads/downloads/2016/11/digital_strategy.pdf.

National School of Public Administration and Local Government (ESDDA): https://www.ekdd.gr/en/the-school/esdda-profile/.

Official Translation Service of the Ministry of Foreign Affairs (Mission): https://www.mfa.gr/en/citizen-services/translation-service/translation-service.html.

Transparency Portal: https://diavgeia.gov.gr/en.

[ELRC, 2017] Gavriilidou, M., Giagkou, M., Pouli, K., Piperidis, S.: *ELRC Workshop Report for Greece*, 2017: http://www.lr-coordination.eu/sites/default/files/Greece/ELRC2-Workshop-Report_Greece%202017-Public_FINAL.PDF.

[ELRC, 2021] Gavriilidou, M., Giagkou, M., Piperidis, S. (2021): *ELRC Workshop Report*, 2021: https://lr-coordination.eu/sites/default/files/Greece/ELRC3_Workshop%20Report%20GREECE_PUBLIC.pdf.

[Gavriilidou et al., 2022] Gavriilidou, M., Giagkou, M., Loizidou, D., Piperidis, S.: *Deliverable D1.17 Report on the Greek Language*, 2022. Project deliverable; EU project European Language Equality (ELE): https://european-language-equality.eu/wp-content/uploads/2022/03/ELE___Deliverable_D1_17__Language_Report_Greek_.pdf.

Annex

Country Profile Hungary

State of Play:

Translation practices and information exchange in ministries and public administrations:

Pursuant to Decree No. 24/1986 (26 June) of the Council of Ministers on Translation and Interpretation and Decree No. 7/1986 (26 June) of the Minister of Justice on the Implementation thereof, attested translation, attestation of translations and making attested copies of foreign language source documents are the exclusive competence of the Hungarian Office for Translation and Attestation Ltd. (OFFI). Apart from meeting its exclusive line of duty and making attested translations, OFFI also provides technical translations, including legal, public administrative translations, translations of laws and revision services with special expertise and terminology expert support. However, it is not obligatory to make use of the OFFI for translations, since other language service providers (LSP) can be contracted as well.

Interesting fact:

The Hungarian Office for Translation and Attestation is a unique institution in Europe in the field of certified (or attested) translations with a history of over 150 years. It has also a unique situation, because since 1994, it is a 100% state-owned shareholding company.

Additional translation activities are also carried out by the ministries and public administration bodies themselves (ad-hoc translations for the ministries' everyday tasks). However, there is no common practice for centralising these ad-hoc translation activities of the Hungarian public administrations. The coordination of translation activity is usually assigned to one department within the particular ministry or public administration. This includes both in-house translations as well as outsourced translations. In some cases, the translation activity is not assigned to one department, but to a secretary of state or a head of department responsible for international communication. The management of the translated documents then follows the specific practices of public administration bodies.

By Act XL of 1995 on Public Procurement, the Public Procurement Authority of Hungary (PPA) was established as a central budgetary organ. The PPA is subordinated to the Parliament. The existing rules for public procurement are based on the Act CXLIII of 2015 on Public Procurement, by the Section 15 (2) and (3) of this act:

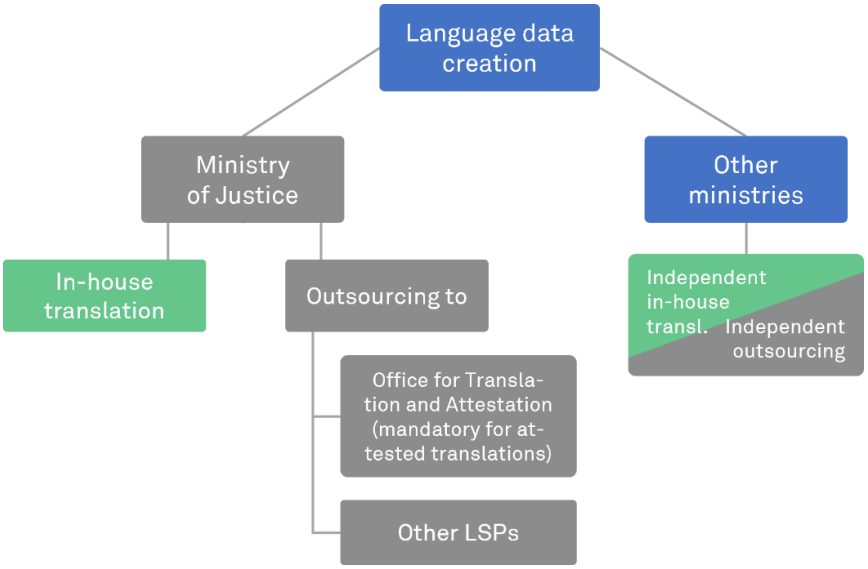
> *(2) The EU thresholds are established and published by the European Commission in the Official Journal of the European Union periodically.*
>
> *(3) The national thresholds applicable to individual subject-matters of procurement shall be specified in the Act on the central budget annually.*

The threshold of the national public administration procurement contracts is 15,000,000 HUF (about 45,000 EUR), but if the public procurement value exceeds 1,000,000 HUF (about 3,000 EUR), at least three different offers must be requested.

Translation needs:

In order to create high quality machine translation solutions today, neural technologies are applied. There are several MT services including Hungarian as well (see for example Google translate), however, combining effective neural technologies with personalised data brings even better results. As in any other fields of neural technology, the amount of data plays a crucial role in the development of the translation application. Although there are several parallel multilingual corpora containing Hungarian as well (and their number is constantly increasing), for lower resourced language pairs it is still not sufficient to develop high quality MT.

The current language data creation and sharing infrastructure in Hungarian public bodies looks as follows:



Language data collection, management and sharing in Hungary

In Hungary, there have been several changes regarding language data creation, management and sharing since the publication of the country profile in the 2019 ELRC White Paper. First of all, several large language resources have been compiled, and some of them have already been contributed to ELRC-SHARE, one of the most prominent ones being the MARCELL Hungarian legislative subcorpus with its more than 30 million tokens. This domain-specific corpus can be used in the development of NMT solutions.

The NLP resource roadmap for the Hungarian language was created in the framework of the European Language Equality project, cataloguing more than 500 language resources. The results of this project show the fields where excellent solutions are available for Hungarian, like multilingual corpora, or tools and toolchains for text analysis. There are also language models built specifically for Hungarian: HuBERT and HILBERT, and several experimental language models developed in the HIILANCO project.

The data collection process also highlighted the gaps, where the lack of data sets or tools mean a significant obstacle for development. There is a significant gap in NLP concerning language data. Although neural methods are used in almost all subfields of NLP, there are not enough data sets in terms of size, annotation and domain. This suggests that researchers and developers need to invest a lot of time and energy in collecting sometimes an incredibly large amount of data. The presenters and the audience agreed that prompt, co-operative actions should be taken, that could even be accompanied by changes in the relevant regulation.

An important development at the policy level is the creation of the Artificial intelligence Strategy of Hungary (cf. National Strategy, 2020). The foundation of two important umbrella organisations was also a salient step to help AI related topics in Hungary. The Artificial Intelligence National Laboratory (MILAB) aims at strengthening the position of Hungary in AI. The research plan of MILAB is built on the National Artificial Intelligence Strategy (2020-2030). It creates the necessary cooperation by connecting the major Hungarian research centres/universities with the industry, society and government.

The other initiative, the Artificial Intelligence Coalition participated in the compilation of the National Artificial Intelligence Strategy, and defines its mission (among three other points) as to “make sure that the government, as a user of AI-powered solutions, should be actively engaged in developing the local AI ecosystem by systematically utilising the national data asset pool” (cf. AI Coalition).

A central task specified by these two organisations on Hungarian Artificial Intelligence is to support the automatisation process of customer service technologies, by strengthening synergies between industry and research, for example through organising common events like workshops and exhibitions.

As for the practices of the creation of translations as multilingual data in the Hungarian public sector, they are the same as indicated in the country profile of 2019. Namely: the translation of foreign

language source documents is still the competence of the Hungarian Office for Translation and Attestation Ltd. (OFFI). However, other language service providers can also be contacted.

There was an opportunity to fill a Country Survey about the usage and prevalence of language technologies in our country after the Workshop. The most interesting point of these answers is that the respondents considered legal issues (the topic of the last panel session) to be the biggest difficulty, namely copyright.

Open Data and data collection in Hungary:

In recent years, there has been an evolving trend considering language resources, namely that a growing number of these resources are made available as open data. Whereas it is mostly true for NLP tools and services, copyright law constrains the use and especially the (re-)sharing of data, thus placing a large burden on researchers when collecting data for neural technologies. However, there are some corpora, see the MARCELL corpus mentioned above, and also a few platforms supporting data collection. HRDA, the Hungarian National Node of the Research Data Alliance of Europe has more than ten members with repositories in the science domain. These repositories are predominantly open source, available under a CC licence. As for open government data in Hungary, unfortunately the progress towards open data is rather slow.

As regards the legal aspect of sharing data, the amendment of the Copyright Act (35/A, 2021) is an important step. This act regulates the ways of data and text mining, mainly for research purposes[39].

eGovernment strategy in Hungary:

Following the Digital Economy and Society Index from 2017, Hungary has exhibited several significant improvements and successes such as eID cards, the Electronic Health Cooperation Service Space, a cloud-based, centralised platform supporting easier communications among the participants, along with highly useful functions for the public, like the e-prescription. Another successful application is the Hungarian Municipality Application Service Provider (ASP), providing modern, integrated shared services for local administrative management, ensuring standardised internal operation and a common platform for e-government service provision on the local government level to the end-users. The electronic form of the personal annual tax declaration is also very popular and available for a growing audience, including e.g. farmers. The Electronic Procurement System is widely used at companies, and is especially useful in logistics where there are swift changes in regulations due to the Covid situation.

Digital policy and language policy in Hungary:

The Prime Minister's Office State Secretariat for the National Policy and its organisations are coordinating the policy and strategy for the Hungarian nation in the Carpathian basin and all over the world. A major contributor of the Hungarian language strategy is the former Institute for Hungarian Language Strategy, the current Institute for Hungarian Studies Research Center for Language Planning.

The main governmental strategies, which contain language policy elements are:

- The National Info-Communication Strategy 2014-2020 (available only in Hungarian).
- Strategies of the Digital Success Programme 2.0 (including the Digital Child Protection Strategy of Hungary, the Digital Export Development Strategy of Hungary, the Digital Education Strategy of Hungary, and the Digital Startup Strategy of Hungary, Digitalisation Strategy of Public Collections)

The main legal acts, which contain elements of the national language policy include:

- Fundamental Law (the status of the Hungarian language)
- Act XLVI of 2012 on the Land Surveying and Cartography (geographical names)
- Act LXII of 2001 on the Hungarian Nation Living in the Neighbouring Countries
- Act CXXV of 2009 on the Hungarian Sign Language
- Developing Strategy for the Public Administration and Public Services 2014-2020
- Act CLXXXV of 2010 on the Mass-Media and Mass-Communication
- Act CCI of 2017 on the Rights of Nationalities

[39] https://www.parlament.hu/irom41/15703/15703.pdf

- Act XCVI of 2001 on Publication in Hungarian Language of Economic Advertisements, Shop Labels and some of Public Statements
- Act LXIV of 2001 on Protecting Cultural Heritage

There are also important resolutions of the Hungarian National Assembly mentioning language policy elements, including the resolutions made on the Day of the Hungarian Language, on the National Heritage Day and on the Day of the Nationalities.

Most interestingly, as a result of a year of research and development work of a team of hundreds of pedagogical experts, psychologists and practitioners, the Education 2030 Learning Sciences Research Group at Eszterházy Károly University made a proposal for the new National Core Curriculum. The National Core Curriculum contains the full language policy for the education system of Hungary.

Stakeholders and major networks:

Within ELRC, around 80 potential stakeholders that are involved in the creation or sharing of language resources, related activities and/or policy setting were identified. 50 of these stakeholders participated in the ELRC Workshop in 2022.

In addition to the Hungarian Office for Translation and Attestation (OFFI), the different ministries and additional certified translators represent the major provider of language resources in Hungary. So far, there are more than 90 language resources containing Hungarian at the ELRC SHARE, most of them bilingual or multilingual corpora.

Main challenges for sustainable data sharing:

- Fundamental internal issues: Public administrations are not aware of the value of language data and do not perceive it as an asset.
- Legal issues related to outsourced translations: Outsourced translations are intellectual property of LSPs, which makes it difficult for public administrations to share outsourced data.
- Continuity issues: Changing government, framework contracts etc.

Action Plan:

Taking into account the main challenges in Hungary, corresponding actions to enable/improve the sharing of language resources in Hungary should focus on:

- **Discovering the document management process of the Hungarian public administration**
- **Discovering the translation practice (tools, competent persons and departments) used and needed during the everyday work in the leading public administrations (ministries, governmental offices)**
- **Raising awareness of language data as Open Data and a valuable asset, including in particular:**
  - Sharing benefits of sharing language data
  - Integrating language data in the national Open Data policy as well as digital agenda
- **Identify and gain access to outsourced translations:**
  - Adapt the procurement process for buying translations in a way that tmx and all usage rights are transferred to the purchasing authority
- **Initiate the institutional organisation and collection of multilingual language data, building a national multilingual public administration terminology database.**

References and links:

Act XL of 1995 on Public Procurement: https://www.kozbeszerzes.hu/english/.

Act CXLIII of 2015 on Public Procurement: https://www.kozbeszerzes.hu/cikkek/hungarian-public-procurement-rules.

AI Coalition: https://ai-hungary.com/en/content/ai-coalition#mission.

[National Strategy, 2020]: Ministry for Innovation and Technology: *Hungary's Artificial Intelligence Strategy*, 2020, https://ai-hungary.com/files/e8/dd/e8dd79bd380a40c9890dd2fb01dd771b.pdf.

[Orbán, 2018] Orbán, Anna: *Open Government Data in Hungary*, 2018. Central and Eastern European EDem and EGov Days 331 (July):373-81. https://doi.org/10.24989/ocg.v331.31.

Annex

Country Profile Iceland

State of Play:

Translation practices and information exchange in ministries and public administrations:

Since 1990, all Acts falling under the EEA agreement are translated by the Translation Centre of the Ministry of Foreign Affairs. In addition, the Translation Centre is responsible for the translation of texts related to the European Economic Area, other international agreements and legal acts. Approximately 35 translators are grouped according to fields such as society, finance, science or technology. In 2016, the Translation Centre's terminology contained about 70,000 entries and it is continuously growing.

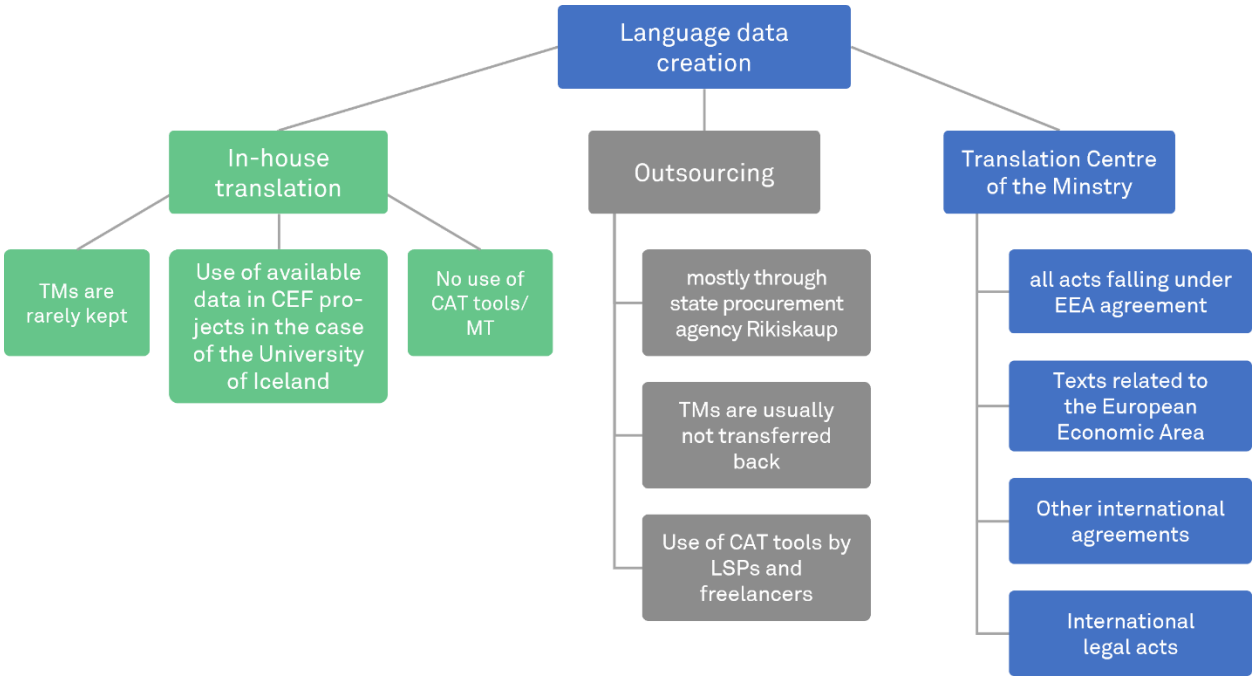
Most other translation and interpreting services for the Icelandic state are procured through the state procurement agency Rikiskaup via a call for tender. Those who are accepted are taken into the so-called framework agreement until the next call for tender has been finished. Most of the signatories are small companies, which are frequently working with freelancers. This makes it difficult to gather data from them, except perhaps from the largest translation agencies. One of them, Skopos, is actually working on a CEF project called "Principle" and is offering bilingual data.

Some other institutions have in-house translation services, such as the RÚV, the Icelandic state broadcaster. Those translations are only for television and are rarely kept in a data base after their use. The University of Iceland has one in-house translator to translate legal regulations and web material into English. The corresponding bilingual data base has been used in the Principle project.

The use of computer-assisted translation (CAT) Tools depends on whether the translations are outsourced or managed in-house. Language Service Providers (LSPs) and freelance translators usually translate their documents with the help of CAT Tools. However, neither CAT Tools nor machine translation (MT) systems are used in Icelandic public administrations and ministries. This may also explain why translation memories (TMs) are usually not transferred back to the public administrations if the translation was outsourced.

Due to the limited number of Icelandic speakers, it is difficult to build and develop costly language technologies. Therefore, the language technology industry in Iceland is relatively small and language technology support for Icelandic is weak (Rögnvaldsson et al., 2012). Various companies developed LT software and systems, such as a spell-checking programme or a text-to-speech system for Icelandic, but neither of them continued their work in the field afterwards. However, in recent years, the Icelandic government has started various actions to improve their position in the digital world and to raise awareness on the importance of language technology.

The current language data creation and sharing infrastructure in Icelandic public bodies looks as follows:



Open Data and data collection in Iceland:

Data privacy in Iceland is legislated by the Data Protection Act. Pursuant to the Icelandic State and Municipal Policy on the Information Society 2013-2016, non-personal information and files stored by the State or municipalities should be accessible to the general public, businesses and stakeholders. The Information Act (No.50/1996) includes conditions on the re-use of public sector information (PSI) and defines both access and restrictions to information. It covers almost all aspects related to the PSI Directive (2003/98/ES), except for the access and re-use of information through electronic means like e.g. databases. The Icelandic Open Data Portal provides access to a growing list of government data and databases[40].

Interesting fact:

In 2018, Iceland signed an agreement with the Nordic Institute for Interoperability to start using Straumurinn. It is based on the Estonian X-Road platform, which will enable standardised, efficient and secure data exchange between public administrations and ministries.

In order to create synergies between different IT systems of public administrations, Iceland has signed an agreement with the Nordic Institute for Interoperability Solutions (NIIS Institute), which also cooperates with Finland and Estonia. By using the Straumurinn data line, processes for data exchange will be streamlined and automated. Together with a comprehensive management plan, the Straumurinn data line is considered the foundation for effective and transparent public services in Iceland.

According to the Icelandic financial plan 2019-2023 (Factsheet, 2019), public sector services should be based on information systems that fulfil the needs and the technical requirements of both public institutions and the industry. They shall be able to access open-source data in one place, monitor discussions on issues in the administration and participate in transparent reporting processes for e.g. draft proposals or policy papers. According to the financial plan 2019-2023, public data "will be free of charge and reusable as much as possible" (ibid.).

The LT Project Plan (Nikulásdóttir et al., 2018) also explicitly states that all resources and tools, which will be developed as parts of the core tasks will be completely open and free. They will be made

[40] https://opingogn.is/

available through CLARIN-IS and maybe other CEF-funded projects. A number of language resources are already available through CLARIN-IS. The majority of them are free and open (under CC licences).

The first phase of monolingual data gathering was collected under the wing of the Arni Magnusson Institute. The data sets include e.g. a large dictionary of the most frequent words in Iceland, the MÍM (2012) corpus including 25 million words, a new giant corpus of 1250 million words and some other smaller corpora. In the meantime, a second phase has begun, which also includes bilingual corpora.

However, more parallel language databases will be required to develop accurate machine translation systems (ELRC, 2018, p.6), which is one of the reasons why Iceland participates in the above-mentioned Principle project together with partners from Ireland, Norway and Croatia. It is dedicated to the gathering of bilingual corpora for the purpose of creating MT engines. The Translation Centre of the Foreign Ministry will provide its TM database of 1.3 million sentence pairs, and other partners will contribute data as well.

Digital policy and language policy in Iceland:

In 2000, a special Language Technology (LT) programme was launched. It aimed to support institutions in creating basic resources for Icelandic language technology, which lead to considerable results, including e.g. a corpus of 25 million words, an isolated word speech recogniser, etc.

After the end of the programme, the Icelandic Centre for Language Technology (ICLT) was established by researchers from the University of Iceland, Reykjavik University and the Árni Magnússon Institute for Icelandic Studies. Their work resulted in a number of projects, leading to the development of various Language Technology tools and resources, e.g. the open source IceNLP package.

The publication of the META NET White Papers in 2012 was a landmark for the Icelandic LT development. Iceland was one of the four countries with the lowest scores in all categories, which highlighted the lack of language technology support in the country. This is why after its publication, extensive propaganda for the development of language technology has started in Iceland. The white paper was even discussed in the Icelandic Parliament in the same year, which lead to the Resolution on the Necessity of Making Icelandic Usable in the Digital Domain in 2014.

The current drive in LT is much better funded by the Icelandic authorities with 2.2 billion ISK (15.7 million EUR) until 2022. Consequently, three language technology experts were commissioned to develop a detailed 5-year project plan for Icelandic and its technology in 2016, which resulted in the Language Technology Project Plan 2018-2022. The non-profit organisation Almannarómur has been contracted to be in charge of the execution and coordination of this LT plan.

The Icelandic Centre for Research has recently awarded grants within the Language Technology programme. Seven projects In machine translation, language learning and other LT projects got the grants.[41] There are both projects for machine translation from EN-IS and PL-IS, In addition to automatic translation of subtitles for television.

The role of LT and language data in Iceland's AI regulations

In April 2021, a committee appointed by the Prime Minister delivered a proposal for an AI Strategy. This proposal has been discussed in the Government and presented to the Icelandic Parliament but has not yet been formally adopted as an AI Strategy for Iceland[42]. Among others, the AI report emphasises the importance of Language Technology for the Icelandic society and Icelandic citizens.

The Icelandic Government is the main funder of the Icelandic National Language Technology Programme which started in 2019 and will end in 2022[43]. The main aim of the LT Programme is to make the Icelandic usable in the digital domain, in all spheres of the society. The Programme includes five core tasks: Development of Speech Recognition, Speech Synthesis, Machine Translation Systems, Language and Style Checking Tools, and Language Resources.

[41] https://www.rannis.is/frettir/uthlutad-ur-markaaetlun-i-tungu-og-taekni-1
[42] https://www.iiim.is/2021/04/iceland-has-a-new-ai-policy/
[43] https://clarin.is/en/links/LTProjectPlan/

According to the LT Project Plan, public data "will be free of charge and reusable as much as possible". The LT Project Plan also explicitly states that all resources and tools, which will be developed as parts of the core tasks will be completely open and free. They will be made available through CLARIN-IS and maybe other CEF-funded projects. A number of language resources are already available through CLARIN-IS. The majority of them are free and open (under CC licences).

Stakeholders and major networks:

The ELRC National Anchor Points for Iceland represent a relevant stakeholder, i.e. the University of Iceland, which is involved in the above-mentioned ICLT and other projects that are relevant to ELRC. Other stakeholders that have already contributed data to ELRC include the Central Bank of Iceland and the European Medicines Agency. In summary, more than 35 organisations showed their interest in the ELRC initiative by participating in local workshops and ELRC conferences.

The first phase of monolingual data gathering was collected under the wing of the Arni Magnusson Institute. The data sets include e.g. a large dictionary of the most frequent words in Iceland, the MÍM (2012) corpus including 25 million words, a new giant corpus of 1250 million words and some other smaller corpora. In the meantime, a second phase has begun, which also includes bilingual corpora. However, more parallel language databases will be required to develop accurate machine translation systems, which is one of the reasons why Iceland participates in the Principle project together with partners from Ireland, Norway and Croatia. It is dedicated to the gathering of bilingual corpora for the purpose of creating MT engines. The Translation Centre of the Foreign Ministry will provide its TM database of 1.3 million sentence pairs, and other partners will contribute data as well.

Main challenges for sustainable data sharing:

- In Iceland, open issues regarding access, copyright and privacy often prevent data holders from sharing their data.
- At the same time, there is a general lack of available parallel language resources, making it hard to train and improve already existing machine translation systems. However, with the new CEF project Principle, greater emphasis will be put on acquiring high-quality bilingual corpora and preparing them for MT engines.
- The limited number of Icelandic speakers makes it difficult to create language resources, since this is also associated with high costs.

Action plan:

Based on the status quo in Iceland and the identified challenges, the following three main objectives were defined:

- **To raise awareness of language data as Open Data:**
  This is to be achieved by e.g. further integrating language data in the national Open Data policy and in the digital agenda. The National LT Project Plan is already an important step in this direction, since it highlights the importance of language data by stating that it is impossible to develop language technology if language resources do not exist (Nikulásdóttir et al., 2018, p.13). At the same time, the project plan aims to ensure that the data created within the programme are not only accessible, but also usable for further development in research or business (ibid., p. 142).
- **To increase interest in MT in public services:**
  In order to increase the public administrations' interest in machine translation, synergies will be established through dedicated projects and initiatives. The project Principle already serves as a good example of these efforts. In addition, best practices and good examples of successful use of machine translation will be promoted.
- **To identify and gain access to outsourced translations:**
  It can be a challenge to gather data from procured translations if the tender was awarded to small companies, which are often working with freelancers. However, larger Icelandic companies may be able to contribute data, which is why they should be involved in projects and initiatives as it was the case in e.g. the Principle project.

References and links:

CLARIN-IS: http://clarin.is/en/resources/.

Debate about “The Icelandic Language in the Digital Age” in the Icelandic Parliament, 2012, https://www.althingi.is/altext/upptokur/lidur/?lidur=lid20121121T153618.

Language Resources for Icelandic: http://www.malfong.is/index.php?lang=en&pg=.

Resolution on the Necessity of Making Icelandic Usable in the Digital Domain, 2014, https://www.althingi.is/altext/143/s/1076.html.

State procurement agency Rikiskaup: https://www.rikiskaup.is/is/english-1/about-rikiskaup/english.

Tagged Icelandic MÍM Corpus: http://www.malfong.is/index.php?lang=en&pg=mim.

[ELRC, 2018] Þorláksdóttir, Arnbjörnsdóttir: ELRC Workshop Report for Iceland, 2018, http://www.lr-coordination.eu/sites/default/files/Iceland/ELRC%20Workshop%20Report%20for%20Iceland_PU_0.pdf.

[Factsheet, 2019] European Commission: Digital Government Factsheet Iceland, 2019, https://joinup.ec.europa.eu/sites/default/files/inline-files/Digital_Government_Factsheets_Iceland_2019.pdf.

[Nikulásdóttir et al., 2018] Nikulásdóttir et al.: Language Technology for Icelandic 2018-2020, Project Plan, 2018, http://clarin.is/en/links/LTProjectPlan/.

[Rögnvaldsson et al., 2009] Rögnvaldsson et al.: Icelandic Language Resources and Technology: Status and Prospects, 2009, https://dspace.ut.ee/bitstream/handle/10062/9670/Icelandic%20language%20resources.pdf.

[Rögnvaldsson et al., 2012] Rögnvaldsson et al.: The Icelandic Language in the Digital Age. In: META-NET White Paper Series, 2012, www.meta-net.eu/whitepapers/e-book/icelandic.pdf.

[Rögnvaldsson, 2019] Rögnvaldsson, Eiríkur: Language Technology News – Iceland, ELG Conference, 2019, https://notendur.hi.is/eirikur/ELG_Brussel.pdf.

Annex

Country Profile Ireland

State of Play:

Translation practices and information exchange in ministries and public administrations:

The Irish language has been highlighted as an under-resourced and therefore a priority language in the context of data collection for improving the EU's automated translation systems. As a result, the Irish <> English language pair (in terms of data collection) has been the current focus of ELRC-related activities to date in Ireland and thus serves as the focus of this report.

Irish is the first official language of Ireland. It is a minority language, with the most recent census[44] reporting 1.7 million speakers, of whom just over 73,000 speak it on a daily basis outside of the education system. The Irish language is a unique minority language in many ways as it has been afforded significant constitutional and legislative protection by the Irish State since its foundation. In addition to the official status of the Irish language in the Constitution, it was recognised as an official and working language of the European Union in 2007.

The Official Languages Act 2003 was signed into law on 14th July 2003, with the primary objective to ensure the improved provision of public services through the Irish language. The Office of An Coimisinéir Teanga (Language Commissioner) was established under the Act in 2004 to monitor compliance by public bodies with the provisions of the Act and to take appropriate measures to ensure such compliance. In the Principal Act, each public body defined their own language scheme, which described the services it proposed to provide either in Irish only, in English only or bilingually. As a result of the Principal Act, all public bodies were obliged to translate official documentation into Irish and therefore large quantities of documentation are available in both English and Irish. This meant that when it came to translation needs in Irish public administration, differentiation should be made between translations for Irish <> English and other language pairs.

The Language Act was updated in 2012 and the Gaeltacht Act was introduced, giving statutory effect to a 20-year strategy for Irish. With respect to Irish <> English translations on a European level, the status of Irish as an official and working EU language came into effect from the 1st January 2007 under Regulation 920/2005 which included derogation on the use of the language, to be reviewed every five years. In December 2015, the Council of the European Union adopted a regulation aimed at eliminating the derogation on an incremental basis by the end of 2021 to eventually provide services through Irish at the same level as the other official EU languages from this date. As of 1 January 2022, the derogation has been lifted, with Irish being granted full official and working status in the EU.

Official Languages (Amendment) Act 2021

The Official Languages (Amendment) Act 2021 was signed into law by the President of Ireland on 22 December 2021. This new legislation is a strengthening of the Official Languages Act 2003 and it is widely recognised that it will make a significant contribution to the quality of services in Irish provided to the public by State bodies.

The enactment of this legislation is the result of a comprehensive public consultation process[45], a pre-legislative examination by the relevant Oireachtas Committee, over 25 hours of debate at Committee stage in Dáil Éireann and over 300 proposed amendments discussed during that time – many of which were accepted. Written records of the debates that took place around the legislation as it progressed through the various statutory stages can be found at oireachtas.ie.

[44] According to the Central Statistics Office (CSO) at a population of 4,761,865 in 2016.
[45] https://www.gov.ie/en/publication/28c94-review-of-official-language-act-2003/

The main goals of this strengthened Act are that:

- 20% of recruits to the public service will be competent in Irish by the end of 2030;
- all public services in the Gaeltacht and for the Gaeltacht will be provided in Irish;
- all public offices in the Gaeltacht will operate through the medium of Irish; and
- a National Plan for the Provision of Public Services in Irish will be developed.

The Department is working to commence all sections and provisions of the Act on an incremental basis.

Another important legislative development since the publication of the previous report was the introduction of Statutory Instrument 230 of 2020 also known as the Official Languages Act 2003 (Public Bodies) Regulations 2019. These important regulations updated the list of over 600 public bodies identified in the Official Languages Act 2003, as some had changed names, been subsumed into other public bodies or they had subsequently become defunct.

Irish Language Services Advisory Committee

Under the relevant provisions of the new Act, the Irish Language Services Advisory Committee[46] was established on 20 June 2022. The work of the Committee will be primarily focused on the preparation of the first National Plan for the provision of public services through the medium of Irish for the first two years to ensure that it is completed before the deadline of 19 June 2024.

Current Infrastructure vs. Goal

At present, each Government Department is responsible for managing its own Irish-language translation requirements. Currently, there is an in-house translator in the Department of Foreign Affairs and Trade; the Department of Justice, the Department of Tourism, Culture, Arts, Gaeltacht, Sport and Media and the Public Appointments Service. In addition, there are five in-house translators in the Office of the Revenue Commissioners and 19 in-house translators in the Houses of the Oireachtas Service. These translators all use computer-assisted translation (CAT) tools. Much of the other Departments' translation work is done by external translation companies or freelance translators.

