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

European Language Resource Coordination
Connecting Europe Facility

Deliverable D3.2.26
Task 3

ELRC Workshop Report for Poland

Author(s): Anna Kotarska

Dissemination Level: Public

Version No.: V3.1

Date: 2023-01-03

Contents

1. Executive Summary

The 3rd Polish ELRC workshop took place on the 21st of September 2022, in a form of a full day hybrid event from 9:00 to 15:30, which was organized locally by the team from the Polish Society for Health Programmes. It took place at the premises of the T-NAP's institution – the Institute of Computer Science, Polish Academy of Sciences (IPI PAN) in Warsaw. The workshop was attended by 120 participants out of whom 20% participated on-site and 80% online via Zoom platform. The attendees were a cross-sectoral representation of public administration, industry (LSPs), academia, national research institutes, LT providers, translators' associations, and freelance translators. Apart from Warsaw-based attendees, more than half of onsite participants came from other cities in Poland (Wrocław, Poznań, and Gdańsk) as well as from abroad (Luxembourg and Latvia – 2 speakers).

In the 11 sessions organized within the predefined format, 9 different presenters (including 1 pre-recorded presentation) appeared, and 5 panellists took part in the discussion panel. The whole event was video recorded. The presentations are available online at the workshop web page whereas the recording was shared with registered participants via a non-public link to the ELRC YouTube channel. The workshop language was Polish and English depending on the nationality of the speaker. Simultaneous interpreting from and into Polish was provided remotely via online Zoom platform and delivered by professional conference interpreters.

The event was focused on the current state of language technologies and Artificial Intelligence (AI) – from the perspective of its state of the art as well as in the context of the Polish language, and how the advancement in AI can stimulate the transformation of digital and multilingual interaction. The keynote presentation was delivered by an expert from academia – Professor Maciej Piasecki, CLARIN-PL Coordinator. Seven of the speakers held at least a PhD degree with the remaining being experts and/or having relevant professional/managerial experience. Major emphasis was given to the significance of language data and their availability and quality as prerequisites for the development of LT solutions, including the latest developments regarding Large Language Models for Polish. Challenges related to collection of data and solutions to improve the situation were discussed in a panel composed of experts representing various sectors. Examples of LT solutions in/for Polish public administration were presented that were supplemented by examples of best practices and solutions in other EU Member States with emphasis on the Baltic Countries. Representatives of the European Commission (EC) gave an overview of the portfolio of the Connecting Europe Language Tools currently available from the EC and the directions of its planned expansion. The history of the ELRC programme in the context of the Multiannual Financial Framework (MFF) 2014-2020 was briefly outlined. The concept of the Common European Language Data Space (LDS) was explained by a DG CONNECT official and presented as the starting point for the Digital Decade. During the MFF 2021-27 the Digital Europe Programme (DEP) will provide funding for deployment of new infrastructure for sharing data across European states, domains, and sectors.

Overall, the 3rd ELRC Polish workshop organised more than 4.5 years after the previous country workshop that took place in December 2017 in entirely onsite form, can be seen as a successful closure of ELRC's local activities.

2. Workshop Agenda

9.00 – 9.10 Welcome and introduction – Anna Kotarska, Polish Society for Health Programs (PTPZ), ELRC Public Services NAP

9.10 – 9.25 Linguistic Light & Magic: how language technology is changing our lives – Associate Professor Maciej Ogrodniczuk, Institute of Computer Science, Polish Academy of Sciences, ELRC Technology NAP

9.25 – 10.05 The potential of Language Technology and AI – where we are, where we should be heading – Professor Maciej Piasecki, Wrocław University of Technology

10.05 – 10.30 Heading for the Future: Common European Language Data Space – Philippe Gelin, PhD, Head of Unit G3: Accessibility, Multilingualism and Safer Internet, DG CONNECT, European Commission

10.30 – 11.00 COFFEE BREAK

11.00 – 11.30 eTranslation latest developments and Connecting Europe Language Tools – Markus Foti, Head of Machine Translation, Directorate-General for Translation (DGT), European Commission

11.30 - 12.10 Language Technologies by/for the public sector: presentation of implemented solutions:

