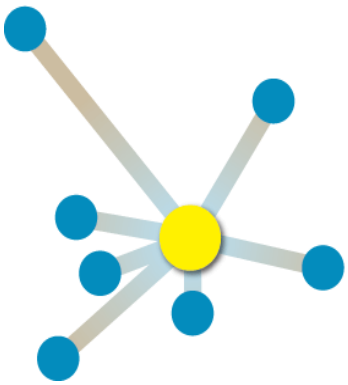
# Deliverable D11.2

# Report on the 6th ELRC Conference



European Language Resource Coordination

Connecting Europe Facility


| | |
|---|---|
| Author(s): | Andrea Lösch (DFKI)<br>Eileen Schnur (DFKI) |
| Dissemination Level: | Public |
| Date: | 2022-04-19 |
| Copyright: | yes |


For copies of reports, updates on project activities, and other project-related information, please contact:


Dr. Simon Ostermann
Stuhlsatzenhausweg 3
Campus D3_2
D-66123 Saarbrücken, Germany

simon.ostermann@dfki.de
Phone: +49 (681) 85775 5310


European Commission

# Contents

# 1 Introduction

This deliverable reports on the 6th ELRC Conference. The virtual event took place on 31 March 2022 via Zoom and was also livestreamed on YouTube.


**Figure 1: Visual 6th ELRC Conference**

The deliverable at hand is structured as follows: First, it will describe the goals and objectives of the conference as well as the target audience and give an overview on the structure and organisation of the event. This will be followed by a summary of the presented contents as well as the questions and discussion points raised during the conference day.

All presentations are available on the ELRC website at https://lr-coordination.eu/6thELRC. The full recording of the conference can be accessed via YouTube: https://youtu.be/ebAbv5KgvrQ.

# 2 Focus and Contents of the Conference

## 2.1 Context

In recent years, research on language-centric Artificial Intelligence (AI) "made in Europe" has become increasingly important. Efforts have been maximised to work towards Europe's independence of non-EU players. This is not only important for the industry and autonomy of the EU but also for the security and privacy of all EU citizens. However, this goal can only be achieved by joining forces and driving the change together – by thinking big for Europe's multilingual future.

Centring around this motto, the 6th edition of the ELRC Conference hence focused on the latest European solutions and activities related to Large Language Models, multimodal language data, and low-resource neural machine translation (NMT). In addition, it put a spotlight on the European Language Data Space, which is foreseen as part of the freshly introduced Digital Europe Programme. Taking into account the present situation in Ukraine, the conference also supported the "Crisis response without borders" initiative and included a discourse on using Language Technologies (LT) for fighting disinformation.

Overall, the conference had the following key objectives:

- Informing the audience about the benefits of the eTranslation services (available at http://language-tools.ec.europa.eu) and promoting its use
- Sharing latest updates on the ELRC activities
- Raising awareness on the importance of European cooperation in the field of language-centric AI
- Raising awareness on the importance of language data and language technology for the future of Europe – and the future of Artificial Intelligence "made in Europe"
- Demonstrating best practice examples of LT solutions for Europe
- Allowing the participants to extend their networks, exchange know-how and ideas

## 2.2 Target Audience

The conference targeted stakeholders who may benefit from the use of language technology (LT) such as the tools provided by the European Commission at https://language-tools.ec.europa.eu as well as actual and potential data contributors. More precisely, the promotion campaign addressed representatives from the public sector (public administrations and public services) as well as small and medium-sized enterprises (SMEs) from all across Europe. In addition, representatives from research and academia, the language technology industry as well as language service providers were targeted.

In order to reach familiar members of the ELRC community, a save the date and corresponding follow-up emails were shared with the participants of previous events, including the ELRC Conference, ELRC Country Workshops and Technology Workshops. Moreover, the event was promoted on the ELRC website, via the ELRC

newsletter and through the ELRC Social Media channels on Twitter, Facebook and LinkedIn.

In total, 531 people registered for the 6th ELRC Conference. Due to the Zoom limit of 300 participants, participants also received a link to join the event via Youtube Livestream upon registration. This option was set up to ensure that everyone who is interested will be able to participate in the conference, even if the Zoom limit is exceeded. During the conference day, more than 257 participants joined the Zoom meeting[1] and at least 20 attendants followed the livestream[2], adding up to a total number of at least 277 participants. The drop-out rate was 52 %. As of 1 April, i.e. one day after the conference, the livestream was already watched 133 times and shared 7 times.

## 2.3 Focus and Contents

Overall, the 6th ELRC Conference was held under the motto “Think big. For Europe’s Multilingual Future”. The event centred around the following key topics:

- Latest achievements and plans of ELRC
- The upcoming Language Data Space
- Large Language Models
- Low-Resource NMT
- Multimodal Data

Due to the Ukraine crisis, the programme was adapted shortly before the conference and a discourse about how language technologies can be used to fight disinformation was included.

The final conference agenda is provided in Figure 2 below. It can also be accessed on the ELRC website at https://lr-coordination.eu/6thELRC.

[1] Ten participants could not be clearly identified because of ambiguities in their user names.
[2] Maximum views at the same time: 21, on average: 12-15 users at the same time. As the users could watch the livestream anonymously, it is difficult to assess the total number of individual viewers.

AGENDA

6TH ELRC CONFERENCE

31 March 2022 | Zoom

**09:30 – 09:45** **Welcome by EC** *(Philippe Gelin, Head of Sector Multilingualism, European Commission)*

**09:45 – 10:00** **Welcome and Introduction** *(Andrea Lösch, ELRC Project Manager, German Research Centre for Artificial Intelligence)*

**10:00 – 11:00** **INSIDE ELRC AND DIGITAL EUROPE**

- **10:00 – 10:20** **ELRC Update** *(Andrea Lösch, ELRC Project Manager, German Research Centre for Artificial Intelligence)*
- **10:20 – 10:40** **ELRC White Paper: Language data sharing & language-centric Artificial Intelligence** *(Eileen Marra, ELRC Communication Manager, German Research Centre for Artificial Intelligence)*
- **10:40 – 11:00** **Quo vadis? The European Language Data Space** *(Philippe Gelin, Head of Sector Multilingualism, European Commission)*

***11:00 – 11:10 Coffee Break***

**11:10 – 12:10** **SPOTLIGHT: LARGE LANGUAGE MODELS FOR EUROPE**

- **11:10 – 11:30** **Why Europe needs large language models. An economic perspective** *(Jörg Bienert, Chairman of the German Federal Association for AI)*
- **11:30 – 11:50** **Possibilities and Limitations of Large Language Models: PAGnol, VLM-4 and Muse** *(Igor Carron, CEO, LightOn)*
- **11:50 – 12:10** **AI Sweden: Language Models for Swedish Authorities** *(Magnus Sahlgren, Head of strategic initiative on NLU, AI Sweden)*

**12:10 – 12:30** **DISCOURSE: Language Technologies for fighting disinformation** *(Nikos Sarris, Senior Researcher, Centre for Research & Technologies Hellas – Maria Bielikova, Director-General, Kempelen Institute of Intelligent Technologies)*

**12:30 – 12:45** **Wrap-up and Outlook** *(Andrea Lösch, ELRC Project Manager, German Research Centre for Artificial Intelligence)*

***12:45 – 13:30 Lunch Break***

**13:30 – 14:30** **SPOTLIGHT: MULTIMODAL LANGUAGE DATA**

- **13:30 – 13:50** **Behind the Scenes: Multimodal Data Analysis** *(Suzanne Little, Associate Professor, Dublin City University)*
- **13:50 – 14:10** **Multimodal Data in the Medical Domain** *(Alexandra König, Neuropsychologist, INRIA)*
- **14:10 – 14:30** **Multi3Generation: Multimodal Data for Natural Language Generation** *(Anabela Barreiro, Researcher, Chair COST Action Multi3Generation )*

**14:30 – 15:15** **BRIDGING THE LANGUAGE GAP: LT SOLUTIONS FOR EUROPE**

- **14:30 – 14:45** **EMBEDDIA: AI Technology for the Media Industry** *(Senja Pollak, EMBEDDIA Coordinator, Jozef Stefan Institute)*
- **14:45 – 15:00** **Microservices at your service** *(Sebastian Andersson, Solution Architect, Lingsoft Oy)*
- **15:00 – 15:15** **ENRICH4ALL** *(Dimitra Anastasiou, Researcher, Luxembourg Institute of Science and Technology)*

**15:15 – 15:30** **Summary and Conclusions** *(Andrea Lösch, ELRC Project Manager, German Research Centre for Artificial Intelligence)*

European Language Resource Coordination
Connecting Europe Facility

**Figure 2: Agenda of the 6th ELRC Conference**

# 3 Synthesis of Discussion Points

## 3.1 Welcome Address by Philippe Gelin

The 6th ELRC Conference was opened by Philippe Gelin, Head of Sector Multilingualism at the European Commission. He started by explaining that even though multilinguality is close to the heart of the European Commission and one of the key values of Europe, it can also create language barriers and avoid the free exchange of ideas and information. Also, he pointed out that the majority of Europeans does not speak a second language well enough to e.g. maintain a conversation, watch a film or sign legal documents. However, this is where AI and LT can be used to support intercultural communication in more and more situations. While it has become possible to translate between languages without major efforts and often with only one simple fingertip, Philippe pointed out that even though current LT can be a useful aid with an obvious economic and social impact, it is not perfect yet and needs further development. Especially when it comes to the coverage of smaller European languages, there is a need for the European Commission to strive towards multilingualism and language equality in the digital world.

The eTranslation language tools provided by the European Commission are one major step towards this goal: The eTranslation machine translation service for example covers all 24 EU official languages, as well as Icelandic, Norwegian, Japanese, Russian, Chinese, Turkish, Arabic and since March 2022 also Ukrainian. The latter was added to the eTranslation offering within only fifteen days after Russia's troops first entered Ukraine in an effort to support cross-border communication between refugees and helpers during the Ukraine crisis. Besides the growing language coverage, there were also extensions in terms of target groups and service offerings. While eTranslation initially targeted the public sector only, it is now also accessible to European universities, research and academia, small and medium-sized enterprises (SMEs) as well as non-governmental organisations (NGOs). In addition to that, the European Commission added numerous LT tools going beyond "simple" machine translation, such as a multilingual tweet creator, speech-to-text or named entity recognition tools. Philippe also explained that with the launch of the Digital Europe Programme and following the EU's strategy for data, the European Commission aims to create a real ecosystem around the data associated to language. The so-called "European language data space" (LDS) that is foreseen in this context will thus not only promote the creation, collection, sharing and reuse of language-related data, it will also help to create, share and reuse computing language models. The speaker concluded his welcome address by saying that further details about the LDS will follow in his presentation that will be part of the thematic block "Inside ELRC and Digital Europe" and by wishing all participants an interesting conference day. No questions or comments were raised after the welcome address.