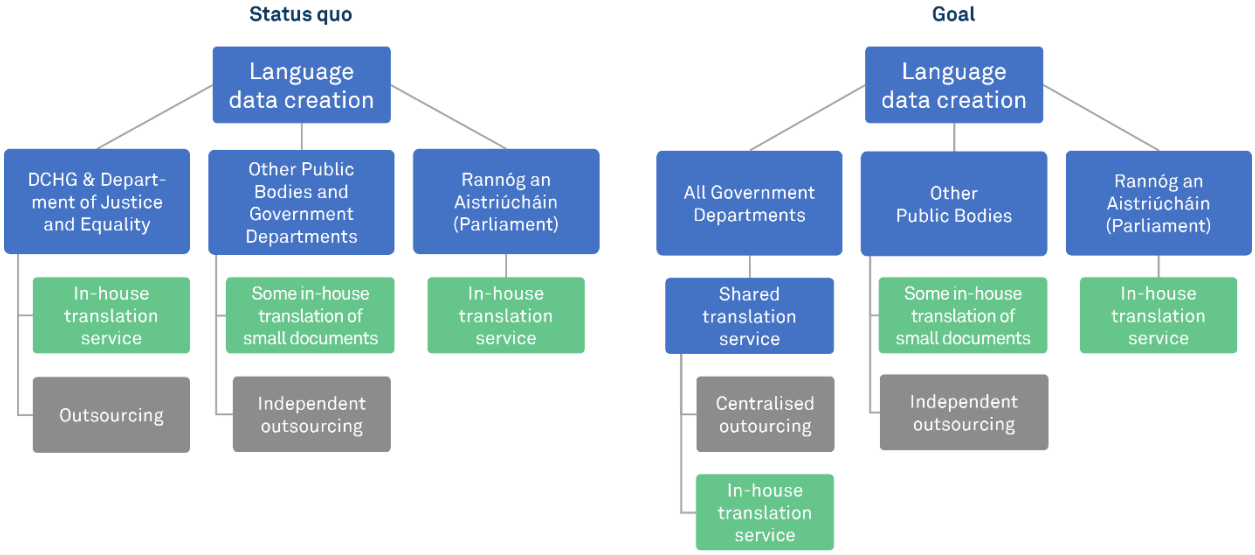
Since 2016, a framework for outsourcing translation work was established which includes a small number of selected accredited language service providers (LSPs) with set rates for translation. Most of the documents outsourced by Government Departments for translation are larger corporate documents. The State Examinations Commission have in-house translators and the Houses of the Oireachtas (the Irish Parliament) have a large translation team called Rannóg an Aistriúcháin comprising 19 translators. The latter team began using CAT tools in 2018 thus generating TMX (translation memory) files for reuse both in-house and at the DGT Irish Language Unit. Many other public bodies, such as universities and county councils, have dedicated Irish language officers who often carry out small translation tasks in-house, without the use of CAT tools. Larger translation tasks are outsourced to LSPs.

In the past it was not common practice for public bodies or government departments to request the return of TMX files from an LSP (which is a by-product of a translation procurement). Since the launch of the European Language Resource Coordination (ELRC) workshops in 2016, however, some members of public administration have begun to do this, while some departments have since reported the inclusion of such a requirement in their translation contracts. As part of the most recent framework agreement agreed by the Office of Government Procurement, a recommendation has been made that TMX files must be made available to the public body. Many Irish language officers or translators in public administration are unaware of technology opportunities or benefits.

In June 2019, a Computer Aided Translation Workshop was held in Dublin City University (DCU) to provide CAT and Machine Translation post-editing training to freelance and public administration translators. Only 16% of attendees had previous experience of CAT tools and only one attendee had ever post-edited machine translation output. The European Commission Representative in Ireland has since worked with DCU to conduct two further workshops of this nature to address the lack of technical skills amongst this translator group.

[46] https://www.gov.ie/en/publication/b2802-irish-language-services-advisory-committee/

The current language data creation and sharing infrastructure in Irish public bodies looks as follows:



For some years, there has been discussion about the establishment of a shared translation service An tSeirbhís Chomhroinnte Aistriúcháin (Shared Translation Service – STS). The proposed model involves developing a centralised point to which all government department translation requests can be submitted and managed, and from which all translation tasks are either handled in-house or outsourced appropriately. Through coordinating a single approach to translation practices and language resource re-use, the STS will assist government departments in complying with their statutory obligations, streamlining and regularising translation services in order to reduce the costs of such services. Other public bodies would continue to translate in-house when possible and outsource larger translation tasks. There has been no progress on this since the last report.

Open Data and data collection in Ireland:

According to Ireland's implementation of the Public Sector Information Directive (PSI) 2003/98/EC, (now known as Open Data Directive (EU) 2019/1024), an individual or a legal entity may make a request in a legible form to a public sector body to release documents for re-use. Every request made in a language other than Irish or English shall be accompanied by a translation of the request into Irish or English.

Open Data listed in data.gov.ie is published by currently 159 Government Departments and public bodies and is operated by the Government Reform Unit of the Department of Public Expenditure and Reform (DPER)[47]. This national Open Data portal is intended to provide easy access to data sets that are free to use, reuse, and redistribute with many of data sets being individually published and updated by public organisations. Ireland ranks highly for Open Data maturity in Europe60 thanks to the Open Data Strategy for Ireland (2017-2022)[48].

With regards to sharing language resources (LRs) for translation technology, users of eSTÓR portal (formerly the Irish National Relay Station (NRS)[49]) can choose to share their data on a national or European level (onward sharing with ELRC-SHARE) either under the Open Data Directive or with specific licencing. Ongoing funding received from the Dept of the Gaeltacht since 2021 for DCU to continue to host the NRS until at least 2025 means that necessary updates and improvements will be applied to the repository. An outreach programme encouraging more public administrators to share their bilingual data on eSTÓR is also underway. The Open Data Unit in DPER are investigating linking the LRs uploaded to eSTÓR (which have been uploaded as Open Data) with the data.gov.ie website.

[47] Figures as of October 2022.
[48] https://data.gov.ie/pages/open-data-strategy-2017-2022
[49] https://estor.ie

In both the European Language Equality project's Report on the Irish Language and the forthcoming Digital Plan for the Irish language, recommendations are made for data sets (language data in digital format, whether bilingual, monolingual, terminology-based, linguistically annotated, etc.) to be made open and freely available where possible (e.g. CLARIN, ELRC-SHARE). While some data sets (e.g. the New Corpus for Ireland) may not be released due to copyright reasons, the recommendation is that such resources are at least shared on a restrictive licence basis for research and development purposes (e.g. training word embeddings for neural-based systems, training language models, etc.).

Digital Policy and Language Policy in Ireland:

Like many minority languages, the relatively low number of speakers of Irish has resulted in little investment from industry to date. As such, language technology support for Irish is weak, with the availability of most existing tools and resources being made possible only through volunteer activities or short-term projects based in universities. The Action Plan 2018-2022 for the Irish language (Action Plan, 2019) highlights the immediate need for further research and development in the area of language technology for Irish. To address this, a Digital Plan for the Irish language is due to be published in November 2022 to outline technological requirements for safeguarding the future of the language. The development of such a plan is crucial to avoid the risk of the language falling behind with regard to technological developments as per 'Report on the Irish Language' published through the European Language Equality project (Lynn, 2022).

Interesting fact:

The CEF-AT funded PRINCIPLE project involved the development of bespoke English > Irish MT systems for use within professional translation environments at Foras na Gaeilge, Rannóg an Aistriúcháin, the Department of Justice and the University of Galway. Not only was the legacy translation data collected through this project relayed to ELRC-SHARE to help improve the eTranslation engines, but newly created TMX files from Rannóg an Aistriúcháin translations were also shared with the Irish translation unit at DGT.

A national Terminology Committee (an Coiste Téarmaíochta) was established to develop, approve and provide authoritative, standardised Irish language terminology. A national terminology database (www.tearma.ie) is then updated accordingly with decisions made by the committee. The database can be downloaded freely in TBX or TXT format. With respect to generating Irish terminology for the European Union, the DTCAGSM also funds the Irish/EU terminology project LEX (GA IATE).

Until recently, Irish public administrations did not exchange language resources such as translation memories or glossaries centrally. Following ELRC promotional and educational activities in Ireland (through workshops, outreach seminars and on-site visits), a number of stakeholders began to contribute existing language resources and change internal data-management processes. These stakeholders included some government departments, county councils, universities, the national broadcaster (RTÉ), dictionary publishers and the language commissioner's office. These practices have become significantly more widespread over the past number of years through the work carried out by the European Language Resource Infrastructure (ELRI) project. As part of this CEF-funded project, Ireland's eSTÓR portal[50] (formerly the National Relay Station (NRS)), available both in Irish and English was developed in 2019: this is a pioneering, online, secure platform where members of public institutions in Ireland can contribute their own language data to a national centralised portal, receive automatically generated TMX versions of their data sets and download shared resources from other public body contributors. According to their sharing licencing or agreements, this data is then "relayed" onto the ELRC-SHARE. The DTCAGSM also provides funding to the eSTÓR project. There are currently 133 registered users.[51]

The role of LT and language data in Ireland's AI regulations

The Department of the Gaeltacht (DTCASM) has funded a number of LT-related projects in Dublin City University and Trinity College Dublin, such as speech synthesis, NLP and machine translation (directly

[50] https://estor.ie
[51] Figures as of October 2022.

through the Irish Language Support Scheme[52]) and corpus/ dictionary development (through Foras na Gaeilge (see AI Strategy, 2021, page 42). Ireland's AI Strategy "AI – Here for Good" is primarily focused on English LT support and makes minimal reference to the need for LT support for Irish (in the public sector only): "To render AI systems accessible to a wider range of our population, as well as to develop services in Irish based on AI for Irish language-speakers, good language technology resources need to be developed". It is expected that the forthcoming Digital Plan for Irish will therefore address the current gap in dedicated funding and planning for Irish language LT.

Main challenges for sustainable data sharing:

Feedback surveys conducted during the ELRC and ELRI outreach workshops[53] reported that many people working in the public sector and dealing with language, translation and data are very enthusiastic about using language technology, sharing data and thus supporting the Irish language. Yet, for a number of reasons, language data sharing is still a difficult endeavour. The main reasons for this are the following:

- Each government department and public body manages their own translation needs, either through in-house translation or outsourcing. There is no regulation around the management of translation data or the requirement for LSPs to return translation memory files with the translated documents.
- Until the establishment of the National Relay Station (eSTÓR), there had been no culture of sharing translation memories or terminologies across departments or public bodies.
- Public servants raise concerns about whether or not they have permission to share their data. This is linked to general lack of awareness or understanding of the Open Data Directive.
- Lack of technical skills with respect to CAT tools amongst translators in public administration.
- General unawareness of the value of language data and leveraging opportunities it presents.
- Unawareness of the need for language data to build translation systems.
- Within public administration, language data management is currently outside the scope of any specific role and therefore can be difficult to ensure follow-through. In addition, staff changeover/ department changes or merges, result in staff not being able to find legacy data.
- Language Technology and Machine Translation is not on top of the priority list of most in public administration, making the efforts more difficult even when people are generally supportive.
- Misuse of free online translation services which leads to scepticism and wariness of MT in general (without understanding the strengths of domain-specific MT systems).
- There is a widespread lack of awareness and uptake of eTranslation within public administration in Ireland and therefore a lack of full understanding of the long-term benefits.
- Language data needs to be identified as a "high value data set" under the Irish Open Data Strategy.

Action plan:

To address the identified challenges, the following five main objectives were formulated, ranked by their importance for the landscape in Ireland:

- **Raising awareness of language data as Open Data and a valuable asset:**
  Awareness has thus far been raised through both the ELRC and ELRI workshops, TV, radio, social media, YouTube, online news articles, public lectures, podcasts and so on. These promotional and awareness-raising efforts require continued support from the DTCAGSM along with the Open Data Unit (currently based in DPER). The continued funding of eSTÓR and appointment of Open Data Liaison Officers in Ireland is expected to positively impact this endeavour.
- **Increasing interest in MT/LT in public services as part of the national digital policy:**
  The forthcoming Digital Plan for Irish highlights the immediate need for upskilling current translators and increasing the uptake of the use of translation technology in public administration, in addition to providing technical training in translation courses. It is expected that through using the NRS, the availability of shared TMX files amongst public administrators will encourage an increased

[52] https://www.gov.ie/en/publication/7547d-language-support-schemes/
[53] Representatives from 41 institutions included but were not limited to bodies such as Department of Foreign Affairs, Department of Culture, Heritage and the Gaeltacht, An Post, County and City Councils, Universities, Health Service Executive, Defence Forces and the Language Commissioner.

use of CAT tools. EU-funded projects such as PRINCIPLE demonstrated the usefulness and cost-effectiveness, bespoke machine translation engines for Foras na Gaeilge, Rannóg an Aistriúcháin, the Department of Justice and the University of Galway.

- **Identify and gain access to outsourced translations:**
  Since the ELRC and ELRI data collection campaigns, a number of stakeholders have updated their contracts with their LSPs in procurement of translations to stipulate the requirement of the return of a TMX file, terminology database or related glossaries. If such a service was established, the Shared Translation Service could also ensure streamlined and centralised access to all outsourced translations from government departments.
- **Establish good data management practices in public services:**
  Outreach workshops, onsite assistance and online training videos are amongst the approaches being taken in Ireland to encourage improvements in data management practices both within Government Departments and other public bodies. The appointment of Open Data Liaison officers and the new Open Data audit requirements will also support the implementation of these practices.
- **Tackle legal concerns:**
  The establishment of the National Relay Station/eSTÓR aimed to address concerns of contributors with respect to data sharing licencing and copyright in Ireland. Administrators of eSTÓR provide a first stop information point to advise a user on the appropriate sharing levels for any given data set. Any queries that require further detailed examination or investigation are referred to the Open Data Unit or the ELRC Helpdesk.

References and links:

20-Year Strategy for the Irish Language 2010-2030, 2018, https://assets.gov.ie/24380/865cef555ba845af8c1e71088f9a6e9d.pdf.

CEF Telecom call – Automated Translation (CEF-TC-2019-1) for PRINCIPLE project: https://euroalert.net/call/3864/2019-cef-telecom-call-automated-translation-cef-tc-2019-1.

ELRI Ireland training video series: https://www.youtube.com/watch?v=xW-sKTTkSzY&t=14s.

European Research Infrastructure for Language Resources and Technology (CLARIN): www.clarin.eu.

Official Languages Act, 2003: http://www.irishstatutebook.ie/eli/2003/act/32/enacted/en/html.

Official Languages (Amendment) Act, 2021: https://www.oireachtas.ie/en/bills/bill/2019/104/.

The New Corpus for Ireland: http://corpas.focloir.ie/.

[Action Plan, 2019] Department of Culture, Heritage and the Gaeltacht: Action Plan for the Irish language (2018-2022), https://assets.gov.ie/223350/0bae47d9-d16d-40de-9339-c30c88d59f58.pdf.

[AI Strategy, 2021] Department of Enterprise, Trade and Employment: *AI – Here for good: National Artificial Intelligence Strategy for Ireland*, https://www.gov.ie/en/publication/91f74-national-ai-strategy/.

[ELRC, 2016] Dowling, Judge, Way: ELRC Workshop Report for Ireland, 2016, http://lr-coordination.eu/sites/default/files/Irleand/ELRC-Workshop-Report_Ireland-Public.pdf.

[ELRC, 2017] Dowling, Lynn, Way: ELRC Workshop Report for Ireland, 2017, http://www.lr-coordination.eu/sites/default/files/Ireland2/ELRC%2B%20Ireland%20Workshop%20Report-Public_0.pdf.

[ELRC, 2021] Lynn: ELRC Workshop Report for Ireland, 2021, https://www.lr-coordination.eu/sites/default/files/Irleand/2021/ELRC3_Workshop%20Public%20Report_Ireland_updated.pdf.

[Guidebook, 2015] Office of An Coimisinéir Teanga: *Official Languages Act 2003*, Guidebook, 2015, https://assets.gov.ie/89742/cce604d6-0bfc-47a3-8a74-0a9d5baaac73.pdf.

[Judge et al., 2012] Judge, Ní Chasaide, Ní Dhubhda, Scannell, Uí Dhonnchadha: The Irish Language in the Digital Age. In: META-NET White Paper Series, 2012, http://www.meta-net.eu/whitepapers/volumes/irish.

[Lynn, 2022] Lynn, Theresa: Deliverable D1.20 Report on the Irish Language, 2022. Project deliverable; EU project European Language Equality (ELE); Grant Agreement no. LC- 01641480 – 101018166 ELE, https://european-language-equality.eu/wp-content/uploads/2022/03/ELE___Deliverable_D1_20__Language_Report_Irish_.pdf.

Annex

Country Profile Italy

State of Play:

Translation practices and information exchange in ministries and public administrations:

In Italian PA, translation is a process that can either be carried out internally, by a unit devoted to translation services, or outsourced to external translation agencies.

The translation process is highly decentralised, meaning that there is no formalised exchange of know-how or language data between public bodies and the procurement process is not centralised either. Only larger ministries have their own translation service but all ministries take care of their own translation needs.

There is no coordination among editorial and translator experts working in different Ministries and Departments. Only a few administrative bodies formally recognise the role of language officer. There is no official network among people working in different institutions in the field of public communication and translation.

Computer Assisted Translation tools are rarely used by public translation units and those instruments are still perceived as **expensive software**, even in terms of training. This evaluation does not consider their positive impact in terms of both quality and efficiency of the translation process. Sometimes CAT tools are also perceived as systems that could potentially replace translators, rather than tools supporting their activities.

There is an "unconfessed" use of open source MT web interfaces underlining the fact that there is little awareness of related copyright issues and potential security breaches through uploading potentially confidential or personal data to these web services. Yet, Italian companies are currently implementing techniques that make MT systems not only competitive with large international players, but in many contexts even superior to them.

Although Italy has only one official language, 12 other languages have co-official status at regional level, out of which only three are official EU languages (French, German, Slovene). Some other immigrant groups form additional linguistic minorities (e.g. Chinese, Arabic) which leads to a high demand for translations. To meet the high demand in translations, all public administrations (PAs) make use of procuring translations. As mentioned above, there is no system or formalised process in place and the produced translation memories (TMs) and any other by-products are not requested back together with the translation.

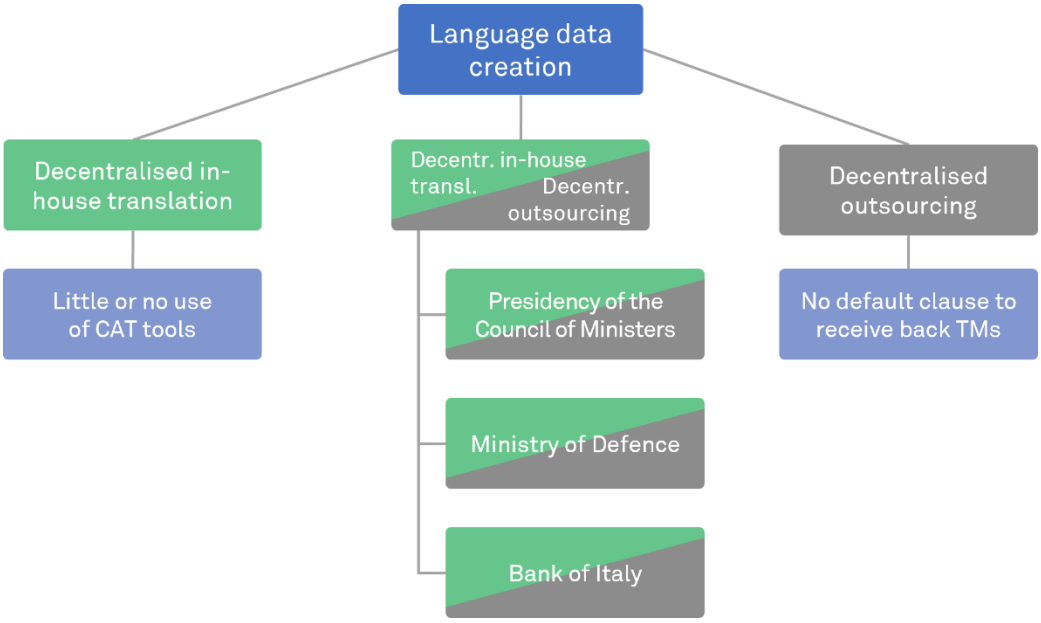
However, there are also some recent developments where some translators in public administrations (Bank of Italy, Department for European Policies at the Presidency of the Council of Ministers, Ministry of Defence) started to use CAT tools. As of recently, the contracts for outsourced translations were rewarded by the Ministry of Defence and the Department for European Policies to claim ownership of the translation memories and any by-products of outsourced translations.

At the second ELRC Workshop in Italy, many participants expressed their wish for an API that can be integrated into websites for automated translation, but concerns were raised with respect to the costs of implementing and maintaining such a service. With respect to a potential eTranslation API, concerns were raised regarding the availability of the service free of charge.

Interesting fact:

The average age of civil servants working in ministries, the Presidency of the Council of Ministers and other public bodies is over 54 in Italy. More than 16% are over 60 years old and less than 3% are under the age of 30[54].

The current language data creation and sharing infrastructure in Italian public bodies looks as follows:



Open Data and data collection in Italy:

Although there is no central repository for language data sharing in Italy and no sustainable infrastructure in place yet, several public administrations and research institutions have already donated language data to the ELRC-SHARE repository.

In the past years, considerable efforts were made to improve publishing of Open Data – making Italy one of the trendsetters for Open Data in Europe. On this front, there has been considerable growth, and according to the Digital Economy and Society Index (DESI) 2018, the country has in fact improved its position in the ranking of 11 places, thus exceeding the EU average for Open Data (cf. DESI, 2021). The initiative "Open Data 200 Italy" aimed at carrying out the first comprehensive, internationally comparable study of Italian companies that are using Open Data to generate business, develop products and services, and create social value. Open Data is widely promoted through public initiatives, but textual data is underrepresented in terms of shared language data and its promotion as a valuable resource.

Digital policy and language policy in Italy:

Italian is spoken by about 58 millions of resident people and is the official language of the Italian Republic, according to National Law no. 482 of 1999. Before this date, an official language had never been explicitly stated. The same law provides the institutional framework for protection of historical linguistic minorities by the Republic protecting the language and culture of the Albanian, Catalan, Germanic, Greek, Slovenian and Croatian populations and of those speaking French, Franco-Provençal, Friulian, Ladin, Occitan and Sardinian.

Italian is largely used for all types of communication in everyday life and is the language of almost all national media, publishing and public administrations of the State. Its use is not explicitly regulated. The use of recognised minority languages is permitted in school education, PA offices and institutions, and public signage.

[54] Source: ForumPA 2019.

The European Charter of Minority Languages has been signed, but not yet ratified.

As the Open Data initiatives already indicated, Italy puts a lot of effort into modernising and digitalising the public sector. To achieve the objectives of the "Three-Year Plan for Information Technology in Public Administration", Artificial Intelligence (AI) methods and tools shall be applied to public services. The White Paper "L'Intelligenza Artificiale – al servizio del Cittadino" published in March 2018 underlines the role AI shall play in the process. The paper is edited by the AI Task Force and is the result of a consultation, synthesis and analysis process that has involved hundreds of public and private subjects dealing with AI (cf. AI Report, 2018). Machine translation is listed as one of the technologies playing a key role to meet the challenges public administrations are facing with the help of AI at the service of citizens. However, the integration of MT tools into everyday translation work carried out by public administrations is still fairly limited. A notable exception is represented by the Dipartimento per le Politiche Europee della Presidenza del Consiglio dei Ministri (Department for European Policies of the Presidency of the Council of Ministers), which is the first Italian PA to integrate eTranslation into their data management flow.

The Italian national strategy was developed by converting the objectives of the European Digital Agenda into initiatives aimed at the digital transformation of Public Administration. The four key pillars are:

- Digital ecosystems
- Immaterial infrastructures
- Physical infrastructures
- Cybersecurity

Yet, there is still a gap between the supply of digital services and their actual use by the population. The digital divide, both at the level of individual users and between producers and possible users of new technologies, is still conspicuous. It is very important that new technologies are perceived as supporting production processes and as a means for developing new, more qualified forms of work.

Stakeholders and major networks:

The Agency for Digital Italy is not only implementing the digital agenda in Italy, the agency is also responsible for the national Open Data Portal. As such, they are one of the key stakeholders for sustainable language data sharing who themselves already contributed language data to ELRC.

Over the past years, more than 50 institutions, among them federal and regional public administrations and agencies as well as research institutions, participated in ELRC events. The national ELRC Workshops provided the opportunity to involve almost all the public institutions that have translation needs as well as those that are currently using embedded eTranslation in the provision of cross-border services. The increased number of attendees of the second ELRC Workshop with new institutions involved (inter alia Bank of Italy, INPS, Ministry of Defence) and of new language resources provided, indicate a growing interest in MT.

Among the active contributors of language data are the Ministry of Interior, the Ministry of Justice, the Province of Bolzano, the Prefecture of Florence, as well as the Universities of Bologna and Pisa and the Institute for Specialised Communication and Multilingualism, EURAC Research, Bank of Italy, and Presidency of the Council of Ministers.

Main challenges for sustainable data sharing:

The challenges Italy is facing in sharing data and restructuring the translation workflow to make it more efficient can be grouped in three categories:

- Public perception of language data
  - Awareness of the importance of language resources for machine translation and other applications of artificial intelligence is growing steadily.
  - Language data is increasingly considered a valuable resource and regulation of its management is starting. An appropriate language data management structure is still lacking at the institutional level.
  - Permanent education re. data sharing and open data is needed in order to increase willingness to share translations.

- Structural issues
  - Little knowledge about automatic anonymisation tools results in valuable data remaining not shareable, in addition to manual anonymisation process being very time consuming.
  - Privacy concerns regarding confidential and personal data (GDPR) are an obstacle for many Ministries (incl. Justice and Interior) to share language data.
  - Strict rules prohibit the use of data conveying confidential information. To keep confidentiality, it is important to identify software solutions for anonymisation or pseudo-anonymisation.
  - The MT field in Italy needs to create strong synergies between excellences at the university, industrial, and research levels: a real national infrastructure in which new researchers, future professors, and new professional translators are trained. This is essential if the necessary conditions to confront ourselves with the big industrial players are to be created and at the same time to develop the collaborations and opportunities that will allow technology to progress further.
  - The representative of the Italian Digital Agency Gabriele Ciasullo referred to Law Decree No. 76/2020, known as the "Simplification Decree" and containing "Urgent Measures for Simplification and Digital Innovation", published in the Official Gazette in 2020. The Law emphasises the importance of public administrative data and the need to share those data for institutional purposes. The decree imposes an obligation on the President of the Council of Ministers to adopt a national data strategy.
- Translation workflow
  - Little or no use of CAT tools and eTranslation (in some public administrations, translators have only recently started to use CAT tools).
  - Unconfessed use of free online translation services.
  - The original text is often on paper or "digitised" through bad quality scans.
  - Resistance to changing the translation workflow due to average age of civil servants being above 54 in ministries.
  - Investing in CAT Tools is only measured by the cost of purchase, the productivity gain or reuse of language resources is not taken into consideration.

Action plan:

Italy runs the risk of falling off on the side of those who do not have enough resources available for competing on the big tech market. It is important to raise the alarm on this concern, including towards political decision-makers.

Italy needs awareness raising activities on the value of language data both with translators and decision makers in public administrations. This should be done through examples of how language data management practices can reduce costs and improve quality.

Legal, privacy and ownership concerns should be addressed and best practices in the use of CAT tools and language data management should be developed, preferably by a central body.

The following specific objectives are suggested to address the challenges Italy is facing when it comes to sharing language data:

- **To increase the interest in MT/LT in public services as part of the national digital policy.**
  Specific actions include:
  - Establishing synergies with national projects and initiatives
  - Diffusing best practices
  - Securing the support of decision makers to adapt the national policy
  - To identify and gain access to outsourced translations
  - Establish practice of receiving any by-product of outsourced translations, especially translation memories
- **To establish good data management practices in public services**
  - Further investigation of data management practices
  - Definition of confidential and personal data that can be used to introduce the practice of clear separation between confidential and personal data from public sector information in the translation process

  - Establish shared language data management practices to reduce costs, improve quality and leverage on existing language assets
  - Establish a network of language experts along the model of the German Bundessprachenamt. An agency for language services, providing language training for civil servants and carrying on all translation and interpretation services could prove very useful for overcoming current fragmentation at the ministerial level.
- **To raise awareness of the importance of language data as a valuable asset and as Open Data**
  - Raise awareness on the value chain of language data and the importance of LR
  - Share benefits of sharing language data
  - Enhance the publishing of Open Data making Italy one of the trendsetters of Open Data in Europe
  - Integrate language data in the national Open Data policy
  - Emphasise the role of digital texts in the digital economy
  - Establish practical guidelines for language data as Open Data
  - Foster knowledge and adoption of data centre services such as CLARIN-IT as an open and secure institutional archive that Italian organisations can entrust their data to.
- **To tackle legal concerns**
  - Develop, share easy to apply guidelines for IPR and privacy issues
- **To help dialogue between users and producers of language data**
  - New professional figures should be developed to help dialogue between users and producers, both to train users in the use of new technologies and to bring user needs to producers.

References and links:

[AI Report, 2018] The Agency of Digital Italy: *Artificial Intelligence at the service of citizens*, 2018, https://ai-white-paper.readthedocs.io/en/latest/

[DESI, 2021] European Commission: *Digital Economy and Society Index (DESI)*, 2021 https://ec.europa.eu/newsroom/dae/redirection/document/80494.

Annex

Country Profile Latvia

State of Play:

Translation practices and information exchange in ministries and public administrations:

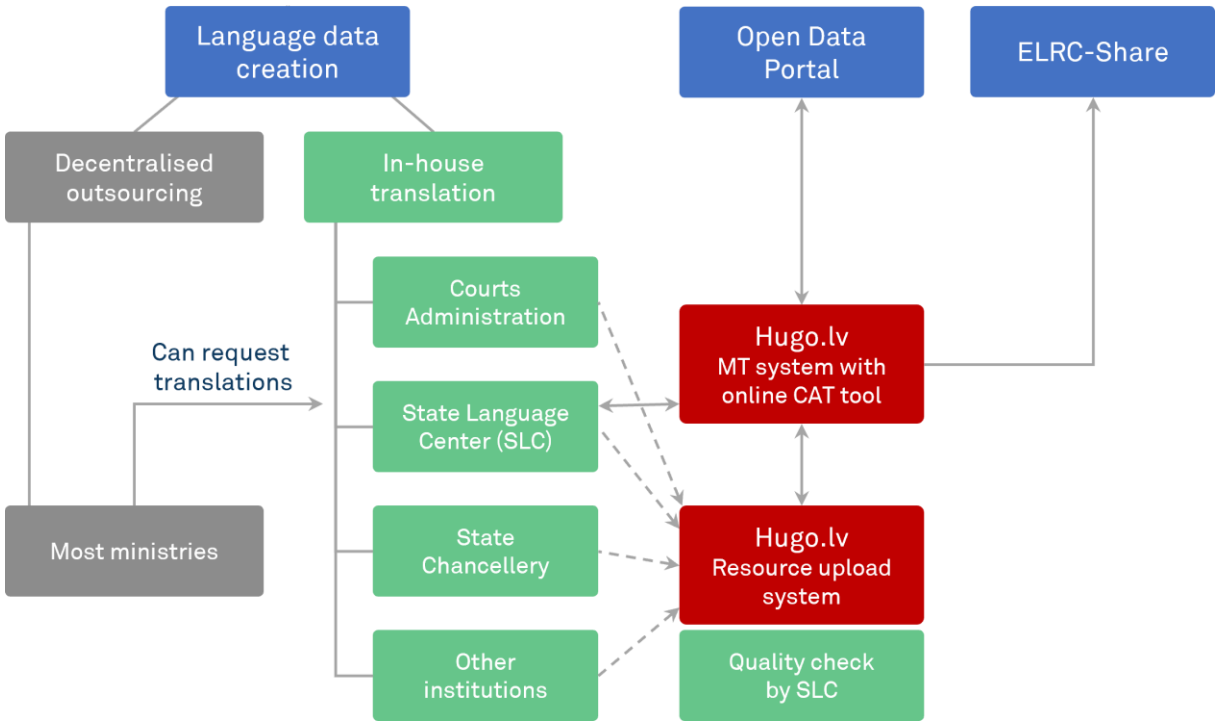
In Latvia, most translations are independently outsourced, and there are only some ministries and public administrations with in-house translation services. If documents are translated in-house or outsourced to freelance translators, computer-assisted translation (CAT) tools are rarely used. Contrary to that, the use of CAT tools is common practice for Latvian language service providers (LSPs) from whom some public administrations request back translation memories (TMs). However, through the language technology platform HUGO.lv, all public administrations and citizens have access to a free machine translation platform (see section "Data sharing infrastructures and Open Data in Latvia"). As regards procurement of translations, there are no specific procedures regulating translation subcontracts in Latvia. Latvia is subject to the EU regulation. Public procurement is currently regulated by two laws, the 2016 Law on Public Procurement "Public procurement Law" (Latvijas Vēstnesis, 254, 29.12.2016.) transposing the EU Directive 2014/24/EU and the Law on the Procurement of Public Service Providers (Latvijas Vēstnesis, 36, 16.02.2017.).

The procedure on the national level prescribes that procurements above the threshold of 10,000 EUR must be published on the website of the Procurement Monitoring Bureau (including the notification and the results of the procurement and if applicable the winner).

Interesting fact:

Every internet user and all Latvian public administrations have free access to the Latvian State administration language technology platform HUGO.lv offering machine translation, an online CAT tool, speech recognition, speech synthesis and other tools free of charge.