- NLP models, data sets: OPI resources and experiences – Marek Kozłowski, PhD, Head of Laboratory of Linguistic Engineering, National Information Processing Institute (OPI PIB)
- COVID hotline chatbot for the Ministry of Health by IBM and customer service at public administration, Michał Gawryś *(pre-recorded)*

12.10 – 12.30 LT requirements and offerings: do they converge? – Q & A with the audience

12.30 – 13.30 LUNCH

13.30 – 13.50 The value of data for Language Technology development: example of CLARIN-PL – Tomasz Walkowiak, PhD , Wrocław University of Technology

13.50 – 14.40 Language data creation, management and sharing: existing practices and challenges – panel discussion with:

- Tomasz Klekowski, expert, Working Group on AI at the Chancellery of the Prime Minister
- Marek Kozłowski, National Information Processing Institute (OPI PIB)
- Tomasz Walkowiak, Wrocław University of Science and Technology
- Marta Bartnicka, Dolby Laboratories
- Wojciech Wołoszyk, POLOT Polish Association of Language Services Providers (LSPs)

Moderator: Anna Kotarska, PTPZ, ELRC Public Services NAP

14.40 – 15.00 COFFEE BREAK

15.00 – 15.25 Establishing National Language Technology Platforms in EU member states – Kaspars Kauliņš, International Business Development Director @Tilde, Latvia

15.25 – 15.30 Summary and closing remarks – Anna Kotarska, PTPZ, ELRC Public Services NAP

3. Summary of Content of Sessions

3.1 Welcome and introduction

The workshop began with the presentation given by Anna Kotarska, the ELRC Public Services NAP (PS NAP) for Poland. Anna Kotarska welcomed the workshop attendees gathered at the Institute of Computer Science, Polish Academy of Sciences (PAS) and online. She briefly outlined the agenda of the workshop and organizational issues. Next, she gave an overview of the ELRC programme consortium composition, its representation in Poland as well as the programme's history and its main objectives in the context of the 2014-2020 Multiannual Financial Framework (MFF) and the EU Connecting Europe Facility (CEF) funding instrument. Anna Kotarska emphasized the programme's relevance for development of the European Digital Single Market (DSM) and its implementation nature. She highlighted the transition to the next MFF and the planned launch of the Common European Language Data Space (LDS) in January 2023.

The introduction was completed by the presentation of dr hab. Maciej Ogrodniczuk, the ELRC Technology NAP (T-NAP) for Poland who sketched the role of Artificial Intelligence for the development of language technology solutions including virtual assistants, chatbots, and machine translation among others. He underlined the need for language data for the development of AI, the ambition of the EU to be a leader in building an open data society. He also mentioned projects financed by the European Commission under CEF, which Poland, represented by the Institute of Computer Science, implemented, including MARCELL and CURLICAT, as well as European Language Grid (ELG) and European Language Equality (ELE). The T-NAP's main message was about the supportive role of language technology and he invited cooperations for the creation of language-centric AI and use of LT and AI in general to its fullest potential.

3.2 The potential of Language Technology and AI: where we are, where we should be heading

The keynote presentation was given by dr hab. inż. Maciej Piasecki, Deputy Head of the Department of Artificial Intelligence, Faculty of Information and Communication Technology of the University of Science and Technology (UST) in Wrocław, Chair of the CLARIN ERIC National Coordinators Forum and CLARIN-PL Coordinator.

The presentation focused on the following topics: natural language in artificial intelligence; language technology and its applications; opportunities and limitations of today's deep learning revolution; challenges in natural language understanding; building a language technology infrastructure in Poland (research and development infrastructure); directions of development, challenges and invitation to cooperation.