## 3.2 Welcome and Introduction by Andrea Lösch

In her introductory talk, Andrea Lösch, ELRC Project Manager and leader of the "Data & Resources" Group at the German Research Centre for Artificial Intelligence (DFKI) welcomed the participants and started her talk with some background information about the European Language Resource Coordination (ELRC), its objectives and mission. She clarified that language data is crucial for the development of LT, which are at the same time a key market in and for Europe. Referring to the recent Slator

2021 “Data-for-AI Market Report”, she highlighted that data is the fuel for Europe’s economy, and that it is ELRC’s mission to provide the urgently required language data. While the initial goal of the initiative was to collect language data to develop and train CEF eTranslation, the scope of ELRC emerged with the EC’s extension of language coverage, service offering and user base as described in 3.1. Andrea also highlighted the importance of the eTranslation services by sharing some of the latest usage statistics: By the end of 2021, the European Commission’s translation tool had more than 18.000 individual users and more than 240 million pages were translated using eTranslation. Considering that eTranslation was launched only about four years ago, this makes on average of 60 million translated pages by year. Andrea Lösch continued her introduction by presenting the key objectives of ELRC, which are:

- To identify the needs of public services (and most recently also SMEs) regarding machine translation and language technologies
- To engage with the public sector and SMEs to identify relevant language resources
- To collect these identified language resources
- To provide technical and legal support to facilitate the sharing of language resources
- And finally, to act as observatory for language resources across Europe.

As it is not possible to achieve the above-mentioned objectives without a strong EU-wide cooperation, Andrea introduced the audience to the ELRC National Anchor Points (NAPs) who support the initiative on national level. In each EU Member State, Norway and Iceland there are two National Anchor Points:

- The Technology NAP, who is a highly regarded language or language technology expert. He or she often has a distinguished academic or research background or represents a national language institution. Technology National Anchor Points support the ELRC from a research and development perspective.
- The Public Services NAP, who is a representative of national public services, the national public administration or a ministry. He or she acts as a liaison contact person to the national, regional and local administrations, and is able to effectively mobilise and spread the word about the importance of language resources among the public authorities in each country.

Altogether, the NAPs constitute the Language Resource Board (LRB), the governance body of the ELRC initiative. Further information and contact details of the ELRC NAPs is available at www.lr-coordination.eu/anchor-points.

To conclude, Andrea Lösch explained what the audience can expect from the 6th ELRC Conference and shared some practical hints to avoid technical issues and disturbances during the event. There were no questions or comments associated to this presentation.

## 3.3 Inside ELRC and Digital Europe

### 3.3.1 ELRC Update

Andrea Lösch, ELRC Project Leader continued with an update on recent activities and achievements of ELRC. What started with about 60 NAPs and a small community in 2015 had grown to a network of several thousand data donors, followers and community members of ELRC. Andrea pointed out that this success was only possible

thanks to the NAPs and all their efforts to promote not only ELRC but also the sharing of language data. She continued by explaining that ELRC focuses on the collection of monolingual corpora, bi- and multilingual corpora, but also language and translation models or lexical and conceptual resources such as terminologies or glossaries. Such data is vital for the development of MT. Now, with the upcoming LDS, this initial focus is about to change, and the future will include a much broader collection and identification of language data than currently exists in ELRC. The future focus of data collection will not only be on supporting the development of MT systems, but also of many other language technologies.

The speaker also presented latest figures regarding ELRC's data collection efforts: As of mid-March, almost 4000 unique language resources have been made available via the ELRC-SHARE Repository. Most of these resources are bi- or multilingual corpora (2803), followed by lexical resources (503) and monolingual corpora (192). Andrea highlighted that the collected resources do not only help the European Commission to train and improve CEF eTranslation, but they are also valuable sources for the entire LT community, including developers, translators or language specialists, because more than 80% of the language data can be freely reused by anyone.

Besides the increase of collected language data, there were further new developments in ELRC. Since 2019, the initiative is also supporting the development of LT specifications for the EC, including

- a WordPress plugin for automated website translation using eTranslation
- a solution for anonymising language resources (available here) and
- an investigation on best practices to support fact checkers and detect fake news

Also, ELRC started a social media campaign on Facebook, LinkedIn and Twitter at the end of 2020. What started as a fallback option during the Corona crisis resulted in a powerful means to communicate with the community and to increase the overall visibility of ELRC and the activities related to CEF eTranslation. Overall, more than 8.4 million users could be reached via Facebook within less than two years. Also in light of the current Ukraine war, social media proved to be a valuable means for crisis response. A few days after the beginning of the Ukraine war, ELRC launched the "Crisis Response without borders" initiative on behalf of the European Commission, and called for Ukrainian language data to and from the EU languages to develop automated translation tools for helpers, public services and refugees that are affected by the crisis. This call resulted in an extensive list of useful hints to language resources, offers of data donations and support. There were no questions or comments from the audience after this presentation.

### 3.3.2 ELRC White Paper: Language data sharing & language-centric Artificial Intelligence (AI)

In the following talk, Eileen Marra, ELRC Communications Manager from the German Research Centre for Artificial Intelligence (DFKI) provided an overview on the latest work on the ELRC White Paper and gave some first insights into latest findings. To set the scene, Eileen presented the first edition of the White Paper, which is titled "Sustainable Language Data Sharing to Support Language Equality in Multilingual Europe – Why language data matters" and which was published in 2019. The initial scope of the White Paper was to report on practices, challenges and recommendations for sustainable language data sharing and to give country-specific insights in the

country profiles provided in the annex. With the second release that is planned for autumn 2022, this focus will be broadened by

- Extending the analysis to European SMEs,
- Analysing the use of language technology tools such as anonymisation, classification, etc., and
- Investigating the role of LT and language data in national AI regulations.

Due to this extension, the second white paper will be titled "AI for Multilingual Europe – Why Language Data Matters". To clarify why the ELRC consortium decided for such an extension, Eileen explained that there is an increasing importance of AI across the EU, Iceland and Norway and quoted part of the Irish AI strategy, according to which "AI is not a technology of the future, it is a technology of the present". The fact that 24 of the 29 CEF-affiliated countries have already published their national AI regulations also underlines this statement. The speaker continued by presenting the foreseen approach to update the white paper. The final version will combine the outcomes of 1) the country workshops that were held in all CEF-affiliated countries, 2) the AI Watch Report, 3) a detailed analysis of national AI strategies to see how LT and language data are represented on national level and 4) the ELRC White Paper Survey, which is collecting insights from the ELRC NAPs, but also from small and medium-sized enterprises, LT industry and language service providers. As the NAPs were already invited to participate in the survey, Eileen presented some first findings and tendencies. In summary, the following conclusions could be drawn:

Translation practices:

- Most translations are still being outsourced, but the overall percentage has decreased from 82% in 2019 to 68% in 2022.
- A massive increase of the use of computer-assisted translation (CAT) tools could be observed, growing from 18% in 2019 to 41% in 2022. This is a great achievement, since CAT tools are critical for the creation of high-quality multilingual data and therefore a huge asset to the LT community.
- When it comes to the outsourced translations, it is still not common practice to request translation memories or by-products back by default. Nonetheless, a decrease in cases where they are not requested back at all could be observed (from 45% in 2019 to 33% in 2022).
- There were no big changes regarding the use of MT in public administrations. Even though MT APIs are not common in public administrations yet, only 5% seem to not use MT at work at all, which leads to the assumption that it has become an essential part of their daily work.
- Due to the increasing importance and the intensified efforts in the field of LT, many other tools gained popularity and some of them are even used on a regular basis. According to the preliminary findings of the White Paper Survey, this applies to e.g. Classification and Speech Recognition.

The value of language data:

- The NAPs mentioned "raising awareness of the value of language data" as key priority in 2019 and also in 2022
- According to the initial results of the White Paper Survey, almost 50% of the contributors' organisations store language data like tmx files, translations, audio

files or video recordings whenever possible. Only 22% seem to never or only hardly store data.

- When asked if the value of language data has been recognised in their countries, 43% of the NAPs agreed and 40% disagreed. 17% were unsure and answered with "maybe". The results of the corresponding live poll which was launched at the 6th ELRC Conference to compare these first results are provided in Figure 3 below and show similar discrepancies.

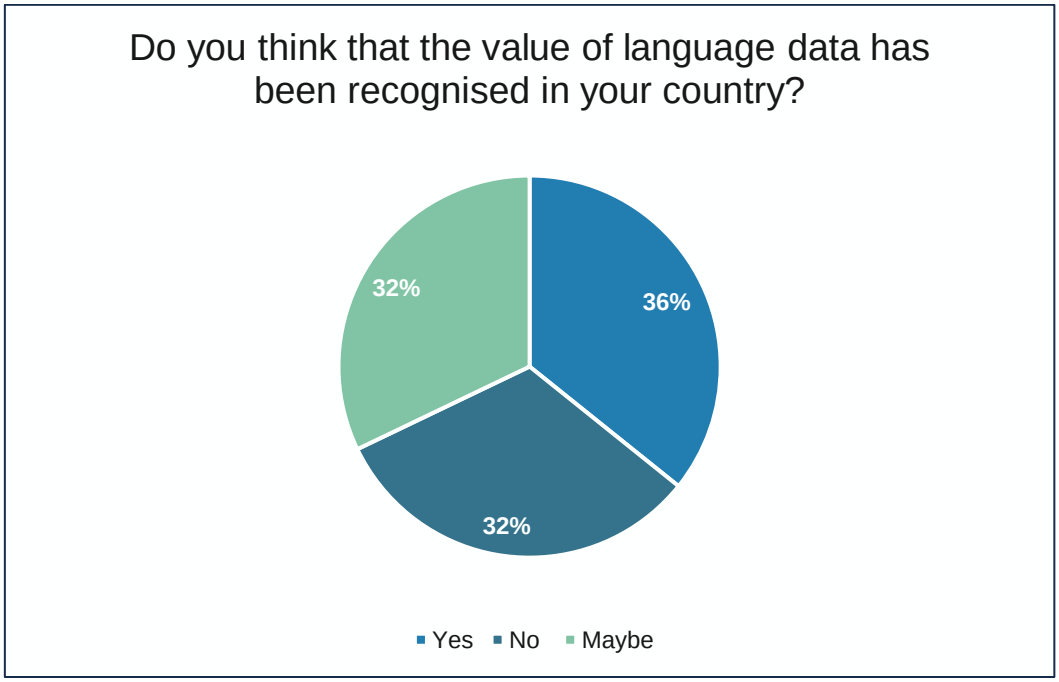



**Figure 3: Results of live poll on value of language data**

- In 11 of the 24 national AI Strategies, Language Resources are explicitly mentioned. Those results underline the overall assumption that the value of language data should become more prominent – within organisations as well as in national regulations.

<u>The role of LT in national policies</u>

- According to the NAPs, "increasing interest in MT/LT as part of the digital policy" is the second most important objective to facilitate language data sharing
- The majority of the NAPs did not agree when asked if LT is appropriately represented in the AI strategy of their country. This also applies to the audience of the 6th ELRC Conference, who were invited to share their feedback via live poll (see below).
- In 21 of the 24 national AI regulations, LT is mentioned but with varying emphasis, ranging from brief side notes to complete chapters or action pillars related to LT[3].

[3] According to the analysis, this applies to Sweden, Estonia and the Netherlands. After the talk, one participant from Estonia pointed out that LT was actually included in the National AI Strategy as well as in the strategic goals of the Language Strategy.