The current language data creation and sharing infrastructure in Latvian public bodies looks as follows:



Open Data and data collection in Latvia:

In Latvia, there are a number of portals and platforms available with the goal to share data and make language technologies available. They include:

- **The Open Data Portal**
  The purpose of the Open Data Portal data.gov.lv is to gather and to circulate data collected by Government institutions and Government organisations in one central place for public use and reuse as this data is valuable for the development of innovations in the State. The Latvian Open Data Portal was created by the project Nr. 2.2.1.1/16/I/001 "Public Administration Information and Communication Technology Architecture Management System" (PIKTAPS) co-financed by the European Regional Development Fund.
- **HUGO.lv**
  Hugo.lv is a Latvian State administration language technology platform that is freely accessible to everyone. It provides machine translation, an online CAT tool, speech recognition and speech synthesis, virtual assistants, as well as a range of tools for supporting multilingual features in e-services. Hugo.lv is customised to the Latvian language and state administration documents, thus, its translation quality is significantly higher than in other online translation services. Furthermore, users of Hugo.lv can enjoy the services of the translation assistant for more convenient translation. One of the functions of the platform is the resource management feature with two main functionalities – "Submit resource" (TMX) and "Search the database". "Submit resource" is an easy and straightforward process – language resources can be submitted by creating a new user and by opening the "Submit resources" section. The user enters the resource name, the e-mail address for communication, describes the areas and languages covered, and uploads files containing text. By clicking the "Upload or drag a file" button, documents can be added or dragged into the enclosed area with a dashed line. The attached files will appear in the list. "Search in repository" – allows the user to search the repository for publicly available language corpora and download the returned results. When the "Resource Search" section is opened, the user can view all available corpora and the languages covered in the drop-down list of the search form. The user enters the keyword, selects the body and language, and then clicks "Search".
- **META-SHARE**
  META-SHARE, the open language resource exchange facility, is devoted to the sustainable sharing and dissemination of language resources (LRs) and aims at increasing access to such resources in a global scale. META-SHARE Latvia node is metashare.tilde.com and it serves as the META-SHARE hub for the Baltic and Nordic region.
- **CLARIN-LV**
  The Common Language Resources and Technology Infrastructure (CLARIN) is a research infrastructure that was initiated from the vision that all digital language resources and tools from all over Europe and beyond are accessible through a single sign-on online environment for the support of researchers in the humanities and social sciences. Latvia joined CLARIN ERIC in June 2016. The coordinating centre of CLARIN Latvia is the Institute of Mathematics and Computer Science (IMCS), University of Latvia. In addition, IMCS develops and provides Latvian language resources (text and speech corpora, treebanks, framebanks machine-readable and computational lexical resources; language processing models, tools and pipelines) as open data and open source for language technology and digital humanities research and development.

  Two particularly popular language resources that are frequently used also by translators are Tezaurs.lv, the major open platform of Latvian lexical resources, and Korpuss.lv, the national corpora collection which currently consists of nearly 30 written and spoken language corpora developed by more than 10 institutions and hosted by several distributed nodes. Most of these corpora have common automatic morpho-syntactic annotation allowing for efficient and uniform federated search through Korpuss.lv.

Digital policy and language policy in Latvia:

The necessity of language technology/ies to support Latvian as Latvia's sole official language in digital means has been recognised by the national government and is reflected in a number of language policy documents. Regulatory enactments governing the use of the official language of Latvia include:

- **Laws:**
  - The Constitution of the Republic of Latvia
  - Official language law
  - Law on Submissions
- **Cabinet regulations:**
  - Regulations regarding the amount of knowledge of the official language and procedures for the verification of the proficiency of the official language for the performance of professional duties, receipt of a permanent residence permit and acquisition of the status of a permanent resident of the European Union and the State fee for the verification of the fluency of the official language
  - Procedures by which translations of documents into the official language shall be certified
  - Rules on the provision of translations into measures
  - Rules on the use of languages in information technologies
  - Rules regarding the formation and use of the names of institutions, public organisations, undertakings (companies) and the names of events
  - Provisions regarding the use of foreign languages on stamps and forms
  - Provisions regarding the State fee for the performance of professional duties regarding the attestation of proficiency in the official language
  - Provisions regarding the spelling and use of personal names in Latvian, as well as their identification
  - Information rules for place names
  - By-law of the meeting of senior officials in matters of the European Union
  - Other cabinet regulations:
- **Cabinet instructions:**
  - Procedures for evaluating, harmonising and correcting translations of European Union documents
  - Procedures for requesting and providing translations
  - By-laws of national regulatory authorities related to the official language:
  - By-law of the Latvian Language Expert Commission of the State Language Centre
  - By-law of the Terminology Commission of the Latvian Academy of Sciences
- **Policy Planning Documents:**
  - National language policy guidelines for 2021-2027
  - Concept for the development of the system of administrative penalties
  - Latvia's Open Data Strategy
  - Language technology is part of Digital transformation guidelines for 2021-2027 (Order of Cabinet of Ministers Nr. 490) as 4.4.7. Action Direction: Machine translation and language technologies with vision: The digital space of EU countries (the single digital market) is accessible to Latvian residents in printed, audio and visual form, as well as the European citizen interacts with Latvia's digital space in his or her mother tongue. The most important language resources are provided for Latvian for sustainable language development and extensive use in digital services.

Stakeholders and major networks:

Overall, more than 60 organisations participated in local ELRC Workshops and conferences including high-level officials, indicating strong interest in the topics covered by ELRC in Latvia.

Interesting fact:

In September 2019, the Culture Information Systems Centre contributed open language data, which was generated by the LT platform Hugo.lv to ELRC. The donation includes monolingual corpora with more than 300 million words, parallel corpora with 15 million words and 19,000 Latvian terms[55].

[55] ELRC News Article: Latvia contributes language data for eTranslation, 2019.

In addition, several data sets were already contributed to ELRC by the Bank of Latvia and other institutions. The key stakeholders of the public sector are:

- **Culture Information Systems Centre (CISC):**
  CISC is the owner and Administrator of HUGO.LV. The CISC operates under the supervision of the Ministry of Culture of Latvia. The objective of the Centre is to provide access to information resources and cultural heritage stored in archives, museums and libraries. CISC implements national and international ICT, provides training and supplies public access to projects and programmes to enable free and equal access to information, resources and cultural heritage stored in libraries, archives, museums and other cultural institutions.
- **The State Language Centre:**
  The centre has two main objectives, i.e. (1) to implement the national policy with regard to supervision and control of the conformity with laws and regulations in the field of the official language use and (2) to supervise the implementation of the Official Language Law.
- **The Latvian Language Agency:**
  The Latvian Language Agency is a direct administration institution supervised by the Minister of Education and Science, and its aim is to enhance the status and promote sustainable development of the Latvian language – the official State language of the Republic of Latvia and an official language of the European Union.
- **Latvian Language Expert Commission:**
  The Latvian Language Expert Commission, on a regular basis, examines the compliance of norms provided for in laws and regulations to the rules of the Latvian language, codifies norms of the literary language, provides opinions on various language issues, for example, the use of capital letters in the names of establishments, the spelling of internationally recognised names of countries and territories, house names and numbers, addresses, languages and language groups in the Latvian language in compliance with the requirements of ISO 639-2 et al. The commission prepared several draft legal acts and participated in the formation of the normative basis for the Official Language Law.
- **Institute of Mathematics and Computer Science (IMCS), University of Latvia:**
  IMCS is the national coordinator and a node of CLARIN in Latvia, and a partner institution of DigitalHumanities.lv. AI Lab at IMCS continuously develops and hosts the open Korpuss.lv and Tezaurs.lv platforms, the open source Latvian NLP pipeline (nlp.ailab.lv) and other language resources and tools. Its mission is to provide state-of-the-art language resources and models for Latvian in the multilingual setting. Together with industry partners, IMCS develops innovative solutions for media monitoring and news production, medical speech transcription, and other domains.
- **National Library of Latvia (NLL):**
  NLL is a partner institution of DigitalHumanities.lv and Korpuss.lv. It is one of the major holders of synchronic and diachronic text and speech resources of Latvian.

Main challenges for sustainable data sharing:

- Lack of awareness and distribution of responsibilities in ministries and other state administration institutions.
- Practices and procedures for subcontracting translations including the return of translation memories are not defined by the state administration.

Action plan:

- Dissemination campaign in all ministries to raise awareness about language data as an important asset for language equality and digital presence is considered vital.
- Together with the Ministry of Environmental Protection and Regional Development changes of procurement procedures for translation subcontracts by state administration should be initiated.
- The state administration should develop good internal practices for language data management.
- The importance of language technologies and benefits for the state administration and citizens from introducing language technologies should be promoted.

References and links:

CLARIN Latvia: http://clarin.lv, https://repository.clarin.lv.

Digital Humanities in Latvia: http://www.digitalhumanities.lv.

HUGO.lv: https://hugo.lv.

Latvian National Corpora Collection: http://korpuss.lv.

Latvian Open Data Portal: https://data.gov.lv.

META-SHARE: http://metashare.tilde.com.

[ELRC, 2015] Berzins, Aivars, Kalnins, Rihards: ELRC Workshop Report for Latvia, 2015, http://lr-coordination.eu/sites/default/files/ELRC-Workshop-Report_Latvia.pdf.

Annex

Country Profile Lithuania

State of Play:

Translation practices and information exchange in ministries and public administrations:

In Lithuania, there is no centralised translation service in the public administrations yet. Consequently, translation practices vary from institution to institution, ranging from decentralised outsourcing to in-house translation. In the case of Lithuania, independent outsourcing clearly dominates in public administrations and there are only single ministries with in-house translation services. If the translations were outsourced, in some public administrations, it is common practice to request the translation memories (TMs) and other by-products of the translations back.

When it comes to translation services, only single language service providers (LSPs) and freelance translators use computer-assisted translation tools (CAT Tools). This also applies to public administrations, where CAT Tools are only rarely used. Public administrations and ministries usually do not use machine translation (MT) APIs, but freely available MT services. However, this is not the case in the Lithuanian Parliament, since its translators are already successfully using the European Commission's MT system eTranslation.

Interesting fact:

In 2018, a new EU project was launched that is developing an MT system for Lithuanian, English, French and Russian to be used nationwide. The project is one of five EU-funded projects in the programme "The Lithuanian Language for Information technologies". It is being implemented by Vilnius University and focuses on the modernisation of previously developed MT systems. Almost 4 million EUR have been allocated for the implementation of this project.

In Lithuania, the Law on Public Procurement of the Republic of Lithuania (LPP) is the main piece of legislation, which governs the implementation of public procurement (NEC TM, 2019). The official procurement portal is called the Central Public Procurement Information System. The one-stop-shop portal for public procurement is managed by the Public Procurement Office and its use is mandatory for all public buyers. The portal includes tender notices, publications of awarded contracts and procurement plans and even allows direct electronic communication between the buyer and the economic operators.

Interesting fact:

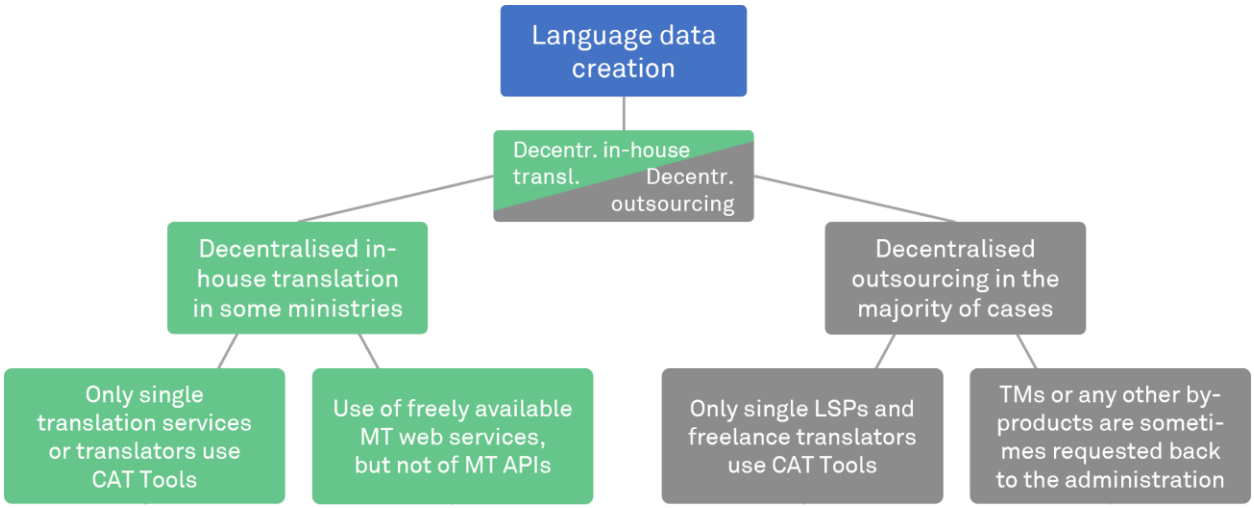
The use of Lithuania's official procurement portal "The Central Public Procurement Information System" is mandatory for all public buyers.

According to the findings of NEC TM, Lithuania spent almost 10 million EUR for translation contracts between 2015 and 2018. Organisations with the highest demand are Lithuania's central purchasing body CPO LT, the Education Exchanges Support Foundation and the State Tourism Department under the Ministry of Economy. This clearly demonstrates that multilingual services are of high relevance in all fields of Lithuanian public services, ranging from finances to education and social affairs.

Interesting fact:

The investment of creation of Lithuanian language resources for AI solutions is included into EU funded "Lithuania's Recovery and Resilience plan (RRF).

The current language data creation and sharing infrastructure in Lithuanian public bodies looks as follows:



Open Data and data collection in Lithuania:

As regards the exchange and collection of data on the national level, Lithuania is part of CLARIN ERIC. The corresponding CLARIN-LT consortium was founded by three partner universities, i.e. Vytautas Magnus University, Kaunas Technology University and Vilnius University and maintains a repository to collect language data.

According to the Vice-Minister of the Ministry of Economy in Lithuania, the country is known for leading public e-government services. Data openness is mentioned as one of the key objectives of the Digital Agenda for Lithuania from 2014-2020, including the development of legal means for opening data from state and municipal authorities and agencies for example. The creation of effective management structures for opening such data also plays an important role in Lithuania's current Digital Agenda.

In addition, the creation and development of publicly available written and spoken digital content in the Lithuanian language and their implementation in Information and Communications Technology (ICT) and eServices are explicitly mentioned. In order to comply with the Public Sector Information (PSI) Directive, the "Law on the Right to Receive Information from State and Local Authorities and Institutions" was adopted, thus increasing the scope of information intended for re-use to e.g. museums and archives and defining the conditions for the open licence to use public sector information based on the Creative Common (CC) Licence. The latter makes it easier for recipients of information to share, re-use, process or translate any received information.

An important portal is the central electronic services portal "eGovernment gateway". It provides public information and e-services for citizens and businesses by redirecting the portal's visitors to appropriate websites. In recent years, the Lithuanian government has aimed to proceed with the centralised digitalisation of public services. Consequently, the Lithuanian Ministry of Economy and Innovations created a strategy for real-time digital government.

Digital policy and language policy in Lithuania:

With approximately four million speakers, Lithuanian is one of the least commonly spoken European languages. Pursuant to the Constitution of the Republic of Lithuania, the Lithuanian language has the status of state language. In order to enforce this status and to protect the language, the so-called Law on the State language was introduced in 1995. Due to the small number of speakers, the Lithuanian government supports a number of different programmes, which aim to promote linguistic research and dissemination. In this context, the Institute of the Lithuanian Language plays a key role, since it is one of the most important centres for research and dissemination of the Lithuanian language.

The Lithuanian language policy consists of a set of principal guidelines, which partly go beyond the national borders of Lithuania. It states that the Lithuanian language must be in line with the language policy of the European Union and that it should be developed as a constituent part of multilingual

terminologies and resources of the EU. In addition, automated translation is considered highly relevant when it comes to language use within the EU (Vaišnienė, 2012).

The development of the Lithuanian language resources and technologies can be divided into three stages, i.e. the first from 2004 to 2012, the second from 2012 to 2015 and the third from 2016 to 2020 (Utka et al., 2016). Whereas during the first phase, Lithuanian was considered highly under-resourced and without any (or only weak) language technology support, the period between 2012 and 2015 was labelled as the breakthrough. This was achieved due to three major actions, i.e. the implementation of the national programme "The Lithuanian Language for Information Society", the preparation of political guidelines for the further development of language technologies for Lithuanian 2016 to 2020 and international collaboration of LT communities and infrastructures. However, since there is only a small market for language technologies in Lithuania, the private sector does not show much interest when it comes to the development of LT.

During the third phase, a new national programme was launched in 2018, i.e. "The Lithuanian Language for Information technologies". The programme is funded by EU Structural funds and in total, more than 17 million EUR have been allocated for five language technology projects. The overall aim of the projects is to produce 21 new electronic language services for the general public and public institutions (i. e. machine translation, speech recognition, automatic speech transcription, automatic summarisation etc.). On the smaller scale, during this phase, the Research Council of Lithuania has funded a number of important scientific LT projects that are important for the overall picture.

The role of LT and language policy in Lithuania's AI regulations

"Guidelines on the Development of Lithuanian Language in Digital Environment and Advancements in Language Technologies (2021-2027)", adopted by the State Commission of Lithuanian Language, provide a thorough overview of the European and Lithuanian strategic documents, funding instruments and institutions regulating LT development in the country[56]. The main goal of the guidelines is to overview and facilitate the full use of the Lithuanian language in the digital environment. Drawing on the information provided in this as well as other relevant documents, this section will briefly highlight the major initiatives, projects and developed language resources.

The Guidelines are the basis for the next phase of the development of the Lithuanian language technologies. The investment of creation of language resources for AI technologies is included into EU funded "Lithuania's Recovery and Resilience plan (RRF)"[57]. The planned Investment for the language resources is 35 million EUR for the 2023-2025 period. Potentially, the investment could boost the development of AI technologies across public and private sectors.

Stakeholders and major networks:

Two relevant stakeholders are represented by the Lithuanian ELRC National Anchor Points, i.e. the State Commission of the Lithuanian Language, Vytautas Magnus University and Vilnius University, who are involved in a number of initiatives that are of relevance to ELRC, such as CLARIN or META-Net. Other important stakeholders are the Office of the Government of the Republic of Lithuania and the Seimas, which has already contributed language data to ELRC. The Research Council of Lithuania and the Institute of Lithuanian language may also play important role in contributing language resources. Overall, more than 30 Lithuanian organisations have participated in previous ELRC events, demonstrating that there is an increasing interest in the topics addressed by ELRC.

Main challenges for sustainable data sharing:

- Potential data contributors are sometimes reluctant to share their language resources due to concerns about their relevance and/or quality;
- Lack of interest at the level of decision makers;

[56] https://e-seimas.lrs.lt/portal/legalAct/lt/TAD/cd0584707b6e11e89188e16a6495e98c
[57] https://ec.europa.eu/info/business-economy-euro/recovery-coronavirus/recovery-and-resilience-facility/lithuanias-recovery-and-resilience-plan_en

- Lack of effective anonymisation procedures, as recent GDPR restrictions made some contributors worry about revealing their sensitive privacy data to outside sources.
- Language resources for further development of language technologies are scarce and require much effort to collect or create them manually or synthetically.
- There is a no legal framework to collect public sector language resources.
- There is a lack of attention and support both from government and institutions to collect, store and share data. This means funding as well as human resources.

Action plan:

Taking the current challenges into account, five objectives could be defined. Ranked by their priority, these are:

- **To increase interest in MT in public services:**
  As it is currently not common practice to use machine translation systems in public administrations, examples of how MT can be a useful asset for institutions and ministries could raise attention and increase interest. These examples should clearly demonstrate how MT can facilitate daily operations and increase productivity. In addition, synergies with other national LT/MT projects and initiatives could have a positive impact on the Lithuanians' general interest in MT.
- **To raise awareness of language data as Open Data:**
  This could be achieved by emphasising the role of digital texts in the digital economy and by making people aware of the benefits of sharing data. As potential data contributors are sometimes concerned about the relevance and quality of their data, it would also be important to spread the message that single translation mistakes do not have a significant impact on the quality of the MT output.
- **To tackle legal concerns:**
  As legal concerns can prevent potential contributors from sharing their data, we would suggest to create and develop effective anonymisation procedures and tools, which are currently not available in Lithuania. In addition, it would be important to develop and share easy-to-apply guidelines, which can help data contributors to overcome issues related to privacy or Intellectual Property Rights (IPR).
- **To identify and gain access to outsourced translations:**
  This could be achieved by e.g. cooperating with the NEC TM Data Project.
- **To establish good data management practices in public services:**
  A first step towards improving data management practices in Lithuanian public services could be the appointment of a data manager who is responsible for the data management practices in the respective ministry or public administration. In order to be able to decide on the practices to be applied, it would be important to investigate potentially useful data management practices on the one hand, but also to introduce clear definitions of confidential and personal data on the other.

References and links:

CLARIN-LT: http://clarin-lt.lt.

European Commission: Digital Government Factsheet Lithuania, 2019, https://joinup.ec.europa.eu/sites/default/files/inline-files/Digital_Government_Factsheets_Lithuania_2019_0.pdf.

Fundings for Lithuanian language in IT: https://www.esinvesticijos.lt/lt/finansavimas/patvirtintos_priemones/veiksmu-programos-prioriteto-igyvendinimo-priemone-nr-02-3-1-cpva-v-527-lietuviu-kalba-informacinese-technologijose.

MT System developed by Vilnius University: https://www.raštija.lt/vu-masininis-vertimas/vilniaus-universiteto-masininis-vertimas-/16.

[NEC TM, 2019] NEC TM Report: *Report on National Translation Contracts in Lithuania Public Procurement Market Research: Process and Findings for Lithuania*, 2019, https://www.nec-tm.eu/wp-content/uploads/2019/03/Lithuania-Report.pdf.

[ELRC, 2016] Špukienė, Renata: *ELRC Workshop Report for Lithuania*, 2016, http://www.lr-coordination.eu/sites/default/files/Lithuania/ELRC-Workshop-Report_Lithuania_public.pdf.

[ELRC, 2019] Špukienė, Renata: *ELRC Workshop Report for Lithuania*, 2019, http://www.lr-coordination.eu/sites/default/files/Lithuania/2019/ELRC%2B%20Workshop%20Report_Public_Lithuania.pdf.

[Utka et al., 2016] Utka et al.: *Overview of the Development of Language Resources and Technologies in Lithuania (2012-2015)*, 2016, https://etalpykla.lituanistikadb.lt/object/LT-LDB-0001:J.04~2016~1569937265781/J.04~2016~1569937265781.pdf.

[Utka et al., 2020] Utka et al.: *Development and Research in Lithuanian Language Technologies (2016-2020).* In: IOS Press: https://ebooks.iospress.nl/volumearticle/55546.

[Vaišnienė, 2012] Vaišnienė, Zabarskaitė: *The Lithuanian Language in the Digital Age.* In: Meta-NET White Paper Series, 2012, http://www.meta-net.eu/whitepapers/e-book/lithuanian.pdf.

Annex

Country Profile Luxembourg

State of Play:

Translation practices and information exchange in ministries and public administrations:

Luxembourg is a multilingual country with three administrative languages (French, German and Luxembourgish) and a multilingual population leading to a considerable need for translation into French, German and English. Luxembourg shows significant efforts into making public services digital and multilingual, the main web portal in this domain being Guichet.lu[58] that is managed by the Government IT Centre[59] (Centre des technologies de l'information de l'État – CTIE). Guichet.lu has its own in-house translation service and regularly exchanges translation memories. In order to fully meet their translation needs, all public authorities outsource at least some translations to either freelance translators or language service providers.

The government.lu[60] website is the information portal of the governmental Information and Press service, SIP. It federates all information and news concerning the Luxembourg government in four languages (German, French, English, and Luxembourgish).

When providing information for its citizens and businesses, the Luxembourg Government follows the principle to adapt the content to a large population by using short sentences and easy to understand, non-specialised language. To ensure that the published information is up to date, a legal team of the competent administration is in charge of checking new laws and procedures, and of adapting and revising the texts accordingly. If the technical solution is available, these changes automatically trigger a new translation job in the CAT tool of the translation unit at Guichet.lu. After the full translation process (human translation, revision, proof-reading and validated finalisation), the translated text is automatically sent to the publishing tool.

Time delays between the French and other versions are generally visualised by an "update" hint on the website informing the users about upcoming changes.

> **Interesting fact:**
>
> **In order to be accessible to as many people as possible, Guichet.lu is offering several documents in plain language. Plain language is a clear and easy-to-understand language. It is helpful for persons with reading and comprehension difficulties. The staff at Guichet.lu has worked on the documents in plain language together with Klaro, the official office for plain language. Klaro is a service managed by the Association pour personnes en situation de handicap (APEMH – association for persons with disabilities)[61].**

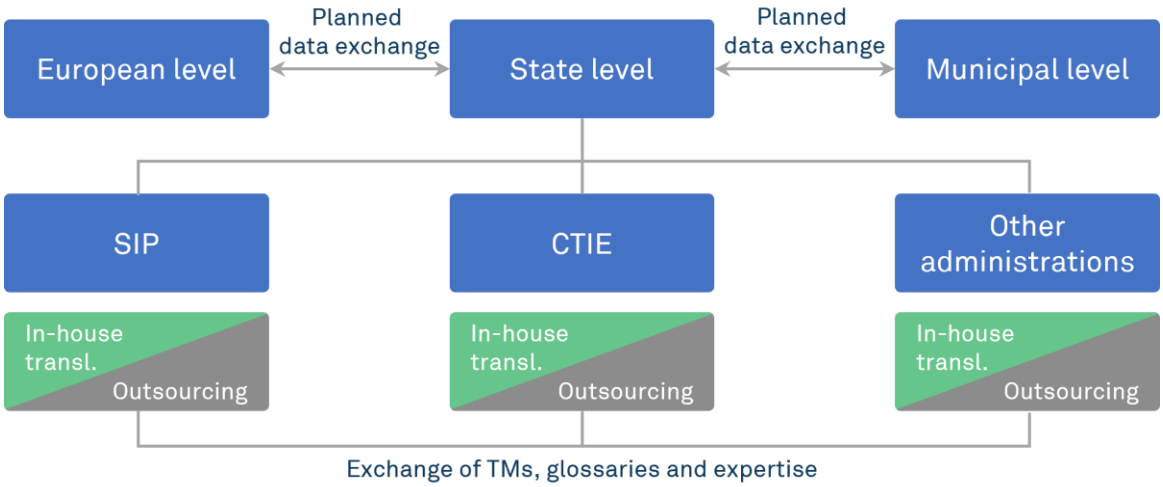
The second Luxembourgish ELRC Workshop showed that in the case of Luxembourg, public and private organisations often do not outsource their translations, simply because most employees already speak 3 or 4 languages. Consequently, translations tend to be handled internally. This is certainly a drawback of multilingualism, since translation management lacks a specific systematic workflow. Nonetheless, with regard to the infrastructures for sharing translations and language data, Luxembourg shows significant efforts in making public services digital and multilingual, as the examples of Guichet.lu and government.lu show. Since recently, there is also an application called GovID[62], with which a user can

[58] https://guichet.public.lu/en.html
[59] https://ctie.gouvernement.lu/en.html
[60] https://gouvernement.lu/en.html
[61] https://guichet.public.lu/en/actualites/2019/decembre/02-langage-facile.html
[62] https://gouvernement.lu/en/dossiers.gouv_ctie%2Ben%2Bdossiers%2Bgouvid%2Bgouvid.html

authenticate an official document issued in Luxembourg, in real time and for free, using a QR code printed on the document.

Talking about current advancements with regard to language data in the Luxembourgish public administration, there were no changes with regard to the preparation and processing of translations/language data in the Luxembourgish public administration. However, following a representative from Guichet.lu, this situation may change in the course of 2021 and corresponding updates will be shared with ELRC. Moreover, other Luxembourgish public administrative entities like the Official Journal of Luxembourg at the Ministry of State (http://legilux.public.lu) (SCL) have started investigating the opportunity of using machine translation tools like eTranslation in their workflows. However, in the case of SCL, the project was put on hold in 2020 due to the added difficulties within the COVID-19 pandemic but may be taken up again in the course of 2021.

The current language data creation and sharing infrastructure in the public bodies of Luxembourg looks as follows:



Open data and data collection in Luxembourg:

The Luxembourgish Open Data Portal (https://data.public.lu/en/) was launched in April 2016. When the first edition of the country profile was published in 2018, the portal hosted more than 800 published data sets. Comparing the number of published data sets from 2018 with 2021, it can be noted that there was a considerate increase from 800 to almost 1400 published data sets. They are in majority numerical data sets, and not textual. In 2019, the Luxembourg Institute of Science and Technology (LIST) published an evaluation of the impact of Open Data in Luxembourg in order to better understand its users and their expectations in terms of content and functionality. The main satisfaction results of the survey follow:

- More communication and advertising about the portal is needed;
- Finding data sets is the main goal of visitors;
- Data sets should be expanded and improved (real time, documentation, harmonisation of data sets);
- An advanced search tool is requested;
- All the domains are validated (with improvement ideas);
- The socio-economic impact of data.public.lu is real.

According to the survey, the main target of the users was to find “data sets for big data, similar to Kaggle.com”, “Raster data, data about Luxembourg” as well as “Roadworks”.

As far as the impact of Open Data is concerned, users had to rate the economic, environmental, and social feedback. The social impact was ranked as the highest with 68% and the environmental and economic equally with 63%. Regarding the social impact, respondents consider that the portal facilitates citizen science initiatives and provides understanding of population need, demand and use. The economic impact focuses on business activities in ITC and data science market, while the economic impact is about land rezoning and urban planning, among others.

For the first time in its history, a Ministry for Digitalisation headed by the Prime Minister, Minister for Digitalisation, Xavier Bettel, and Minister Delegate for Digitalisation, Marc Hansen was created on 11th December 2018, when the government programme was presented by Prime Minister Xavier Bettel to the Chamber of Deputies. Within this context, there is a recent initiative of the Ministry for Digitalisation and the Government IT Centre (CTIE), called GovTech Lab, which is an innovation laboratory that uses open innovation to work with internal (ministries, administrations, public actors) and external actors in the development of innovative solutions (technological or conceptual).

Language Technology in Luxembourg

AI is indeed a strategic vision for Luxembourg and Luxembourg intends to remain at the forefront of AI by collaborating across borders, boosting investments, enabling skills training and optimising its data market. The document "Artificial Intelligence: a strategic vision for Luxembourg." has been published by the Government of the Grand Duchy of Luxembourg and digital Luxembourg. It includes a foreword by the Prime Minister, Xavier Bettel, a description of the vision for AI in Luxembourg, the human-centric focus, the regional cluster of AI research in Luxembourg and the focus areas (e.g. data, ethics, skills & lifelong learning, etc.) The strategic vision is not intended as a one-off strategy, but rather the first edition of a policy vision, to be updated on a regular basis and further defined where needed. This policy vision is built on Luxembourg's ambitions as a digital front-runner:

- Ambition #1: – To be among the most advanced digital societies in the world, especially in the EU
- Ambition #2: – To become a data-driven and sustainable economy
- Ambition #3: – To support human-centric AI development

During the second ELRC Workshop in Luxembourg, the participants were invited to contribute to live polls, aiming to get more information about their use and satisfaction level when it comes to LT in Luxembourg. Most interestingly, almost 80% indicated that their organisation was either using or planning to use eTranslation. In addition, 93% answered that their organisation was using/intends to use other language technologies or services. This illustrated that LT is already a well-known topic in Luxembourg and that the vast majority is well-aware of its usefulness.

Digital policy and language policy in Luxembourg:

In the Grand Duchy of Luxembourg, Luxembourgish (Lëtzebuergesch), a Moselle Franconian dialect, is the national language. According to the provisions of the Languages Law of 1984, the three languages of the Grand Duchy, Luxembourgish, French and German, are the languages used in administration and the judiciary system. Legislative documents are in French and an important consequence of this on a judicial level is that only the French language text is deemed authentic for all levels of public administration. The Grand Ducal Regulation of 30 July 1999[63] reformed official spelling in Luxembourgish.

According to the National Institute for statistics and economic studies (STATEC)[64], as of 01.01.2019, the total population of Luxembourg is 645,397 with the native Luxembourgish people 341,230 and total foreigners 304,167. Based on another study published by STATEC in 2013, 70.5% of the population use Luxembourgish at work, at school and/or at home, 55.7% use French, and 30.6% German. On average, 2.2 languages are used. At the same census, 55.8% – a large majority of the country's inhabitants – gave Luxembourgish as their 'principal language'. Portuguese and French followed in second and third positions (15.7% and 12.1% respectively).

Stakeholders and major networks:

After the election of the new Government in October 2018, the Government's political programme has placed digitalisation at the centre of its policies. The importance of this topic is thoroughly discussed in the coalition agreement and, for the first time, a Ministry only dedicated to Digitalisation has been created showing that Luxembourg sees digitalisation as a core element of its development. The position of Minister for Digitalisation is filled by the Prime Minister, Xavier BETTEL, who is also Minister for

[63] http://legilux.public.lu/eli/etat/leg/rgd/1999/07/30/n2/jo
[64] https://statistiques.public.lu/dam-assets/catalogue-publications/en-chiffres/2022/demographie-en-chiffre-22.pdf

Administrative Reform, which underlines the central importance given to digitalisation and the reformation of the public administration. Marc HANSEN is appointed as Luxembourg's Minister Delegate for Digitalisation and works on a daily basis to enhance Luxembourg Government's efforts to support citizens and businesses on the road to digitalisation. One of the main objectives of the **Ministry for Digitalisation**[65] is to make the lives of the citizens and businesses of Luxembourg easier and to act as a 'facilitator' and a 'coordinator' of all activities related to digitalisation and eGovernment across all ministries.

A central player in this process is the Government IT Centre (Centre des technologies de l'information de l'État, CTIE), directly subordinated to the Ministry for Digitalisation and in charge of the setting up and development of eGovernment. Its main mission is to accompany the digital transition of the Grand Duchy's administrations, so that each of them may take full advantage of the opportunities offered by the information and communication technologies (ICTs).

Guichet.lu is managed by CTIE and is a web portal run by the Luxembourg Government to facilitate users' access to information and online services pertaining to any life event or administrative procedure they may have to deal with as private citizens or representatives of businesses, and to simplify administrative procedures. It was launched in November 2008 and offers step-by-step guidance on some 1600 administrative procedures.

In the context of the exchange of data (TMs, glossaries, best practices, etc.), the Information and Press Service (SIP)[66] and Guichet.lu have collaborated closely for many years. Together they maintain and regularly update the official glossary containing the names of Luxembourg administrations (FR, DE, EN, LB) which is available to every State administration as well as to the public on the Data.public.lu portal[67].