Dr hab. inż. Maciej Piasecki began by noting that the field of Language Technology (LT) encompasses Natural Language Processing (NLP), Natural Language Engineering (NLE), and Computational Linguistics (CL) among others. Next he presented most typical uses of LT by individual users, researchers in the humanities and social sciences, public institutions and the industry, drawing upon CLARIN-PL experience. Dr hab. inż. Maciej Piasecki discussed the barriers for the development of LT (reference to Report 3.4, 2022 ELE Project) emphasising factors contributing to the great progress of NLP in recent years and various degree of advancement of LT for different languages. He gave an overview of eras in the development of LT: symbolic, statistical, and the era of deep neural models

such as a transformer type BERT. He discussed the limitations of DNM using the example of GPT-3, pointing to the environmental impact of technology due to high carbon footprint resulting from very high computing power needed. Dr hab. inż. Maciej Piasecki next listed numerous challenges that include the availability of data sets relevant for practical tasks; Deep Natural Language Understanding, i.e. a more structural and in-depth analysis of linguistic statements; small, often random and noisy data sets; reproducibility of research and experiments; multi-modal processing; the ever-increasing cost of computing for increasingly complex neural language models; explainability of NLP systems; domain adaptation and personalisation.

He subsequently moved on to present the local environment for the development of the Polish language that encompasses CLARIN-PL (Language Technology Infrastructure for the Polish language) and its support for DARIAH.PL and DARIAH.Lab projects as well as CLARIN-PL -biz, which provides R&D infrastructure and support for AI in all areas of science and business applications. Also, he presented the offer of CLARIN and CLARIN-PL for research and scientific community especially from the area of humanities and social sciences as well as for business applications. Dr hab. inż. Maciej Piasecki gave an overview of corpora and language resources available from CLARIN-PL. Subsequently he presented the CLARIN-PL-biz consortium and its business supporters as well as its Technology Centre organisation, data resources and language tools available, and services offered (applications and analysis). Finally he invited interested entities to take advantage of the offer and cooperate.

3.3 Heading for the Future: Common European Language Data Space

Philippe Gelin, Head of Unit G3: Accessibility, Multilingualism and Safer Internet, DG CONNECT, European Commission, delivered a presentation on the Common European Language Data Space. In his online talk he presented the latest developments at the end of the Connecting Europe Facility Programme and its transition to the Digital Europe Programme (MFF 2021-27). The concept of data spaces, resulting from the realisation of the potential and value of data and integrated with the initiatives for a data-based EU economy, was introduced. Particular emphasis was put on the Common European Language Data Space (LDS) and its future role in the European LT landscape. Also, the concept behind the Centre of Excellence for Language Technologies (CELT) and Centre of Excellence in LT Plus (CELT+) was explained. A CELT will be established in order to facilitate the deployment of the LDS. On a strategic level, a CELT will coordinate the efforts of collection and creation of multimodal language data and platform bringing to consumers (users) a variety of additional language technology based services. It is envisaged that the CELT will be formed by Member States ministries representatives whereas CELT+ will comprise research and industry representatives, with strong emphasis on the latter group of stakeholders (60% minimum as per the text of the call closed on 22nd September 2022).

The deadline for submitting the calls for tenders for a Common European Language Data Space had been still open at the time of the workshop, hence the evaluation process had not yet started. No particular details of the results of the evaluation were known at the time of writing this report. The first calls for grants are envisaged for 2023. Depending on the interest and actual involvement of Polish entities from various sectors, and cooperation between them, as well as with foreign partners, funding might be obtained to develop LTs also for Polish.

3.4 eTranslation latest developments and Connecting Europe Language Tools

The eTranslation platform was introduced by Markus Foti, from the Commission's Directorate General for Translation (DGT). Apart from the history of the EC's machine translation tool, he also presented NLP tools (available at https://language-tools.ec.europa.eu/) which are offered to European public administrations, local and regional authorities, small and medium-sized enterprises (since March 2020), EU Freelance Translators, universities, non-governmental organisations (since 2022), and Connecting Europe Facility (CEF) projects. The initiatives and supporting actions under the CEF funding instrument are to be continued in the 2021-27 Multiannual Financial Framework under a new programme called Digital Europe (DEP). Services currently available through the language tools website include eTranslation, Multilingual Tweet, Speech-To-Text (currently for English, French, German, and Spanish), text classification, anonymisation, Interactive Terminology for Europe (IATE), European Language Resource Coordination (ELRC), and services to integrate the eTranslation with an organisation's online services. The CEF eTranslation services can be accessed via web upon registration (EU login needed) or via API.

