The Language Data Space (LDS) is a service contract operating under the DIGITAL Work Programme 2021-2022.

CNECT/2022/OP/0026
Common European Language Data Space (LDS)

EUROPEAN
LANGUAGE
DATA SPACE

D10.4.6 Slovakia Workshop Report

Author(s):	Miroslav Zumrík and Kristína Bobeková (Ľ. Štúr Institute of Linguistics, Slovak Academy of Sciences)
Dissemination Level:	Public
Version:	1.0
Date:	2024-12-05
©:	2024 – European Union. All rights reserved.

Service contract no.	CNECT/2022/OP/0026
Project acronym	LDS
Project full title	Language Data Space
Type of action	Service Contract
Coordinator	Prof. Georg Rehm (DFKI)
Start date, duration	19.01.2023, 36 months
Dissemination level	Public
Contractual date of delivery	05/12/2024
Actual date of delivery	05/12/2024
Deliverable number	D10.4.6
Deliverable title	Slovakia Workshop Report
Type	Report
Number of pages	10
Contributing partners	ELDA, DFKI
Task leader	ELDA
Author(s)	Miroslav Zumrík and Kristína Bobeková (Ľ. Štúr Institute of Linguistics, Slovak Academy of Sciences)
EC project officer(s)	Philippe Gelin (DG CONNECT)

For updates on project activities, and other LDS-related information, contact:

Prof. Georg Rehm	coordination@language-data-space.eu
DFKI GmbH	Phone: +49 (0)30 23895-1833
Alt-Moabit 91c	Fax: +49 (0)30 23895-1810
D-10559 Berlin, Germany	

Table of contents

1 Introduction

The Slovak ELDS (European Language Data Space) workshop took place on 7 November 2024 from 9.30 to 16.00 CET at the premises of the Slovak European Commission (EC)'s Representation in Bratislava, Slovakia. It was organised by the Department of Slovak National Corpus at the Ľ. Štúr Institute of Linguistics, Slovak Academy of Sciences. The Slovak organisational unit for the workshop were Miroslav Zumrík and Kristína Bobeková, in close cooperation with ELDA representatives Victoria Arranz and Fernanda González Campo.

The main objective of the workshop was to raise awareness on a) the ELDS project, and b) the state-of-the-art, possibilities and challenges for the Slovak Artificial Intelligence (AI) and Natural language Processing (NLP) research and industry communities and Language Technology (LT) producers.

The content was aimed at Slovak publishing and media houses, libraries, archives, press agencies, but also academia and, most importantly, all industrial branches where language and other data are massively produced. In this context, the agenda provided a presentation of the ELDS initiative, offered by the representatives of the (EC) and the DFKI (German Research Center for Artificial Intelligence).

Last but not least, the workshop has served to increase mutual awareness between the various activities, projects, challenges, needs and efforts of the Slovak AI, NLP and LT-related stakeholders. The general purpose of the event can thus be characterised as another initiative for creating a more functional Slovak research and industrial network in this area, following projects such as ELDS.

2 Workshop Agenda

09:30 – 09:45	**Registration and coffee**
09:45 – 09:55	**Welcome and Introduction** *Miroslav Zumrík, Ľ. Štúr Institute of Linguistics of Slovak Academy of Sciences (in Slovak)*
09:55 – 10:00	**Welcome by the representative of the European Commission's Directorate General for Translation in Slovakia** *Barbora Tricoire (in Slovak)*
10:00 – 10:30	**The Digital Europe Programme and The Common European Language Data Space** *Mateusz Szoturma, Representative of the European Commission (in English)*
10:30 – 11:00	**Common European Language Data Space: Developing a Market for Language Data and Services and Benefitting From a Joint European Effort** *Simon Ostermann, Representative of the Language Data Space & DFKI (in English)*
11:00 – 11:30	**Coffee Break**
11:30 – 12:00	**Importance and Advances in Language Data Processing in Slovakia** *Radovan Garabík, Ľ. Štúr Institute of Linguistics of Slovak Academy of Sciences (in Slovak)*
12:00 – 12:30	**Horizon Europe Project Disai** *Marián Šimko, Kempelen Institute of Intelligent Technologies (in Slovak)*
12:30 – 13:15	**Language Data and Language Technologies in Slovakia and for Slovak – Panel session (in Slovak)** *Miroslav Zumrík (Moderator)* *Michal Gregor, Kempelen Institute of Intelligent Technologies*