In order to extend data exchange and ensure terminology consistency at Luxembourg level, Guichet.lu created a platform for regular data and knowledge exchange that also serves as a discussion forum. This platform is available upon request and under certain conditions for Luxembourg administrations, which already have a multilingual website or aim to translate their website.

The Luxembourg City Municipality (VDL) and Guichet.lu collaborated closely during the time the VDL set up their multilingual website. Guichet.lu provided them with glossaries and best practices. This cooperation is continued through sporadic exchanges on best practices and translation problems.

As for data exchange with DG CONNECT, the Guichet.lu translation memory (FR, DE, EN) has been shared in its entirety with the European Commission in the framework of the ELRC project[68].

Digital Lëtzebuerg[69] (Digital Luxembourg), founded in 2014, is a multidisciplinary government initiative working with public, private and academic players to harness digitalisation for positive transformation. It approaches digitalisation holistically, focusing on five key areas: skills, policy, infrastructure, ecosystem and government. Executing the Luxembourg government's digitalisation strategy, Digital Luxembourg enables new projects, supports existing ones & boosts the visibility of nationwide efforts.

Zesummen digital[70] This newly formed digital inclusion portal of Luxembourg includes the National Action Plan, which aims to facilitate the emergence of a digitally inclusive society, the Actors and Actions who are committed to digital inclusion in Luxembourg.

Main challenges for sustainable data sharing:

- Raising motivation and awareness about language data and machine translation in the public administration;
- Small country with many multilingual people – need for translation.

[65] https://digital.gouvernement.lu/en.html
[66] https://sip.gouvernement.lu/en.html
[67] https://data.public.lu/en/datasets/liste-dadministrations-et-dorganismes-de-letat-luxembourgeois-en-francais-allemand-anglais-et-ou-luxembourgeois/
[68] https://elrc-share.eu/repository/search/?q=guichet
[69] https://digital.gouvernement.lu/en/dossiers.gouvernement%2Ben%2Bdossiers%2B2014%2Bdigital-letzebuerg.html
[70] https://zesummendigital.public.lu/en.html

- Participants at the second ELRC Workshop in Luxembourg also indicated that legal issues often prevent them from sharing their data
- Inadequate practices for language data management were also mentioned to be a key issue.

Action plan:

Taking the current challenges into account, the most important objective for Luxembourg seems to be to raise awareness that language data should be considered as Open Data and a valuable asset. The role of digital texts in the digital economy is the first step in the terms of this kind of awareness.

As a further objective, even better data management practices could be established in public services. In the European Union, the legislative framework of the Open Data movement is set out in Directive 2003/98/EC and Directive 2013/37/EU on the reuse of public-sector information. In the Grand Duchy, these Directives have been transposed by the Law of 4 December 2007, as amended, on the reuse of public-sector information.

It could also perhaps be helpful to integrate the use of MT and Language Technology more strongly in the national digital policy. As Luxembourg is a small country with many foreigners, there is a clear need for translations of information available to the citizens and businesses.

References and links:

Digital Lëtzebuerg: https://digital.gouvernement.lu/en/dossiers.gouvernement%2Ben%2Bdossiers%2B2014%2Bdigital-letzebuerg.html.

Centre des technologies de l'information de l'Etat (CTIE): https://ctie.gouvernement.lu/en.html.

Guichet.lu: *Legal notice*, 2019, https://guichet.public.lu/en/support/aspects-legaux.html.

Luxembourgish Open Data Portal: https://data.public.lu/en/.

[CTIE, 2017] Centre des technologies de l'information de l'Etat (Press Release): *English-language version of the Citizens Portal on Guichet.lu*, 2017, https://ctie.gouvernement.lu/en/support/recherche.gouvernement%2Ben%2Bactualites%2Btoutes_actualites%2Bcommuniques%2B2017%2B11-novembre%2B07-guichet-anglais-en.html.

[ELRC, 2021] Anastasiou, Dimitra: *ELRC Workshop Report*, 2021: https://lr-coordination.eu/sites/default/files/Luxembourg/ELRC3_Workshop%20Report%20Luxembourg_Public.pdf.

[Gautier et al., 2019] Gautier, Martin, Turki: *Impacts of Open Data in Luxembourg and the Greater Region*, 2019, https://download.data.public.lu/resources/study-impacts-of-open-data-in-luxembourg-and-the-greater-region-2019/20190510-143345/impacts-of-open-data-in-luxembourg-and-the-greater-region-2019-final.pdf.

[Pundel, 2018] Pundel, Lynn: *Guichet.lu as a prime example for enabling and sustaining multilingualism*, 2018, https://ec.europa.eu/cefdigital/wiki/pages/viewpage.action?pageId=61932141&preview=/61932141/73543847/2_05_Lynn%20Pundel_Guichet.lu.pdf.

[STATEC, 2022] STATEC: *Populations by nationalities in detail*, 2022, https://statistiques.public.lu/dam-assets/fr/actualites/population/population/2022/04/stn-population-04-22.pdf.

Annex

Country Profile Malta

State of Play:

Translation practices and information exchange in ministries and public administrations:

Each Ministry caters for its own translation needs, with translations being carried out either internally (exclusively) or both internally and through outsourcing (to individual translators or translating companies). The frequency of requests for outsourced translations may vary, with demands being sometimes made on a monthly or quarterly basis. Although most translations are required from MT to EN or vice-versa, some Ministries have also pointed out other language combinations, namely: FR to EN and vice-versa; IT to EN and vice-versa; DE to EN and vice-versa.

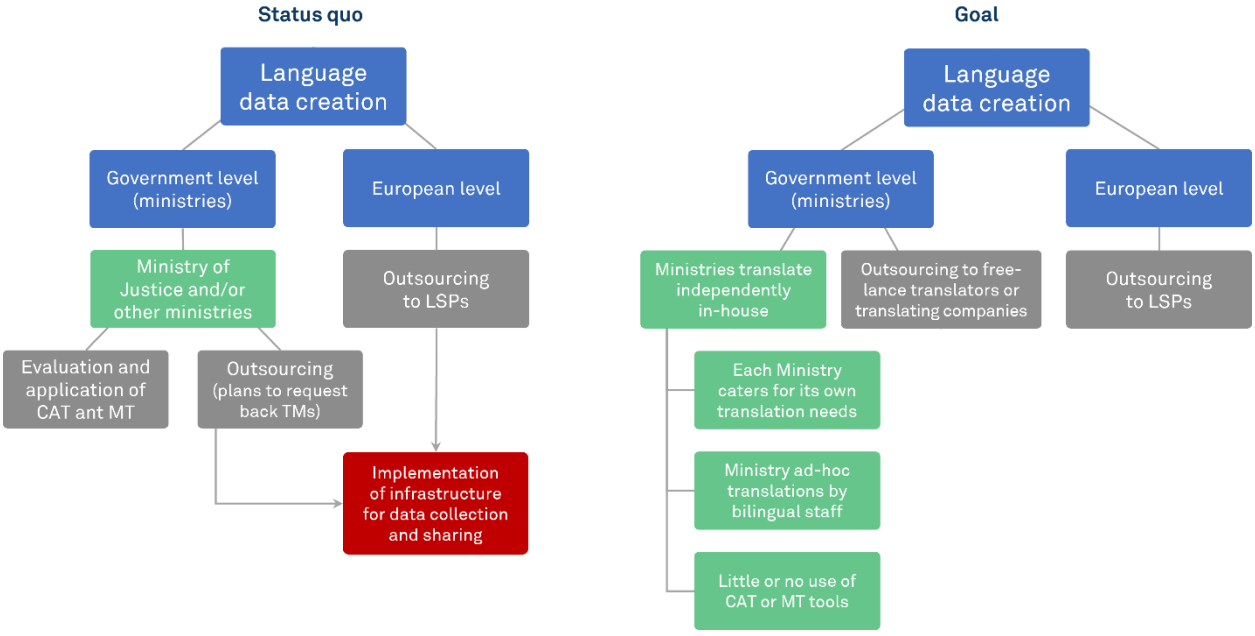
The kind of documents required to be translated also varies, with the following examples having been quoted by Ministries: legal documents, press releases, reports, speeches, calls for applications and recruitment calls, official websites, memos, promotional content, letters and forms sent to new or existing pensioners, job descriptions, parliamentary questions, notifications, and conference set-ups. No CAT Tool is used when translating in-house, although commercially available translation platforms are sometimes employed.

There is yet no government policy for sharing or pooling of translations. Instead, there is a tendency for Departments and offices to adopt the same silo mentality that applies to untranslated documents whereby once produced remain in situ. As regards pooling of translation resources, there is no contractual obligation for LSPs carrying out translations to deliver translation resources alongside the translations themselves.

Interesting fact:

Maltese is the only EU official language with Semitic roots, which however is written in the Latin alphabet.

The current language data collection and sharing infrastructure in Maltese public bodies looks as follows:



The goal is to implement an infrastructure for language data collection and sharing.

Open Data and data collection in Malta:

The Maltese Open Data Portal[71] is designed to enable a shared platform for the management and support of the Foundation Data Layer and its main Administrative Registers. It will eventually serve as the Data Governance workbench for the management of metadata, requests for data and data sharing and re-use authorisations in respect of the national official registers emanating from the Laws of Malta. The contents of the portal are currently work in progress and unless otherwise indicated should not be considered as providing official records.

Above this, the Maltese Language Resource Server (MLRS)[72] currently serves as a central repository for Maltese language resources and tools built by the UM, such as a tokeniser, a part-of-speech (POS) tagger and corpora, such as the Korpus Malti.

Ġabra[73], an open lexicon for Maltese, gathers information from several different lexical resources into one common database.

Finally, the National Language Technology Platform (NLTP)[74], which is currently under development as a CEF-financed project, will also serve as a data repository where existing translations are accumulated for the purposes of Artificial Intelligence machine learning.

Digital policy and language policy in Malta:

The delivery of AI research, development and adoption can only be achieved in a well-developed and cutting-edge data and ICT infrastructure. As highlighted in Malta's strategy, an enabling infrastructure should be part of a holistic AI strategy and contains various dimensions such as data and connectivity resources, compute capabilities and data sharing platforms for all institutional stakeholders, ranging from research institutes, regulatory authorities, and start-up ecosystems to innovation hubs. Among others, the strategy proposes to set up the following initiatives:

- Investing in Maltese language resources to foster language technology solutions;
- Supporting data centres to meet the growing needs of computing power and storage. Malta Entreprise, Malta's economic development agency, will offer incentives and support measures in this regard;
- Increasing access to open data with the launch of Malta Data Portal, an open data repository developed by the Malta Information Technology Agency (MITA);
- Providing cost-effective access to computing capacity through various initiatives, such as supercomputing cluster A.L.B.E.R.T, and Malta's participation in the European Initiative EuroHPC to develop a pan-European supercomputing infrastructure;
- Enabling access to cloud platforms for the public and private sector by means of initiatives such as Malta Hybrid Cloud procured by MITA.

The Maltese Language Council was established in April 2005 to promote a suitable language policy and strategy for the Maltese islands. It promotes the Maltese Language in Malta and in other countries by engaging actively to foster recognition and respect for the national language. It maintains regular contacts with local, national, and international organisations which have similar functions. The Council is a full member of the European Federation of National Institutions for Language (EFNIL) and works to:

- update the orthography of the Maltese Language as necessary and establish the correct manner of writing words and phrases entering from other languages;
- develop, motivate and enhance the recognition and expression of the Maltese language;
- promote the dynamic development of such linguistic characteristics as identify the Maltese;
- adopt a suitable linguistic policy backed by a strategic plan;
- establish a National Centre of the Maltese Language offering printed and audiovisual resources to members of Maltese language associations, institutions and other interested persons;

[71] https://open.data.gov.mt
[72] https://mlrs.research.um.edu.mt/
[73] https://mlrs.research.um.edu.mt/resources/gabra/
[74] https://www.nltp-info.eu/

The role of LT and language data in Malta's AI regulations:

In April 2021, three projects led by University of Malta researchers in the field of Artificial Intelligence were collectively funded by the Malta Digital Innovation Authority, as part of the undertaking of Malta's National AI Strategy. The second and third projects are being led by the Department of Artificial Intelligence within the Faculty of ICT and will upgrade existing text processing resources and tools for them to be used by any industry wanting to process Maltese text, and training people in the annotation of speech data so it can be computationally processed.

Stakeholders and major networks:

The Ministries and the Office of the Prime Minister are currently being identified as the main stakeholders, both as contributors of language resources as well as potential eTranslation and NLTP users, with interest being shown both for "individual to machine" use as well as "machine to machine" use (potential integration of translation solutions in public digital services). "Ministries" are to be understood as inclusive of the most granular level, thus covering specialised agencies, authorities and other entities falling thereunder, that encounter specific domain information and that could therefore populate the pool of resources with specialised corpora and terminologies.

The interest is tangible and very promising. The ELRC Malta seminar organised in January 2022 reached over 240 registrations and was attended by over 170 participants, from the Maltese public administration, European Institutions [Court of Justice of the European Union, European Parliament, European Commission], Local Councils, SMEs, academia and students.

Main challenges for sustainable data sharing:

- **Lack of awareness of the importance of language resources:**
  The main challenge is to raise awareness and to put the message across that the documents produced daily by the public administration constitute invaluable language resources, which, like any other resource, should not be left dormant but should be re-used, with the aim of maximising their value.
- **Rare re-use of translated data:**
  The current situation wherein a document, once translated, dies a natural death, should become a thing of the past. For eTranslation and NLTP purposes, the life of that document is just about to start and may indeed be given an eternal existence if injected into the pool of resources and used to train the engines and fine-tune translation results.
- **Lack of awareness concerning the benefits of data contribution:**
  Currently, potential data contributors are not necessarily always aware of how they can benefit from translation tools and language technology tools generally. Consequently, it is important to convey the right message: use translation tools + donate to translation systems: it's a win-win situation.
- **Weak overall support for the Maltese language with respect to applications in speech processing, machine translation, text analysis:**
  The support that currently exists is fragmented, being mainly oriented towards access to text corpora. Facilities for collecting and managing language data systematically with an eye to different application areas have yet to be put into practice.

The time is also right, since AI is currently at the forefront of Malta's national digital policy, with two documents in this regard having been published in October 2019, i.e. The Ultimate AI Launchpad, A Strategy and Vision for Artificial Intelligence in Malta 2030 and "Malta: Towards Trustworthy AI, Malta's Ethical AI Framework. Further information about the documents can be found in the references.

Action plan:

Based on the identified challenges, a number of actions could be defined:

- **To increase interest in Machine Translation in public services and ministries:**
  Based on the understanding that every public official is a potential eTranslation/NLTP user and therefore a potential contributor of data, one-to-one sessions with ministries at all levels, including more granular levels (authorities, agencies, departments, units, etc.) should be organised. During such meetings, eTranslation/NLTP will be presented and a live demo will be given on the spot.

In the past, this proved to be very effective, triggering off genuine interest and forming the basis for strong, lasting working relationships.

- **To tackle technical and legal issues:**
  Public administrations should be informed about the technical and legal support that ELRC may provide, including advice about data management generally. The plan is simple, but the message will be strong: "Data is power".
- **To update translation policy:**
  Further steps need to be taken, so that the Maltese government recognises the importance of language data. These efforts should result in an updated translation policy prescribing that any by-products of translations, e.g., translation memories and terminological equivalencies should be included amongst the deliverables of all translation contracts between third parties and the government.
- **The development of a National Language Technology Platform (NLTP)**
  A National Language Technology Platform[75] is currently under development. NLTP will be provided, *inter alia*, with an inbuilt CAT tool, a terminology database, a website translator, as well as a data repository where previous translations are accumulated for the purpose of Artificial Intelligence machine learning. The platform will also provide for the potential provision of speech technology services at a later stage. NLTP, which will be available to the Maltese public administration and the general public, therefore constitutes an effort at national level to contribute towards the current niche of language technology support for the Maltese low-resource language, across multiple computational linguistics fields.

References and links:

Digital Language Resources and Tools for the Languages of Malta: a Roadmap, http://www.kunsilltalmalti.gov.mt/file.aspx?f=309.

Draft National Data Strategy, https://mita.gov.mt/wp-content/uploads/2020/07/Data-Driven-Public-Administration-Malta.pdf.

Innovative Technology Arrangements and Services Act, https://legislation.mt/eli/cap/592/eng.

Judicial Proceedings (Use of English Language) Act, Chapter 189 of the Laws of Malta, https://legislation.mt/eli/cap/189/eng/pdf.

Malta AI Strategy, https://malta.ai/wp-content/uploads/2019/11/Malta_The_Ultimate_AI_Launchpad_vFinal.pdf.

Malta Digital Innovation Authority Act, https://legislation.mt/eli/cap/591/eng.

[AI Framework, 2019] Parliamentary Secretariat for Financial Services, Digital Economy and Innovation, Office of the Prime Minister: *Malta: Towards Trustworthy AI, Malta's Ethical AI Framework*, 2019, https://malta.ai/wp-content/uploads/2019/10/Malta_Towards_Ethical_and_Trustworthy_AI_vFINAL.pdf.

[AI Launchpad, 2019] Parliamentary Secretariat for Financial Services, Digital Economy and Innovation, Office of the Prime Minister: *Malta: The Ultimate AI Launchpad, A Strategy and Vision for AI in Malta 2030*, 2019, https://malta.ai/wp-content/uploads/2019/10/Malta_The_Ultimate_AI_Launchpad_vFinal.pdf.

[Cortis et al., 2021] K. Cortis, J. Attard, and D. Spiteri: *Malta National Language Technology Platform: A vision for enhancing Malta's official languages using Machine Translation (2021)*, https://aclanthology.org/2021.mmtlrl-1.3.

[75] https://www.nltp-info.eu/

Annex

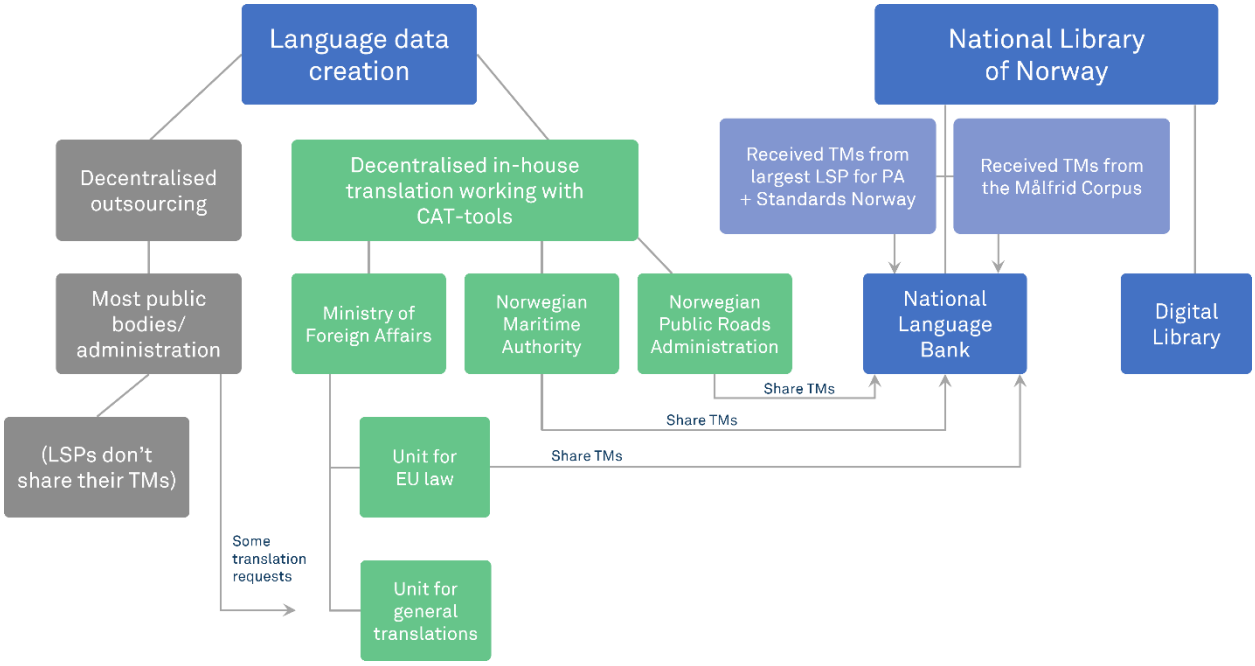
Country Profile Norway

State of Play:

Translation practices and information exchange in ministries and public administrations:

In Norway, each public administration is responsible for adapting its information to different language users. There is no central translation service or procurement contract, and no systematic exchange of translations and/or knowledge between public administrations. Translations for public administration bodies are mainly procured from commercial translation companies or done in an ad hoc manner and without use of CAT tools by the administrations' own employees. Only very few ministries and public administrations have in-house translation services generating Translation Memories. These include the Ministry for Foreign Affairs that have two translation units: one unit for the European Economic Area (EEA) and trade law and another unit for general translations, from whom the government and other ministries can also request translations. The only other public administrations that have in-house translation services are the Norwegian Maritime Authority, the Norwegian Public Roads Administration and the EFTA Secretariat in Brussels.

The current language data creation and sharing infrastructure in Norwegian public bodies looks as follows:



As indicated above, most Norwegian public administrations do not exchange Translation Memories or expertise in translation procedures with each other. When it comes to sharing textual data in order to improve eTranslation, it is important to remember that the European Commission does not translate any texts to or from Norwegian. Hence, the collection and provision of Norwegian language resources for eTranslation in the context of ELRC is not a supplementary activity, but constitutes the only source for Norwegian language data in eTranslation. Until recently, most Norwegian language resources in eTranslation were provided by the Norwegian Ministry for Foreign Affairs. These data have been supplemented with translations of anonymised complaints provided by the European Consumer Centre Norway. The Norwegian Maritime Authority, the Norwegian Public Roads Administration and the EFTA Secretariat in Brussels have also shared their TMs with the National Language Bank (Språkbanken) and ELRC.

In addition to data delivered by the above-mentioned public institutions, the National Library of Norway has concluded agreements with commercial language services providers that led to the transfer of Translation Memories derived from public contracts. This includes data from Doffin, the Norwegian web-based database for notices of public procurement and procurement in the utility sector. The agreements permit the reuse of Translation Memories and multilingual terminology lists on the condition that the memories are submitted to random scrambling in order to prevent automatic reconstitution of the translated texts.

More recently, additional language resources have been generated from parallel texts published on the web by various Norwegian public institutions. The National Library of Norway and the Language Council have created Målfrid, a service which harvests web pages from all Norwegian public institutions. From these data the Language Bank has mined parallel corpora of Bokmål and Nynorsk and of Bokmål and English. These data will be harvested on a yearly basis. In the foreseeable future, the Målfrid corpus will be the most important source of parallel texts from the Norwegian domain.

According to the Norwegian Education Act, textbooks and teacher manuals for Norwegian schools must be published in both official forms of written Norwegian. On the basis of these publications, the National Library of Norway has created a parallel corpus within the EU-funded PRINCIPLE project. The corpus is made available on ELRC Share. Another important source of parallel texts in Nynorsk and Bokmål is the Nynorsk Press Office, which has contributed a corpus of news texts translated from Norwegian Bokmål into Norwegian Nynorsk covering various subject areas.

While terminology resources have been shared through the Language Bank, the National Terminology portal, Termportalen, is planned to appear in a new version In 2023, catering mostly to research and higher education. Public and private bodies may share their terminology through APIs on the National Data Catalogue (data.norge.no).

Digital Policy and Language Policy in Norway:

Norway has two official languages, Norwegian and Sami. Norwegian exists in two written varieties, Bokmål and Nynorsk that must be treated as equal. Norwegian Bokmål is employed by the major part of the population, whereas Norwegian Nynorsk is used by approximately 12% of the population. To ensure equal treatment, both official language forms must be represented by at least 25% of publications published by state level agencies. Textbooks for primary and secondary schools must also be available in both language forms. The legal requirements for the use of Nynorsk in public administrations, however, only apply to published text. Consequently, the amount of Nynorsk in internal documents (e.g. translations) is significantly lower than for Norwegian Bokmål. Still, citizens can address public administrations in either variety and will receive an answer in the same language form they used.

> **Interesting fact:**
>
> **Language equality is applied to both written varieties of Norwegian. Norwegian Bokmål and Norwegian Nynorsk have to be treated as equal and therefore both forms must be represented in at least 25% of publications published by state level agencies.**

In 2016, a report was published stating the need for automated translation in the Norwegian public sector (cf Oslo Economics, 2016). Since then, the Norwegian Ministry of Culture, which is in charge of language policy, has decided that both official forms of written Norwegian must be available in eTranslation.

A new Language act entered into force in 2022. One of its purposes is to ensure that “public bodies take responsibility for using, developing and strengthening Bokmål and Nynorsk”, with “a special responsibility for promoting Nynorsk, as the least used written Norwegian language” (cf. Language Act, 2021, Section 1. Purpose). The Language Council of Norway, subordinated to the Ministry of Culture, is given the task to oversee the Norwegian language policies and to “provide guidance to public bodies concerning the rules in this Act.” The tasks include ensuring that language technology, including machine translation, works for both varieties of the written language. There are ongoing conversations with the Norwegian Digitalisation Agency (Digdir) on how to secure data for Nynorsk.

The National Library of Norway, also subordinated to the Ministry of Culture, hosts Språkbanken – the Norwegian Language Bank – an initiative to ensure the development of language technology solutions for the Norwegian language, thereby preventing domain loss of Norwegian in technology-dependent

areas. Språkbanken offers digital language resources to the language technology industry, to linguistic research and education, and to public administration.

Role of LT and language data in national AI regulations

The Norwegian strategy for artificial Intelligence (cf. National Strategy, 2020) outlines the need for language data in artificial intelligence as well as general principles for data sharing and privacy and data protection. The Data Protection Authority has established a regulatory "sandbox"[76] – a project environment for artificial intelligence that makes use of personal data. The Digitalisation Agency has established the Norwegian Resource centre for Sharing and Use of Data.[77] It provides, among other things, legal advice for data sharing. Neither the sandbox nor the resource centre focus on how to deal with privacy issues that are particular to language data, such as automatic anonymisation of text resources for sharing.

Stakeholders and major networks:

The Norwegian National Anchor Points represent two key institutions (Norwegian National Library and the Language Council of Norway) related to language policy and language data collection, which underlines the interest and importance of this topic in Norway. This is also exemplified by the fact that both commercial language service providers and national public administrations have contributed language data to ELRC-SHARE. The local ELRC events were attended by representatives from more than 30 institutions, thus ensuring that a significant number of national stakeholders were informed about the importance of collecting, managing and sharing language data to ensure language equality in Norway and beyond. The collection of language data is also strongly supported by the Norwegian Digitalisation Agency (Digdir).

Main challenges for sustainable data sharing:

- One of the main challenges for using and sharing language data for Norwegian Nynorsk is the fact that there is still not enough language data available at this point.
- Additionally, even less data is available for parallel texts in Norwegian Nynorsk and English as existing parallel corpora are made up almost exclusively of Norwegian Bokmål and English.
- A third challenge Norway is facing, applies to both varieties of Norwegian and results from the fact that language data "produced by public administrations, whether published online or for internal use, and more specifically the value of translations done internally or outsourced" are not considered valuable, worth managing, processing and sharing.
- "The need for awareness raising also applies to the privacy and confidentiality of documents sent out for external treatment, for example when such information is kept in the form of Translation Memories by external executives without the client being aware of this. Either the documents must be anonymised before dispatch, or the contract with external executives must ensure that Translation Memories are returned and/or deleted." (ELRC, 2019)

As the 2021 Country Workshop has shown, many of the challenges listed in the ELRC White paper from 2019 have already been addressed through the defined action points: The amount of data has increased substantially for both Bokmål and Nynorsk, there are growing corpora of parallel Bokmål-Nynorsk texts as well as translation to and from English. The National Library, the Language Council and The Norwegian Digitalisation Agency have produced a guide for how to identify translations and other language data, and to make them available for reuse.

This does not mean that there is enough data. As non-EU member, Norway has had less translation memories generated from EU documents than the EU countries, and has therefore made a substantial effort to harvest translation memories from other sources and to make them available for use in machine translation. Nevertheless, raising the awareness of what language data is, its value and its uses in language technology must be continued, as well as when it comes to area specific data.

[76] https://www.datatilsynet.no/regelverk-og-verktoy/sandkasse-for-kunstig-intelligens/
[77] https://www.digdir.no/datadeling/norwegian-resource-centre-sharing-and-use-data/2766

Action plan:

For Norway, six objectives could be defined that will help to address the identified challenges. In the order of their priority these are:

- Increase the number of bilingual resources in English – Norwegian (Bokmål or Nynorsk), including terminology
- Increase the parallel corpus in Norwegian Bokmål and Nynorsk
- Informing public bodies of their responsibilities according to the Act relating to Language
- Raise awareness of language data as Open Data and establish good management practices for language data in public services
- Increase interest in MT in public services
- Tackle legal concerns

The most important objective for Norway is to continue the harvesting of bilingual resources in both varieties of Norwegian and English in order to ensure equal treatment of both varieties of written Norwegian and to enhance the quality of translations from or into other European languages provided by eTranslation. Ongoing projects such as Målfrid, and the agreement with the Nynorsk Press Office have a direct impact on the perception of the value of language data. In addition, by opening up language data and making it freely available in the National Language Bank, the visibility of language data and its value is further supported. Next to dissemination activities, the collection and sharing of language data itself helps to raise awareness of language data as Open Data.

The language Council of Norway works closely with the National Library to ensure that the language policies can be implemented successfully. It has started conversations with the Agency for Public Management and eGovernment on how it can gather different kinds of language data automatically from the public sector and thereby create sustainable infrastructures for sharing language data and at the same time increase the interest in machine translation in public services. These activities will improve the overall management of language data, including privacy and confidentiality of documents as well.

References and links:

[Datatilsynet, 2018] Datatilsynet: *Artificial intelligence and privacy*, report, January 2018, https://www.datatilsynet.no/en/regulations-and-tools/reports-on-specific-subjects/ai-and-privacy/.

[Digital Agenda, 2016] Norwegian Ministry of Local Government and Modernisation: *Digital agenda for Norway in brief*, https://www.regjeringen.no/contentassets/07b212c03fee4d0a94234b101c5b8ef0/en-gb/pdfs/digital_agenda_for_norway_in_brief.pdf.

[ELE, 2022] ELE Report on the Norwegian Language, https://european-language-equality.eu/wp-content/uploads/2022/03/ELE___Deliverable_D1_26__Language_Report_Norwegian_.pdf.

[ELRC, 2019] Olsen, Jon Arild: *ELRC Workshop Report for Norway*, 2019, http://lr-coordination.eu/sites/default/files/Norway/ELRC%2B%20Workshop%20Report_public.pdf.

[Language Act, 2021] Ministry of Culture and Equality: *Act relating to Language*, 2021, https://lovdata.no/dokument/NLE/lov/2021-05-21-42.

[National Strategy, 2020] Ministry of Local Government and Regional Development: *The National Strategy for Artificial Intelligence*, 2020, https://www.regjeringen.no/en/dokumenter/nasjonal-strategi-for-kunstig-intelligens/id2685594/.

[Oslo Economics, 2016] Oslo Economics: Kartlegg*ing av behovet for automatisk oversettelse I statlig sektor/2016-15*, 2016, https://www.regjeringen.no/contentassets/61298b7ccab04fddb2c2b3bb9465cf38/automatisk_oversettelse_oe.pdf.

[Teknologirådet, 2018] Teknologirådet (Norwegian Board of Technology): Artificia*l Intelligence: Opportunities, Challenges and a plan for Norway*, Nov. 2018, https://teknologiradet.no/en/publication/ai-and-machine-learning-possibilites-challenges-and-a-plan-for-norway/.

Annex

Country Profile Poland

State of Play:

Translation practices and information exchange in ministries and public administrations:

Polish ministries, public institutions and state-owned enterprises (such as the Industrial Development Agency, the Polish Press Agency etc.) translate their texts either by using their internal resources or by outsourcing the services to external language service providers (LSPs), with outsourcing being the prevailing trend. The estimated volume of outsourced translations amounts to 80 to 90% of the total volume. There is no central translation authority nor office to coordinate the related activities. As a result, there is no central terminology base nor organised management of translation memories.

Due to their specific requirements (e.g. confidentiality), some ministries, such as the Ministry of Foreign Affairs or the Polish Financial Supervision Authority (KNF) have small in-house units providing translation (2 to 7 people). There are also individual in-house translators working e.g. in the departments responsible for international cooperation or implementing international projects or delegated to foreign representation offices or embassies. Part of the translations hence is delivered internally by the staff having appropriate knowledge of the language(s); however, in the majority of the cases, they are specialists/experts in other fields and therefore have to rely on external LSPs. This policy is also due to the translation volumes and the need for highly specialised (legal, medical, technical) translations and/or certification of documents by a sworn translator.

> **Interesting fact:**
>
> **The Unit of Sworn Translators and Interpreters at the Ministry of Justice deals with formal licencing of translators and interpreters by e.g. arranging appropriate examinations and keeping a register of persons who have passed the state exam. This unit is also responsible for the recognition of professional qualifications acquired by sworn translators and interpreters in other Member States.**

The CEF eTranslation tool is also used, but its use is not widespread yet due to the low awareness of its existence and its potential applications among the public sector employees. The expected human translation quality is still an important factor. Adding Ukrainian language to the eTranslation portfolio early in 2022 has increased the interest in the EC service on behalf of individual translators and NGOs, however, its potential has not been fully exploited by public services. European Interoperability Framework (EIF) recommendation on multilingual character of public services is mentioned among the interoperability principles for the Information Architecture of the State (in Polish: Architektura Informacyjna Państwa: AIP-P-09) on the governmental website[78], however, it is not mandatory. Using the eTranslation plug-in for the translation of the website was considered but given low priority at this stage. Simultaneously, the State Commission for Investigation of Aircraft Accidents (Polish acronym: PKBWL) expressed an interest in building a domain translation engine in the language pair PL>EN for aviation domain based on its parallel resources.