Machine translation is available for all 24 official European languages and several other languages of social and economic significance such as Russian, Arabic, Chinese (Mandarin), Japanese, Turkish, and most recently – Ukrainian – added within a very short period of time in 1H2022 as a European response to the urgent communication needs resulting from the outbreak of war in Ukraine. However, it should be noted that the service is not available to non-Member States, and Ukrainian bodies and entities or individuals cannot benefit from it directly. The goal is to add more non-EU languages in the next phases.

Two important features that differentiate EC's eTranslation services from other, commercial services were underlined. Firstly, confidentiality of data processed within the system. The anonymization tool developed under the CEF-funded project MAPA (Multilingual Anonymization for Public Administrations) utilizing anonymisation techniques like obfuscation and replacement was mentioned.

Secondly, domain coverage – in addition to general text, the following domains are included: EU formal language, Court of Justice Case Law, Cultural, Deutsche Bundesbank, IP Case Law, Ministère des Finances (France), Public Health, Technical Regulation Information Systems, Valtioneuvoston Kanslia. The best eTranslation results can be expected for formal texts related to EU policies. Translation of non-standard, new or creative text, single words or expressions, and basically anything that highly depends on the context is usually of lower quality. In the next phases of developing and broadening the eTranslation services it is planned to extend domain coverage (e.g. adding scientific text, social media), add more language technologies and language coverage (i.e. add more languages for Speech-to-Text, anonymisation, named entity recognition (NER), and a basic CAT tool).

3.5 Language Technologies by/for the public sector: presentation of implemented solutions

The session was split into two parts. In the first one, Dr. Marek Kozłowski, Head of Laboratory of Linguistic Engineering at the National Information Processing Institute (OPI PIB) outlined NLP evolution in the context of the Institute's data sets for Polish and language models from 2013 up till present. He discussed data as the key element of the unsupervised and supervised learning, mentioning OPI PIB's dedicated inhouse repositories and raw text corpora used for language model pretraining. He moved on to present Neural Language Models published by OPI PIB including transformer based NLMs, such as RoBERTa, BART, GPT-2 (2020-22); sentence transformer paraphrase models, or long-formers (2022) as well as simpler shallow language representations (context-free ones), such as Polish Word2vec,

Glove, FastText. Next Dr. Marek Kozłowski described how the Polish RoBERTa models were trained and tuned. He presented systems with NLP components developed at the Institute including Uniform Antiplagiarism System (JSA); ARBUZ, an intelligent tool capable of scanning, detecting and analysing consumer contracts for provisions that violate consumer rights; ANSI which automatically assesses the quality of products based on the information available on the Internet supporting various languages (with the emphasis placed on Polish, English, and German), developed for the Polish Office for Competition and Consumer Protection (UOKiK). He closed the presentation by inviting participants to download and use the freely available OPI PIB's resources (link: Swagger UI (opi.org.pl)

In the second part of the session the history of a COVID hotline chatbot developed in 2020 for the National Health Fund and the Ministry of Health by IBM company was presented with a pre-recorded video (recorded in November 2020 for purposes of a conference on e-health solutions, Forum e-Zdrowia). Michał Gawryś, the then IMB high-level manager gave a summary of the origins and implementation of the tool utilized to provide customer service in health crisis situation, which was among the first use cases of chatbots for automation of communication at public administration. Practical details of cooperation between industry leaders and public administration, and its voluntary, non-commercial character were described as well as the effectiveness of the tool. It was mentioned, for instance, that at peak the chatbot was able to answer approx. 50 thousand inquiries per day, thus considerably reducing the workload for the hotline staff.

3.6 LT requirements and offerings: do they converge? (Q/A session)

Due to a slightly prolonged duration of the previous session(s), the moderator, Anna Kotarska, decided to utilize the remaining time for commenting on other LT solutions for Polish for public administration in place that could not have been mentioned due to time limitations of the workshop. One of the examples was the Neurosoft anonymisation tool for court judgements published on the portal of the Ministry of Justice. The solution dating back to 2011 was presented at 1st webinar on the use of AI in the justice field in March 2021 (Anonymisation and pseudonymisation of judicial decisions) organized by DG JUST (Link: Conferences and events (europa.eu)).