	Marek Šuppa, Sli.do & Faculty of Mathematics, Physics and Informatics of Comenius University Bratislava
13:15 – 14:15	**Lunch Break**
14:15 – 15:00	**Language Data Production, Management and Market Development: Overcoming Obstacles – Panel Session (in Slovak)** *Miroslav Zumrík (Moderator)* *Filip Bednárik, Essential Data* *Andrej Greguš, Nettle.ai*
15:00 – 15:15	**Conclusions** *Miroslav Zumrík, Ľ. Štúr Institute of Linguistics of Slovak Academy of Sciences*
15:15 – 16:00	**Coffee and networking**

3 Summary of Session Content

Workshop Structure

The event started with a brief *Welcome and Introduction* by the representative of the local organisers, Miroslav Zumrík from the Ľudovít Štúr Institute of Linguistics, who gave an overview over the context for the workshop. The event was presented as a loose follow-up to the series of ELRC (European Language Resource Coordination) workshops, which also took place on the very same premises, back in 2016 and 2018 (as physical events). This fact alone creates a certain sense of continuity which is very much needed for creating a fruitful environment for synergies between people, projects and technologies. In the introduction, Miroslav Zumrík presented the structure of the event, with all programme posts, and gave organisational instructions to participants. The introductory part was completed by the *Welcome by the Slovak EC Representation* given by Mrs. Barbora Tricoire. She expressed her appreciation for the fact that events like this are being organised, so that the role and possibilities of language technologies and artificial intelligence are given their well-deserved attention. Mrs. Tricoire especially mentioned the projects concerning the question of how to deal with a wealth of often misleading information spread throughout the internet and labelled dis- or misinformation in Slovakia.

After the introductory parts followed the presentation named *The Digital Europe Programme and The Common European Language Data Space*, given by the EC representative Mateusz Szoturma. The presentation shifted to a European level and focused on the Digital Europe Programme and the rationale behind the European Language Data Space. Mr. Szoturma accentuated the revolutionary potential hidden in the sheer amount of (language) data surrounding us, which also motivated the creation of the 2030 Digital Decade, based on cutting-edge technologies for people, rules and democracy. He then presented the basic LT-related landscape, where language technologies can be considered crucial for preserving identity and competitiveness. The importance of LTs is reflected in the EC's actions focused on AI talent, computing and data. The EC's actions resulted in the creation of Data Spaces (such as ELDS), European Digital Infrastructure Consortia (ALT-EDIC) and the DIGITAL Programme (Digital calls 2024). These results are meant to be a bridge between the public and private sector. Mr. Szoturma proceeded to explain the what, why, who, how and when of the ELDS project and commented on the basic architecture of the system, founded on data, metadata, its transfer and exchange through data space connectors. Attention was also given to the what, why, who, how and when for the Alliance for Language Technologies EDIC EDIC (ALT-EDIC), where Slovakia currently has the status of an observer. As for the Digital Work Programme of 2024, there were three already closed calls with a total budget of approximately 49 million euro.

The follow-up and tightly connected presentation was given by Simon Ostermann from ELDS and DFKI, titled *Common European Language Data Space: Developing a Market for Language Data and Services and Benefitting from a Joint European Effort*. Mr. Ostermann provided a contextual overview of the European Language Data Space in the current data landscape, starting with the rise and unprecedented capabilities of Large Language Models (LLMs). It is to be said, however, that Europe somehow lags in harvesting the potential of innovative technologies, as compared to US and China. Then, there are several European initiatives to overcome the lag, while facing challenges like availability of data for European languages, HPC facilities, as well as the speed of the big technology players in the US and Asia alone. As the data shows, practically all European languages, except for English and, to a certain degree, German, Spanish and French, are massively under-resourced in terms of both data, tools and technologies. Moreover, the gap between English and the other languages seems to get bigger, not smaller with time. The European Union (EU) has employed a data strategy to resolve this problem and has now a long history of language data sharing, also in the form of data spaces. Mr. Ostermann presented several data initiatives at EU level, ELDS being one of them. The ELDS is a procurement type of action, involving four core partners and together building a consortium with its subcontractors. Then, the different classes of data and the ALT-EDIC were presented, as well as the implementation of the ELDS. Here, both the architecture of the LDS connectors and data were once again brought to the forefront. The process of getting involved with the project was shown step by step (via the LDS Connector Management Panel): this allows creating new data sets, creating and adjusting policies, as well as creating and publishing offers. Furthermore, a historical outline of creating language resources and data products was sketched, and then the relationship between LDS and data products explained. Finally, the value proposition for LDS was described, including parameters for data products in LDS, and the next steps were proposed, also to be taken within the LDS user group.