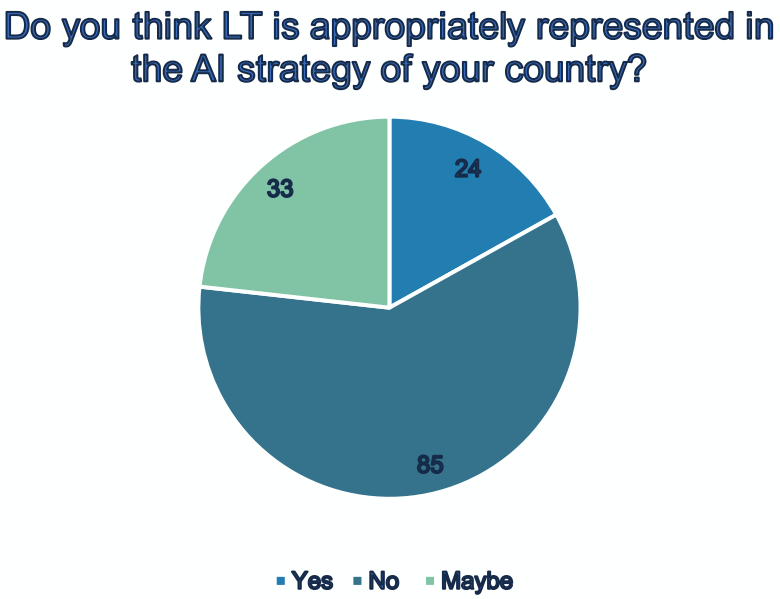



**Figure 4: Results of live poll on value of LT in national AI strategies**

Finally, Eileen concluded her talk with a selection of the NAPs' feedback with regard to related achievements and success stories since 2019. One representative called data collection campaigns like ELRC workshops a major breakthrough with an impact on the perception of digital language data. In addition to that, it was mentioned that MT made enormous progress and that language corpora have become more easily accessible. Furthermore, it was mentioned that even though Covid had serious effects on people all across the world, it also seems to have driven digitalisation forward. Last but not least, Eileen invited the audience to share their views, experiences and visions for the multilingual future of Europe by participating in the ELRC White Paper Survey. No questions were raised by the audience.

### 3.3.3 Quo vadis? The European Language Data Space

This presentation was held by Philippe Gelin, Head of Sector Multilingualism at the European Commission. It centred on the Language Data Space, one of the Common European Data Spaces outlined in the Digital Europe Work Programme 2021-2022. The presenter set the scene by giving some background information on the funding programme that the language data space will be embedded in, i.e. the Digital Europe Programme (DIGITAL). DIGITAL focuses on bringing digital technology to businesses, citizens and public administrations and provides funding for projects in five key areas, i.e.

- Supercomputing or HPS
- Cloud, data and artificial intelligence
- Cybersecurity
- Advanced digital skills and
- Ensuring the wide use of digital technologies across the economy and society

According to the speaker, the vision of DIGITAL is rooted in one of the 'Priorities 2019-2024' set by the EC, i.e. "A Europe Fit for the Digital Age" which focuses on building the digital capacities of the European Union and facilitating the wide deployment of digital technologies in businesses and public administrations. The programme was specifically designed to bridge the gap between digital technology research and market deployment, while supporting Europe's objectives of a green transition and digital transformation, thus strengthening the EU's resilience and technological sovereignty. After presenting the indicative budget foreseen for the creation of data spaces (roughly €205 Mio), Philippe explained the concept behind the "Common European Data Spaces", which are the centre piece of the "European Strategy for Data". These infrastructures bring together relevant data infrastructures and governance frameworks, and aim to facilitate data pooling, sharing and exchanging. The data spaces will include:

- the deployment of data sharing tools and services for the pooling, processing and sharing of data by an open number of organisations, as well as the federation of energy-efficient and trustworthy cloud capacities and related services;
- data governance structures, compatible with relevant EU legislation, which determine the rights of access to and processing of the data in a transparent and fair way,
- improving the availability, quality and interoperability of data – both in domain-specific settings and across sectors.

Together with 14 other sectoral data spaces[4], the language data space will be established as part of DIGITAL. With an indicative budget of 6 million EUR, the call for the LDS procurement action will be opened in 2022. Overall, two work strands are foreseen, i.e. 1) the establishment of the Centre of Excellence for Language Technologies (CELT) and 2) the deployment of the LDS. While the first work strand will focus on aggregating stakeholders and developing a multi-stakeholder data and services governance scheme as well as on elaborating a blueprint of the LDS, the second work strand will centre around the actual deployment on governance, technical as well as on promotional level. The speaker moved on by providing a global visualisation of the LDS, illustrating the stakeholders that should be involved (see Figure 5, left side) and the expected outcomes for the different groups (see Figure 5, right side).

On a final note, Philippe Gelin presented the EDIC, the European Digital Infrastructure Consortium, which was introduced by the Path to the Digital Decade programme. Comparable to a European Research Infrastructure Consortium (ERIC)[5], EDIC will define and operate multi-country projects, which are part of the Digital Decade Strategy. However, according to the speaker, the establishment of an EDIC should be quicker, and its implementation should be more flexible than that of an ERIC. An EDIC needs to be composed of at least three member states and funding will be provided by

[4] The other sectoral data spaces are: Green Deal, Smart communities, Mobility, Manufacturing, Agriculture, Cultural Heritage, Health – Genomics, Health Cancer Images, Media, Financial, Skills, Public Procurement, Security and Law Enforcement, Tourism

[5] An ERIC is a specific legal form facilitating the establishment and operation of Research Infrastructures with European interest.

the member states on the one hand and by directly managed EU instruments on the other. Illustrating the relation between the EDIC and the LDS, Philippe explained that one of the areas of activity of an EDIC is the European common data infrastructure and services, meaning that the LDS would be implemented through an EDIC.

In consequence, the EDIC would both bring together European and member state investments, while coordinating the actual deployment of the LDS. According to the speaker, precise indications on when and how it could come into force were not known yet and further details will follow.

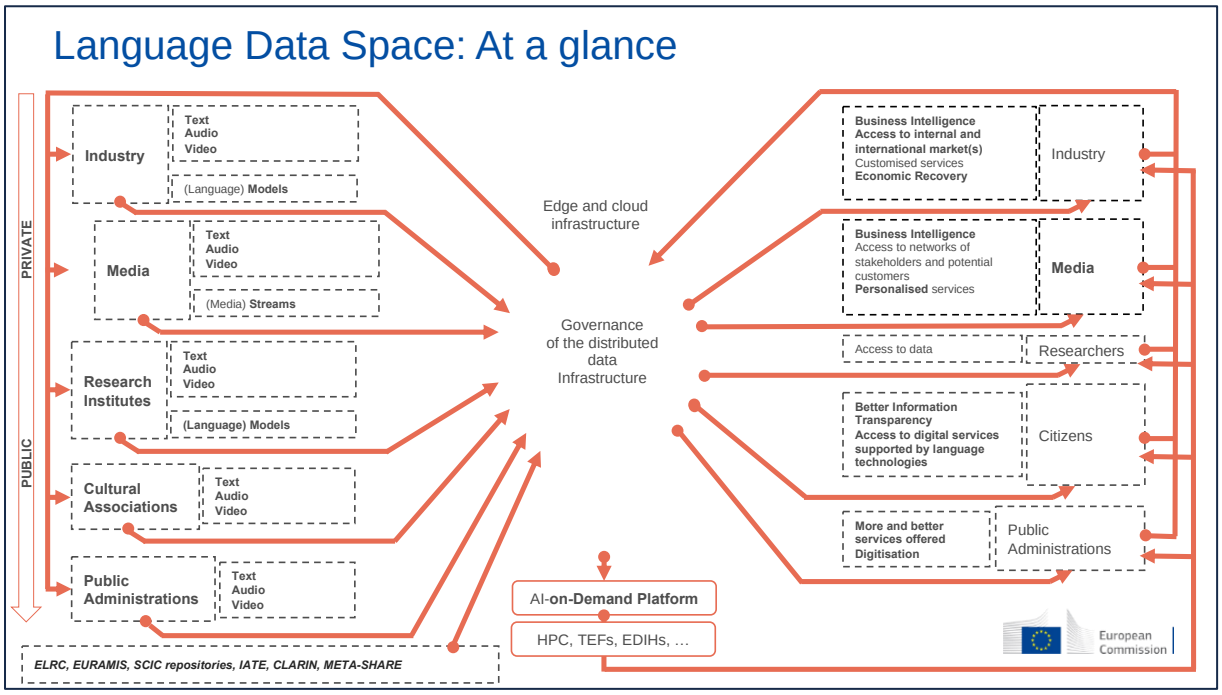



**Figure 5: Language Data Space at a glance**

The presentation was followed by several questions and comments from the audience:

- One of the participants wanted to know when the call for the LDS will be launched. According to the speaker, this is going to happen between June and September 2022. He added that the call will be issued for both work strands, as they will run in parallel. Of course, the EC would promote the call once it has been published.
- Following up on that, moderator Andrea Lösch was interested in the foreseen timing for the deployment of these two work strands, saying that the first phase would need to be completed before the second one. Philippe Gelin hinted on budgetary and timely constraints, at the same time indicating that at least for the language data space and the cultural heritage data space, most of the structures already exist, which is why there will be no need to start from scratch. Of course, the scope will be larger and the LRB would need to be extended, but he highlighted that there is already an existing community, which allows both activities to run in parallel and to work hand in hand, at the same time enabling a fast deployment of the data space.

- One participant wanted to know more about the relation between the LDS (covering all of Europe) and the EDIC (with at least three member states). The speaker explained that timewise, the LDS will be launched without an established EDIC, but that it needs to be adapted once the EDIC structure comes into force. He added that since the creation of an EDIC depends on the member states, the European Commission is not in control of the timeline, but that he would expect it to take longer than three months. Philippe concluded that hopefully, the EDIC will comprise far more than three countries.
- One participant came to the conclusion that the LDS will be developed and launched without a CELT and without an EDIC. In response to that, Philippe explained that the EC aims for a smooth transition to reduce the break in terms of technological support, data deployment, etc., once more highlighting that the LDS will not start from zero, but rather slowly progress and evolve. However, he pointed out that there is a certain need for flexibility, as the LDS will need to adapt to the CELT governance requirement afterwards.
- Moderator Andrea Lösch pointed out that the LDS visualisation contains a section on media even though there will be a separate data space dedicated to the topic. Philippe Gelin explained that the media market is far more complex than language technologies and that in media, there are large producers and users of LT. He added that there will be some overlap with the media data space, but that the focus of the LDS will basically be an extended version of ELRC.
- Closely related to that, one participant was wondering if ELRC would remain or become incorporated into the EDIC. Philippe answered that he sees ELRC as the starting point and that he can imagine the LRB being incorporated into the CELT or EDIC. However, he stressed that it would be in the member states' responsibility to decide on their delegates. He summarised that the integration of the LRB into the CELT or EDIC would make sense, but that the CELT will not only focus on public administration but also on the private sector. In summary, two shifts need to occur: The incorporation of the private sector into the CELT and the extension of the member states' contact points, i.e. not focusing on cultural ministries or the digitalisation ministries only, but having a combination of both.
- Following the question on which industry the CELT should target, the speaker answered that basically, the CELT should approach the media industry and all stakeholders with an interest in LT. This would mean that not even though LT industry will be a key component in this exercise, the scope should not be limited to them, and that major users of LT should also be approached. Research centres and research deployment will also be important, even though DIGITAL will not support research per se but rather the actual deployment.
- In response to the previous questions, the speaker admitted that the structure may look complex, as it needs to match all different sectors. He explained that in the context of the LDS, it may sound like reinventing the wheel, because of the achievements and work within the last 20 years, but for other sectors, the creation of a data space will be totally new and require much more effort. Therefore, he sees a great advantage and assumes that the LDS can be deployed a lot faster than other data spaces. He underlined this statement by saying "the faster you go, the more you can define your own rules".

- Philippe concluded the discussion round by inviting the audience to share the information presented during his talk and by expressing his willingness to present the topic and to promote the LDS at other occasions. Referring to the conference motto “Think big. For Europe’s Multilingual Future”, he once more highlighted that this is also true for the data collected within the LDS, as language resources will only be one part of it. One participant answered that it needs to be big for AI to be as representative as possible and that he likes the element and hopes that sampling measures will be implemented along the way.