Next she commented on the CEF-funded MAPA project (2020-21) in the context of Polish input to the project implementation (beginning with providing letters of intent from the Ministry of Foreign Affairs, Medical University of Gdańsk and the Polish Society for Health Programs to support the tender, to the tests on Polish carried out in the final phase in December 2021 in cooperation with NASK National Research Institute, followed by reporting to the EC on the project results and their subsequent dissemination). She pointed out several aspects brought to light in the course of the project: low interest of Polish PA entities in submitting grant applications for financing of joint projects under CEF, low involvement of CEF POC (Point of Contact) at the then Ministry of Digital Affairs (and other competent ministries) in promoting calls for tenders in the Automated Translation category. Moreover, legal constraints combined with conservative attitude of the PA legal departments considerably limited the size and scope of data available for tests of the MAPA tool.

Apart from the above examples, Anna Kotarska mentioned yet another solution for anonymisation, (called "Anonimizator"), developed within the CLARIN-PL-biz project at the Wrocław University of Science and Technology for potential users from the industry.

3.7 The value of data for the development of top-quality LT

The value of data for the development of top-quality LT was discussed by Dr. Tomasz Walkowiak from the Wrocław University of Science and Technology based on the example and experiences from the projects implemented at the University, mainly CLARIN-PL. He pointed to the language data as being at the core of the project which focuses on language data (resource) creation, maintaining a language data repository; data being input information for CLARIN-PL infrastructure, and serving as a basis for building language tools. Dr. Tomasz Walkowiak described the process of machine learning (ML) emphasising the significance of properly annotated data in the ML process, mentioning typical issues and limitations encountered such as quantity of data available, high cost of preparation of annotated data, and mislabelled data. Next, he discussed knowledge transfer as a remedy for some of the these problems, including: transfer learning, using large language models (LLM), SOTA models/AutoNLP, and cross-language transfer. Dr. Tomasz Walkowiak mentioned personalized prediction applied to various subjective natural language processing (NLP) tasks (emotions, hate speech). He moved on to talk about extracting relationships between fragments of texts including semantics and sentence similarity. Strong emphasis was put on data cleansing and the need to remove personal data from the utilised data sets. The speaker pointed to the fact that the extracted textual data are often "noisy" in the sense that they contain many (source-specific) contaminations, are distorted or corrupted, as well as to the need to minimize "noise" in order to obtain usable, quality language data/corpora.

3.8 Language data creation, management and sharing: existing practices and challenges (Panel session)

The panel session was moderated by Anna Kotarska, Public Services NAP. Five panellists representing various sectors (policy makers/advisors/strategists, academia, state research institutes, industry, LSPs) and regions (Warsaw, Wrocław, and Gdańsk) accepted the invitation for the discussion:

- Tomasz Klekowski, Working Group on AI at the Chancellery of the Prime Minister (GRAI) (policy subgroup), C-suite level manager in IT industry
- Dr. Marek Kozłowski, Head of Laboratory of Linguistic Engineering, National Information Processing Institute (OPI PIB)
- Dr. Tomasz Walkowiak, Wrocław University of Science and Technology
- Marta Bartnicka, Dolby Laboratories (formerly IBM)
- Wojciech Wołoszyk, POLOT Polish Association of Language Services Providers (LSPs)

The moderator opened the panel by briefly introducing the panellists and requesting them to extend their introduction by providing more details of their professional background, current roles and fields of interest. The discussion focused around two major questions: "What are the main challenges for sharing language data in Poland?" and "What would need to be done to improve the situation and facilitate the sharing of language data? Each panellist commented from her/his own perspective. Their key messages can be summarised as follows:

Tomasz Klekowski: Although data is fundamental for AI development and its importance is reflected in the structure of GRAI – there is a "Data Subgroup", language data collection/processing is not financed in a satisfactory degree. Suggested investment by the state could be justified by the fact that language data are a common, horizontal, national asset, relevant to all citizens and to all domains. GRAI expert group monitors the implementation of "Policy for the development of AI" published in December 2019. The document contains references to building of Polish language corpora, LT

solutions for Polish, and in particular machine translation. However, there exists a need to translate policy into specific actions, setting specific goals and priorities for opening data that should be guided by their usefulness. Barriers to opening data result from a siloed character of public organisations, and (more) coordinated governance by the state is needed. Poland should also focus on building up its potential for the digital market by putting more emphasis on local, national projects as currently Polish AI sector delivers mainly for foreign markets and international projects. Cooperation with local (provincial) authorities to obtain language data seems to have a large potential due to high agency and involvement of local authorities.