The ministries as well as their subordinate units outsource translation services individually, which results from both budget regulations and potential limitations concerning the organisation and coordination of joint procurement procedures. Typically, the amount of a procurement contract for translation with a ministry varies between 0.4 million PLN[79] to 0.7 million PLN with the highest ones reaching up to 1 million PLN. The contracts are usually concluded for a period of one year and with a single LSP, which does not allow for developing good practices in cooperation between the contracting authority and the LSP, including e.g. terminology management.

[78] https://www.gov.pl/web/ia/pryncypia-architektoniczne
[79] 1 PLN = 0.20841 EUR as of 19 Oct. 2022 (OANDA Currency Converter)

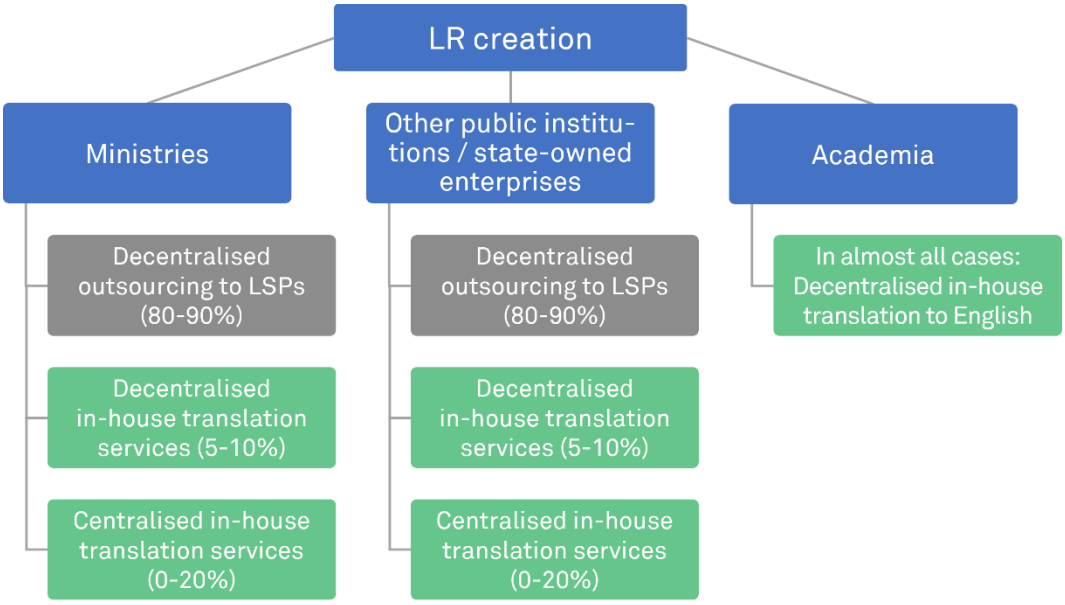
Interesting fact:

The translation contracts most frequently include both translation and interpreting services, which is often problematic, as LSPs are typically specialised in providing either translation or interpretation services.

The procurement process is often deficient as most emphasis is still put on the price criteria while not taking into account (or not taking into account to a sufficient degree) the quality criteria such as e.g. the experience of an LSP of providing translations in a particular subject area. The requirements regarding the translation industry norms, the use of computer-assisted translation (CAT) tools and the delivery of translation memories (TMs) to the contracting authority are, therefore, most frequently omitted. It should be noted here that this is not necessarily a deliberate action, but often a consequence of the limited knowledge of the language service market and language technologies (LT) on behalf of the procurement units.

Recently, however, a more positive trend can be observed, including the involvement of the Public Procurement Office (Polish acronym: UZP) and the development of corresponding procurement practices: In January 2019, the UZP published the first set of documents describing good practices in proceedings for outsourcing translation services in its Knowledge Base (Industry Practices Forum). A revision is planned with the support of the Polish Association of LSPs POLOT. As a consequence, public entities tend to use non-price criteria more often and pay more attention to the actual expertise of the language services provider and guidelines prepared by industry associations, also with regard to interpreting services (cf. PSTK, 2019). Technological awareness and the use of CAT tools also gradually improve thanks to various actions and assessments by e.g. the Ministry of Foreign Affairs, which started to use a selected CAT tool for in-house use or the Polish Air Navigation Services Agency (Polska Agencja Żeglugi Powietrznej), which intended to procure software for computer-assisted translation (cf. TED, 2019). Whilst first positive examples with regard to the above aspects can be observed, the delivery of TMs along with the translation is practically non-existent.

The current language data creation and sharing infrastructure in Polish public bodies looks as follows:



As the large language models (LLMs) have been gaining importance in the last few years, it is also the case of Poland. The ones reaching the highest results in evaluation tasks such as KLEJ Benchmark[80] are Polish RoBERTa[81], made available by the National Information Processing Institute – Public Re-

[80] https://klejbenchmark.com/
[81] https://github.com/sdadas/polish-roberta

search Institute (Polish acronym: OPI PIB)[82] supervised by the Ministry of Science and Education, HerBERT[83] and plT5[84] pre-trained on the National Corpus of Polish[85] by Allegro and Linguistic Engineering Group at the Institute of Computer Science, Polish Academy of Sciences.

Another initiative worth noting is the PolEval evaluation campaign for natural language processing tools for Polish started in 2017 to advance the state-of-the-art with a series of tasks in which submitted tools compete against one another using available data and are pre-established evaluation procedures. The contest integrated the Polish NLP community and resulted in the development, enhancement and public release of reference data sets for NLP tasks such as sentiment analysis, speech recognition or question answering.

Open Data and data collection in Poland:

The Ministry of Digital Affairs (MoDA)[86] has prepared the Public Data Opening Programme (Open Data Programme) specifying data sharing standards adopted by resolution of the Council of Ministers on 20 September 2016, No. 107/2016. The implementation of the programme is coordinated by the Minister of Digital Affairs. One of the goals of the Open Data Programme consists in building and coordinating a cooperation network, including institutional plenipotentiaries for Open Data[87].

The tasks performed by the MoDA also include the development of standards for public data opening with regard to legal, security, technical and API issues, training and workshops for administrative staff on data opening, as well as knowledge dissemination. The project Open Data Plus, which is the successor of the Public Data Opening Programme, aims to e.g. build new APIs for a number of public databases, opening an analytical central Open Data Laboratory that supports the development of relevant policies in offices and ministries. Related educational activities are carried out by the Open Data Academy.

In July 2019, the Ministry of Digital Affairs also published a corresponding "Data Opening. Good practice guide". The good practice guidelines are part of the project "Open Data – access, standard, education", which aims to increase the availability and quality of Open Data and its reuse. The guide describes the basic framework for the process of opening data by referencing relevant legal acts, identifying desired institutional settings, and presenting practical scenarios for data opening in government offices. It focuses on covering the most important legal regulations affecting the opening and usage of public data and shows inter-institutional and non-institutional cooperation models that have worked best in this context. Furthermore, the guide shows how to implement the data opening process effectively and provides guidance on the standards for data openness.

Digital policy and language policy in Poland:

Poland established a governance centre for the national AI strategy, located at the Chancellery of the Prime Minister and under the chair of the then Minister of Digital Affairs and the Council of Ministers Committee for Digital Affairs. The centre includes the Task Force on AI Policy enforcement, the Scientific Council for AI, the AI Observatory for the Labour Market, the Observatory of international AI Policy, as well as the Legal Task Force for changing regulations.

In December 2020, the Council of Ministers adopted the Polish national AI strategy (cf. AI Strategy, 2020). The document specifies short-term goals until 2023, including the development of projects adapted to Polish needs and challenges, such as machine processing of the Polish language and its translation into foreign languages.

[82] https://opi.org.pl/
[83] https://huggingface.co/allegro/herbert-large-cased
[84] https://huggingface.co/allegro/plt5-large
[85] http://nkjp.pl/
[86] Since Oct. 2020 incorporated into the structure of the Chancellery of the Prime Minister (KPRM).
[87] Plenipotentiaries are civil servants working in the Ministries, the units subordinated to ministries, the Chancellery of the Prime Minister, and Central Statistical Office appointed for permanent cooperation in the implementation of the Open Data Programme, responsible for the scope and deadlines of data provision by individual offices.

A number of tools to support this goal have been identified: The Foundation for Polish Science (FNP), the National Science Centre (NCN) and the National Centre for Research and Development (NCBiR) grants and scholarships for projects related to Polish language processing based on world-leading algorithms, elimination of legal barriers to the exploration of Polish language text corpora under copyright protection, and provision of architectures, trained models & training data sets for common use.

The progress and milestones in developing the national AI strategy were highlighted in a roadmap[88] released by the Ministry. In 2H2021 a Working Group on Artificial Intelligence (Polish acronym: GRAI), an advisory body at the Chancellery of the Prime Minister was established.[89] The Group's members recognise the importance of language data for the development of language technologies in the national language and will advocate for giving a higher priority to this topic in the Group's further works, in particular within its sub-group on data.

AI-related LT projects and initiatives

- CLARIN-PL-Biz (https://clarin.biz/) is the most-intensely funded (with the budget of over 100 million PLN) LT infrastructure offering advanced computing services and data storage with particular emphasis on the use of NLP technology in industrial research.
- DARIAH.Lab (https://lab.dariah.pl/) is another prominent DH infrastructure with an LT component, based on the biggest humanities consortium in Poland.

Following the Polish Language Act of 1999 of 7 October 1999, the Polish language is the only official language in the territory of the Republic of Poland. Generally, the provisions of the Act shall apply to the protection of the Polish language, the use of the Polish language in implementation of public tasks and the use of the Polish language in the course of trade and implementation of the provisions on the use of the scope of labour law (in the territory of the Republic of Poland). There is a corresponding Council for the Polish Language at the Presidium of the Polish Academy of Sciences. The Council's main task is to provide valuations and assessments on all matters concerning the use of the Polish language in public communication.

The issue of regional languages is regulated on the EU level by the provisions of the European Charter for Regional or Minority Languages ratified by Poland in 2009. In Poland, the Kashubian ethnolect enjoys the status of the regional language, which is regulated by the Act of 6 January 2005 on National and Ethnic Minorities and Regional Language. In some municipalities of the Kashubian region, officials are legally obliged to respond to the letters of the interested parties in the respective ethnolect, if the interested parties wish to do so. There are also repeated efforts to recognise such status for Silesian, however, unsuccessful up till present as they are perceived by the current government as related to attempts at gaining regional autonomy by the Silesians.

Major AI Networks and Collaborations

- The Polish strategy proposes various policy initiatives to encourage a culture of collaborations in AI developments. The Future Industry Platform[90], the Virtual Research Institute[91] and the GovTech[92] programme have recently been created to respond to the traditional lack of cooperation.
- In addition, an Innovation Map[93] has been established to monitor the scale and deployment of newly applied technologies in local government, scientific research centres and public administration. The data registry contains innovations based on new technologies, such as the Internet of Things (IoT) and AI. The map is a collection of good practices that can be a source of inspiration for other economic players and potentially lead to research collaborations.

[88] https://www.gov.pl/web/cyfryzacja/ai
[89] Summary of the Group's activities: https://www.gov.pl/web/ai/podsumowanie-konferencji-grai?utm_source=newsletter&utm_medium=email&utm_campaign=elrc_newsletter&utm_term=2022-10-19.
[90] https://ec.europa.eu/growth/tools-databases/dem/monitor/content/poland-%E2%80%9Cinitiative-polish-industry-40-%E2%80%93-future-industry-platform%E2%80%9D
[91] https://wib.port.org.pl/en/homepage/
[92] https://www.gov.pl/web/govtech-en
[93] https://www.gov.pl/web/cyfryzacja/mapa-innowacji

- Poland takes part in the Global Partnership on AI (GPAI)[94], an international initiative to spur a responsible development and use of AI in full respect of human rights, inclusion, diversity, innovation and economic growth.
- Poland is represented in the High Level Expert Group on AI for EU (AI HLEG)[95] acting as a steering committee for European AI Alliance – a multi-stakeholder forum engaged in a broad and open discussion of all aspects of AI development and its impact on the economy and society (European AI Alliance Platform)

Stakeholders and major networks:

Within ELRC, more than 80 potential stakeholders that are involved in the creation or sharing of language resources (LR), related activities and/or policy setting were identified, including in particular LR holders and creators (public bodies as well as language service providers). Thirty of these stakeholders participated in the last ELRC Workshop. The actual four pillars of Polish NLP are:

1) **Research institutions:**
   a) technical and non-technical universities (e.g. University of Warsaw, Warsaw University of Technology, Wrocław University of Science and Technology, Technical University of Gdańsk)
   b) institutes of the Polish Academy of Sciences (e.g. Institute of Computer Science PAS, Poznan Supercomputer and Networking Center)
   c) research infrastructures (CLARIN-PL, DARIAH-PL)
2) **Industrial players:**
   a) global companies (e.g. Samsung, IBM, Amazon)
   b) big local companies (e.g. Allegro, Summa Language Technologies)
   c) mid-size local companies and start-ups (e.g. Applica, VoiceLab, SamuraiLabs)
3) **Governmental and publicly funded institutions:**
   a) ministries (e.g. Ministry of Digital Affairs)
   b) National Research Institutes (e.g. NASK, National Information Processing Institute)
4) **Skilled individuals, i.e. NLP and data science enthusiasts.**

In addition to certified and specialised translators (via LSPs/translation agencies), major providers of language resources in Poland include, for instance, the Chancellery of the Prime Minister, Ministry of Foreign Affairs, the Ministry of Justice, the Ministry of Culture and National Heritage, the Polish Press Agency, the Ministry of Entrepreneurship and Technology, the Ministry of Health, and the Polish National Bank. These institutions could also be considered as the main beneficiaries of the use of eTranslation.

Main challenges for sustainable data sharing:

ELRC identified several challenges and issues that need to be overcome to enable sustainable data sharing, namely:

- General lack of transparency among public institutions related to opening their language data for further reuse.
- The lack of a public-private sector initiative for developing guidelines and good practices for contracting translation services in the public sector to be approved by the Public Procurement Office.
- The shortcomings of the procurement process (in particular the lack of the requirement to deliver the full rights and TMs to the contracting authorities).
- Individual procurement of translation services by each public entity, usually for a short period of time (one year).
- The lack of awareness-raising actions on the purpose of ELRC in relevant public sector events and focused meetings.
- The lack of technological expertise in the public sector especially with regard to the use of CAT tools. As a result, sharing of language resources by public entities is limited to non-TMX formats in almost all cases.

[94] https://gpai.ai/
[95] https://futurium.ec.europa.eu/en/european-ai-alliance/pages/about

- Technical problems concerning data delivery and the public institutions' lack of awareness on how ELRC can help to solve such issues.

Action plan:

Based on the identified challenges, the following objectives could be defined for Poland. In the order of their priority, these are:

- **Raising awareness of language data as Open Data and a valuable asset. This is to be achieved with the help of the following actions:**
  - Raising the interest in language data among the members of the relevant public bodies and sharing benefits of sharing language data. This is an ongoing activity of the Polish PS NAP (presentations and publications). In addition, ELRC Workshops should be conducted on a yearly basis(targeting also respective IT and procurement personnel).
  - More OSA (on-site assistance) cases funded under successive tenders as well as greater support for the promotional work and communication activities of the NAPs.
  - Reaching out to the MoDA and integrating language data in the national Open Data policy, digital agenda etc.
  - Proposing to the MoDA to appoint a plenipotentiary for open language data and an eTranslation /Language Data Space national contact point at the ministry.
  - Establishing practical guidelines for LRs as Open Data. In this context, the Ministry of Digital Affairs has established a corresponding Open Data Programme and Good practice guide.
  - Establishing cooperation with GRAI (Working Group on AI)
  - Emphasising the role of digital texts in the digital economy (data as the main source for Artificial Intelligence), also illustrating the value and application of Natural Language Processing tools (e.g. for preventing online violence, fake news/disinformation).
  - Emphasising the role of digital texts as part of national cultural heritage.
- **Increasing interest in MT/LT in public services as part of the national digital policy by:**
  - Establishing synergies with national projects/initiatives
  - Promoting the eTranslation API and Connecting Europe Language Tools
  - Providing best practices and use cases of MT applications in public administrations in other EU countries, particularly in trans-border DSIs
  - Educating about LT and CAT tools and their benefits
  - Co-organising Info Days of LT projects financed from European funds
  - Organising regular communication activities (e.g. a newsletter in local language)
  - Educating NAPs (enabling them to participate in related events of educational value, such as META FORUM or other EU-funded events related to LT-
- **Cooperating with the National CEF Contact Point at the MoDA**
- **Tackling legal concerns by:**
  - Developing and sharing easy-to-apply guidelines for IPR and privacy issues
  - Investigating an idea to implement rights management along with data management (in collaboration with the MoDA)
  - Providing clear anonymisation guidelines and advocating for the use of Connecting Europe Language Tools
  - Informing and involving the Personal Data Protection Office in the above actions
  - Including presentations on IPR and privacy issues in the agenda of national Translating Europe Workshops and other events for translators
- **Identifying and gaining access to outsourced translations by:**
  - Establishing cooperation between translators/translation units and public procurement units to include relevant clauses for retaining TM and corresponding rights
  - Promoting clear licencing guidelines
  - Promoting clear recommendations on procurement of translation services by public sector bodies in cooperation with the Public Procurement Office.
  - Obtaining financial support (e.g. more OSA cases or Generic Services projects in this area as well as financial support of communication activities such as writing articles or newsletters)

  - Establishing cooperation with local government (voivodeship) bodies as language data owner
- **Establishing good data management practices in public services by:**
  - Extending the scope of available public Open Data categories to language data
  - Involving the plenipotentiaries for open data in activities supporting their sharing of language data on behalf of their mother institutions
  - Identification of data managers
  - Investigation of data management practices
  - Establishing Data Management Plans based on ELRC findings

References and links:

Act on National and Ethnic Minorities and Regional Language, 2005, https://isap.sejm.gov.pl/isap.nsf/DocDetails.xsp?id=WDU20050170141.

Complete list of plenipotentiaries: https://www.gov.pl/web/cyfryzacja/pelnomicnicy-ds-otwartosci-danych.

Complete list of state-owned enterprises: https://nadzor.kprm.gov.pl/spolki-z-udzialem-skarbu-panstwa.

HackYeah: https://hackyeah.pl/.

Information Campaign to raise awareness on public Open Data: https://www.wirtualnemedia.pl/artykul/resort-cyfryzacji-chce-popularyzowac-w-zakresie-otwartych-danych-publicznych.

Open Data Practice Guide: https://dane.gov.pl/media/ckeditor/2019/07/04/open-data-good-practice-guide.pdf.

Open Data Programme, Ministry of Digital Affairs: https://www.dane.gov.pl/media/resources/20171201/Program-EN.doc.

Polish Association of Language Service Providers: http://www.polot.org.pl/.

Polish Language Act of 1999: https://isap.sejm.gov.pl/isap.nsf/DocDetails.xsp?id=WDU19990900999.

Polish Open Data Portal: https://dane.gov.pl/.

Polish Personal Data Protection Office: https://www.uodo.gov.pl/

Project “Open Data Plus”: https://www.gov.pl/web/cyfryzacja/otwarte-dane-plus.

Project “Open Data – access, standard, education”: https://www.gov.pl/web/cyfryzacja/otwarte-dane-dos-tep-standard-edukacja2.

Regulations of the Polish Language Council: http://www.rjp.pan.pl/index.php?option=com_content&view=article&id=194&catid=40&Itemid=73.

[AI Strategy, 2020] Committee of the Council of Ministers for Digitisation: *Policy for the Development of Artificial Intelligence in Poland from 2020*, https://wp.oecd.ai/app/uploads/2021/12/Poland_Policy_for_Artificial_Intelligence_Development_in_Poland_from_2020_2020.pdf.

[PSTK, 2019] Leaflet by Polskie Stowarzyszenie Tłumaczy Konferencyjnych (PSTK), 2019, http://pstk.org.pl/wp-content/uploads/2019/01/PSTK-Ulotka-A4.pdf.

[TED, 2019] Tenders Electronic Daily (TED), Contract Notice: *Poland-Warsaw: Software package and information systems*, 2019, https://ted.europa.eu/TED/notice/udl?uri=TED:NOTICE:496078-2019:TEXT:EN:HTML&src=0.

[Wołoszyk] Wojciech Wołoszyk: *First set of documents describing good practices in proceedings for outsourcing translation services* (published by UZP, prepared by Employers of Pomerania), https://www.uzp.gov.pl/baza-wiedzy/dobre-praktyki/forum-praktyk-branzowych/uslugi-biznesowe-prawnicze,-marketingowe,-konsultingowe,-rekrutacji,-drukowania-i-zabezpieczania/dobre-praktyki-w-postepowaniach-na-uslugi-tlumaczen-pisemnych-pracodawcy-pomorza.

Annex

Country Profile Portugal

State of Play:

Translation practices and information exchange in ministries and public administrations:

In Portugal, most translations are outsourced independently by public administrations, but there are also institutions with small in-house translation services. Given the specific areas of activity in each institution, in many cases, public authorities tend to hire the same companies or freelancers in order to maintain consistency across translations. The efforts of recent governments to reduce public spending are reflected in the reduced number of civil servants and increasing difficulties in hiring new staff. Consequently, it is often easier for public administrations to outsource services than to set up a permanent team of translators. When the documents are translated in-house, computer-assisted translation (CAT) tools are rarely used in public administrations. On the contrary, the use of CAT tools is widespread among language service providers (LSPs) and freelance translators. As regards language technology (LT), there is still no or weak language technology support for Portuguese as stated in the META-NET White papers series (Branco et al., 2012).

Some public administrations are using free translation services that are available online, but translation practices vary from one institution to another. In the Parliament for example, documents are frequently translated using computer-assisted translation software. In addition, terminological databases including Portuguese, English and French terms are created (ELRC, 2016). According to representatives of the Parliament and the Ministry of Foreign Affairs, their institutions frequently use the European Commission's machine translation (MT) system eTranslation, instead of other online translation services. The Portuguese Institute of Registration and Notary Affairs, on the contrary, decided to have their texts translated by a private MT system obtained by the institution. The output translations are revised and post-edited by the employees afterwards.

Due to the above-mentioned restrictions in hiring public servants, the reduction of translation costs is not always a strong argument for those managing translations in the public services. Political changes, staff instability and other priorities in terms of digital transformation make it increasingly difficult to establish a national strategy for procuring and outsourcing translations.

In Portugal, public procurement data is available on the BASE Portal. The portal gathers all relevant information on public procurement in Portugal and makes it available to citizens in an open and transparent way. Further information is available on www.base.gov.pt/Base/en/Homepage.

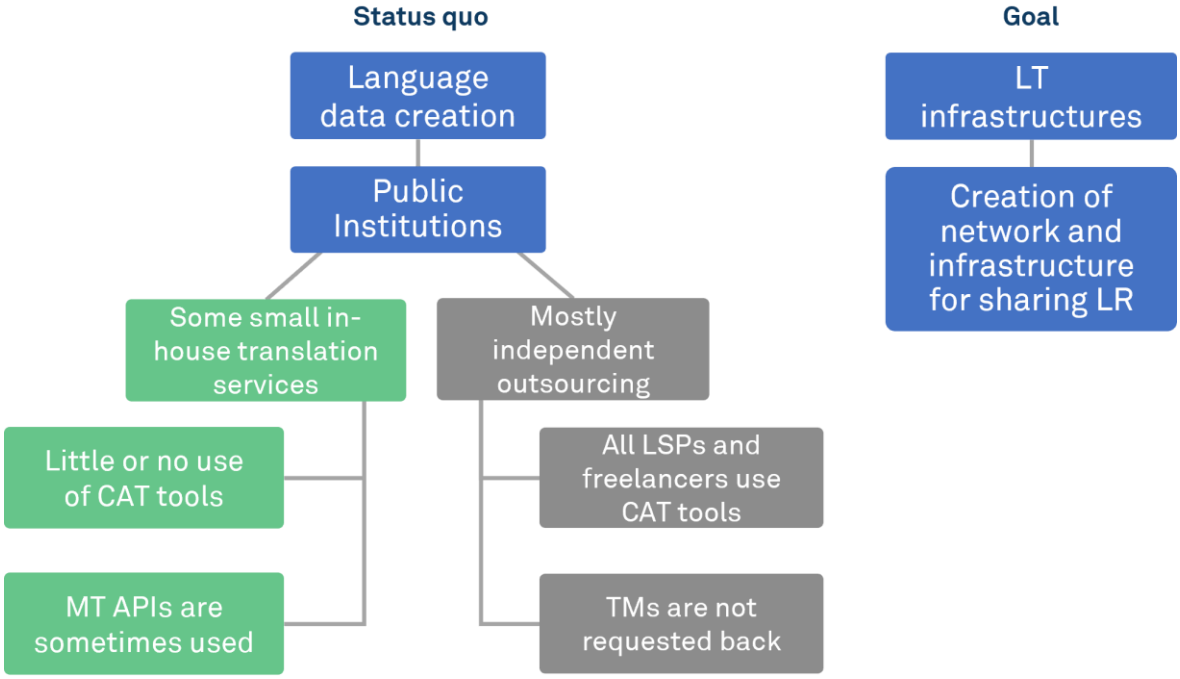
> **Interesting fact:**
>
> **The Portuguese public procurement portal is called "BASE" and collects information about all contracts concluded under the Public Contracts Code (PCC).**

When it comes to outsourcing translations, according to the NEC TM Country Reports, there was a significant increase of awarded contracts for translation and interpretation services from 2015 (63 awarded contracts) to 2018 (110 awarded contracts) in the public sector. This could be an indicator of the increasing importance of multilingual content in Portuguese public administrations. The report also states that the Portuguese market is very fragmented and that contracts are awarded to Portuguese companies, but also to independent freelance translators. In the majority of the cases, the documents' source language was Portuguese, which had to be translated into English, followed by Spanish. The areas for which most translation services were requested include Education and Communication, Social Questions, but also Finance and International Relations.

In Portugal, there is currently a low utilisation rate of translation support technology, resulting in the perception that there is no obvious benefit in adding a clause to new contracts, requiring the delivery of translation memories along with the translated work. The almost non-existent use of CAT tools also contributes to the fact that even within each organisation, there is no common practice for sharing

resources. Translated work is considered relevant only for the purpose for which it was produced, but the potential benefits of its reuse are currently not taken into account.

The current language data creation and sharing infrastructure in Portuguese public bodies looks as follows:



Open Data and data collection in Portugal:

The Portuguese Republic shares reusable data under the 2013/37U Public Sector Information (PSI) directive, which has been transposed into the Portuguese legislation by the so-called "Lei de Acesso aos Documentos Administrativos" (LADA, Lei n.º 26/2016 de 22 de Agosto) in 2016.

In Portugal, there is a GOV data portal. This portal has existed since 2011 and has been revised three years ago to be better prepared for the challenges in terms of data, both in volume and typology, and in the management of the open data community and ecosystem in Portugal. It is recognised as the official portal for Portuguese public administration, as a repository of open data, and can gather information from other open data portals that already exist in Portugal, namely sectoral portals, such as justice, health and environment, and more local portals, such as municipal councils. Lisbon Municipal Council is a successful example of what has been done in this domain. The GOV data portal aims to gather information on open data in Portugal, not only at the Portuguese public administration, but it also has the ambition to become a portal that allows sharing data from the private sector as well.

When it comes to sharing Open Data, the reuse licence is probably one of the essential conditions. The licence which is used by default and which is of recommended use at the dados.gov portal is the Creative Commons Attribution 4.0 – CC BY 4.0. In some public administrations, such as the Lisbon municipality, the Creative Commons Zero (CC 0), which is the most permissive CC licence, is used. Data sets with a CC 0 licence can be used as if they were public domain.

With respect to Language Technology (LT) and ELRC, AMA aims to support citizens by facilitating translations and by setting up a national repository of digital services available for all services of the public administration. As an outcome of the ELRI project, AMA made available the National Repository for Translation Resources, known as eTradução, where language resources are collected, prepared and shared between public institutions and research centres. This web platform has been active since May 6, 2019.

eTradução enables the collection, processing and sharing of language resources, namely translations to and from the Portuguese language that can be used to improve machine translation services. eTradução is an online platform with restricted access (registration required), and complements ELRC-SHARE repository and offers its own features, among which:

- Web-based platform
- File upload for authenticated users and accepted formats only
- Downloading resources according to authorisation level
- Resource search including filtering criteria
- Access to the upload history.

The steps for resource sharing process are:

- Uploading of resources (language pair can be uploaded simultaneously)
- Sharing conditions:
  - Only eat the institution itself + AMA (level 1)
  - With other national institutions (level 2)
  - With European institutions (level 3)
  - Public (level 4)

The resources are automatically processed for alignment, formatting, language, cleaning, and conversion to TMX format. The resources are available for download once they have been processed. Then the resource can be used with any machine translation tool (Trados, MemoQ, eTranslation).

The eTradução repository is a safe place for storing translated resources. It can be used in machine translation of digital services – making some digital public services multilingual and contributing to the development of Information and Communication Technology tools. With this repository, security, confidentiality, and compliance with Intellectual Property Rights are ensured. Finally, it is a great tool to promote the Portuguese language: every document shared will help to support the presence of the Portuguese language on the European scene, preventing its digital extinction, and it will improve the quality of machine translation services for everyone working in the Portuguese public administration.

Portugal is represented in 18 EU-funded projects, which cover the three building blocks eID and eSignature, eInvoicing and eTranslation. Three projects are related to Open Data, i.e. the Open Waste Compliance, the Cross-Nature project and the Urban Co-creation Data Lab project. The latter started in October 2019 and aims to develop a new generation of public services in the context of smart cities exploiting supercomputing facilities and public and private data to analyse complex combinations of large data sets in areas of public interest. In addition, the country is actively involved in two initiatives of the eTranslation building block, since AMA and Lisbon University are not only part of the ELRI consortium, but also represented by the National Anchor Points for Portugal in ELRC (ELRC, 2018).

Digital policy and language policy in Portugal:

Portuguese is the official language throughout the country and also the official language in nine other countries globally. With 10 million speakers in Portugal and around 220 million speakers in total, it is the third most spoken European language in the world (Branco et al., 2012).

In 2017, the Portuguese government approved the ICT2020 Strategy, which is also known as the Portuguese Digital Transformation Strategy. It aims to facilitate the cooperation between public administrations and focuses on the creation of new eGovernment services and the reduction of public sector costs.

The strategy is built on three main pillars (cf. HIllenius, 2017):

- Promotion of integration and interoperability;
- Innovation and competitiveness;
- Resource sharing and investment in digital competences.

In 2018, Portugal also launched two policy initiatives on digital competences and digitisation of the economy. One of them is INCoDe.2030, the National Initiative on Digital Competences, which aims to enhance and foster digital competences by educating young people and requalifying available human resources. In the same year, the so-called “Indústria 4.0” was launched, which focuses on the development of industry in the digital area (cf. EC, 2017).

Portugal has made significant progress over the past years in the field of digital public services, but there is still room for improvement when it comes to Open Data (EC, 2018, p. 12) and the availability of language resources. There are already several digital services available like automatic tax declaration

or electronic authentication through the public administration web portal, but the provision of multilingual digital services is not a common practice yet.[96]

Recently, on 10 September 2021, the Resolution of the Council of Ministers No. 129/2021, approved the procedure for the coordination of the Digital Transition initiatives of the Public Administration integrated in the Recovery and Resilience Plan.

Under this Recovery and Resilience Plan, Portugal has defined a set of reforms and investments around three structuring dimensions: Resilience, Climate Transition and Digital Transition. Regarding the Digital Transition dimension, the instrument aims to overcome constraints and accelerate the digital empowerment of people, the digital transformation of the businesses and the digitalisation of the State.

On the other hand, the Resolution of the Council of Ministers No. 55/2020 of 31 July, which approves the Strategy for Innovation and Modernisation of the State and Public Administration 2020-2023, establishes three strategic objectives under the "Exploiting Technology" axis: i) strengthen the global governance of technologies; ii) improve interoperability and integration of services; and iii) manage the data ecosystem with security and transparency. The main challenge of this axis is to use digital technology to provide citizens and businesses with secure, accessible and effortless services, facilitating and reducing interactions, making available and reusing data and promoting the efficiency, sustainability and simplification of the Public Administration's operating processes.

In turn, the Action Plan for Digital Transition, approved by the Resolution of the Council of Ministers No. 30/2020, of 21 April, highlights, under the "Digitalisation of the State" pillar, digital public services; an agile and open central administration and a connected and open regional and local administration, with particular focus on developing and expanding the supply of public services available online and promoting the simplification and efficiency of internal processes of the State as a whole, encompassing not only the Central Administration, but also the local and regional authorities.

With the approval by the Council of Ministers of the coordination procedure for Public Administration Digital Transition initiatives integrated in the Recovery and Resilience Plan, it is determined that all investments with an impact on the Digital Transition of the Public Administration, must meet the principles of digital government contained in the common model for the design and development of digital services, published on tic.gov.pt. One of these principles determines the availability of services and content in at least the Portuguese and English languages. This resolution may significantly increase the services available in bilingual format.

Stakeholders and major networks:

The ELRC National Anchor Points for Portugal represent two relevant stakeholders, namely the Administrative Modernisation Agency (AMA) and the University of Lisbon. Another important stakeholder is the Parliament, since it has already contributed a significant amount of language data to ELRC.

Related to language technology and translation, the key network is PORTULAN CLARIN Research Infrastructure for the Science and Technology of Language:[97] belonging to the Portuguese National Roadmap of Research Infrastructures of Strategic Relevance, and part of the European research infrastructure CLARIN ERIC, it encompasses over twenty organisations and centres related to research innovation and the Promotion of the Portuguese language. Its mission is to support researchers, innovators, citizen scientists, students, language professionals and users in general whose activities resort to research results from the Science and Technology of Language by means of the distribution of language data sets and other scientific resources, the supplying of technological support, the provision of consultancy, and the fostering of scientific dissemination.