## 3.4 Spotlight: Large language models for Europe

### 3.4.1 Why Europe needs large language models. An economic perspective

The spotlight on large language models for Europe was opened by Jörg Bienert, chairman of the German AI Association (KI Bundesverband e.V.), who provided first-hand insights into why Europe urgently needs large language models. The German AI Association was founded in 2008 to foster the usage of AI. Now, with more than 380 members, including SMEs, start-ups and experts focusing on the development and application of AI technology, the German AI Association is the largest AI entrepreneur network in Germany. At the same time, it is broadly connected within Europe thanks to the “European AI forum”, a network of eight national AI associations from the EU.

Moving to the main topic of the spotlight, speaker Jörg Bienert explained that AI will shape the world of tomorrow and have a great impact on innovation and future economic growth. According to him, the next stage of AI has approached, which was mainly caused by the release of the OpenAI GPT-3, the first very large AI model with more than 175 billion parameters. This large neural language model raised attention in the business area and led to increasing investments, new NLP applications and a growing ecosystem. According to the speaker, however, GPT-3 was just the beginning of the race on large AI models, as other international players from China and the United States intensified their efforts shortly afterwards. He also mentioned an example of Germany, i.e. the company Aleph Alpha which is focusing on building large language models, at the same time admitting that the efforts made within the EU are far away from achieving the complexity and size of the models developed by key players from the outside. He added that large AI models have the potential to outperform and gradually replace other AI solutions in a variety of areas, such as text or picture generation, chatbots, translation, business processes or fake news detection. Jörg Bienert also explained that Europe is lagging behind due to the lack of computing power and capabilities to build such large models, which may put its digital sovereignty in the field of AI at risk for various reasons:

1) GPT-3 was not released as open source and can only be accessed through an API provided by Microsoft Azure, which may cause data protection issues.
2) The Azure cloud does not support all EU languages.
3) Due to the lack of transparency, it is impossible to control bias.

He stated that instead of finetuning and building the front ends for the AI models that are running in the United States, it would be important to have the capability to train and build large AI models within Europe. In consequence, Europe needs to keep its leading position in AI development and preserve its digital sovereignty by creating large AI models according to European values, i.e. models that

- are open source,

- support all EU languages,
- ensure high data protection and transparency in algorithms to reduce bias,
- are Co2-neutral

However, Jörg mentioned numerous challenges that need to be tackled to not fall behind, namely the lack of available infrastructures for AI development and the speedy developments in HPC due to continuous investments by global players. The presented status quo and the obstacles led to the motivation of LEAM[6], an initiative of the German AI Association, which aims to build large AI models that enable Europe to take on a leading position in AI development. As illustrated in Figure 6 below, the capacity to train such large AI models will be built in five steps, i.e. the establishment of a high-performance supercomputing centre dedicated to AI, data collection, the development and improvement of algorithms and finally the integration into European AI ecosystem.

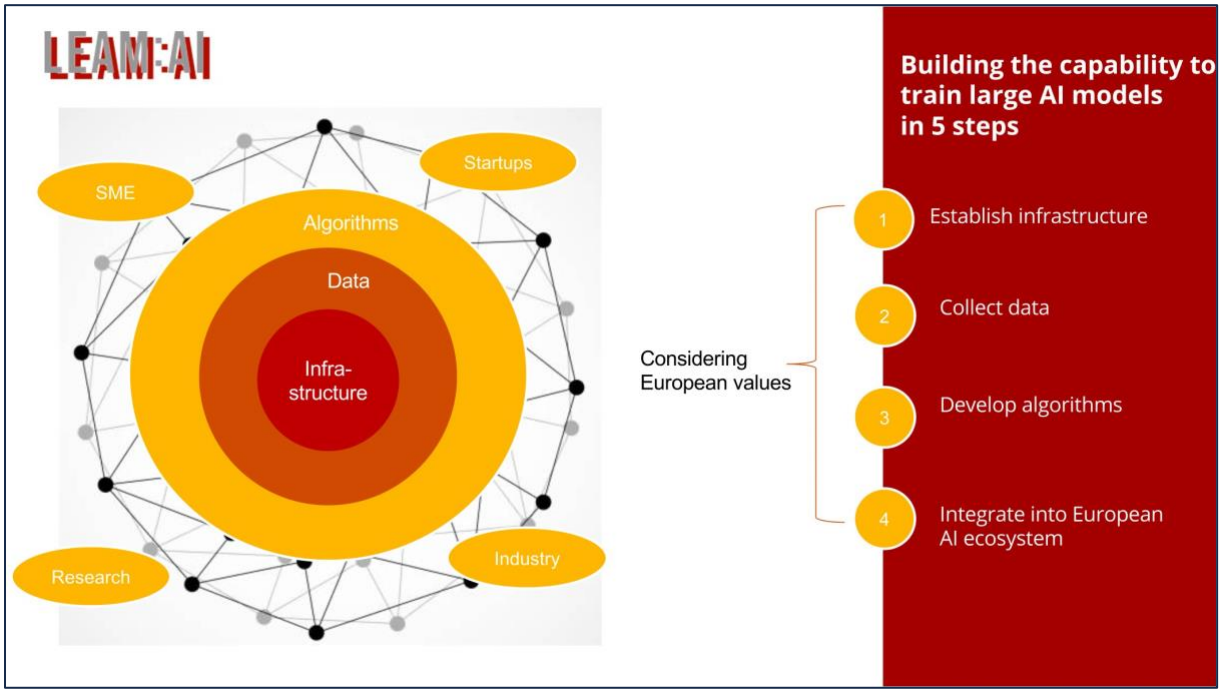



**Figure 6: LEAM capacity building for large AI model training**

The initiative is supported by German research centres like the German Research Centre for Artificial Intelligence, Fraunhofer Institute and Darmstadt University but also by German companies such as Deutsche Telekom, RTL or Rewe Digital. According to the speaker, the current focus is on finding additional supporters and corporates to jointly drive the initiative forwards. In terms of operations, LEAM foresees three clusters, i.e. 1) the core model development, where the large models will be developed, 2) model tuning, where applications will be built according to certain requirements or use cases and 3) inference, i.e. providing computing power and storage to enable customers and companies to make use of these models. Jörg concluded his talk by presenting the first pilot project Open GPT-X. Funded by the German Ministry of Economy, the project aims to build a model comparable to GPT-3, which will cover

[6] Short for "Large European AI Models"

German and other European languages and which is based on Europe's largest supercomputer, the JUWELS, developed by the "Atos Forschungszentrum Jülich".

The presentation was followed by a number of questions and comments from the audience:

- One participant referred to the distribution of the LEAM operations (see slide 25 of the presentation) and asked whether the speaker considers the 90:10, 60:40 and 10:90 cost distribution between public and commercial customers as a general model or rather something specific to the LEAM initiative. The speaker explained that there may be more focus on research when it comes to building the large base models, while inference and usage are expected to be used more by the industry and corporates. However, he added that this was just an estimate to demonstrate that there may be differences in usage efforts.
- When asked how the different supercomputing centres in Europe are going to collaborate and if the initiative should not be at European level instead of national level, Jörg confirmed that the major goal is to bring LEAM.AI to European level. According to the speaker, starting in Germany might have worked faster than it would have been possible in other countries, but the overall aim is to strengthen the discussion with EU partners and to collect further ideas also from other countries. He concluded that being at an event like the ELRC Conference would be a perfect occasion for that.
- Another participant referred to presentation by Philippe Gelin and asked if Jörg sees any connections between LEAM and the planned EDIC. Jörg answered that he sees great potential in the cooperation with other initiatives. He clarified that the goal is to have all major players on board and that there will be no competition between EU research institutes for example, because the competition would be outside the EU.
- When asked about how to reduce Co2 when building large language models, the speaker mentioned two options, i.e. increasing the efficiency of the algorithms on the one hand and relying on hosting providers that focus on environmental-friendly hosting on the other.
- One participant stressed that the point of foundation models is to enhance them with different knowledge and that it is a crucial to find ways to get high-quality machine-ready knowledge. That is why he wanted to know if there are any related initiatives in this respect. Jörg Bienert explained that this should be part of the research in this area and that currently, LEAM is collecting all kinds of dimensions that need to be considered in terms of research. He mentioned efficiency, looking at bias and quality, fostering the usage of different languages as example areas requiring intense research efforts.

### 3.4.2 Possibilities and Limitations of Large Language Models: PAGnol, VLM-4 and Muse

The next speaker, Igor Carron, CEO of LightOn presented the possibilities and limitations of large language models. Referring to zero-shot or few-shot learning, Igor Carron explained that large language models can tackle new natural language tasks and that larger models generally score higher, generalise better and train faster. In 2021, his company LightOn released PAGnol, which is one of the largest models in French (1.5 billion parameters). PAGnol was trained by LightOn in cooperation with the ALMAnaCH team of Inria Paris, and used the JeanZay supercomputer through

a GENCI allocation. In April 2022, LightOn will launch a Muse API, allowing the interaction with some of these large language models called VLM-4. They are going to be available in French, Italian, Spanish, German and Arabic. He continued with a number of examples where large language models can be used, e.g. French text generation, Italian review classification, Spanish keyword detection or question answering in German. While those language models are trained for a certain language only, Igor Carron raised the question if it would not be possible to use GPT-3 in English and then translate into French, Italian, Spanish or German. He pointed out that this may, however, lead to mistranslation of local knowledge and language. He moved on by listing some of the major limitations for building and training large language models:

- Limited computing power: As already mentioned by Jörg Bienert in his opening session, the requirements for building large language models are going much beyond of what was expected from current AI workloads (see Figure 7 below).

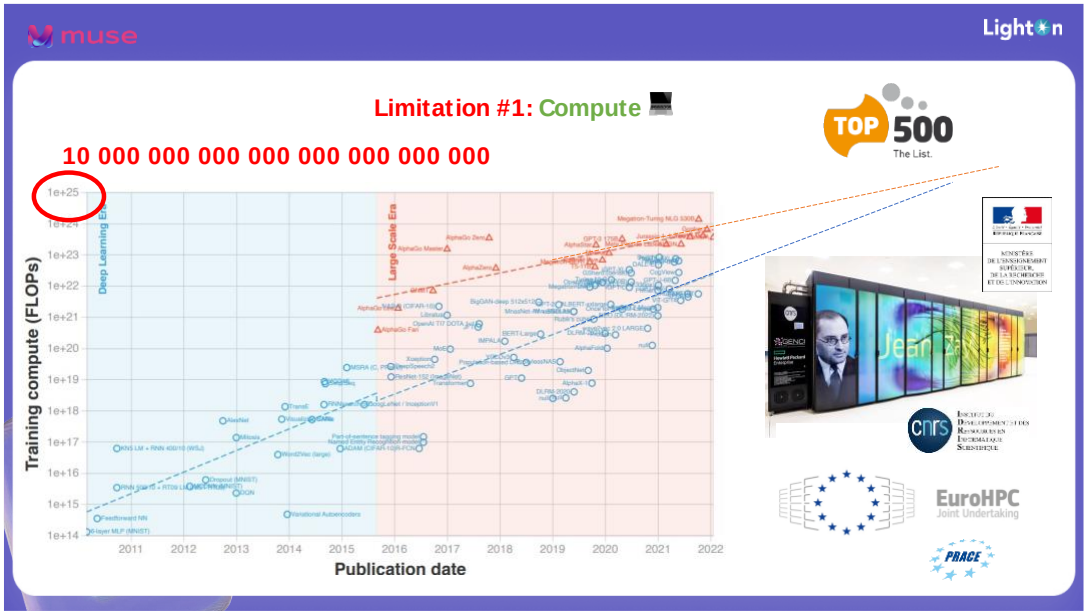



**Figure 7: Limitation of computing power**

- Data quantity: Even though there is enough data available to go beyond English, restricted data access is still a major limitation to building large language models.
- Data quality: According to the speaker, this does not only refer to getting the correct translation for an expression, but also to considering the diversity of speech, which is much more complex.
- Public high-quality data sets for downstream task evaluation: Igor highlighted that compared to data sets which are openly available in English, those for EU languages such as Czech or Slovene are very scarce, which limits the ability of downstream task evaluation and the largeness of some of the models.