Dr. Marek Kozłowski: Unless there are specific restrictions to making data open for re-use such as, for instance, national defence or energy security issues, data should be open by default. Barriers to opening data may result from their unclear ownership status or foreign ownership.

Dr. Tomasz Walkowiak: Research materials such as texts of PhD theses or dissertations should be open by default and higher education institutions should make them available as a rule. Resources created with the use of public funds should be available for re-use (mandatory requirement), and it is the role of the state to provide relevant legal frame for data sharing.

Wojciech Wołoszyk: Despite the PSI directive being in place, internal ministry public procurement procedures constitute a main obstacle to acquiring language resources in the form of re-usable translation memory (TM) files. This results from poor or outdated knowledge of LT and CAT tools on behalf of procurement units staff. Only the Ministry of Foreign Affairs has an internal translation unit and has a good understanding of how CAT tools work. In almost all cases in public tenders LT or CAT tools use is practically non-existent, and frequently the use of machine translation (or "translating machines"), is forbidden, which is contrary to the actual market practice, including that of the EC. The EC's (DGT's) practices and interpretations could be used as an example to follow. Delivering TMs along with translation should be regulated under a contract for language services.
The translation of legal texts and judgements into Polish (from English and French), was given as an example of where such requirement could be introduced to acquire not only bilingual parallel texts but also high-value language resources in the form of TMs coming from human translation.

Marta Bartnicka: Public sector could follow the path taken by the industry, where the issue of language resources (both translations and TMs) generated in the course of translation) was addressed many years before CAT tools and MT became popular. The issue was regulated already in the tender specifications, specifying that using CAT tools was obligatory. Delivering TMs along with translation was a formal requirement. Similar clauses could be used in contracts for language services commissioned by public administration, while retaining some flexibility of requirements (i.e. in case of less popular languages or express services it might be impossible to find a CAT-proficient translator). Translation memory management should be implemented due to, e.g. confidentiality issues, and unrestricted TMs should be separated from those that need anonymisation. High quality of data coming from human translation (90% usable) vs. data coming from crawling (10% usable) is a strong argument for adding this requirement.

3.9 Establishing National Language Technology Platforms in EU member states

In the final session the National Language Technology Platform and other solutions and applications developed by a Latvian company Tilde, ELRC consortium member, were showcased. Kaspar Kauliņš, company Business Development Director, presented the key language technologies developed by Tilde including custom neural machine translation, natural language understanding/processing (NLU/NLP), automated speech recognition (ASR) and synthesis, and conversational AI. He stressed language diversity in Europe, limited knowledge of English among European citizens and its consequences for consumer behaviours, as well as other challenges related to minority and under-resourced languages resulting in the threat of their digital extinction.

As the European Parliament Resolution states, the EP supports the development of multilingual public e-services in European, national [..] administrations with innovative, inclusive and assistive LTs, which will reduce inequalities among languages and language communities, promote equal access to services, stimulate the mobility of businesses, citizens and workers in Europe and ensure the achievement of an inclusive multilingual Digital Single Market. To address the multiple language-related challenges, Tilde developed the EU Council Presidency Translation tool (including its variant: the German EU Council Presidency Translator). Examples of the projects for the Finnish Prime Minister Office followed – of a government-wide machine translation, as well as an MT system for the Parliament of the Republic of Lithuania.

Next the speaker presented hugo.lv, a language technology portal for Latvian e-gov, available to all citizens for free, and similar solutions for other Baltic Countries inspired by the success of the hugo.lv platform. The platform encompassing several LT services, was described in detail including its development process. The experience collected with launching and deploying the platform was also presented. This platform served as one of two predecessors to a current National Language Technology Platform (NLTP) project. In his talk Kaspar Kauliņš described the NLTP project, the collection of mono- and multilingual data from five partnering countries (Latvia, Estonia, Croatia, Iceland and Malta), the development of the platform that will encompass a number of LT services, predominantly MT and CAT, but also terminology management and speech processing modules (ASR, TTS) for selected languages.