The presentation on *Importance and advances in Language Data Processing in Slovakia* was given by Radovan Garabík from the Ľudovít Štúr Institute of Linguistics, the local expert on Slovak NLP. He started with the historical roots of the NLP explorations back in the 1950s and continued up to today's corpora and corpus tools, also for Slovak. The talk was not strictly scientific, as there was given space to some language- and corpus-related curiosities as well (such as the search for the longest word in the Slovak language, etc.). Mr. Garabík explained the differences in corpora sizes and what the respective sizes enable linguists and data analysts to do, starting with 1 million token-corpora for lemmatization, up to 13 trillion (ten to the power of twelve) token-corpora, the foundation for GPT-4. Then, there were presented use cases concerning lexicography and other linguistic research, such as the examination of the loanwords percentage in Slovak language. Subsequently, the topic was shifted towards use cases within information retrieval, machine translation, named entity recognition, sentiment analysis, text classification and other tasks in NLP. A special attention was given to the "breaking point", that is, the emergence of LLMs. The talk was concluded by a comment on individual corpora provided by the Ľudovít Štúr Institute of Linguistics, also with respect to less resourced languages, such as Slovak. Then, the LLM-related use cases were listed up, such as detection of spam or fake news, as well as the approaching new level of human-computer interaction (to be utilised in information retrieval, AI entertainment, etc). Last but not least, Mr. Garabík mentioned the unresolved, yet important issues within LLM, such as vendor lock-in, copyright question and non human-creative work. In this way, a new question is to be answered and regulated – the precedence of AI as a juridical person. He gave the overview status of data, artificial intelligence and language technology, emphasising their impact at a social and economic level.

Tying in with some of the issues in the previous talk followed the presentation on the *Horizon Europe project DisAI*, which is aimed at improving scientific excellence and creativity in artificial intelligence and language technologies to fight disinformation. The talk was given by Marián Šimko from the Kempelen Institute of Intelligent Technologies, which was presented as a non-profit research institute focusing on AI, established in 2020 in Bratislava, Slovakia and currently hosting more than 50 researchers and research engineers (including PhD students) and still growing. The institute has connections to 27 countries and more than 35 industry partners at the moment. The aim of the institution can be summarised as to propel the economic transformation in Central Europe. The institute conducts basic research closely linked to applied research and knowledge transfer, which is fully in line with the objectives of the ELDS initiative. The institute's know-how was inspired e. g. by pillars for the non-profit science system in Germany – it does not reinvent the wheel, so to speak. As for the DisAI project in particular, it aims to enhance the scientific excellence of KInIT and the consortium partners in trustworthy AI, multimodal natural language processing and multilingual language technologies to combat disinformation. It is a CSA (Coordination and Support Action) type of project within the Twinning scheme, with research component included as its special feature. The project has partners abroad (DFKI in Germany, CERTH in Greece, and the University of Copenhagen in Denmark) and it approaches "disinformation detection" by using IR/NLP tasks, such as claim detection/generation, claim matching, classification of document stance towards the claim. Multilingual Language Technologies open up for a few issues, such as:

- How can claim matching performance be improved by creating a multilingual machine learning model, i.e., a model using training data from multiple languages at the same time?
- Would there be a positive transfer of knowledge between languages?

and several others. Because of the relative novelty of the AI-related "terrain", development of a trustworthy AI is a responsible task, even here with open questions such as:

- How to explain the predictions of claim matching models understandably and objectively correctly to the users?
- How can we use explanations to increase robustness of neural models?
- How can we detect and mitigate unwanted biases in claim matching models?

The rise of GenAI and LLMs has definitely changed the NLP landscape, and it can be concluded that: LLMs are state-of-the-art for a vast number of NLP tasks, while, at the same time, capabilities of LLMs for low-resource languages are limited. Up to present, the Institute has managed to a) connect the Slovak NLP community (so far, in its academic part, but moving on), b) start an initiative that aims to improve the large language models and their usefulness for the Slovak language. It also aims to fine-tune several foundation models on Slovak data. Presence of suitable and efficient language technologies is crucial, while it is critical to develop and maintain high-quality language resources, especially for low-/limited resources languages, such as Slovak. As with other programme posts, there was room for questions from the audience, concerning for example the notorious issue of how to reliably discern and legally regulate discrimination between information and (dis/mis) information.