ELRC events like the local workshops and the annual conferences were attended by more than 20 different institutions from Portugal, which clearly demonstrates that there is an interest in the topics dealt with by ELRC.

[96] https://files.dre.pt/1s/2021/09/17700/0000200005.pdf
[97] https://portulanclarin.net

Main challenges for sustainable data sharing:

The current situation in Portugal regarding the practices for multilingual data creation and sharing has changed significantly with the availability of the platform eTradução, established and run by AMA the national Agency for Administrative Modernisation. However, challenges remain.

- **Lack of awareness and information flow:**
  According to a representative from the Ministry of Justice, a compelling message on eTranslation needs to be conveyed to be able to convince public administrations to share their data and to use the eTranslation services. Currently, potential mutualisation is hindered by the lack of information flow among the entities that would have the required competences. There is generally a lot of willingness to participate, but impediments remain.
- **Lack of professionals:**
  Whereas Open Data initiatives are usually perceived positively, the biggest barrier that prevents Portuguese administrations from sharing their data is the lack of skilled staff (i.e. computer scientists), who are capable of identifying, preparing and uploading data on the CEF platform (ELRC, 2018, p.7).
- **Financial issues:**
  Preparing the data for eTranslation does not only require skilled personnel, but also financial resources, which hinders Portuguese institutions from sharing their data.
- **Lack of available resources to train MT systems:**
  More language data will be required to achieve high-quality machine translation output. At the same time, dissatisfying translation results may lead to growing scepticism among Portuguese public institutions.
- **Political changes:**
  Political changes can complicate the establishment of a national strategy for procuring and outsourcing translations.
- **Legal concerns:**
  Data protection is often seen as an obstacle to data sharing, although it is not always true. For instance, some open data that do not contain personal data nor sensitive and confidential information, can be published under specific licences created for this purpose. On the other hand, many Portuguese public administrations are legitimately concerned about the protection of personal information and fear that it might still be possible to identify people even after anonymisation.

Action plan:

In order to tackle the identified challenges mentioned above, this major objective could be defined:

- **To raise awareness of language data as Open Data and valuable asset:**
  As mentioned above, public administrations and ministries are currently not aware of the value of language data. Therefore, it would be important to emphasise the role of digital texts in the digital economy and to clearly illustrate the benefits of sharing and reusing language resources.

References and links:

Administrative Modernisation Agency (AMA): https://www.ama.gov.pt/web/english.

eTradução: https://etraducao.gov.pt.

Guides to Implementation of the (Revised) PSI Directive in Portal: https://data.europa.eu/sites/default/files/report/country_portugal.pdf.

ICT strategy 2020: https://tic.gov.pt/pt/web/tic/-/estrategia-tic-2020.

INCoDe.2030: https://www.incode2030.gov.pt/en/incode2030.

Portuguese Open Data Portal "Dados.gov": https://dados.gov.pt/en/dashboard/.

[Branco et al., 2012] Branco et al.: *The Portuguese Language in the Digital Age*. In: META-NET White Paper Series, 2012, http://www.meta-net.eu/whitepapers/e-book/portuguese.pdf.

[EC, 2018] European Commission: *Digital Economy and Society Index (DESI) 2018 Country Report Portugal*, 2018, http://ec.europa.eu/information_society/newsroom/image/document/2018-20/pt-desi_2018-country_profile_eng_B440E073-A50F-CF68-82F6A8FB53D31DE5_52232.pdf.

[EC, 2017] European Commission: *Digital Transformation Monitor, Country: Portugal "Indústria 4.0"*, 2017, https://ec.europa.eu/growth/tools-databases/dem/monitor/sites/default/files/DTM_Ind%C3%BAstria%204.pdf.

[eGovernment, 2018] European Commission: *eGovernment in Portugal*, 2018, https://joinup.ec.europa.eu/sites/default/files/inline-files/eGovernment_in_Portugal_2018_0.pdf.

[ELRC, 2018] Branco, Oliveira: *ELRC Workshop Report for Portugal*, 2018, http://www.lr-coordination.eu/sites/default/files/Portugal/ELRC%2B2%20Workshop_Public_Portugal.pdf.

[ELRC, 2016] Querido, Carvalho: *ELRC Workshop Report Portugal*, 2016, http://www.lr-coordination.eu/sites/default/files/Portugal/WorkshopELRCPortugal_PublicReport.pdf.

[Hillenius, 2017] Hillenius, Gijs: *Modernisation the focus of new ICT strategy Portugal*, 2017, https://joinup.ec.europa.eu/collection/egovernment/news/modernisation-focus-ne.

Annex

Country Profile Romania

State of Play:

Translation practices and information exchange in ministries and public administrations:

In Romania, there are different scenarios how translation needs are met in the public administration. In general, translations are regarded as a secondary activity and most translations are outsourced when a need arises. The Ministry of Culture, the Superior Council of Magistracy and the Ministry of Justice for example outsource most of the needed translation to language service providers.

When other public administrations need translations, they are either done in-house by whoever knows a foreign language and without CAT tools, or they are outsourced. Only very few institutions in Romania have dedicated translation departments or employees whose main task is to translate, these institutions are the European Institute of Romania, the National Bank of Romania, the Constitutional Court and the Romanian Standards Association (ASRO). The most common language combination for in-house translations is Romanian <> English and sometimes Romanian <> French (French is mainly used for the legal field). Although the EIR operates under the coordination of the Ministry of Foreign Affairs, translations in the ministry itself are mainly outsourced because of the volume and the variety of languages needed. At the ministerial level, there is no administration with their own in-house translation service.

Some professional translators are still sceptical about the quality of machine translations. Their opinions are propagated upstream to the decision makers and thus lower the interest for this approach.

Public procurement and the use of CAT tools:

Public institutions that outsource translations were obliged by law to search the electronic public procurement system first (SICAP[98]). If public institutions could prove that they could not find a suitable offer or if the offers were more expensive on SICAP than on the open market, they were free to choose any LSP on the market. However, this is now no longer an obligation but a recommendation.

A regulation that is still in effect is a fix price for the translation of one page in the legal field. According to Order no 2907/2020, the Ministry of Justice and other institutions operating in the legal field pay a certified translator with 44,82 lei per page (that is approx. 8,97 euro) – at this price, the translation is usually accepted by inexperienced language service providers and the result usually lacks in quality. Even when the translator can provide a certificate for translating legal texts, this is not necessarily a seal for quality because obtaining certificates for translating legal texts is a formality and does not require specific legal training. The certificates can be obtained upon request to the Ministry of Justice by any person who graduated from any language university regardless if the respective student has followed any course for translation of legal texts. In addition, by taking an exam organised by the Ministry of Culture, a person without a university diploma in foreign languages may obtain such a certificate. This causes a serious issue when it comes to the translation quality of legal texts.

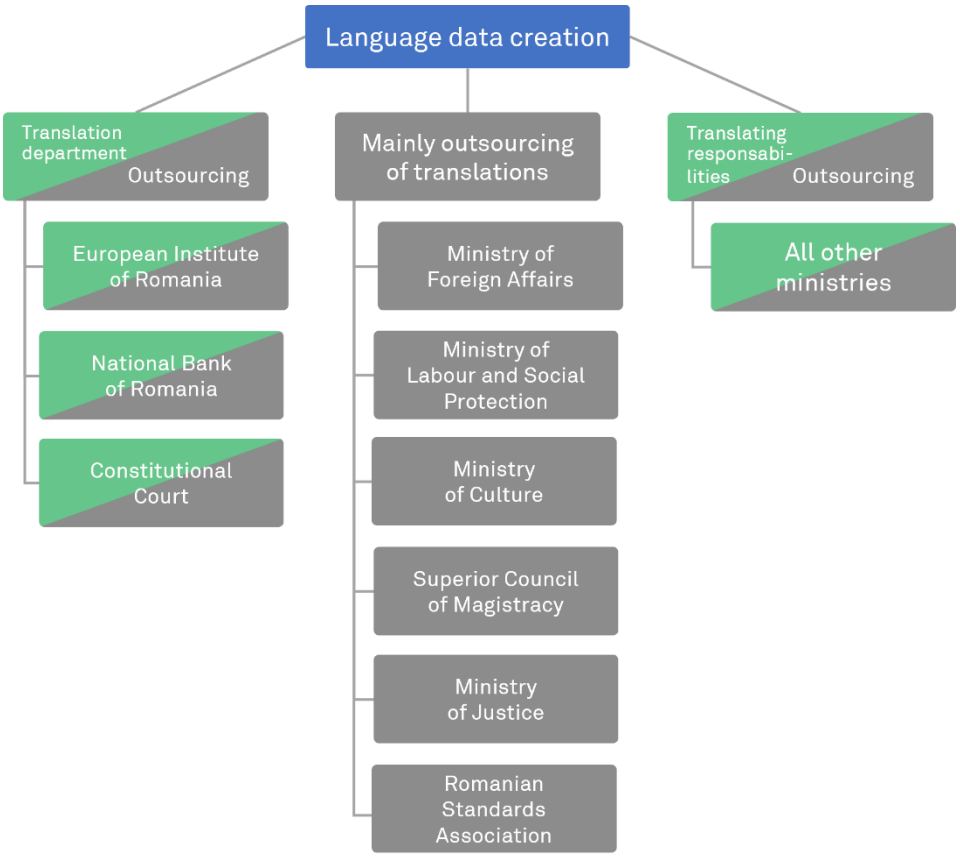
> **Interesting fact:**
>
> **It is not yet standard practice for language service providers to use CAT tools. Usually, only LSPs which work for foreign clients invest in CAT tools following suit to clients' requests.**

Another challenge arises from the fact that the majority of those who request translations (public institutions or not) usually do not request additional services like revision/review, terminology lists, glossaries or translation memories. The awarding of contracts is based on the number of pages, pair of languages, sometimes the domain, the deadline and certificates of the translators. CAT tools are considered to be very expensive and sometimes too complex. Therefore, only very few LSPs can afford

[98] https://www.licitatii-seap.ro

them. LSPs that work with European institutions or foreign clients however, usually do use CAT tools. CAT tools are also used by the translators in the in-house translation services mentioned above, whereas only the European Institute of Romania uses a server-based translation memory system.

The current language data creation and sharing infrastructure in Romania public bodies looks as follows:



In Romania, there is no proactive exchange of translation memories, terminology or expertise on the national or inter-ministerial level. During the translation of the acquis, a terminology network was put in place and although it was considered very useful, it proved to be difficult to manage for a number of reasons. The terminology experts in the various ministries were not paid for this activity and therefore could only allocate limited time to this task. In addition, a high fluctuation of human resources in the ministries and their constant reorganising led to information loss. Another initiative that was started by the Romanian Language Department in the DGT was the RO+Network, a Linguistic Network of Excellence for Institutional Romanian that allowed for information exchange between the DGT and experts of the Romanian language regarding linguistic or terminological questions. The experts provided advice pro bono but the network is no longer active.

Currently, there are no concrete plans for the organisation of a terminology network on the national level.

Open Data and data collection in Romania:

Most people in public institutions do not consider or do not know that language resources can also be Open Data. Hence there is no bi- or multilingual data sets in TMX format provided by public institutions on the national Open Data Portal[99]. Language data is also not sought after as only numerical data is considered useful, especially for decision makers.

[99] http://data.gov.ro/

Digital policy and language policy in Romania:

Romania's official language is Romanian, which belongs to the group of Romance languages. It is spoken by about 25 million people in Romania and abroad. The most spoken minority languages are Hungarian, German and Romany although 20 languages have minority status in Romania. Education is also provided in the languages of the minorities and learning foreign languages is included in the compulsory school curriculum. According to Law no 500/2004, the Romanian language is to be used in all official documents. The law does not address the use or planned use of language technologies to protect and support the Romanian language in the digital age. Some other provisions state that any technical manual or instructions regarding the use of a foreign product must also be translated into Romanian and that all TV productions in a foreign language must be subtitled into Romanian.

Stakeholders and major networks:

Although the use of machine translation and computer aided translation tools are not a common practice in Romanian public administration and in the public sector yet, an increased interest in past ELRC events was shown.

The second ELRC Workshop was attended by almost 100 participants representing various ministries and language service providers. In 2021, the third ELRC Workshop was held online due to the pandemic. Out of almost 100 participants who registered, 72 actually took part, showing that the interest in language data and tools remains strong among the Romanian public and private sectors, including administrations and services, the national radio, universities, and research centres, SMEs, and freelance translators.

Among the data donors of multilingual language data to ELRC are the Romanian Parliament, the European Institute of Romania, and RACAI. The research sector in Romania is instrumental in producing and sharing the language data and tools that result from individual institutional or European infrastructure initiatives such as META-SHARE, European Language Grid, ELRC-SHARE, LLOD Cloud.

The European Institute of Romania, and RACAI are the two institutions represented by the ELRC National Anchor Points and are critical institutions for language data collection in Romania.

Main challenges for sustainable data sharing:

- Legislation regarding the use of the Romanian language is not observed by all public institutions (according to Law no 500/2004, public institutions have to use the Romanian language in all official documents, that is with special characters (ă, â, î, ș, ț) and observing the directives of the Romanian Academy ("sunt/sînt", writing with "â/î ").
- There is a general tendency to disregard the quality of the Romanian language for a number of reasons (e.g. the message is considered more important, lack of time, low speed when typing with Romanian special characters), which also affects the quality of the Romanian language on the internet. This low quality also affects the collection of language resources in various ways. Ignoring the official orthography, i.e. not using diacritics, or using non-standard ones (or combining these practices) can result in a different meaning which turns textual documents into low-quality data that is not accurately representing the Romanian language and is thus avoided by data collectors.
- **Educational issue:** Poor use of CAT tools (they are considered to be very expensive and more of a luxury) and of computers (still not a norm to use Spell check, Track Changes, advanced text formatting); Proper (education) on data management is a major challenge
- **Quality issue:** It is still not a standard for public institutions to ask LSPs to also provide revision/review for their translations and to return the translation memories (the decision criterion for contracting LSPs is usually the lowest price not quality).
- **Financial issue:** CAT tools and the respective training are very expensive for the public sector.
- **Interoperability issues:** For example, Romanian characters were not initially supported by CAT tools.
- **Fundamental issue:** Language data is not considered a valuable asset and is not managed adequately.
- **Continuity issue:** Decision makers change frequently, however, proposed changes must be top-down, creating an even bigger challenge.
- **Political leadership:** The Secretariat General of the Government could/should lead the reform consequently, as the institutional level is not very relevant.

Action plan:

To support the Romanian language, to improve the translation workflow and to make data sharing in the future easier, the following actions are recommended:

- **Provide comprehensive access to CAT tools:**
  This objective addresses the need to raise awareness of the productivity gain through CAT tools and MT but also the procurement of necessary funding to make them available to staff translators. Special attention needs to be paid to facilitating training for efficient and purposeful use of CAT tools, including managing TMs in a way that allows for uncomplicated future language resource sharing.
- **Raising awareness of language data as Open Data and a valuable asset:**
  To achieve this objective, EU legislation would be most effective and would help to increase interest in language technology related issues.
- **Establishing good data management practices in public services:**
  This goal includes actions such as creating databases with translated documents and their metadata (e.g. date of translation, information whether the document was translated in-house or outsourced, IPR holder etc.).
- **Identifying and gaining access to outsourced translations:**
  This objective could be addressed through a high level decision with a clear mandate for public institutions to collect language data and make it available respectively.

References and links:

3rd ELRC Workshop Report for Romania (2021): https://lr-coordination.eu/sites/default/files/Romania/2021/ELRC3_Workshop%20Report%20Romania.pdf.

2nd ELRC Workshop Report for Romania (2018): https://www.lr-coordination.eu/sites/default/files/Romania/2018/ELRC%20Workshop%20Report%20ROMANIA_Public_FINAL.pdf.

Survey on the translation needs in public institutions (2017): https://ier.gov.ro/wp-content/uploads/newsletter/newsletter_noiembrie_2017_en.pdf.

1st ELRC Workshop Report for Romania (2016): https://lr-coordination.eu/sites/default/files/Romania/ELRC-Workshop-Romania-Public_Report.pdf.

Annex

Country Profile Slovakia

State of Play:

Translation practices and information exchange in ministries and public administrations:

In Slovakia, there are still no centralised operations for translations as each ministry has its own translation practices. The applied translation practices are diverse, ranging from in-house translation to decentralised outsourcing or a mixture of both.

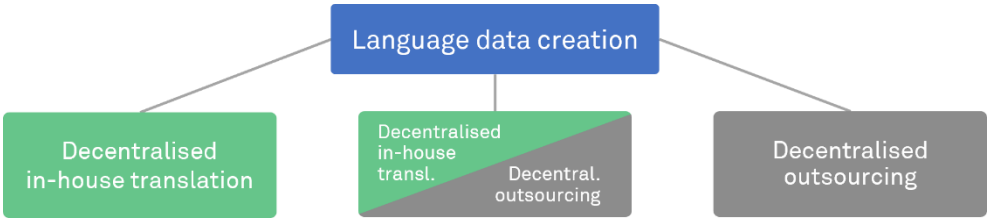
While language service providers (LSPs) and freelance translators build on the support of computer-assisted translation tools (CAT tools) for their translation activities, their use is not generally a common practice in public administrations. Since in the majority of Slovak public institutions, there are no specialised language and translation services, the use of outsourcing is often inevitable (ELRC, 2016). Public procurement data is openly available on the national procurement portal[100]. According to the NEC TM Report (cf. NEC-TM, 2019), more than 1.3 million EUR were spent for outsourced translation services between 2015 and 2018. Most of these outsourced translations were contracted by the Slovak administrative sector. The results of the above-mentioned report demonstrate that there is a need for multilingual content within public administrations and ministries.

As it was stated by a representative from the National Agency for Network and Electronic Service (NASES) at the second Slovak ELRC Workshop in 2018, the need for more multilingual digital services is one of the biggest challenges in Slovakia. However, according to the META-NET White Papers published in 2012, language technology industry is not sufficiently developed and the quality of Slovak language technologies and resources is not satisfactory yet (Šimková et al., 2012). Machine translation (MT) systems are hardly used in Slovak public administrations and ministries. However, a Slovak institution that has already translated documents with the help of machine translation is the Social Insurance Agency. They used the European Commission's machine translation system eTranslation, but not all results were satisfactory, since some of the texts had special requirements concerning term accuracy, which could not be fulfilled by the MT system.

Discussions at the third Slovak ELRC Workshop in May 2021 revealed that since the publication of the first version of this country profile, the situation has not changed dramatically. Slovak belongs to the less-resourced languages, which complicates language processing, thus making the development of LT solutions like chatbots more complex and time-consuming. As a consequence, it is difficult for Slovakia to keep pace with the global developments in the field of Artificial Intelligence. On the other hand, this fact does not mean that the Slovak representatives of LT based industry do not show effort to engage in technology development. Based on the experience of companies like Nettle.AI or Xolution, it can be said that the Slovak stakeholders and representatives in this domain provide viable and valuable LT-based solutions for Slovak language, although mainly through cooperation with foreign or transnational companies. At the same time, the LT industry in Slovakia is more spread and does not constitute a functioning network of stakeholders, in contrast with the neighbouring Czech republic, to name just one example. Here, the room and need for creating and maintaining a functioning network is very palpable. But, according to the representatives at the workshop, the nature of the Slovak language may also be an advantage: The fact that Slavic languages generally show less ambiguity has the potential to lead to future research activities in the field of systematic and comparative linguistics, thus boosting the development of language technologies.

[100] https://www.uvo.gov.sk

The current language data creation and sharing infrastructure in Slovak public bodies looks as follows:



At the third Slovak ELRC Workshop, it was mentioned repeatedly that it is important for the Slovak research community to participate in international infrastructures, such as CLARIN or DARIAH to be able to join forces and create synergies on an international level. With respect to this need, the signals from the political milieu are positive: after several years of attempts, the Road Map for Research Infrastructures (SK VI Roadmap 2020 – 2030) was published in March 2021, allowing for an active membership of Slovakia in the foreseeable future. In addition to that, there is also a clear need to improve networking on a local level: Currently, the Slovak research community is widespread, and networking is largely absent. According to the workshop participants, better networking is required, so the Slovak community can jointly work on improving language-centric AI for the Slovak language.

Data sharing infrastructures/Open Data in Slovakia:

In Slovakia, the Act on Free Access to Information is an important legal document when it comes to sharing data. Pursuant to this law, public institutions and ministries are required to provide information upon request. When it comes to Open Data, the Slovak government has already put a lot of effort in maintaining the national Open Data portals. The two main national Open Data portals are Data.Gov.sk and Slovensko.sk. They are maintained by the National Agency for Network and Electronic Services and include information about all state/public institutions and all Open Data. The electronic services offered by Slovensko.sk are diverse and aim to address the Slovaks' needs in their daily lives. In 2018, there were approximately 1700 services available in the national Open Data portal. It was used by about 480,000 registered users primarily from Slovakia, but also from other EU countries. Besides Data.Gov.sk and Slovensko.sk, there are also other local and regional Open Data portals such as Crime Map, which is based on criminal statistics of the Slovak police. In addition, Slovakia supports the Free Flow of Data initiative by the European Commission, which turned into effect in May 2019. The initiative "aims at removing obstacles to the free movement of non-personal data across Member States and IT systems in Europe" (cf. EC, 2022).

As pointed out by Martina Slabejová, who attended the second Slovak ELRC Workshop back in 2018 as the Slovak Digital Leader, collecting Open Data needs to go hand in hand with raising awareness of how the gathered data can be used within the public administrations. There are several projects focusing on improving eGovernment services and making data available to the public. However, these initiatives primarily focus on the provision of e-services, not on sharing language data. One of them e.g. aims at integrating Slovak public bodies into the government cloud, where both public and private information can be stored. Nonetheless, the public part could deliver valuable Open Data material that is usable for e.g. training and improving the European Commission's machine translation service eTranslation (cf. ELRC, 2018).

Digital Policy and language policy in Slovakia:

Pursuant to the Act on the State Language of the Slovak Republic published in 1995, Slovak is the official language of Slovakia. It is currently spoken by about 4.5 million inhabitants of the country, followed by Hungarian with more than 450,000 speakers.

> **Interesting fact:**
>
> **In 2009, the Slovak language policy was modified and preferential use of the state language was mandated. This was criticised by the Hungarian community, which makes up 10% of Slovakia's population.**

Since July 1, 2020, the agenda for digitalisation of public services in Slovakia, originally in the gestion of the Deputy Prime Minister's Office for Investments and Informatisation (UPVII), has been overtaken by the newly founded Ministry of Investment, Regional Development and Informatisation of the Slovak Republic (MIRRI), which is a central public authority body. The Ministry is a budgetary institution of the government, the revenues and expenditures of which are bound to the state budget. The main tasks of the Ministry include participation in creation and implementation of the uniform state policy in the field of the use of European Union funds, as well as informatisation of the society, and investment. As part of its powers, the Ministry performs tasks concerning management, coordination and supervision of the use of European Union funds in the area of informatisation of the society, as well as in the field of investments. The Ministry also performs tasks that stem from the membership of the Slovak Republic in international organisations (European Union, United Nations Organisation, Organisation for Security and Co-operation in Europe, Organisation for Economic Co-operation and Development, World Bank, Visegrad Group (V4)). The Ministry also provides for performance of obligations resulting from international treaties and conventions (United Nations Organisation, Organisation for Security and Co-operation in Europe, Council of Europe) that are binding on the Slovak Republic and fall under the scope of competences of the Ministry (ELRC, 2018). Through this central public body, Slovakia further aims to cooperate with European institutions, which is why the MIRRI also encourages administrations to participate in open CEF Calls. The Slovak government also created a number of operational programmes, which focus on the implementation of digitalisation of public administrations, thus helping to build a Digital Single Market.

Interesting Fact:

In Slovakia, the Ministry of Investment, Regional Development and Informatisation of the Slovak Republic is the main institution, which is responsible for the digitalisation of public services.

Data management is indispensable for building eGovernment services. Since eGovernment services require cleaned, structured and categorised data, several Slovak initiatives have been started with the aim of cleaning data and connecting public bodies. According to the Slovak Digital Leader Martina Slabejová, there were (as of 2018) approximately 100 projects that were crucial for the mission of the then-central body, UPVII, which was to offer better e-services. In 2017, the "Detailed Action Plan on Digitisation of Public Administration" was published in Slovakia. The goal of this action plan is the development of an eGovernment system, which serves the needs of Slovak citizens, public administrations, businesses and academia (EC, 2018).

The digitalisation agenda in Slovakia continues within The 2030 Strategy for Digital Transformation of Slovakia. This is a government strategy for the years 2019 to 2030 that needs to be seen as a key and decisive material for Slovakia at the beginning of the 21st century, when an inevitable transformation of the industrial society to the information one takes place. The strategy represents a means for Slovakia to succeed in the digital transformation brought not only by integration into the European digital single market, but also by the digital age in broader understanding. The document provides a beyond-departmental strategy to accelerate the digital transformation measures that have been already launched, define the new measures resulting from global digital trends and the European Union's priority policies, and transform them into a unique vision of Slovakia's digital transformation.

The complexity and severity of this issue requires a thoughtful view of the system to address it. This is reflected by the logical structure of the Strategy. Three assumptions were made – the resources for the digital transformation of society – i.e. human capital, infrastructure and regulatory framework. At the same time, five priority areas have been defined in the State, in which the individual transformation priorities are to be directed between 2019 and 2030, i.e. Economy, Society and Education, Public Administration, Territorial Development, and Science, Research and Innovation.

Consequently, the assumed or expected priority areas are derived from this vision; they are divided into two time horizons in terms of the current status of preparation and difficulty, i.e. to the short and long term time horizons. It is necessary to understand the process of digital transformation of Slovakia even in a wider context, as part of the wider process of building the 21st century information society in the context of respecting digital humanism.

The Strategy represents a key and decisive document for Slovakia at the beginning of the 21st century, at the time of necessary transformation of industrial society into information society. It covers the time period from 2019 to 2030 and it has been prepared as part of already launched and partially managed processes of digitalisation, informatisation and agenda of the single digital market of the European Union, as well as in the context of global priorities of a broad digital transformation. Thus, the Strategy puts primary emphasis on current innovative technologies such as Artificial Intelligence, Internet of Things, 5G Technology, Big Data and Analytical Data Processing, Blockchain and High-Performance Computing that will become the new engine of economic growth and strengthening of competitiveness. Therefore, at the national level, it will be necessary to accelerate already launched processes, connect national strategic measures with global trends as well as implement new policies based on the latest cross-cutting priorities of the EU and specific needs of Slovakia.

The strategy is a follow-up of the creation of new Multiannual Financial Framework of the EU for 2021-2027, including Cohesion Policy instruments as well as directly managed programmes (including Digital Europe Programme 1 and Connecting Europe Facility – digital part 2), where the need for development of the digital economy is given special attention. Besides the aforementioned facts, it also directly reflects conceptual materials and recommendations of international organisations, in particular, the Organisation for Economic Cooperation and Development and the United Nations Organisation, which consider the process of digital transformation as a key factor for achieving sustainable and inclusive growth. At the same time, the strategy was inspired by digital policies of developed countries such as Finland, France, Singapore and the United Kingdom. The strategy also analyses the current starting point of Slovakia – in particular, it is based on the current situation, specific priorities and the most important needs of the country that have been evaluated also on the basis of prestigious international documents, including Country Report Slovakia 2019 prepared by the European Commission. At the same time, the strategy respects and works with existing national strategies and action plans, in particular, it is based on the Action plan for smart industry. All that knowledge was summarised and incorporated into the vision of digital transformation of Slovakia with a list of recommendations for measures of short-term and long-term horizon that will turn visions into reality. Based on that, the vision of digital transformation of Slovakia has been defined as follows: By 2030, Slovakia will become a modern country with innovative and ecological industry built on knowledge-based and data economy, with effective public administration ensuring smart use of the territory and infrastructure and with information society whose citizens use their potential at full and live high-quality and secure lives in the digital era.

The vision of the strategy is materialised in assumed priority areas for short-term (3Q/2019 – 2Q/2022) and expected priority areas for long-term horizon (3Q/2022 – 4Q/2030) (cf. 2030 Digital Transformation Strategy for Slovakia).

As the third Slovak ELRC Workshop in May 2021 showed, two important developments were mentioned with regard to digitalisation: 1. the approval of the Road Map (cf. SK VI Roadmap 2020-2030), which was created by the Slovak Ministry of Education, Science, Research and Sport and published in March 2021 and 2. the Action plan for the digital transformation of Slovakia for 2019-2022.

1. The SK VI Roadmap 2020 – 2030 is a key document for the domain of Slovak research infrastructures. It not only monitors development of infrastructures so far and its current situation, but also describes its interconnectedness with the Slovak economy, international cooperation within the ESFRI, as well as the EU's research and innovation programme for 2021 – 2027, Horizon Europe. The document briefly describes the research infrastructure environment at the national and transnational level, identifies established international research infrastructures, where Slovak Republic acts as member (ECRIN ERIC – European Clinical Research Infrastructure Network; INSTRUCT ERIC – Integrated Structural Biology Infrastructure; European XFEL – European X-Ray Free-Electron Laser Facility; HL-LHC – High-Luminosity Large Hadron Collider; ILL – Institut Max von Laue – Paul Langevin; CESSDA ERIC – Consortium of European Social Science Data Archives; ESS ERIC – European Social Survey) or an (unofficial) observer (CLARIN ERIC – Common Language Resources and Technology Infrastructure; DARIAH ERIC – Digital Research Infrastructure for the Arts and Humanities). The Road Map was created in cooperation with World Bank experts and ESFRI representatives at the national level. The aim of the document is to point

at the significance and potential of the existing infrastructure and its function of being a motor for further development and innovation in Slovakia toward knowledge-based society (ibid.).

2. As for the Action plan for the digital transformation of Slovakia for 2019 – 2022, it contains concrete steps to build a sustainable, human-centric, and trustworthy AI ecosystem within the long-term Strategy of the digital transformation of Slovakia 2030, mentioned above. One of the proposed projects of the Action plan is the development of a tool for natural language processing to accelerate the development of AI in the private sector and improve the quality of public services. In detail, the Action plan specifies the following:

   "... it will be necessary to remove barriers in the use and development of text and voice corpus of the Slovak language with specific regard to safe and practical application of such technologies in the field of public services. It will be possible to use the methods of natural language processing for monitoring of priority holistic goals, i.e. increasing transparency of the Slovak regulatory framework. Subsequently, it will be possible to use features of semantic text and voice analysis for automation and electronisation of subset of services in the contact with authorities, medical facilities and schools, which will make opportunities for developing innovative packages of services and products also for commercial sector, e.g. in IT sector, in the field of data transfer security, in automobile industry as well as in other fields of the commercial sector."

Additionally, the Action plan foresees the preparation of a new Act on Data to better define regulations on data protection, disclosure principles, data access and open data regulations. The proposed measure

"... will result from precisely defined categorisation and classification of data based on their information value and required level of protection. It means that there will be a precise definition of rules and processes for reference data, open data and the manner in which it will be possible to analytically process data (including rules for anonymisation and pseudonymisation of data)".

In addition to that, a large part of the Pandemics Renovation Plan is devoted to digitalisation and language development.

Stakeholders and major networks:

The ELRC Anchor Points for Slovakia represent two important stakeholders, i.e. the Ministry of Culture and the udovít Štúr Institute of Linguistics, Slovak Academy of Sciences.

As for the major LR providers in previous years, motivated by the first Slovak ELRC Workshop in 2016, the Ministry of Justice and the Ministry of Culture of the Slovak Republic contributed with more than 1 million tokens of raw mono- and bilingual texts in different language combinations, mainly English-Slovak, covering a number of fields, e.g. laws, reports, letters, brochures, invitations, etc. This contribution led to the creation of two parallel corpora in English and Slovak plus two monolingual data sets that were delivered to the ELRC-SHARE repository. In addition, the Ministry of Economy contributed a significant amount of English-Slovak parallel data after the second ELRC Workshop. Overall, more than 40 organisations participated in ELRC events such as local workshops or conferences. In 2021, almost 60 people joined the third ELRC Workshop in Slovakia. They represented a variety of fields, including the public sector, research and academia, SMEs as well as industry LT providers.

It is of great importance that a new and fast growing stakeholder within AI and NLP appeared in the last two years: the Kempelen Institute of Intelligent Technologies (KInIT). The non-profit institute KInIT, founded in 2020, is dedicated to **intelligent technology research**. It brings together and nurture experts in AI and other areas of computer science, with connections to other disciplines, such as information security, web and user data processing (including false information and malicious behaviour modelling), processing and comprehension of natural language, data analysis for green energy, ethics and human values in intelligent technologies. The institute has quickly become one of the most notable and active figures within the Slovak AI community and managed, amongst other projects, to develop

SlovakBERT, the Slovak Masked Language Model, the first Slovak-only transformers-based model trained on a sizeable corpus[101].

The Slovak Research Centre for Artificial Intelligence (Slovak.AI) was created in 2019 to **support excellence in the field of AI** by bringing together all relevant stakeholders such as businesses, research communities, and governmental institutions. In this endeavour, the platform **highlights the increasing importance of AI in solving major societal challenges such as climate change, safety, health, and food security**.

In July 2020, the President of the Republic of Slovenia was on a visit to Bratislava to strengthen the relationship between Slovakia and Slovenia. During his visit, a Slovenian-Slovak business forum was held with individual business meetings between participants from both countries. Both Slovenia and Slovakia attach **great importance to sustainable development**, with an emphasis on the green agenda and digitalisation and in particular on the circular economy, smart technologies, e-mobility and AI. Slovak companies will exchange views with Slovenian companies on breakthrough solutions in these areas.

The Slovak strategy also has a **dedicated policy to increase the international visibility of AI education** by making it more accessible to foreign students.