Tilde's response to war in Ukraine and resulting communication (translation) needs is also worth mentioning. The company launched an open MT platform available for non-EU residents. Languages offered include Polish, which, given the geopolitical context is a highly useful alternative to the eTranslation service.

3.10 Summary and closing remarks

Anna Kotarska, PS-NAP, gave a summary of the workshop, stating its goal, which was to provide a comprehensive overview of the latest trends in LT solutions and their dependence on the language data, as well as to show practical examples of use cases both on a national and international scale. She stressed that different perspectives of various stakeholders were taken into account during presentations and discussions, including that of the translators delivering translations for, among others, the public sector. She mentioned numerous suggestions made by representatives of academia, industry and LSPs aimed at improving language data sharing. She pointed to the fact that cross-sectoral cooperation between the industry, academia and public sector, raising awareness actions and knowledge and best practices sharing were key issues raised by all presenters and panellists during presentations and the panel discussion.

The moderator expressed her hope for another onsite or hybrid meeting with the audience in the following year, already under a new heading of the Common European Language Data Space. She summarised the timeline of the call for tenders with regard to creation and governance of the LDS and reminded the date of the envisaged launch of the new programme – that is January 2023. She also called for the involvement of Polish entities, including public administration, in the activities to be carried under the LDS and cooperation with the new consortium to be announced.

Anna Kotarska asked the participants once again to give their opinion on the workshop by filling in the evaluation form as well as encouraged them to provide additional descriptive feedback. She thanked the participants for attending the workshop, and the speakers and panellist for their substantial input and closed the workshop.

4. Synthesis of Workshop Discussions

The take-home message could be summarized as follows:

- A noticeable progress has been made in the area of language technologies in recent years, mainly thanks to the progress in technology (AI, computational capacities, data resources), which is reflected in the activities and projects carried out by key national research institutes and universities as well as the launch of GovTech initiatives. The number of LT use-cases is growing, also in public administration, nevertheless, the quality of CEF Language Tools for Polish is below those available on the market;
- Despite an over 4.5 year time span between the 2nd and the 3rd ELRC Workshop in Poland, the latter one gathered the largest number of participants since the launch of the ELRC programme. Over the past MFF the "brand" ELRC has been gaining visibility and is recognisable to a wider and more diversified audience. Higher frequency of local (country) events could be considered to upkeep the positive trend.
- Adopting a first national policy on the development of AI in December 2019 has laid formal grounds for commencement of a number of actions related to LT. Establishing a Working Group on AI in 2021 with a subgroup on data shows that the understanding of value of data has been increasing, however, more recognition is still needed with regard to language data. Also setting specific goals and monitoring progress and delivery is necessary.
- To support research and development of language technology, the ongoing funding for related projects and infrastructure is much needed and should be secured.
- There is a need for inter-institutional and cross-sectoral cooperation to be stimulated and formally guided by a competent ministry (The Chancellery of the Prime Minister) to break the silo modus operandi.
- Legal constraints remain a major barrier in data collection and re-use thus hindering the development of language technologies.
- Educating public administration staff about LTs, CAT tools as well as developing guidelines for public procurement taking into account the current state of the art of the LT seems necessary to break the impasse with regard to translation memories collection, sharing, and re-use.
- The geopolitical developments (war in Ukraine), have shown a need for communication support in crisis situations, including support in the form of language technology. This potential may be only used to its fullest if relevant language resources of appropriate quality are available.
- The launch of the Common European Language Data Space and implementation instruments such as EDICs and MCPs provide an opportunity to take language technology for Polish to the next level via international cooperation and funding. Intensified direct communication by/supported by the EC and providing an area for interaction could be instrumental in involving national policy makers and stakeholders in activities throughout the Digital Decade.

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

An updated profile for Poland with respect to language data creation and management has been recently published in the 2022 ELRC White Paper (see: https://lr-coordination.eu/sites/default/files/LRB/LRB-12/ELRC-White-Paper.pdf).