- Critical mass for certain languages and markets: As illustrated in Figure 8 below, Igor Carron sees the EU in the middle between the key players focusing on the English language (such as OpenAI, Microsoft, Google, etc.) and those focusing on Chinese or Korean (e.g. Naver, Alibaba, Huawei).

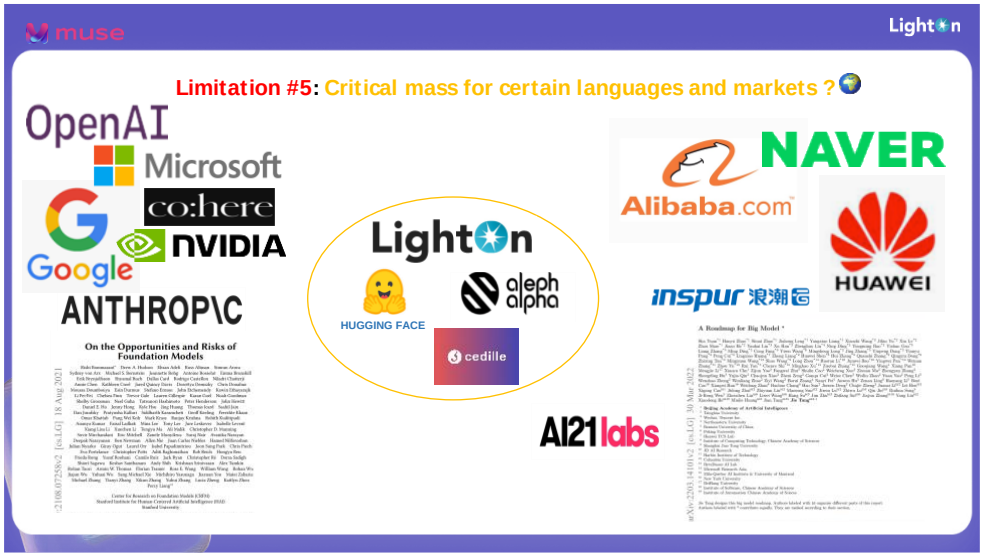



**Figure 8: Limitation of critical mass for languages and markets**

In summary, Igor Carron pointed out that large language models offer a new and seamless way of interacting with machines and the world around us. According to him, they do not only have an impact on our scientific activities, but will also change the way of our work, of doing business and of how we learn. He concluded that Europe can make a powerful move for a huge standardisation training and data sets evaluation for all EU languages to strengthen the internal market and enhance the EU's competitiveness in the world. Following Igor Carron's talk, a number of questions were raised by the audience:

- On the question how Igor would describe the LightOn business model, he answered that LightOn sees itself as a start-up finding its path and that their clients are for example interested in text generation, language assessments or building models for them. He added that LightOn is in the process of self-discovery and that they have a lot of good ideas and are talking to a lot of different people. According to Igor, the already mentioned Muse API will also help to reach a larger cross-section of people who are interested in trying it and who may get back to LightOn on how it is going to be used. He concluded that the current focus is on the size of the large language model compared to the business use case for some of the different customers.
- One participant asked if the PAGnol model was trained on Proust. The speaker explained that PAGnol was trained on JeanZay. As for the other models, he did not want to share details about their approach and size due to international competition.
- Following the question about the data quality assessment of PAGnol, Igor explained that the model was first trained on common crawl without much filtering, followed by intense work on the data quality pipeline.

### 3.4.3 AI Sweden: Language Models for Swedish Authorities

The third presentation of the spotlight on large language models was held by Magnus Sahlgren, Head of Research for Natural Language Understanding at AI Sweden. AI Sweden is focusing on language models for the public sector in Sweden and aims to promote the use of AI in the Swedish society. According to the speaker, the focus of AI Sweden is on language models for NLU because they see it as a powerful technology, which already now provides added value, and which will be dominating the future development of LT. Before diving into the specific context of the Language Models for Swedish Authorities project, Magnus explained what language models are. Language models are statistical models that learn the probability of distribution over language through self-supervised learning. Leveraging from basic linguistic knowledge, this allows to solve many different types of language processing tasks.

Magnus pointed out that the public sector in Sweden mainly works with textual data, which is why it is highly important that they are capable of handling language data properly. According to the speaker, this would open great opportunities for NLP, which is also why the Language Models for Swedish Authorities project was initiated. The project was funded by the Swedish Innovation Agency Vinnova and the consortium consists of AI Sweden, the research centre RISE and three Swedish authorities, i.e. the Swedish Tax Agency, the Swedish Employment Agency as well as the Swedish Agency for Regional and Economic Growth. Magnus explained that the idea behind this project is to provide tools and prerequisites for actors in the public sector in Sweden to be able to use language models to solve natural language processing tasks. More precisely, the project provides new types of algorithms for solving tasks using language models, concrete implementations, code, data, trained models or applications based on language models. Major application areas the project has been working on are e.g. text similarity, named entity recognition, machine translation or text categorisation. Magnus also explained that Sweden is under-resourced when it comes to language model evaluation frameworks, which is an issue, because although large language models are already built in Sweden, there is currently no way to evaluate them and to choose the proper model for a specific scenario. Therefore, the Spin-off project SuperLim[7] was started. In this project, a suite of test data for language models was already produced. In the next step, it is foreseen to create a suite of training data for all the tests and to set up everything that is required to actually compare language models, including e.g. a leader board or baseline results. The speaker highlighted that one of the key lessons learnt is that an evaluation framework is essential to build language models in a certain country or language, because otherwise, it is not possible to validate the developed technology.

Besides that, the project launched a second spin-off, the so-called “Data Readiness project” or “Data Readiness factory”, which aims to improve the data readiness in the public sector and in Sweden at large to allow for the use of language modelling techniques in specific scenarios. The project is on the one hand working on the data readiness framework and on providing tools to assess data readiness in organisation, but also dedicates itself to other aspects related to data readiness, like how to annotate data, which tools to use or topics like anonymisation or pseudonymisation. Another lesson learnt from the work with large scale models for the Swedish authorities was

[7] “Superlim” is the literal translation of “SuperGlue”, an English evaluation framework for language models

that the focus should be on large scale models for Sweden right now. The speaker proved his point with several arguments:

- As already illustrated by Jörg Bienert, large language models perform well on basically all tasks
- There is a need for in-house technical competence on how to train, fine-tune and deploy language models, which can be really challenging
- There is a need for computing resources to be able to deploy the techniques, which also means increasing costs. Not all public agencies can afford the required hardware.

That is why the Language Models for Swedish Authorities project aims to build large-scale foundation models for Swedish and to provide the public administrations with an API solution to leverage on the pre-trained predict paradigm, essentially skipping the fine-tuning phase and going straight for zero-shot or few-shot learning. Also, the project is exploring the use of prompt tuning and p-tuning, which is, according to the speaker, a potentially powerful way of solving a more general applicability of these models. Regarding data, Magnus highlighted the lack of openly available data sets in Sweden. In their initial experiments on large-scale Swedish models, they tried to find as much web data as possible and ended up with 100 GB to build a first version of the GPT model, the so-called GPT-SW3, which is currently the best performing generative model for Swedish. At the moment, this is being extended and a second, much larger version will be launched in the second half of 2022 (see **Error! Reference source not found.**).

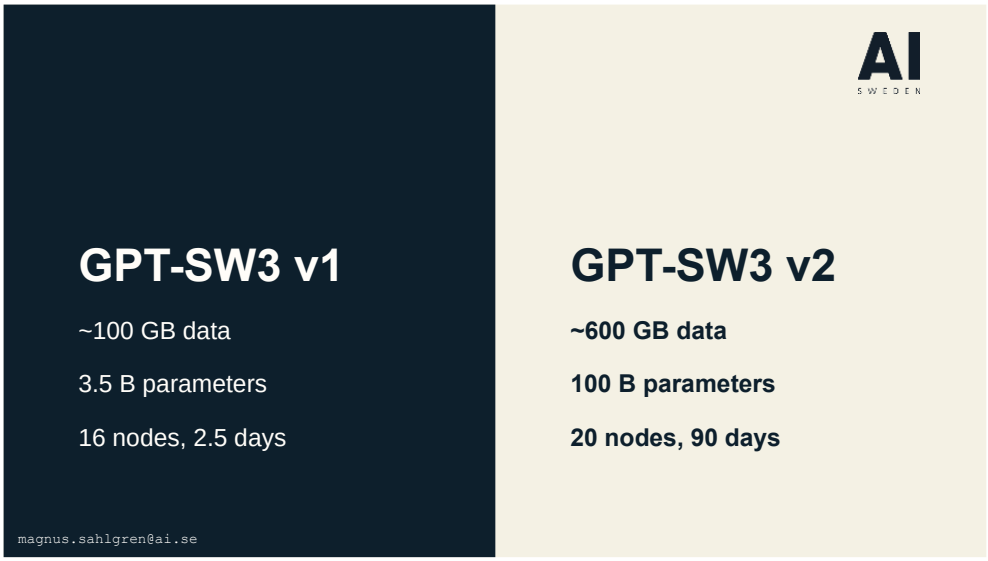



**Figure 9: Overview of GPT-SW3 version 1 and 2**

On a final note, the speaker stressed that data constraints are especially challenging for small countries like Sweden. However, similarities in the North Germanic languages can be a true benefit, which is why according to Magnus, pooling resources and building multilingual models is the way to go. The following questions were raised after Magnus Sahlgren's presentation:

- One participant was wondering if training on other Germanic languages in addition to Swedish could be useful for the application development and asked

if there were any IP concerns about the web crawled data. Magnus confirmed that it is a great idea to benefit from typologically similar EU languages and that it makes sense to train the models on all languages from the Nordic region. He also admitted that there are concerns related to GDPR and copyright. However, the National Libraries have been storing data for ages, which makes accessing textual data easier. When it comes to online data, it is much more difficult. That is also the reason why the data sets used for the preliminary model have not been released.

- Another participant wanted to know how it will be possible to serve a model with 100 billion parameters. Magnus agreed that this will be a challenge that requires a couple of DGX nodes. Currently, the project is exploring different options to find the best way to handle that.
- The final question was related to the project's plans to also include spoken language. The speaker answered that they are not planning to do this at the moment but possibly in the future.
- Finally, one participant commented that there are similar challenges in the Irish context.

## 3.5 Discourse: Language Technologies for fighting disinformation

Due to the situation in Ukraine, it was decided to add a special discourse on the fight against disinformation. This discourse was presented Nikos Sarris (Senior Researcher at the Centre for Research and Technology Hellas) and Maria Bielikova (Director at the Kempelen Institute of Intelligent Technologies). First, Nikos gave a general overview on the state of disinformation in Europe. This was assessed as part of a study analysing ethical standards, political interests and policy recommendations related to disinformation. On the conference day, the analysis results had not been published yet, but the speaker indicated that they will be available soon.