The MIRRI collaborates with the Slovak Ministry of Economy and the IT Association of Slovakia, on a **feasibility study to create the European Digital Innovation Hub** in **Slovakia**; it also launched a web portal[102] **to collect AI project proposals from public sector institutions**, so as to measure the dissemination and uptake of AI in Slovakia. Soon the portal will also include analyses and summary details of the submitted projects. In addition, there will be a survey to measure AI uptake by companies and companies' attitude towards AI.

The **Security Council will establish a working group on disinformation and fake news** which is going to use AI technologies against disinformation and fake news.

Main challenges for sustainable data sharing:

- There are few multilingual digital services available.
- Public services are often not aware of the value of Open Data.
- The lack of accessible multilingual data. This is also partly due to technical issues that prevent institutions from making Open Data accessible. As for the monolingual (Slovak) textual resources, there are some large and valuable resources available, e. g. the Slovak portal of judicial decisions (https://otvorenesudy.sk), collecting respectable volumes of data (texts in pdf files), as well as metadata (information on courts, judges, procedures etc.).
- Furthermore, language technology support still needs to be improved to be able to better serve the public administrations' requirements.
- Slovak public administrations are generally not aware of the range of technical solutions that could support them in their daily operations.
- There is a lack of awareness concerning the possibilities of the European Commission's funding mechanisms amongst Slovak public bodies.
- There is a lack of national and international networking
- According to the results of a SWOT analysis published within the Slovak 2030 Digital Transformation Strategy, the Slovak (data) infrastructure is challenged by the missing Digital Innovation Hubs (DIHs) in Slovakia, low level of digitalisation of the economy and ineffective functioning of the public sector. As for the eGovernment, results of the DESI index for 2018 show that Slovak manufacturing and services and, above all, the public sector still report low level of informatisation. A specifically serious problem is the low quality of eGovernment services. Other challenges include: inadequate infrastructure for data economy in the public administration, low level of engagement in international initiatives that provides countries with innovations and know-how, low number of pro-investment approaches of the public administration into infrastructure, only low amount of investments into infrastructure and its stabilisation.

[101] https://kinit.sk/publication/slovakbert-slovak-masked-language-model/
[102] https://datalab.digital/dopytove-vyzvy/prehlad-schvalenych-studii-uskutocnitelnosti/prve-hodnotiace-kolo-pre-lepsie-vyuzivanie-udajov-instituciami-verejnej-spravy-opii-2019-7-10-dop/

- As for the opportunities, the SWOT analysis identifies: European ecosystem of Digital Innovation Hubs (DIHs), social acceptance of AI and other technologies in order to improve functioning of the private and public sector, increasing the quality, effectiveness and subsequent enhancement of eGovernment services. The analysis also mentions the position of Bratislava as a big innovation hub in the V4 region.
- As for the opportunities within the Slovak regulatory framework, there are new possibilities to accelerate and increase the efficiency of building and sharing infrastructure, digital transformation as the new engine of the economic growth of the country, public administration reform in the right direction in order to increase competences and accelerate processes, room for policy for creating favourable environment for small, medium and big enterprises.

Action plan:

Based on the current situation in Slovakia and the identified challenges, five objectives could be defined. Ranked by their priority, these include:

- **To intensify networking on a national and international level:**
  The third Slovak ELRC Workshop revealed that there is a clear need for Slovakia to participate in international research infrastructures such as CLARIN or DARIAH. In addition to that, it was stated that Slovak actors often have to rely on themselves due to a general lack of networking opportunities and/or national initiatives. Consequently, intensified networking and the creation of national and international synergies may be useful for Slovakia to catch up with the global developments in AI and LT.
- **To increase interest in MT in public services:**
  As the use of MT is currently not common in public administrations and ministries, it is necessary to further promote the benefits of using machine translation. This can be achieved by creating synergies with national projects and initiatives on the one hand, but also by securing the support of Slovak decision makers on the other. In addition, it is important to provide more information on how MT systems work and to communicate how much data will be required to improve an MT system.
- **To raise awareness of language data as Open Data:**
  Since language data are currently not included in eGovernment initiatives, it is important to raise awareness of language data as Open Data. A first and important step towards this goal would be to identify and contact an Open Data officer.
- **To identify and gain access to outsourced translations:**
  As already mentioned, there is a general lack of openly accessible language data in Slovakia. At the same time, a high number of translations needs to be outsourced to LSPs or freelance translators. Consequently, one way to increase the number of accessible data would be to identify and gain access to outsourced translations. This could be achieved by making it a common practice to receive any by-products of the outsourced translations back.
- **To tackle legal concerns:**
  Since MT systems need to be improved to serve the public administrations' needs, as much language data as possible should be made available. However, legal issues often prevent potential contributors from sharing their data. Easy-to-apply guidelines for Intellectual Property Rights (IPR) and privacy issues could help overcome these issues and provide data holders with the necessary expertise.
- **To establish good data management practices in public services:**
  Last but not least, it would be useful to identify a data manager, supporting the creation and development of data management practices in public services.
- Since 2021, the course Elements of AI has been offered even in Slovak language, provided by AIslovakIA, with Comenius University as the scientific supervisor[103]. The KInIT Institute has published together with Gerulata Technologies the Slovak language model (SlovakBERT). In 2021, it also started a project aimed to create the first Slovak-only language model trained with a large text corpus.

[103] https://aislovakia.com/elementsofai/

- In the EU Recovery Plan, component Digital Slovakia, there is an explicit item "Creation of Data Sets for Language Models".
- The key organisation for digitalisation of public services in Slovakia is the Ministry of Investment, Regional Development and Informatisation of the Slovak Republic (MIRRI) as a central public authority body. The Ministry is a budgetary institution of the government, the revenues and expenditures of which are bound to the state budget. The main tasks of the Ministry include participation in creation and implementation of the uniform state policy in the field of the use of European Union funds, as well as informatisation of the society, and investment. As part of its powers, the Ministry performs tasks concerning management, coordination and supervision of the use of European Union funds in the area of informatisation of the society, as well as in the field of investments. The Ministry also performs tasks that stem from the membership of the Slovak Republic in international organisations (European Union, United Nations Organisation, Organisation for Security and Co-operation in Europe, Organisation for Economic Co-operation and Development, World Bank, Visegrad Group (V4)). The Ministry also provides for performance of obligations resulting from international treaties and conventions (United Nations Organisation, Organisation for Security and Co-operation in Europe, Council of Europe) that are binding on the Slovak Republic and fall under the scope of competences of the Ministry. Its goal is the centralisation of informatisation. Slovakia aims to cooperate with European institutions, which is why the UPVII also encourages administrations to participate in open CEF Calls. The Slovak government also created a number of operational programmes, which focus on the implementation of digitalisation of public administrations, thus helping to build a Digital Single Market.
- The Slovak Government will put in place digital data platforms to let high-quality and trustworthy data accessible for the needs of AI.
- The strategy includes the following policy initiatives for the data economy:
  - Creating an Institute for trustworthy data to provide open access to high value databases from the public administration after controlling validity, constancy and credibility of the data;
  - The MIRRI will provide public administration with analytical tools for data management. So, the public administration will receive user-friendly SQL and machine learning tools for data simulations, visualisations and statistical calculations to facilitate policy making. With this help, end-users in the public sector can run data analytics without technical issues on data management;
  - Setting up a Personal Information Management System (PIMS), a centralised data repository with data collected by the public administration about citizens. The PIMS will comply with data protection and data sharing regulations by allowing citizens to give their consent on these issues;
  - The Ministry of Environment is setting up a platform for sharing harmonised spatial data in compliance with the INSPIRE directive.
- Lastly, the Slovak strategy envisages actions to boost the digital and telecommunication infrastructure:
  - Setting up a national high-performance computing competence centre, and participating to the European EuroHPC that pools European resources to develop supercomputers;
  - Supporting the completion of a gigabit fibre infrastructure and the 5G for Europe Action Plan. Both initiatives aim to increase internet connectivity and achieve the goals of the EU gigabit society.

References and links:

2030 Digital Transformation Strategy for Slovakia. Strategy for transformation of Slovakia into a successful digital country. The English version available at: https://www.mirri.gov.sk/wp-content/uploads/2019/10/SDT-English-Version-FINAL.pdf.

Crime Map: https://mapazlocinu.sk/.

Freedom of Information Act: https://www.ustavnysud.sk/en/zakon-o-slobode-informacii.

G-Cloud: https://www.sk.cloud.

SK VI Roadmap 2020 –2030. Roadmap for the Slovak research infrastructures. The Slovak version available at: https://www.minedu.sk/data/files/10600_cestovna-mapa-vyskumnych-infrastruktur-sk-vi-roadmap-2020-2030.pdf.

Slovak portal of judicial decisions: https://otvorenesudy.sk.

The Department of Slovak National Corpus, Ľ. Štúr Institute of Linguistics, Slovak Academy of Sciences: https://korpus.sk/.

The Kempelen Institute of Intelligent Technologies: https://kinit.sk/.

The "Open Courts" – Slovak Portal of Judicial Decisions: https://otvorenesudy.sk.

[EC, 2018] European Commission: *Digital Economy and Society Index (DESI) 2018*, Country Report Slovakia, 2018, http://ec.europa.eu/information_society/newsroom/image/document/2018-20/sk-desi_2018-country-profile_eng_B4415E7E-9154-E26E-7B403212919F3F7C_52238.pdf.

[EC, 2022] European Commission: *Free Flow of Non-Personal Data*, https://ec.europa.eu/digital-single-market/en/free-flow-non-personal-data.

[ELRC, 2016] Zumrík, Levická: *ELRC Workshop Report for Slovakia*, 2016, http://www.lr-coordination.eu/sites/default/files/Slovakia/ELRC-Workshop-Report_SLOVAKIA-public.pdf.

[ELRC, 2018] Zumrík, Miroslav: *ELRC Workshop Report for Slovakia*, 2018, http://www.lr-coordination.eu/sites/default/files/Slovakia/2018/ELRC%2BWorkshop%20Slovakia%20Public%20Report_FINAL.PDF.

[ELRC, 2021] Zumrík, Miroslav: *ELRC Workshop Report for Slovakia*, 2021, https://www.lr-coordination.eu/sites/default/files/Reports%202021/ELRC3_Workshop%20Report%20Slovakia_public_FINAL.pdf.

[NEC-TM] NEC-TM Report: *Slovakian National Contracts Report: Process and Findings*, 2019, https://www.nec-tm.eu/wp-content/uploads/2019/05/Slovakia-Report.pdf.

[Šimková et al.] Šimková et al.: *The Slovak Language in the Digital Age*. In: META-NET White Paper Series, 2012, http://www.meta-net.eu/whitepapers/e-book/slovak.pdf.

Annex

Country Profile Slovenia

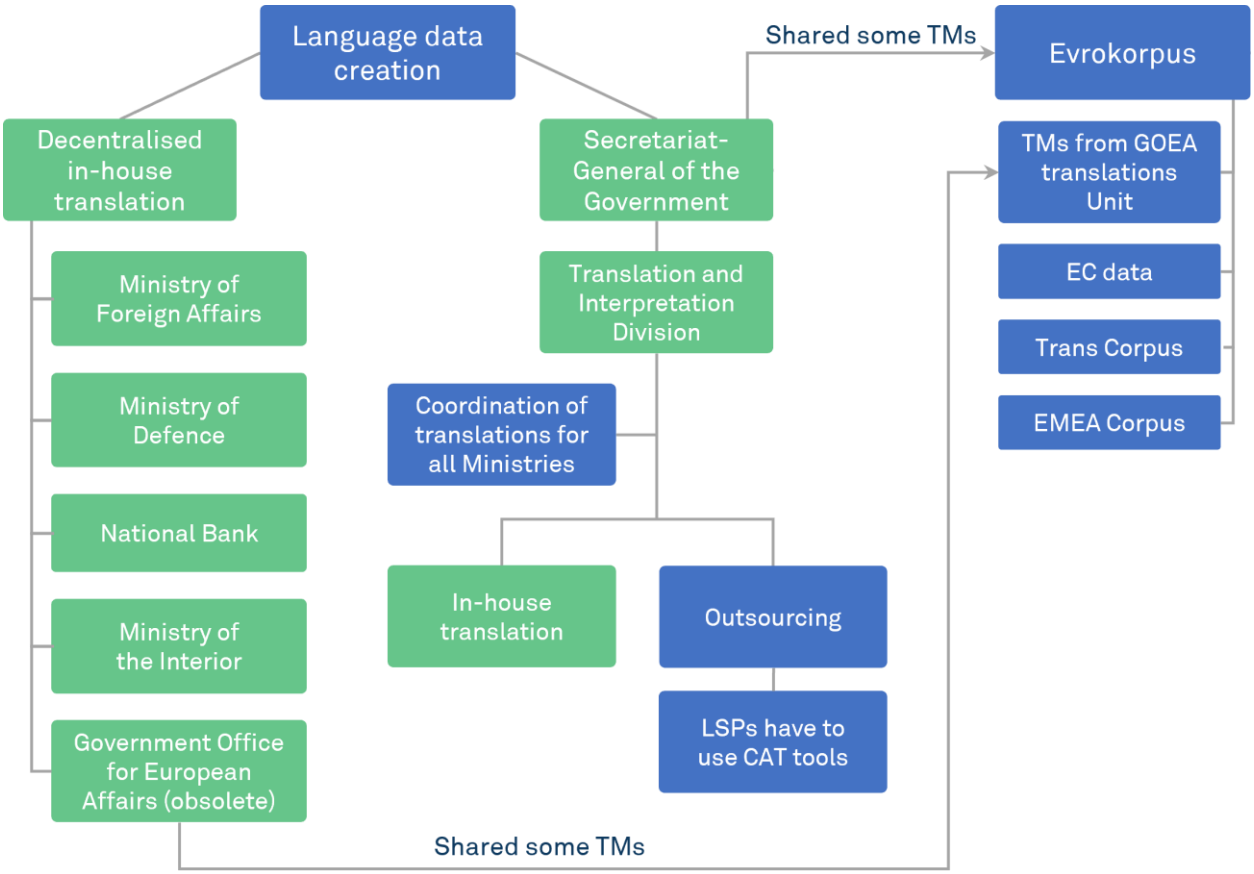
State of Play:

Translation practices and information exchange in ministries and public administrations:

In Slovenia, the majority of the translation demands in public services are handled centrally through the Translation and Interpretation Division (TID) at the Secretariat-General of the Government who manages translation and interpretation demands for most ministries. The TID handles 26% of the translations in-house, whereas the other translations are outsourced to language service providers. The Translation and Interpretation Division uses CAT tools and requires LSPs to do the same when they generate their translations. TMs are not shared, but the LSPs are required to return bilingual files for English, German and French translations.

A few ministries handle their translation needs independent from the TID, among them are the Ministry of Foreign Affairs, the Ministry of Defence, the Ministry of the Interior and the National Bank. These public bodies use CAT tools but they do not exchange data or know-how with other ministries or the TID. However, they also outsource part of their translations to the TID.

The current language data creation and sharing infrastructures in Slovenia public bodies looks as follows:



Open Data and data collection in Slovenia:

The Evrokorpus is a dedicated portal for parallel language resources for Slovene <> English, German, French, Italian, Spanish. It contains data from the (former) translation unit at the Government Office of European Affairs, data from the European commission, the Trans Corpus and the EMEA corpus, and is

maintained and updated by the TID. Due to anonymisation issues, however, only parts of the corpus could be shared with ELRC so far. Evrokorpus has a companion terminology database called Evroterm which contains English, Slovene, German, Italian and Spanish terminology. The Evroterm database has been released on the Slovene Open Data portal under the CC BY-NC-ND licence and is available in the ELRC-SHARE repository. The full Slovene legislation, however, with more than 100 million tokens was made available in the Slovene Open Data portal as an open access database in JSON format which is considered an important achievement of the ELRC data collection task in Slovenia.

Interesting fact:

Full Slovene legislation with more than 100 million tokens was made available in the Slovene Open Data portal in JSON format.

The Slovene Open Data Portal (https://podatki.gov.si/) has a dual function. The first one is to provide a central catalogue of all the records and databases of Slovenian public bodies. In this catalogue the metadata about all the Open Data from state authorities, municipalities and other public sector bodies is made available. The second function of the portal is to be the single access point for data in a machine-readable format and with an Open Data licence. This includes Open Data collections which had already been published on different websites, such as Evroterm.

Digital policy and language policy in Slovenia:

According to the Constitution, Slovene is the only official and state language of the Republic of Slovenia (Nećak Lük, 2017, p. 57). However, in municipalities, where the Italian and Hungarian speaking population resides, these languages have official status as well (ibid., p. 62). Romany languages have minority status in Slovenia (ibid., p. 60). In 2004, the Act on Public Usage of Slovenian Language came into effect monitored by the Ministry of Culture. The act determines the use of the Slovenian Language in public communication and in specific areas and resulted in the first resolution on the National Programme for Language Policy (NPLP) for a period of 4 years (2007-2011) (ibid., p. 62 ff.). With the resolution, a budget of 12 million EUR was allocated to language policy and language planning for the first time. However, only about 300,000 EUR were actually spent by the Ministry of Culture in the framework of the resolution. The Ministry of Education on the other hand, spent over 3.2 million EUR from structural funds for the development of modern language technologies and resources for the Slovene language between 2008-2013.

It took three years to adopt the second Resolution on the National Programme for Language Policy (2014-2018), After the resolution passed, an action plan for Language Infrastructures was initiated in 2015 budgeted with 11 million EUR. The financing bodies are the Ministry of Education, Science and Sports, the Ministry of Culture, the Slovenian Research Agency (subordinated to the Ministry of Education) and the publicly funded Slovenian Academy of Sciences and Art. Another result of the language act was the foundation of the Council for Continuous Monitoring of the Development of Language Resources and Technologies for Slovene representing several ministries, government offices and agencies who produce a yearly progress report.

According to this report, less than 250,000 EUR were spent in 2015 for the realisation of the resolution but it was announced in 2018 that the Ministry of Education had allocated up to 2 million EUR on the *Promotion of flexible and innovative learning techniques with the development of language resources and technologies* (Call, 2019). The funding for the CLARIN infrastructure was increased from 42,000 EUR to 100,000 EUR per year, the funding for the Centre for Language Resources and Technologies at the University of Ljubljana were raised to 55,000 EUR per year compared to 11,000 EUR before, and a new research group for “Language Resources and Technologies for Slovene” at the University of Ljubljana received resources to fund 2,5 full time equivalent personnel per year from 2019-2024.

A third resolution for the timeframe 2019-2024 passed in 2018 followed by a public consultation on “Development of Slovene in digital environment – language resources and technologies” conducted by the Ministry of Culture, for which structural funding in the amount of four million EUR was available. The DSDE (Development of Slovene in a Digital Environment) project runs from 2020-2023. All programming code and databases produced for this project will be publicly available under an open licence from December 2022. All applications (speech recognition, transcription, machine translation,

terminology extraction, and a terminology portal) will be made available on the public DSDE portal, where anyone will be able to try and use them. The project Development of Slovene in a Digital Environment is co-financed by the Republic of Slovenia and the European Union from the European Regional Development Fund.

The role of LT and language data in Slovenia's AI regulations:

According to the National programme (NpUI) promoting the development and use of AI in the Republic of Slovenia by 2025, the Slovenian Government will invest a total of 10 million Euro to support technological research projects in various AI-related fields (Measure 3.2), including language technologies. The funding will be available to consortiums consisting of public institutions and private companies. Additionally, measure 5.7 allocates 150,000 Euro to funding the Slovenian CLARIN infrastructure.

Stakeholders and major networks:

The Council for Continuous Monitoring of the Development of Language Resources and Technologies and all its members play a crucial part in all activities related to language data and language technologies in Slovenia. As one of the main language data creators of parallel data for Slovene public administrations, the Translation and Interpretation Division subordinated to the Secretariat General of the Government is also a key stakeholder to create sustainable language data sharing infrastructures in Slovene public administrations. Many activities related to language resource collections are supported by the Josef Stefan Institute, represented by the Technology NAP, and the Centre for Language Resources and Technologies at the University of Ljubljana. So far, more than 30 institutions from the public sector, academia and industry have attended ELRC events and the Secretariat-General of the Government is one of the data contributors that shared language resources with ELRC.

Main challenges for sustainable data sharing:

The main challenges for sustainable language data sharing in Slovenia are the following:

- One of the central issues is the fact that the implementation of the Resolution on the National Programme for Language Policy is often dependent on high-ranking individuals and their disposition regarding language technology and language resources, which makes continuous and sustainable progress difficult.
- Another issue is the lack of efficient cooperation between stakeholders, although a lot of expertise is available.
- The unawareness of the value of language resources and Open Data results in reluctance to share language data as the benefits and incentives are not evident.
- In addition, concerns about personal or confidential data are holding many potential data donors back from sharing their language data.

Action plan:

Several objectives should be targeted to address the identified challenges. The first two objectives are suggested recommendations, whereas the last two objectives are partly addressed in the language resolution but their implementation should be reinforced as they are considered very important.

- **Tackle legal concerns:**
  To address the concern of confidential or personal data potentially included in language resources, legal experts are needed that can advise each public administration if their data needs any kind of pre-processing before it can be shared. This also includes copyright and IPR-related issues. One venue worth exploring could be the practice of differentiating between non-personal open government data and texts that contain personal or confidential data which could help establish good data management practices in public services. In addition, appropriate guidelines for the creators of language data are needed that can be followed during or after the translation process to make data sharing easier.
- **Raising awareness of language data as Open Data and a valuable asset:**
  This objective includes activities such as integrating language data in the national Open Data policy and establishing practical guidelines for language resources as Open Data.

- **Identify and gain access to outsourced translations:**
  By establishing the value of language resources, changes in the procurement process can be initialised. As the procurement of language service is not centralised, each public administration that has a demand for translation would have to change its procurement process and ensure the provision of translation memories as well as any other by-product of translation including the transfer of copyright mandatory for outsourced translations.
- **Increasing interest in MT/LT in public services as part of the national digital policy:**
  To increase the public interest in machine translation and language technologies, it is important to create synergies between different initiatives and address the needs of the public sector. This includes e.g. the dissemination of use cases and best practices.

References and links:

Centre for Language Resources and Technologies: https://www.cjvt.si/en/.

Development of Slovene In a Digital Environment: https://www.slovenscina.eu/en.

Evrokorpus: https://evroterm.vlada.si/evrokorpus.

Evroterm: https://evroterm.vlada.si/evroterm?clang=en.

Slovene Open Data portal: https://podatki.gov.si/.

Slovenian CLARIN: https://www.clarin.si/info/about/.

[Call, 2019] Call for proposals: Innovative and flexible forms of teaching and learning, 2019, https://www.gov.si/zbirke/javne-objave/inovativne-in-prozne-oblike-poucevanja-in-ucenja-2/.

[Nećak Lük, 2017] Nećak Lük, Albina: *Slovene Language Status Planning*. In: Revista de Llengua I Dret, Journal of Language and Law, 2017, https://www.raco.cat/index.php/RLD/article/download/329938/420648.

Annex

Country Profile Spain

State of Play:

Translation practices and information exchange in ministries and public administrations:

Spain, being a multilingual country, in addition to the need to translate from Spanish to English and other international languages, there is the need to translate to and from Spanish and the regional languages. Regional languages, i.e. Catalan, Basque and Galician, are co-official only at the regional level, thus it is the regional administrations, which traditionally have a more principled approach to translation practices and use technology more intensively. They also tend to have more dedicated translation services and more in-house translators.

In general, most state-level public bodies, and many regional ones too, tend to outsource their translations to language service providers.

Both translation and procurement processes are run in a highly decentralised way, sometimes even within one institution. Only very few ministries have in-house translation services; among those: the Ministry of Interior, the Ministry of Defence, and the Language Interpretation Office (OIL) at the Ministry of Foreign Affairs. According to a "White Paper on Institutional Translation and Interpretation", published in 2011, only one translator at the OIL used computer-aided translation (CAT) tools in 2011. The same survey provides further information about number of translators in the Spanish Administration:

- OIL (Ministry of Foreign Affairs: 17 translators, 1 used CAT as of 2011)
- Ministry of Interior (230 translator no CAT as of 2011)
- Ministry of Defense (30 translators,. no use of CAT as of 2011)
- Ministry of Presidency (11 translators, no use of CAT as of 2011)

Since no similar surveys have been done since, we cannot provide more updated information, but, to our knowledge there have not been major changes in the approach to translation in the Public Administration, after the aforementioned White Paper was published.

We do have a more recent report, which is the Spain Country Report, compiled in 2019 by the NEC-TM project. This report details findings from 2015-2018 on translation contracts from public administrations to the private sector in Spain. Not considering translation contracts below 10,000 EUR, the translation costs for outsourced translations amounted to approximately 44 million EUR in that period. The report reveals that most of the contracts did not include requesting back the resulting translation memories. This can be attributed to lack of awareness for the intrinsic value of TMs, aggravated by the fact that most public entities do not use CAT tools, with a few notable exceptions, such as Segittur, the digital agency for Tourism. According to this study, requesting translation memories from language service providers could reduce translation costs by up to 10%.

In the course of another study commanded by the State Secretariat for Telecommunications and the Information Society (SETSI) in 2016, called "Inventory of linguistic resources of the Public Administration for automatic translation" (cf. Aguado de Cea et al., 2016) several sectors were identified that have a particularly large demand for translations. All the related public bodies outsource translations. Those sectors are:

- Police (state and autonomous communities)
- Administration of justice (both state and regional level)
- Tourism
- Social security
- Tax agency

To support the need for translation, the Spanish government runs their own Machine Translation Platform called PLATA[104] that is used as an API to translate Spanish Government web pages, but is also serving other customers such as MUFACE, and the Spanish Agency for Data Protection, to translate to and from Spanish and the co-official languages and English. To support more language pairs and provide a better service, it has been recently connected to eTranslation.

Interesting fact:

Due to the need to translate between co-official languages and Spanish, translation procedures at the regional level have a longer tradition of using CAT tools and machine translation than those at national level.

Overall, it can be said that most public administrations at the state level, but also at the regional level (with the possible exception of the Basque administration) rely heavily on outsourcing their translations, without a standard procedure to request back translation memories. Generally speaking, there is no systematic use of translation memories but this may be starting to change.

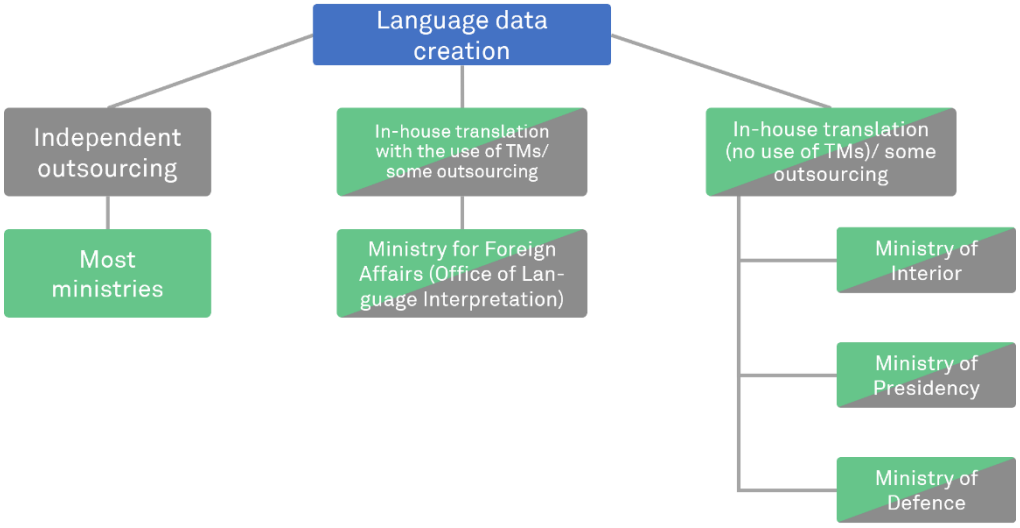
Moreover, so far there has been little coordination and management of language data in terms of protocols for consistent archiving, using metadata standards, differentiating between confidential documents and documents that contain personal information, or keeping original documents aligned with their translations. Again, hopefully this may be starting to change with a greater awareness of the value of data in general, and language data in particular.

After the last ELRC Workshop conducted in May 2021, where several representatives of the Spanish Administration participated, it became clear that there is an increased awareness of:

- language technologies, which are acknowledged as an important aspect of Artificial Intelligence
- language data, which is starting to be valued as a technological asset that needs to be reused and shared
- the need to go further in the initial wave of digitalisation that has taken place along the last decade
- the crucial importance of technology to enhance the services to the citizens, and that, far from posing a threat to humans, it is able to optimise their work.

With the new AI strategy starting to be put in place and the targeted funding from the EU Next Generation funds, we expect to see important developments at all levels in the Spanish Administration soon.

The current language data creation and sharing infrastructure in Spanish public bodies looks as follows:



[104] https://administracionelectronica.gob.es/ctt/plata#.Y0PwoXbP1D8

Open Data and data collection in Spain:

The web portal http://datos.gob.es federates the Open Data from different public administrations: government, municipalities and regional autonomous governments. The richest language data comes from the Basque Administration in the form of translation memories. Other text collections can be found, most of them in PDF format.

Digital policy and language policy in Spain:

In Spain, the implementation of the digital agenda is coordinated by the State Secretariat of Digitalisation and Artificial Intelligence (SEDIA). SEDIA strongly supports the use of language technologies in the public and private sector through a National Plan of Language Technologies, underlining the importance of collecting and sharing language data (cf. Advancement Plan, 2015, p.8).

This plan focuses on three main points:

- "Increasing the amount, quality and availability of linguistic infrastructure in Spanish and in Spain's co-official languages.
- Fostering the language industry by promoting knowledge transfer from the research field to industry. Bolstering internationalisation of companies and institutions in the sector. Improving the reach of current projects.
- Improving the quality and capacity of public services, by integrating natural language processing and machine translation technologies, while simultaneously driving market demand. Supporting creation, standardisation and distribution of language resources, created by the management activities performed by the public administrations." (ibid., p.7)

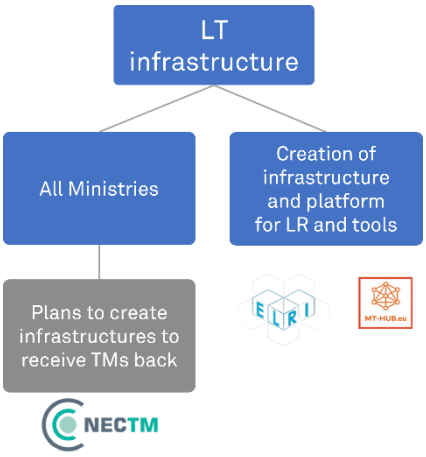
The governance bodies of the Spanish language are the Royal Spanish Academy (RAE) and the Association of Spanish Language Academies in Ibero-America (ibid., p.6).

Main challenges for sustainable data sharing:

- Undervalued perception of language data leads to a number of issues:
  - Translations are not filed consistently which makes it difficult to match the original text with the translation
  - Text and translations produced for single occasions are not stored in TMs as it seems unlikely that they will be reused
  - Most documents are stored in PDF format, which is the less suitable format for building translation memory files, while the source format is lost
  - General lack of data plans and protocols and resistance to modify internal document management practices (cf. ELRC, 2018, p. 11)
- Translation needs are met very decentrally, even within ministries
- Legal uncertainty (unclear authorisation chain to decide what can be shared)
- Most translation contracts are outsourced

Action plan – ongoing projects and future plans:

Following the strategy laid out by the national Plan of Language Technologies, an appropriate infrastructure to facilitate language data reuse and sharing is ready to be deployed thanks to the synergies with several CEF-funded projects participated by the SEDIA (see figure).



The national Relay Station resulting from the ELRI project (https://elri.plantl.gob.es/es-es/) can be leveraged as a data storage and processing solution, with complex sharing capabilities. It has already been used as the main data portal to gather language data from public institutions.

The MT-hub platform, an outcome from the CEF project iADAATPA, can function as a translation portal for the public administration, and an access point for eTranslation. MT-HUB is

compatible with multiple content management tools, and helps the user to select the best domain-adapted MT engine for their document in need of translation.

A central translation memory server is available, as a result of the NEC TM project. The server is a secure place to store translation memories coming from outsourced translation contracts, directly by the LS providers. It will allow sharing of TMs between administrations themselves, and administrations and providers, and is compatible with commercial CAT tools. An open CAT tool integrated into it is also provided.

References and links:

Plan of Impulse of Language Technologies: https://www.plantl.gob.es/Paginas/index.aspx.

[Advancement Plan, 2015] Plan for the Advancement of Language Technology, 2015, https://plantl.mineco.gob.es/tecnologias-lenguaje/PTL/Bibliotecaimpulsotecnologiaslenguaje/Detalle%20del%20Plan/Plan-Advancement-Language-Technology.pdf.

[Aguado de Cea et al., 2016] Aguado de Cea et al.: Inventario de Recursos Lingüísticos de la Administración Pública para Traducción Automática, 2016, https://plantl.mineco.gob.es/tecnologias-lenguaje/actividades/Estudios%20tcnicos%20y%20de%20gobernanza/Inventario%20de%20recursos%20para%20traducci%C3%B3n%20autom%C3%A1tica/inventario-recursos-traduccion-Retele.pdf.

[ELRC, 2018] Bel, Núria, Melero, Maite: ELRC Workshop Report for Spain, 2018, https://lr-coordination.eu/sites/default/files/Spain/2018/ELRC-Workshop-Report_Spain-2018-public-final-EC_.pdf

[Libro Blanco, 2012] Ministerio de Asuntos Exteriores y de Cooperación: Libro Blanco de la traducción y la interpretación institucional, 2012, https://cpage.mpr.gob.es/producto/libro-blanco-de-la-traduccion-y-la-interpretacion-institucional-2/.

[Melero et al., 2012] Melero, Badia, Moreno: The Spanish Language in the Digital Age. In: META-NET White Paper Series, 2012, http://www.meta-net.eu/whitepapers/e-book/spanish.pdf.