More concretely, the study examined three different dimensions, i.e.

- Codes/principles and authorities that monitor and supervise the ethical application of journalism at a European level
- Patterns of disinformation as means to serve the strategies and interests of political groups, examining involved stakeholders and ways in which disinformation operates at the expense of democracy
- Proposal on measures that could be taken at European and national level, against online disinformation with a view to strengthen democracy

The study also used a classification framework, which summarises the six most prominent areas when it comes to disinformation policy, which are illustrated in Figure 10 below.

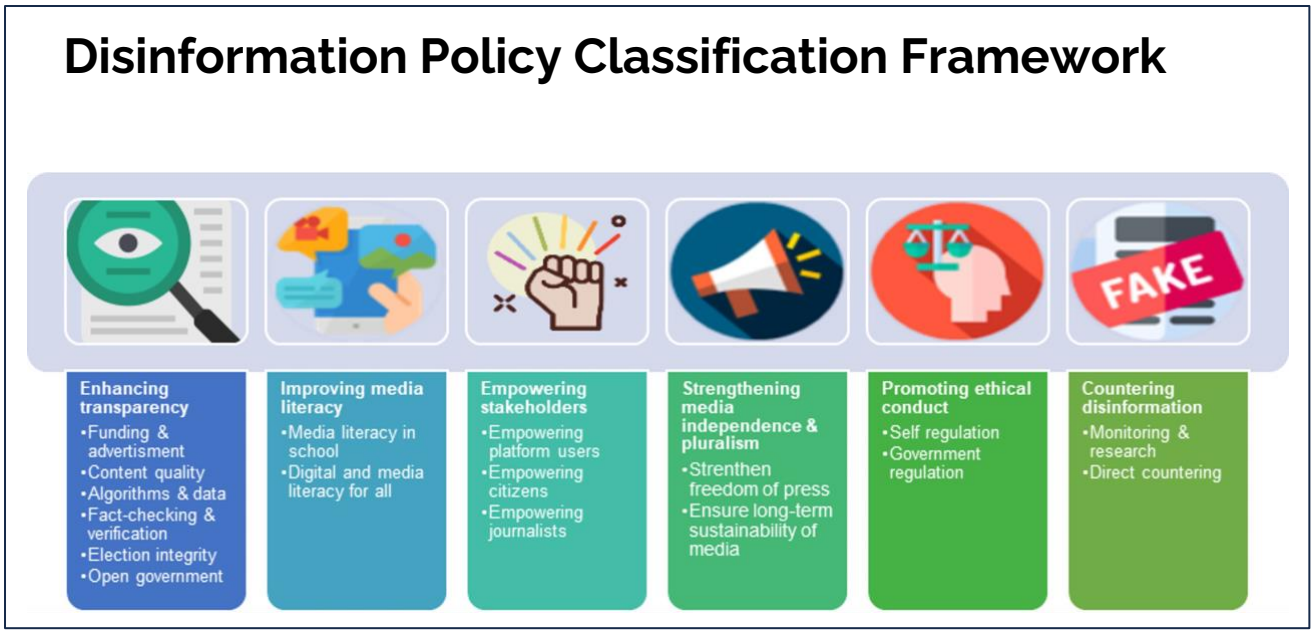



**Figure 10: Disinformation Policy Classification Framework**

Nikos pointed out that disinformation is an ever-growing issue, which is spread by politicians, corporations and the media to achieve their goals. Recent advances in technology can be both a blessing and curse, as they can be either used to produce and share fake news, but also to create solutions to counteract disinformation. The speaker pointed out that the issue of disinformation needs to be tackled using a multi-dimensional, multi-faceted, multi-stakeholder policy framework which assigns fair responsibility to, and which requires decisive action from all relevant stakeholders. He explained that there is a clear need for initiatives that can help move towards this direction.

One of them is the European Digital Media Observatory (EDMO), which supports independent community working to combat disinformation. The overall idea of EDMO is to have national and multinational hubs that will act on a national level and be coordinated by the central EDMO project. Currently, there are already such hubs in Ireland, France or Italy. According to the speaker, at least seven additional hubs will follow in the course of this year. The project itself supports training activities for fact-checkers as well as academic activities on disinformation and organises policy recommendations, among others. Especially with regard to the fact-checking activities, EDMO unites fact-checking organisations from all EU member states, which cooperate with each other. Within the context of the Ukraine war, the EDMO community relies on numerous activities to counter the spreading of disinformation, as they

- Gather fact-checks from all across Europe on a daily basis;
- Identify disinformation narratives on a weekly basis;
- Publish and disseminate corresponding analysis articles;
- Share resources among the European fact-checking communities and
- Repel continuous attacks against the EDMO.

The speaker then handed over to Maria Bielikova, Director at the Kempelen Institute of Intelligent Technologies (KInIT), who gave some insights into how language technologies can help in this fight against fake news. According to the speaker, the

primary interest of the Central European Digital Media Hubs (CEDMO), which is one of the EDMO hubs mentioned by Nikos is in fact verification, i.e. investigating if a certain claim is true, using a fact verification model. An overview on the fact-checking tasks is provided below.

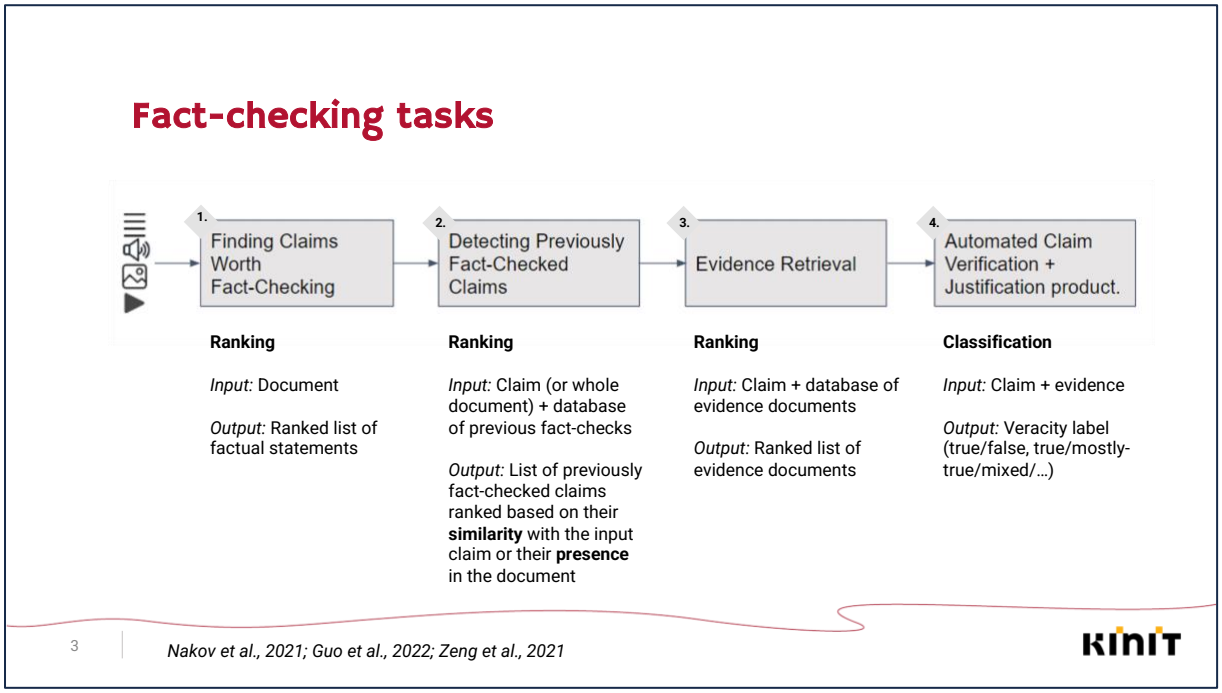



**Figure 11: Overview on fact-checking tasks**

From the perspective of NLP, there are three underlying main tasks, i.e.

- claim matching, the most basic task which involves the identification of semantically equivalent or similar occurrences of a given claim in a larger unit of text
- stance detection, the detection of the author's position towards a specific target and
- textual entailment, the most complex task, which refers to the investigation on whether the retrieved evidence supports the hypothesis or not, i.e. the verification.

Maria highlighted once more that language resources are crucial for any machine learning task, and this also applies to the domain of fighting disinformation. She pointed out that even though there are quite a lot of data sets available, their quality is often very low and their focus very narrow. In addition, she explained that there is a need for multilingual data sets, because fake news are spread across languages and borders, especially when it comes to global events such as the COVID-19 pandemic or the war in Ukraine and currently, disinformation is only fact-checked in some of the languages. Another challenge is multimodality, because disinformation often combines text, images and videos, which makes fake news detection even more complex and requires intense research in many different areas. The speaker concluded by highlighting that fake news detection requires the collaboration between fact checkers and technology, admitting that from a technological perspective, there is still much to do until detection can run fully automatically. However, she stressed that LT can already be a major support for fact checkers to fight disinformation.

Following this discourse, one question was raised by the audience:

- One participant wanted to know what the speakers think about fake news spreaders and checkers detection and how they can be used to counter disinformation. One of the speakers answered that it is an important source of research, which can either help in a fully automatic or semi-automatic way. An example would be a visualisation which provides names of popular “negative influencers”.

## 3.6 Wrap-up and outlook

Due to the intense discussions which caused some timely delay, Andrea Lösch briefly thanked the speakers for sharing their insights and expertise and the audience for their numerous contributions and questions. Before the audience left for lunch, she announced the second spotlight of the day, i.e. Multimodal Language Data and then wished everyone a nice break.

## 3.7 Spotlight: Multimodal language data

### 3.7.1 Behind the scenes: Multimodal Data Analysis

The spotlight was opened by Suzanne Little, Associate Professor at Dublin City University and SFI Principal Investigator at the Insight SFI Research Centre for Analytics, who gave insights into multimodal data analysis. She set the scene by explaining that the meaning of an image is highly variable, as it depends on multimodal information such as associated text, context or perspective of the audience. She used the example of memes, which typically consist of text and image and explained that if considered in isolation, the two components do not indicate the ultimate interpretation. Therefore, if a computer vision model would have to label this meme, it may only come up with descriptive elements about the content of the image and the text statement, which does, however, not necessarily reflect its actual content or message. In consequence, a number of challenges need to be faced, e.g. the lack of labelled data or the challenge of multiple meanings and context. Suzanne added that technical research in this area is focused on how to codify the text and images and the context, using pre-trained models built on much larger and more generic data sets and in particular investigating different ways to fuse data in deep learning models. To do this, a certain amount of labelled tuning and evaluation data is required. She pointed out that while people are very good at identifying patterns like for instance faces or implications, this is really hard for computer vision, which is why there is a great scope for research on its improvement and evaluation. This does not only apply to modern materials such as memes, but also to archival and historical materials like caricatures or comics. She summarised that classifying offense or hate speech is a real challenge for computer vision models.

Suzanne concluded that the underlying difficulty is getting data for classifications. Even though there are some useful data sets, getting a sufficient volume to train the machine learning models, having enough variety of the class, source or type and being able to label or create the required ground truth annotation is therefore the main challenge. Within the projects Suzanne is involved in, this data is collected through crowdsourcing, creating synthetic data or augmenting small samples and doing transfer learning or by manual labelling.