[NEC TM, 2018] NEC TM Country Report: Report on Spanish National Translation Contracts, 2018, https://www.nec-tm.eu/country-reports-nec-tm/country-report-spain/.

Annex

Country Profile Sweden

State of Play:

Translation practices and information exchange in ministries and public administrations:

In Sweden, most public administrations outsource translations through public procurement. The process itself is coordinated by each public administration or agency independently, but they are obliged to order translation services through Kammarkollegiet, the Legal, Financial and Administrative Services Agency that is in charge of the framework agreements for translation services in Sweden. This framework agreement includes, inter alia, that the supplier must be able to use translation memories provided by the public administration and that they have to deliver the produced translation memories upon request of the public administration without extra charge (Kammarkollegiet, 2005, p.3). Public agencies that outsource translations under the legal framework hold the Intellectual Property Rights to the translation and the translation memories by default and can also include more explicit formulations about their specific needs in the framework agreement. If a public administration decides not to procure translation services through the framework agreement, they are obliged to notify the National Procurement Services of the reason.

Public procurement in Sweden in general follows five fundamental principles. These are: the principle of non-discrimination, the principle of equal treatment, the principle of transparency, the principle of proportionality and the principal of mutual recognition (Public Procurement, 2014). Most of the regulations for public procurement are implementations of the EU directive 2014/24 (European Parliament, 2014).

> **Interesting fact:**
>
> **According to the framework agreement for procuring translation services, TMs must be transferred upon request to the contracting authority.**

Although most translation services are outsourced, some institutions have in-house translation services such as the Swedish Police who exclusively translates in-house and the Ministry for Foreign Affairs who outsources part of their translations. The main reason for the in-house translation services is the nature of the texts that need to be translated, i.e. texts that contain confidential information, which is the reason why these translation memories are not shared with e.g. the National Language Bank. Generally, it is very common to use computer-aided translation (CAT) tools in the translation process, this applies to both in-house translation services as well as most freelance translators and language service providers. Some large commercial language service providers also have integrated machine translation systems.

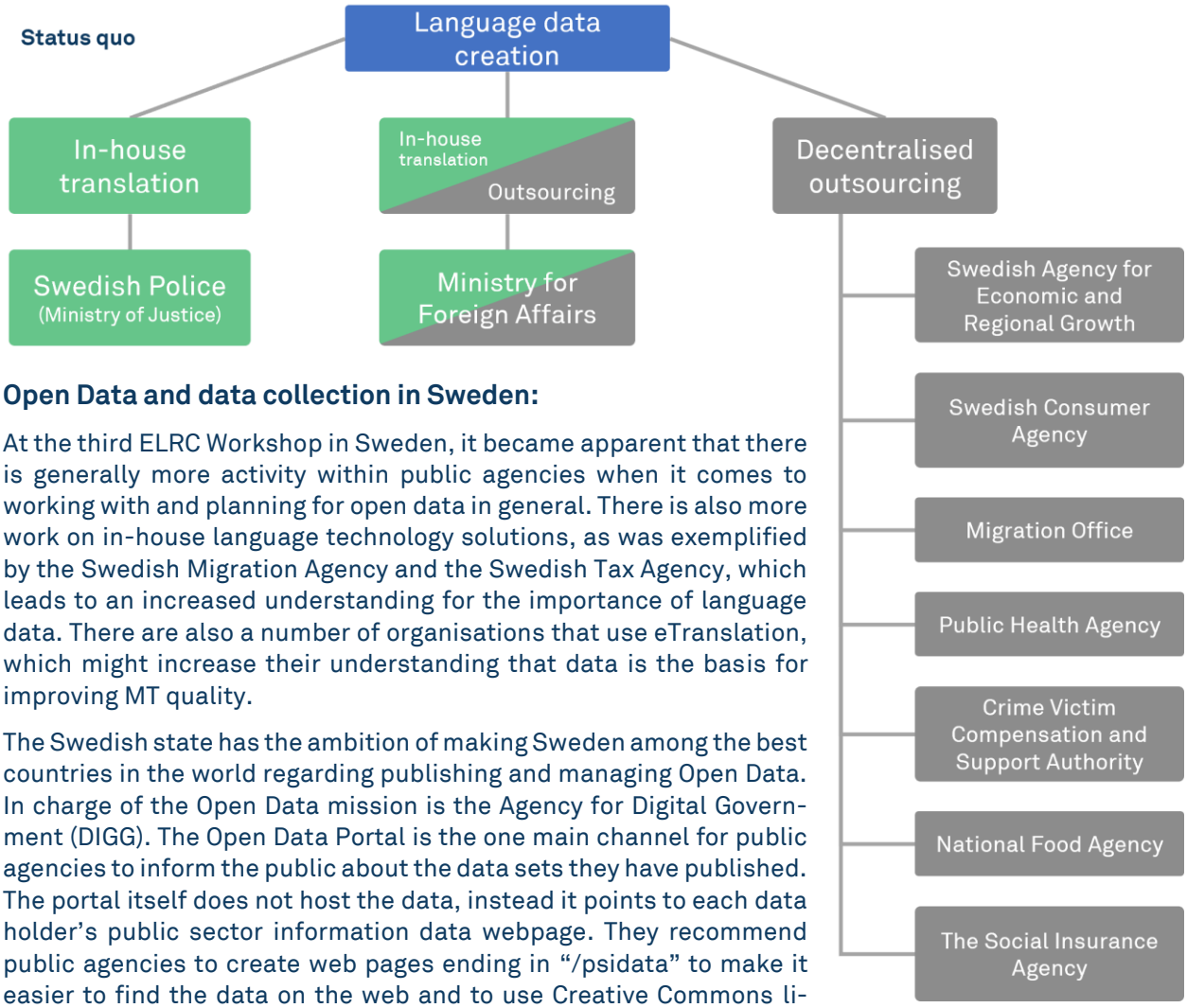
Although the National Food Agency does not have an in-house translation service, they have access to eTranslation, i.e. machine translation on the institutional level to skim texts and documents in other languages or draft texts in languages other than Swedish.

Language data sharing infrastructures in Sweden:

The Swedish Language Council, Språkrådet, subordinated to the Swedish Institute of Language and Folklore (Ministry of Culture), plays a central role in collecting language data in Sweden in order to promote the development of language technology and terminology, as stated by the instruction in the regulation for the Institute of Language and Folklore (Sveriges Riksdag, 2007). For that purpose, texts and terminologies are regularly fed into Nationella Språkbanken, the National Language Bank of Sweden, who then share the data with the ELRC-SHARE repository. In its function as national coordinator of terminology, the Language Council manages the national termbank and also supports other public agencies in their terminology management. The Language Council envisions that this pipeline will be employed for any language resources, including translation memories.

A network consisting of representatives from a few different public agencies has been established to actively put this vision into practice and to discuss measures that will make language data sharing in the future easier.

The current language data creation and sharing infrastructure in Swedish public bodies looks as follows:



Open Data and data collection in Sweden:

At the third ELRC Workshop in Sweden, it became apparent that there is generally more activity 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 a number of organisations that use eTranslation, which might increase their understanding that data is the basis for improving MT quality.

The Swedish state has the ambition of making Sweden among the best countries in the world regarding publishing and managing Open Data. In charge of the Open Data mission is the Agency for Digital Government (DIGG). The Open Data Portal is the one main channel for public agencies to inform the public about the data sets they have published. The portal itself does not host the data, instead it points to each data holder's public sector information data webpage. They recommend public agencies to create web pages ending in "/psidata" to make it easier to find the data on the web and to use Creative Commons licences for sharing data.

Interesting fact:

Public agencies are encouraged by the national Open Data Portal to use the ending "/psidata" for web pages containing public sector information.

Digital policy and language policy in Sweden:

Although the Swedish Language Act was established fairly recently, Sweden has a long tradition of language planning. The Swedish Academy for example was founded in 1786 to "advance the Swedish language and Swedish literature"[105] and the precursor to the current Language Council has been working with language planning and cultivation for Swedish since 1944.

In 2005, the Swedish Parliament adopted a bill addressing that the four main objectives of its concerted language policy are:

[105] Cf. The Swedish Academy Website.

- Swedish is to be the main language in Sweden.
- Swedish is to be a complete language, serving and uniting society.
- Public Swedish is to be cultivated, simple and comprehensible.
- Everyone is to have a right to language: to develop and learn Swedish, to develop and use their own mother tongue and national minority language, and to have the opportunity to learn foreign languages (Lindberg, 2007, p.74).

The importance of language technology and language data collection for the Swedish language is also acknowledged in the government bill:

Central to promoting good development in the language technology area is to systematically build up large text and speech databases and to develop software. Text and speech databases store very large amounts of authentic spoken and written language in a way that makes it accessible for computerised, linguistic analysis. Such an analysis, in turn, is a prerequisite for developing programmes for automatic translation, for transmitting text to speech (and vice versa), for computerised speech recognition, etc. The construction of text and speech databases is costly and labour-intensive and requires long-term planning and is about creating basic language technology resources to develop well-functioning language technology. We therefore believe that a function for coordination of language technology should exist with the new language care organisation so that resources can be better coordinated and the conditions for participating in major collaboration programmes in the Nordic countries and the EU is improving (Sverigs Rigsdak, 2005, p. 30).

It was recognised that to meet these objectives, a coordinated effort was needed and consequently, the Swedish Language Council, the national language planning authority, was founded in 2006 (cf. Lindberg, 2007, p.74). In 2009, the Swedish Language Act entered into force, establishing Swedish as "the principle language in Sweden" that must be usable and therefore have specialist terminology in all different areas of society (Swedish Language Act, p. 1). Five other languages have been granted minority status, these are: "Finnish, Yiddish, Meänkieli (Tornedal Finnish), Romany Chib and Sami." (ibid., p. 2).

The Language Act has 15 sections addressing the Swedish language and its status, national minority languages, Swedish sign language, the use of language in the public sector, Swedish in international context and Individuals' access to language (cf. Swedish Language Act, p. 1 f.). The sections addressing the public sector impose responsibility on the public sector to use, develop and cultivate Swedish while focusing on simple and comprehensible language. The status of Swedish as an official EU language is also addressed in the language act and is deemed important to be safeguarded (ibid., p. 3). At the same time, the public sector is also obliged to "protect and promote the national minority languages" (ibid., p.2 f.) and to ensure that "the individual is given access to language" including official minority languages as well as other first languages spoken by residents in Sweden (ibid., p3). The Language Act is also monitored by the Swedish Language Council.

In 2008, the National Language Bank became a national research infrastructure (2017-00626) funded by the Swedish Research Council by about 1.5 million EUR per year until 2025. The overall budget including co-financing is about 3 million per year. The Language Bank is divided into three divisions: Text, Speech and Sam (for Society) and supports all fields of research related to language data including e.g. language technology, digital humanities or artificial intelligence. The SWE-CLARIN consortium including 10 organisations is part of the research infrastructure. In 2011, the Swedish government stated in the Digital agenda for Sweden (cf. Digital Agenda, 2011) that a National Language Bank is an important infrastructure for the development of language technology.

The role of LT and language data in Sweden's AI regulations

The importance of language technology and language data collection for the Swedish language is acknowledged in the government bill. The Swedish strategy emphasises the need for a digital infrastructure to harness the opportunities that AI can provide, including both a high-quality data infrastructure and a well-developed digital and telecommunication infrastructure in terms of computer power, connectivity and network capacity. The AI Sweden programme covers both the development of the data infrastructure – by improving data quality, data availability and data sharing opportunities, and the setting up of the IT infrastructure.

The AI Sweden programme makes data sets accessible via the Data Factory[106], which aims to provide horizontal resources to all research partners, ensuring that data sets are available across industries and application areas in order to accelerate AI innovation and applications.

With regard to AI in the public sector, the Swedish Agency for Digital Government (DIGG) supports AI uptake and deployment in public administrations. In the policy report from the 14th of January 2020[107], the agency points out that the economic gains resulting from AI could be potentially large for the Swedish public sector. The report includes a mapping exercise on how the Swedish public sector currently uses AI, and it presents suggestions to increasingly use AI in the future. The DIGG is also supporting open data policies to foster data-driven innovations and technology developments.

Stakeholders and major networks:

Among the main stakeholders for language data collection and sharing are the Agency for Digital Government, subordinated to the Ministry of Infrastructure who is responsible for Open Data collection in Sweden and the Institute of Language and Folklore, subordinated to the Ministry of Culture, who is in charge of monitoring the language policy and who is the national terminology coordinator. As such they closely cooperate with the National Language Bank as well. Over 40 institutions have participated in ELRC events and several public administrations that continuously create language resources through outsourcing translations have already contributed language data to ELRC. Some of the data donors are the Swedish Agency for Economic and Regional Growth, the Public Health Agency, the Migration Office, the Consumer Agency, the National Audit Office, and the Crime Victim and Support Authority. It should be noted that not all data shared with ELRC is Open Data.

Language Technology in Sweden:

The third ELRC Workshop, which took place in November 2020 showed that a lot has changed since the first workshop in 2016. It used to be the case that representatives from public agencies were informed about the 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 licences, 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 multilingual environment. Representatives from the Swedish Food Agency stated that using eTranslation is a time-saving solution to e.g. share information with international project groups, translate content relating to the EU Rapid Alert System for Food and Feed or to translate letters related to export issues. In the case of the Swedish Tax Agency, chatbots, automatic classification of incoming emails and machine translation are already in use. However, all 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.

Main challenges for sustainable data sharing:

In the past years, several projects and initiatives aimed at collecting and sharing data in Sweden have increased the awareness of the value of language data for promoting the languages of Sweden and improving the efficiency of public services. Yet, there are still some challenges that need to be overcome:

- During the third Swedish ELRC Workshop, it was stated that there is a general lack of knowledge when it comes to copyright and licencing of language data created at public agencies
- Closely related to that, a second challenge was identified in 2020: Text data requiring machine translation is very often sensitive data, making it already difficult to share it within the organisation. As a consequence, it is considered impossible to share the data outside of the organisation.

[106] https://www.ai.se/en/data-factory
[107] https://www.digg.se/download/18.79c61f7c17db5871992f0ad/1647952779554/framja-den-offentliga-forvaltningens-formaga-att-anvanda-ai.pdf

- Although a considerable number of public administrations have already shared their language resources, language data are still undervalued.
- The benefits of sharing data should be more tangible and address the needs and efforts of public agencies, these however have to be identified first.
- Since most translations are outsourced, guidelines and expertise are necessary on how to request translations with maximum mutual benefit for both the contracting authority and the contractor.
- Although significant progress has been made in reaching out to public administrations and convincing them to share their language resources, the processes of continuous data sharing from public administrations to a central data bank is not yet defined. This includes the licencing of data sets. Currently, different authorities use different licences without consensus on which ones to use.
- Current translation practices do not allow for language data sharing as translations that contain personal or confidential data are not separated from translations that fall under the public sector information directive and can be safely shared.

Action plan:

The formal recommendations for sharing and licencing 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. The recommendations initially lacked (i) concrete examples of common types of text data produced at agencies and their recommended licencing, (ii) information on licencing of terminology resources. In accordance with our feedback, the recommendations now include examples of different information types produced at public agencies, including different types of documents and databases. We also plan to supplement the recommendations with even more detailed information on texts and terminologies, when contacting agencies regarding resource 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 licence it, might be a first step towards creating procedures for how to handle this separation.

To make data sharing easier, it would be important to keep following the objectives below:

- **Promote Language Technologies for the languages in Sweden including minority languages according to the national language policy:**
  This is the main objective for Sweden and all other objections, actions and goals are subordinated to this main objective.
- **Raising awareness of language data as Open Data:**
  In order to encourage more public administrations to share their language data, practical guidelines have been published that will be disseminated to the agencies.

- **Increasing interest in MT in public services:**
  By identifying specific needs that can be addressed through machine translation or language technologies and creating synergies between different actors and initiatives, the potential and benefits can be showcased.
- **Tackle legal concerns:**
  Public administrations are uncertain about how to handle legal concerns relating to language data. In practice, it is not clear what licence to use. This has been clarified in the national guidelines that will be disseminated to the agencies.
- **Identify and gain access to outsourced translations:**
  The first step is to further clarify the nature of the translation contracts and to collect best practices. On that basis, changes can be discussed and introduced. Employees of contracting authorities that outsource translations also need to be advised on how to procure translation services.
- **Establish good data management practices in public services:**
  The current data management practices need to be further investigated along with resources. These activities have already started. There have also been discussions in the network of representatives from public administrations about how to introduce separation between confidential and private data from public sector information.

References and links:

Avropa: Framework agreements, https://www.avropa.se/topplankar/In-English/.

National Agency for Public Procurement: https://www.upphandlingsmyndigheten.se/en.

National Language Bank of Sweden: http://www.sprakbanken.se/eng.

SWE-CLARIN: https://sweclarin.se/eng/about.

Swedish Open Data Portal: www.dataportal.se.

The Swedish Academy: https://www.svenskaakademien.se/en.

[Borin et al., 2012] Borin et al.: *The Swedish Language in the Digital Age*. In: The META-NET White Paper Series, 2012, http://www.meta-net.eu/whitepapers/volumes/swedish.

[Digital Agenda, 2011] Ministry of Enterprise, Energy and Communications: ICT for Everyone – A Digital Agenda for Sweden, 2011, https://de.scribd.com/document/89103979/A-Digital-Agenda-for-Sweden.

[Domeij et al., 2018] Domeij et al.: *Enhancing Information Accessibility and Digital Literacy for Minorities Using Language Technology – the Example of Sámi and Other National Minority Languages in Sweden*. In: Perspectives on Indigenous writing and literacies, 2018, https://brill.com/view/title/31954.

[ELRC, 2021] The Institute for Language and Folklore: ELRC Workshop Report for Sweden, 2021, https://lr-coordination.eu/sites/default/files/Sweden/ELRC3_Workshop%20Report%20Sweden%20Public.pdf.

[European Parliament, 2014] Directive 2014/24/EU of the European Parliament and of the Council of 26 February 2014 on public procurement and repealing Directive 2004/18/EC: https://eur-lex.europa.eu/legal-content/EN/TXT/?uri=CELEX:32014L0024.

[Kammarkollegiet] Kammarkollegiet: *Procurement agreements for Translation and Language services*, https://www.avropa.se/ramavtal/ramavtalsomraden/Ovriga-tjanster/oversattning-och-spraktjanster.

[Lindberg, 2007] Lindberg, Inger: Multilingual Education: a Swedish Perspective, 2007, https://www.srii.org/public/documents/transactions/transaction-18/66f9f6e9-c2d7-47ab-8b37-c37c6fb1471a.pdf.

[Public Procurement, 2014] National Agency for Public Procurement: Sustainable Public Procurement, 2014, https://www.upphandlingsmyndigheten.se/en/publicprocurement/about-the-public-procurement-rules/.

[Swedish Institute, 2016] The Swedish Institute: The Swedish Language, 2016, https://sharingsweden.se/app/uploads/2016/10/The-Swedish-language_high-res.pdf.

[Swedish Language Act] Ministry of Culture: Swedish Language Act, https://www.regeringen.se/contentassets/9e56b0c78cb5447b968a29dd14a68358/spraklag-pa-engelska.

[Sveriges Ricksdag, 2007] Sveriges Riksdag: Instruction 2007:1181, https://www.riksdagen.se/sv/dokument-lagar/dokument/svensk-forfattningssamling/forordning-20071181-med-instruktion-for_sfs-2007-1181.

[Sveriges Ricksdag, 2005] Sveriges Riksdag: Kulturutskottets betänkande 2005/06:KrU4, https://www.riksdagen.se/sv/dokument-lagar/arende/betankande/basta-spraket---en-samlad-svensk-sprakpolitik_GT01KrU4.

Annex

Country Profile The Netherlands

State of Play:

Translation practices and information exchange in ministries and public administrations[108]:

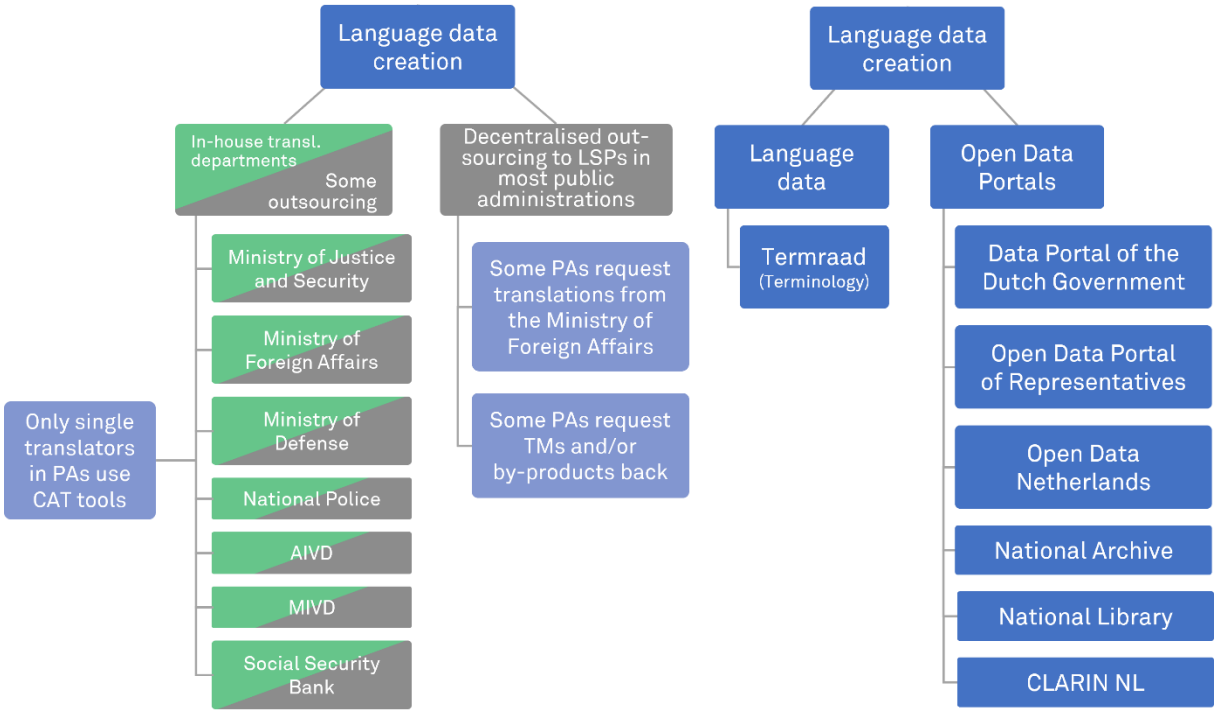
In public administrations in the Netherlands, the creation of translations is predominantly decentralised and often, the translation process is not centralised even within one ministry. Overall, eight ministries and executing bodies have their own in-house translation department, namely: The Ministry of Foreign Affairs, the Ministry of Justice and Security, the Ministry of Defence, the National Police, the General information and Security Service (AIVD), the Military Information and Security Service (MIVD), the Social Security Bank (SVB), and the Employee Insurance Agency (UWV). The translation department of the Ministry of Foreign Affairs sometimes translates documents for other ministries, but only if the documents are considered confidential at the moment of translation (they may become public at a later stage).

For translating, a computer-assisted translation software suite is often used in the Dutch Ministries. With regard to machine translation (MT), CEF eTranslation and other freely available online translation tools are used by the Ministry of Foreign Affairs. The use of CEF eTranslation is not widely spread though in the Dutch Ministries.

Most documents that need to be translated are speeches, speaking notes, memos, diplomatic cables, as well as some proposals for tender (especially for embassies). Language pairs which are frequently requested include NL<>EN (45%), NL<>FR (30%) and NL-Other EU languages (ES, IT, FR – 20%). Emails are also often translated by individuals using common online translators or other freely available MT services. Until now, there is no organised or centralised exchange of language data on the national level. However, within the inter-institutional Termraad, attempts for direct collaboration on the terminological harmonisation have been made between the translation services of the various European institutions and the Dutch, Belgian and Flemish authorities.

[108] The information in this section is taken from the 2019 Country Profile for the Netherlands. Lack of a Public Sector NAP in the Netherlands makes it difficult to find and address the right people. Consequently, it has not been confirmed whether the text about translation practices and information exchange in ministres and public administrations still reflects the current situation

The current language data creation and sharing infrastructure in Dutch public bodies looks as follows:



Open Data and data collection in The Netherlands:

In the Netherlands, KOOP (Kennis- en Exploitatiecentrum Officiële Overheidspublicaties – Knowledge and Expertise Centre Official Government Publications), which is placed under under the umbrella of the Ministry of Interior and Kingdom Relations, serves as the official publisher of the central and local governments of the Netherlands. These publications can be found online on overheid.nl, data.overheid.nl, and officielebekendmakingen.nl. Law texts are available through wetten.overheid.nl. The platform data.overheid.nl is the data register of the government of the Netherlands. On the one hand, it provides access to a catalogue of all data sets (open and closed), and on the other hand, it functions as a data broker to help finding and disclosing hidden data sets. Most data is, however, not natural language text.

So far, no systematic attention was given to multilinguality of Open Data or the availability of language data as Open Data in the Netherlands. Dutch is the only official language for official publications in the Netherlands and the need for translated national legislation is often underestimated. If translations are made, often by third parties, they are not easily retrievable (e.g. there is no clear link between the translation and the publications on the websites of the central and local governments). Two legislations are important in the context of Open Data. First, "Wet open overheid" (Open Government Act) which implies an active disclosure of "everything" via PLOOI (PLatform Open Overheid Informatie by KOOP). Second, "Wet Elektronische Publicatie" (Electronic Publication Act) which implies that all legislation and regulations become available online.

In addition to Open Data initiatives, the Netherlands place great emphasis on the digitisation of public administrations and services. In addition to PIANOo (the Dutch Public Procurement Expertise Centre), the Routing Institute for National and International Information Streams (RINIS) is the hub for fully-automated electronic data exchange in the public domain and tries to harmonise the exchange of information and corresponding infrastructure on the national level.

The Netherlands also play an important role in CLARIN and data, tools and services to support research on language resources are available through repositories at CLARIN B-centres in the Netherlands[109]. Another important source for Dutch language materials can be found at Taalmaterialen[110] from the

[109] https://www.clarin.eu/content/certified-b-centres
[110] https://taalmaterialen.ivdnt.org/

Dutch Language Institute. This catalogue contains resources, data and tools for linguistic research and language and speech technology within the Dutch language area.

During the last ELRC Workshop it was noted that collaboration between different platforms and infrastructures is key to future service provision and data sharing.

Digital policy and language policy in The Netherlands:

The language policy of the Netherlands has been outsourced to the Dutch Language Union (the same is true for Flanders and Surinam). The Dutch Language Union decides on the official spelling of words, for example.

In 2018, the State Secretary for Economic Affairs and Climate Policy, the Minister of Justice and Security and the State Secretary for the Interior and Kingdom Relations presented the Dutch Digitisation Strategy 2018-2021. With this strategy, the government wants to maintain the position of the Netherlands as a digital frontrunner in Europe. The Strategy is updated annually and in 2021 a Parliamentary Standing Committee on Digital Affairs has been established. The aim of the committee is to create an overview and to ensure connections in the handling of digital affairs in the various policy areas.

The role of LT and language data in Netherlands AI regulations:

In autumn 2019, the NL AI Coalition has been set up to substantiate and stimulate AI activities in the Netherlands. It is a public-private partnership in which the government, the business sector, educational and research institutions, as well as civil society organisations collaborate to accelerate and connect AI developments and initiatives in the Netherlands. NAIN (Netherlands AI for the Dutch language) is one of the use cases specifically focusing on language and speech technology. The aim of the project is to join forces and not to be dependent on the arbitrariness of large foreign commercial parties. Another initiative is the Nederlandstalige Spraak Coalitie (Dutch Speech Coalition) which aims to develop speech technology in the Dutch language area, together with various organisations, companies, universities and institutions, as a public-private partnership.

Stakeholders and major networks:

So far, more than 120 potential stakeholders have been identified for the Netherlands, most of them being holders and creators of language resources. 65 of them participated in the last ELRC Workshop on 11 June 2021. So far, 145 language resources have been contributed to the ELRC-SHARE repository including Dutch[111]. Main potential beneficiaries include: UWV, SVB, Ministry of Foreign Affairs, Ministry of Justice and Security. In the area of research, the Netherlands are involved in ELG, ELE, CLARIN, DARIAH, and various COST actions.

Main challenges for sustainable data sharing:

- **Legal concerns/lack of explicit mission to share language resources:**
  There is a hesitation in ministries and public administrations to share translations for various reasons including translator's rights to the texts.
- **Unavailability of translated texts:**
  When creating a translation within the public administrations, a lot of the work is not so much to actually translate texts, but rather to create a new text in a different language with contents similar to the source but without direct translation of this source. Moreover, the sources are not even text in all cases. This process generally involves a good portion of localisation.
- Last but not least, the vast majority of translations are being outsourced without transfer of respective translation memories.

[111] For application and development of LT tools at national level, the distinction between Dutch as used in Belgium and Dutch as used in the Netherlands is important. This distinction is needed as both countries have their own terms for specific concepts. This distinction may not be important at European level, but it is important at the national level. At the last ELRC Workshop it was noted that it would be good if data repositories could include this information in the metadata to increase reusability of the data.

With regard to language data creation, management and sharing practices, this situation has not changed significantly since 2019. Legal issues as well as the absence of data management practices (or even guidelines governing the sharing of language data) in (public) services remain the main barriers hindering the sharing of language data in the Netherlands.

Action plan:

Taking into account the main challenges in The Netherlands, corresponding actions to enable / improve the sharing of language resources focus on:

- Establishing good data management practices in public services
- Tackling legal concerns that may prevent the sharing of language resources and
- Identifying and gaining access to outsourced translations.

Especially with regard to the latter, the procurement policy needs to be changed and TMs need to be transferred to the contracting authority. As regards the tackling of legal concerns, a corresponding EU-wide initiative may help.

Another important action is to increase interest in MT/LT in public services as part of the national digital policy. This includes on the one hand establishing synergies with related national projects and initiatives (which is part of national implementation of Regulation (EU) 2018/1724). On the other hand, it involves securing the support of decision makers to change/adapt national policy. National initiatives for the active promotion of CEF eTranslation (and other CEF tools) need to be installed. In this context, an investigation and investment in a national Proof of Concept where the functionality of the CEF eTranslation building block for a department of the national government will be tested seems advisable. It should be noted that important criteria for using LT tools in the public sector are security and privacy. These aspects should be emphasised in promoting CEF eTranslation as an alternative to e.g. Google.

It is of utmost importance to raise awareness of language data as Open Data and a valuable asset. This includes, above all, the integration of language data in the national Open Data policy/digital agenda, with accompanying relevant metadata (e.g., language of the text, explicit relation between source and translated texts). Various steps into this direction are taken.

References and links:

Data Portal of the Dutch Government ('Open Data van de Overheid'): https://data.overheid.nl/.

Dutch Public Procurement Expertise Centre: https://www.pianoo.nl/en/public-procurement-netherlands.

Dutch Speech Coalition ('Nederlandstalige Spraak Coalitie'): https://www.spraakcoalitie.nl/.

Employee Insurance Agency (UWV): https://www.uwv.nl/overuwv/wat-is-uwv/index.aspx.

General information and Security Service (AIVD): https://www.aivd.nl/.

Knowledge and Expertise Centre Official Government Publications ('Kennis- en Exploitatiecentrum Officiële Overheidspublicaties – KOOP'): https://www.koopoverheid.nl/.

Military Information and Security Service (MIVD): https://www.defensie.nl/onderwerpen/militaire-inlichtingen-en-veiligheid.

National Archives ('Nationaal Archief'): https://www.nationaalarchief.nl/en.

National Library Lab ('Koninklijke Bibliotheek' – KB Lab): https://lab.kb.nl/.

National Open Data Portal: https://data.overheid.nl.

NL AI Coalition ('NL AI Coalitie'): https://nlaic.com/.

Netherlands AI for the Dutch language ('Nederlandse AI voor het Nederlands' (NAIN)): https://nlaic.com/use-cases/nain-nederlandse-ai-voor-het-nederlands/.

Open Data Netherlands ('Open Data Nederland'): https://opendatanederland.org/.

Open Data Portal House of Representatives ('Open Data Portaal van de Tweede Kamer'): https://opendata.tweedekamer.nl/.

Routing Institute for National and International Information Streams (RINIS): https://www.rinis.nl/en/.

Social Security Bank (SVB): https://www.svb.nl/en/.

Wettenbank Overheid: https://wetten.overheid.nl.

[ELRC, 2018] Odijk, Tiberius: *ELRC Workshop Report for the Netherlands*, 2018, https://www.lr-coordination.eu/sites/default/files/Netherlands/2018/ELRC%2B%20Workshop%20Report%20Dutch_Public_final.pdf.

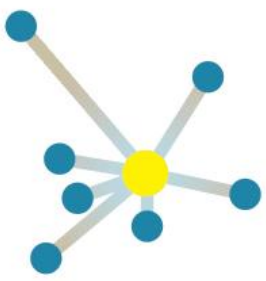
[ELRC, 2021] Tiberius, Odijk: *ELRC Workshop Report for the Netherlands*, 2021, https://www.lr-coordination.eu/sites/default/files/Netherlands/2021/ELRC3_Workshop%20Report%20The%20Netherlands%20Public_vfinal_updated.pdf.

[Taalunieversum, 2017] Taalunieversum: *Veelgestelde vragen over ons taalbeleid (Frequently Asked Questions about our Language Policy)*, http://taalunieversum.org/inhoud/veelgestelde-vragen-over-ons-taalbeleid#t560n4260.

[Verhagen, 2019] Verhagen, Michel: *Language services at the Ministry of Foreign Affairs – Hurdles for sharing data*, 2019, http://lr-coordination.eu/sites/default/files/LRB%20Nice/2019/8th%20LRB%20meeting_Language%20services%20at%20MFA_Verhagen.pdf.

Language Data Matters

ISBN-13: 978-3943853070



European Language
Resource Coordination

Connecting Europe Facility