Another challenge in computer vision is related to bias, e.g. when the background blurring in Zoom fails with certain skin types. According to the speaker, these issues cannot be solved by increasing the number of examples, because even the currently best performing models do still show these problems. She added that by using these pre-trained models, there is a risk to inherit or even amplify the underlying biases, at the same time explaining that this does not mean that pre-trained models should not be used at all. She pointed out that they provide a very useful backbone and can be successfully used in many applications, but that research on how to recognise and mitigate bias will be required before they are deployed for widespread use.

The speaker continued by briefly introducing the Crowd4Access project, an Irish citizen science project to capture information and examples of accessibility and to assess mobility in local areas. She pointed out that the project comes with several advantages: It does not only provide Suzanne's research group with interesting research data to work on computer vision models, it also raises people's awareness on the possibilities of science application in their daily lives. On a last note, Suzanne mentioned additional examples which can benefit from multimodal data, including scene description for video for accessibility or knowledge extraction and sensors for eye tracking or brain waves.

After the presentation, there was one question from the audience:

- One participant wanted to know whether there is a way to detect manipulated images using computer vision models. The speaker answered that there are some very good methods for detecting manipulated images, but that it is also kind of a race between the developers of the detectives and the creators of fake image generators. She added that there is no universal solution to counteract on that, because it depends on the motivation and context behind the manipulation. While in some cases, manipulated images are easy to identify and can be detected using automated systems, others are undetectable deep fakes. She concluded that this is a challenge to be tackled, but that there is already a lot of interesting work in progress to address this.

### 3.7.2 Multimodal Data in the medical domain

The opening talk was followed by a presentation from Alexandra König, Neuropsychologist and clinical researcher at INRIA, who presented ongoing research on multimodal data in clinical practice. To lay the grounds, Alexandra started by explaining that there is a need for diverse data and dimensional approaches as fundamental principles for better diagnosis and care in the medical domain. According to the speaker, the aim would be to rely more on a measurement-based care instead of the clinical judgement only and to opt for continuous preventative care, which would allow for an earlier detection of diseases. She also introduced the term "digital phenotyping", which has become very popular in the medical field. It describes the "moment-by-moment quantification of the individual- level human phenotype in situ using data from personal digital devices, especially smartphones and wearable sensors ". Once the background was explained, Alexandra presented two research projects that were conducted by INRIA, i.e.

- The Deep Speech Analysis for Cognitive Assessment in Clinical Trials (DeepSpa) which was funded by the European Institute of Innovation and Technology and

- MePheSTO – Digital Phenotyping 4 Psychiatric Disorders from Social Interaction, a collaboration with the German Research Centre for Artificial Intelligence (DFKI).

As the speaker is mainly working in a memory clinic, she presented a brief discourse on Alzheimer's and highlighted that most patients detect the disease in very late stages only, i.e. when their memory deficits. However, there are other symptoms that appear earlier and even though they are very important, they are currently not detectable with the available instruments. According to Alexandra, detecting Alzheimer's at an earlier stage may allow for a more efficient and beneficial treatment, but this would require new tools and measurements.

One method which was investigated as part of the DeepSpA project was telephone-based screening. The team built an automatic platform for cognitive testing, which makes it possible to collect information on the linguistic features (e.g. sentence complexity or semantics) as well as on the paralinguistic features (e.g. acoustic features or prosody) via phone and to extract results, scoring and other additional features afterwards. Most interestingly, when comparing the results of this automatic scoring with face-to-face scores obtained in the clinic to assess their validity, it turned out that there was a high agreement between both methods. Alexandra added that combining the analysis of facial movements and speech leads to an even better accuracy in detecting e.g. signs of mental health issues such as depression but also apathy, which is often also observed on Alzheimer's patients. Even though a clinician is usually capable to detect such signs, too, this may provide a useful and objective measurement to back up his/her impression.

Another use case presented by Alexandra was remote cognitive monitoring via telemedicine, targeting patients living in "medical deserts" who have no access to health care. The main idea behind this was to merge data coming from audio and video, using a telemedicine platform. Also in this case, the comparison of face-to-face and remote methods showed similar results and the patients' feedback was very positive, as some of them even felt more comfortable talking about their feelings from distance instead of in front of the clinician.

The speaker concluded by presenting the second project, MePheSTO, which does not only analyse the audio and video of the patient, but also of the clinician to get an idea on the quality of the actual interaction. The study involves 150 patients with a major depressive disorder or Schizophrenia. Overall, four research cases are in focus, i.e. 1) supporting differential diagnosis for major depressive episode aetiology, 2) quantifying therapeutic alliance by means of social synchrony, 3) treatment outcome/Relapse prediction from negative symptoms in schizophrenia and 4) robust and objective measurement of formal thought disorder in schizophrenia.

After the presentation, the following questions were raised:

- One participant was interested in whether AI can detect if an expression is forced and not natural. Alexandra explained that many training data for emotion recognition is based on videos of actors, which makes it hard to find "real" emotion data sets and which limits the detection capabilities in this respect. She added that differentiating between natural and forced expressions is a difficult topic, even though certain traits like "forced smiling" may already be detectable.
- Another participant wanted to know what kind of questions were asked during the recruitment exercise via phone and if the patients knew that this was part of

the study. Alexandra confirmed that they were informed about being recorded because they had to give their consent. As for the type of questions, the patients were asked to tell a negative or positive story.

- Last but not least, a question was raised on whether there are any attempts to relate the results to the brain status/activation. The speaker answered that they did some studies on comparing the speech features with brain atrophy, i.e. brain-imaging data, and added that they are also looking into the comparison with CSF, amyloid deposition in the spinal fluid (a biomarker for Alzheimer's disease) and speech features. According to Alexandra, it would be very interesting to relate to fMRI data. However, they are not doing it at the moment.
- Also, one of the other speakers showed special interest in the contents and offered to do some experimental tasks together.

### 3.7.3 Multi3Generation: Multimodal Data for Natural Language Generation

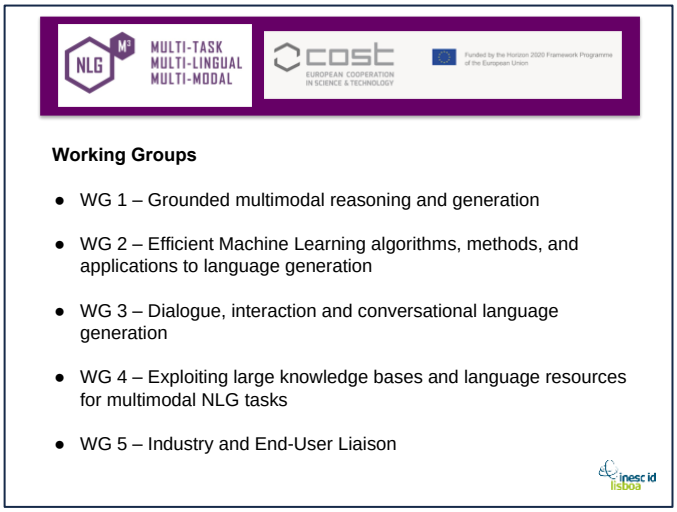



**Figure 12: Multi3Generation Working Groups**

The last presentation of the spotlight on multimodal data was held by Anabela Barreiro, Chair of the COST Action Multi3Generation and researcher at the Human Language Laboratory at INESC-ID. The speaker showcased the Multi3Generation action, a networking program which is funded by the European Cooperation in Science and Technology (COST) and focusses on multimodal data for natural language generation. Overall, the action consists of five working groups (see Figure 12) and aims to foster an interdisciplinary network of research groups working on distinct aspects of natural language generation (NLG) with an emphasis on any combination of multilingual models, multitask learning and multimodal uses.

Anabela explained that Multi3Generation focuses on four core challenges, i.e.

- The representation of data and information
- Machine Learning (ML) when applied to NLG: inputs to be mapped to different correct outputs (e.g. structured prediction and representation learning)
- Interaction: applications challenges for NLG (e.g. Dialogue Systems, Conversational Search Interfaces and Human-Robot Interaction)
- Knowledge Base exploitation: structured knowledge is key to NLP tasks, including NLG and supporting ML methods that require expansion, filtering, disambiguation or user adaptation of generated content

Following a brief overview, the speaker also gave some insights into recent outcomes of the different working groups, including among others

- a survey on recent advances in Natural Language Generation, which was conducted by WG2 and which will soon be published in the Journal of Artificial Intelligence Research (JAIR),
- WG3's work on Human Computer Interaction tasks in multilingual and multimodal settings, applying NLG models to conversational agents. This led to new answer generation techniques, techniques for conversational quality estimation and sentiment analysis as well as to multilingual datasets for low resourced languages
- WG4's in-depth analysis and review of methods to efficiently incorporate and enrich structured multimodal knowledge bases into NLG models.

Multi3Generation also organises training schools, events and offers funding for young researchers and short-term scientific missions, i.e. short research visits where one person travels and visits a host in an EU institution to work on a short-term research project (for up to 3 months) related to the Multi3Generation Action. The speaker concluded by presenting two future research projects, i.e. the generation of monolingual/multilingual paraphrases and the digitalisation of humanities and social sciences. Further information about the project is available at https://multi3generation.eu/. No questions were raised after the presentation.

## 3.8 Bridging the language gap: LT solutions for Europe

### 3.8.1 EMBEDDIA: AI Technology for the Media Industry

This talk was held by Senja Pollak, coordinator of the H2020 project EMBEDDIA, which ran from 2019 to 2022. Short for "Cross-Lingual Embeddings for Less-Represented Languages in European Media", EMBEDDIA leverages the technologies in cross-lingual embeddings and deep neural networks with a specific focus on less-presented EU languages, which are morphologically rich, such as Estonian, Croatian, Finnish, Slovenia or Latvia.

Senja pointed out that in news industry, there are dominant languages and, in several countries, there are already advanced tools available which can support the news media industry. However, when it comes to countries like Slovenia, the situation is different. In terms of news media applications, the project focused on 1) comment analysis (e.g. filtering hate speech or fake news spreader detection), 2) news analysis (e.g. detection of topics and viewpoints or summarisation) and 3) news generation (e.g. text generation from structured data or headline generation). The project consortium consists of six academic partners with a background on NLP and machine learning (Jozef Stefan Institute, University of Ljubljana, Queen Mary University of London, University of Helsinki, University of La Rochelle and University of Edinburgh) as well as three news media industry partners (Trokoder, Ekspress Meedia, Finnish News Agency STT), who helped to define the application tasks, provided data and tested the developed tools. Last but not least, TEXTA OÜ, an Estonian SME from text mining industry, was involved in wrapping up the tools and exploiting them for further use in their text toolkit. Senja explained that in terms of the technological background, the project is leveraging on recent trends in embeddings in deep neural networks, especially the transformer structure. EMBEDDIA aims to benefit from the large multilingual language models that can then be fine-tuned on specific task. As there are many data sets available for well-resourced languages, the project tries to adapt the tools that were e.g. trained on English data to the less-resourced languages and to

create cross-lingual applications using zero-shot learning or few-shot learning. She added that if data is available, they can directly fine-tune the models on the target language data.

Senja continued by presenting a selection of results:

- Release of new data sets in agreement with the media partners for languages including Russian, Finnish, Croatian or Estonian. They are available through CLARIN.
- Release of new multilingual models with selected language combinations, such as Croatian, Slovenian and English or Finnish, Estonian and English.
- Release of monolingual models
- Participation in 20 shared tasks and organisation of various events

Besides that, as part of the project, numerous applications were developed, such as systems that can recognise keywords in articles or sentiment analysis applications, classifying whether a news article is negative, neutral or positive. Other selected news analysis applications referred to topic modelling, article retrieval and linking as well as viewpoints analysis.

Referring to the second pillar, news analysis, Senja explained that comment moderation was an important task, as media houses have to moderate comments with offensive or toxic contents, which often needs to be done manually. As a consequence, EMBEDDIA started several experiments on cross-lingual offensive speech, using solutions that are specific to the media company, but also trying a more systematic evaluation across different languages, using five hate speech datasets in Arabic, Croatian, German, English and Slovenian and two different models, mBert and cseBERT. Other tasks on user-generated content include multilingual sentiment analysis, user demographics characteristics and fake news detection.

When it comes to the third pillar, text generation, EMBEDDIA generated reports from structured data and worked with solutions for headline generation and text summarisation. Senja also highlighted that the project put a lot of effort into making the developed tools freely available, so that researchers and the media industry can benefit from the solutions, for example. They can be found at https://embeddia.texta.ee. The speaker concluded her talk with a demonstration of the keyword extraction and invited the audience to try it themselves at https://embeddia-demo.texta.ee/. There were no questions associated to this session.

### 3.8.2 Microservices at your service

The CEF-project "Microservices at your service", which aims to fill the gap between academic NLP research and industry, was presented by Sebastian Andersson, Solution Architect at Lingsoft. The speaker started by introducing the partners involved in the project, i.e. Gradiant, Lingsoft, Reykjavik University and the University of Tartu. The project will run from March 2021 to February 2023 and is looking into the EU speech and language technology landscape. Sebastian clarified that even though many tools are being offered by large international organisations, smaller languages are not in focus of the big players and better tools can be found in the local communities. He continued that within Europe, there is a strong speech and language technology research community with a long tradition of sharing the tools as open source and that the European Language Grid also tries to establish itself as an alternative platform for LT in Europe.

Moving on to the benefits and challenges of open-source tools, Sebastian pointed out that they grant access to the "latest and greatest" from the EU research community and that there are many trusted and well-established tools, that can be freely used and adapted, depending on their licenses. Regarding the challenges, Sebastian listed a few examples:

- It can be difficult to find the right tool,
- The documentation can be of poor quality,
- There may be a lack of technical support, as this is not the core interest of many researchers,
- Making the tool run and/or using it on another machine can be very time-consuming

"Microservices at your service" aims to address these challenges, which is why they dived into the European research community and tested available open source tools, contacted research groups who are known for producing good solutions and additionally used a bottom-up search for tools they were hoping to find. The identified tools will be provided as dockerised solutions plus APIs and shared via the European Language Grid platform and ELRC-SHARE. In summary, the objective of this CEF project is to make at least 40 tools available as easily integrable microservices covering a total of 11 languages, i.e. Finnish, Swedish, Norwegian, Northern Sami, Estonian, Latvian, Lithuanian, Spanish, Portuguese, Icelandic, and Faroese. An updated list of tools and previous workshops is also available on the project website.

After Sebastian's presentation, the following questions were raised:

- One participant was wondering if there is any non-obvious downside to this approach to develop NLP tools, like performance, for example, but the speaker answered that from his experience, there are no drastic performance issues.
- Another participant wanted to know more about the security risks coming from open source. Sebastian answered that there are no obvious risks, but that of course, one needs to keep in mind that if he/she shared an image, he/she basically shares everything, i.e. models, data, code. However, this would often not be a problem with open source.
- Following the question whether a recording of the mentioned Docker and API workshop would be available, Sebastian shared the following link in the chat: https://www.lingsoft.fi/en/microservices-at-your-service-bridging-gap-between-nlp-research-and-industry.
- On a final note, one participant hinted on the Latvian public administrations' extensive use of chatbots (by more than 20 institutions). He added that they use a government LT platform with MT and chatbot services from hugo.lv and provided the following links: https://hugo.lv/en/tools, https://va.hugo.lv/directory.

### 3.8.3 ENRICH4ALL

The last presentation was held by Dimitra Anastasiou, researcher the Luxembourg Institute of Science and Technology (LIST) and coordinator of ENRICH4ALL. The project is funded under the Connecting Europe Facility and aims to lower the language barriers in the EU, Iceland and Norway with the help of an eTranslation-based multilingual chatbot in public administrations. Dimitra explained that ENRICH4ALL is planning to deploy this multilingual chatbot in public administrations in Luxembourg,

Romania and Denmark, but that there also plans for extension, because the project strives for a unified, broad EU-wide eTranslation-based chatbot.

According to Dimitra, chatbots have numerous benefits, such as the capability to handle huge numbers of requests, a constant availability, and the delivery of up-to-date information without timely delay. By implementing a chatbot, public administrations could reduce costs and employees would be able to focus on more complex requests. The speaker highlighted that personal data security is an essential requirement for the deployment of e-Government chatbots, which makes eTranslation a perfect fit, as the data stays within Europe. The project is also working on a user authentication service, allowing the bot to store some basic information about the user he is interacting with to avoid repetitive conversations, such as the address or health insurance number. Providing examples of government agencies that already use chatbots, Dimitra pointed out that most of them are located in the United States or e.g. in India. While there are some agencies that use chatbots in Europe, some of them do not belong to the EU and most of them are not in public domain either. The speaker also presented the eTranslation integration, which is illustrated in Figure 13 below.

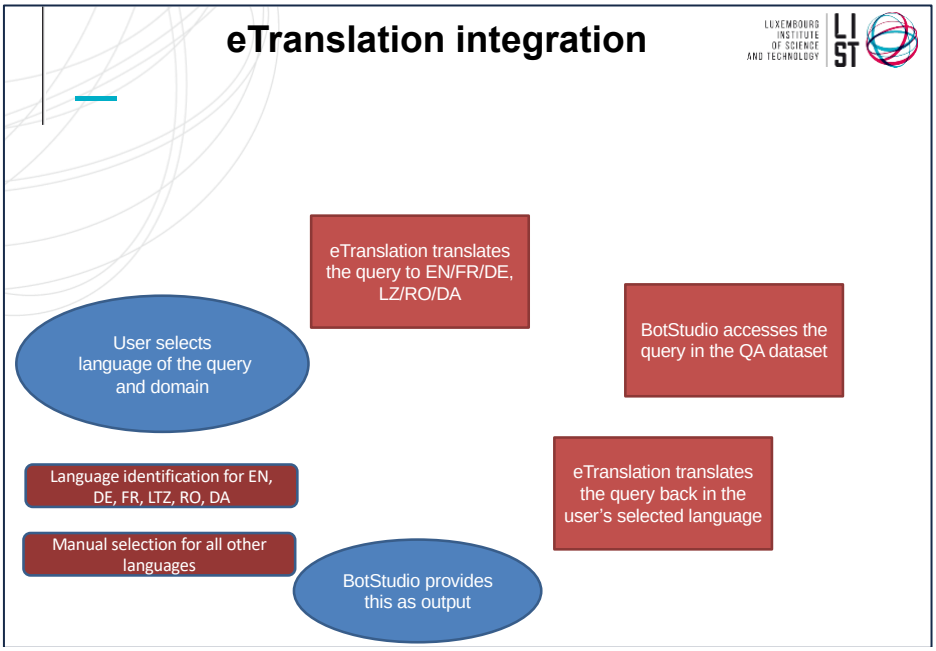



**Figure 13: eTranslation integration**

She highlighted the importance of NLU for this endeavour, as it understands context, typos and synonyms. That is why BERT models have been tested for question labelling and question similarity.

Referring to the datasets used, Dimitra pointed out that the corpus and question answering data set is based on Guichet.lu, the administrative platform for citizens and business. Additional data sets were from Romanian construction permits and COVID-19 corpora. The speaker also briefly mentioned the limitations of chatbots, taking the example of COVID-19 regulations and illustrating that, depending on the questions asked, a domain-specific customised conversation agent is required to be able to provide the correct answers (see Figure 14 below).

To conclude her presentation, the speaker gave some insights into the future of chatbots. According to her, chatbots will soon move away from "simple" question answering that is driven by keywords and gain the capacity to handle questions "human-like" and to cope with complex interactions thanks to the advancement of AI and ML. Further details about the project as well as the available data sets is provided at https://www.enrich4all.eu.

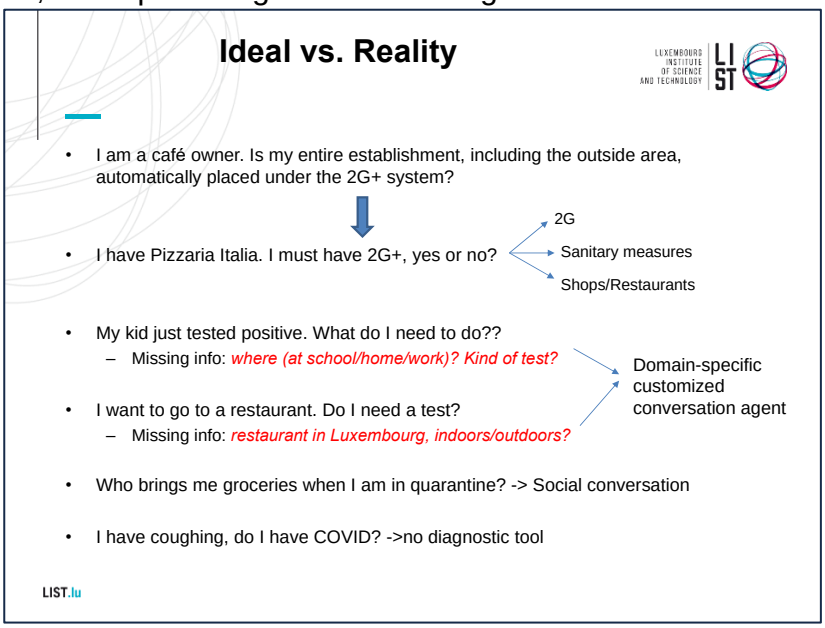



**Figure 14: Ideal vs. Reality of Chatbots**

No questions were raised after the talk, but one participant called it a great and impressive initiative.

## 3.9 Summary and conclusions

At the end of the conference, Andrea Lösch summarised the key takeaways of the event, which are:

- Large language models will outperform and gradually replace other AI solutions. They are critically important to ensure Europe's Digital Sovereignty which is why future data collection efforts need to include large language models.
- The availability of multimodal language data offers enormous opportunities for the development of language-centric AI. Consequently, future data collection efforts need to account for language data of all types – and not be limited to text-based data.
- Despite intense efforts to spread disinformation and fake news, there are solutions that can be used for detection and language technologies play a central role in this respect.
- There are a variety of CEF and Horizon 2020 projects with the goal to bridge language gaps of whatever sort. This ranges from using cross-Lingual embeddings to support less-represented languages in European News Media or promising attempts to bridge the gap between academic NLP research and industry, to the successful integration of eTranslation to an already existing AI-based chatbot technology.

To sum it up, Andrea highlighted that there were a lot of important insights and lessons learnt, which will be beneficial for the future work and activities of each and everyone who joined the conference on that day. On a last note, Andrea expressed her thanks to the speakers, the participants, the EC representatives as well as the conference organisers and invited the audience to complete the online feedback form and to stay in touch by following ELRC on social media and by attending one of the upcoming ELRC events.


**Figure 15: Screenshot of "Thank you" Visual**

# 4 Annex: Conference Presentations

All presentations are available on the ELRC website: https://lr-coordination.eu/6thELRC. In addition, the full recording of the 6th ELRC Conference can be accessed via YouTube at https://youtu.be/ebAbv5KgvrQ.