The presentations were followed by the first panel discussion, on *Language Data and Language Technologies in Slovakia and for Slovak.* The session was hosted by Miroslav Zumrík and the guests were Michal Gregor, representing the Kempelen Institute of Intelligent Technologies, and Marek Šuppa from the company Sli.do & the Faculty of Mathematics, Physics and Informatics of the Comenius University in Bratislava. The panel session was thus smaller as to the number of participants, the rationale of which being to provide more space for individual speakers, and resonance toward the public. The choice of panelists was motivated by the idea of bringing together the domains of research (at the Kempelen Institute and Comenius University) and industry, the latter represented by Sli.do. The moderator started by asking both participants about the general profile of the company and/or organisation they work for. This was done so that the public would get a precise picture of what their NLP and AI activities consist in. The questions were also focused on the issue of new generations of NLP and AI specialists. Both at the Kempelen Institute and Comenius University, this future-oriented task was considered to be very important. At the moment, there are several PhD-students at the Kempelen Institute, while Comenius University has a long tradition of offering AI- and NLP-related courses. The Sli.do company specialises in developing solutions for how to engage event participants with live polls, Q&A, quizzes and word clouds, be it during business meetings, webinars, conferences or educational settings. The participants agreed that while conditions for educational and professional development for future generation might face some limitations as to governmental regulation and backup, the will to overcome the obstacles and create a functioning environment between present and future stakeholders in the Slovak AI and NLP remains strong. Here, the initiatives like the one organised by Kempelen Institute (with the aim of bringing together Slovak researches within AI and NLP) earlier this year, or the ELDS initiative itself, was approved by both panelists. Some time was reserved for questions from the audience.

The final discussion took up a similar topic, as it discussed *Language Data Production, Management and Market Development* and how to *overcome obstacles* in these processes. This segment was also hosted by Miroslav Zumrík as moderator, and the participants were Filip Bednárik from Essential Data and Andrej Greguš, representative of Nettle.ai. Both (Slovak) companies work on developing innovative and state-of-the-art digital solutions. Essential Data are primarily oriented on software for the retrieving and processing of Slovak data, creating user-friendly agenda systems, as well as optimising processes in the public administration (using Java and Java Script). The ongoing projects deal with e. g. making Slovak judicial decisions and Slovak legislation more accessible and easily searchable for users. Here, one of the major technological challenges would be to anonymise decisions in a reliable way, or to offer effective tools for navigating the users in the complex text corpus of legislation. While the company itself might not be big in terms of the team size, the applied solutions are rather sophisticated. The same can be said about the Nettle.ai, which works within bringing “digital humans to life”. The idea is to blend lifelike digital humans and mixed reality technology to deliver exceptional “phygital” (physical and digital) experiences. In other words, the research shows that the more smoothed out the human-computer interaction is, the more comfortable and willing to buy the customer is. Here, one could say that despite being a relatively small Slovak-based company, the products are on a par with technologies abroad. AI-Powered Digital Humans: the company’s in-house avatar engine is a “cornerstone” of the “browser-based and phygital experiences”: it features “customisable digital humans in real-time 3D scenes with facial animations and gestures”. The engine has “extensive configurability and scalability” and offers “complete control via purpose-built APIs, including lip synchronisation based on freely generated speech audio”. A fluid conversation with a realistic digital persona is enabled by an advanced natural language processing (NLP), which means that this company as well could be interested in becoming a partner of the ELDS initiative. The company also offers an end-to-end engagement platform: The Nettle “BOT Platform” consists of a “proprietary Conversational AI engine” built entirely by the company itself, and “leveraging industry standard tools sets, custom learning algorithms and highly modelled business domains”. The robust and flexible platform can work in tandem with any openly-available LLM, including the major ones worldwide. Last but not least, the company developed the “Nettle Voicebot Platform”, aimed at “utilising automated voice interfaces in the engagement process”, and it also incorporates “vendor agnostic, industry available Speech-to-text (S2T) capabilities”. Although both companies, similarly to those in the previous panel, face some, not insignificant (legislative) difficulties, they have been able to work on overcoming these with the aim of building reliable and respectable Slovak AI and NLP solutions in the future as well.

In the brief *Conclusions* segment, Miroslav Zumrík summarised the programme and some of its main findings (commented below). He also expressed hope that the participants experienced the workshop as meaningful and inspirational, as well as a step in the right direction. The feedback provided by the guests (see attachments) shows a respectable interest of attendees in the discussed topics. The polls also show that the overall message has been rather successfully conveyed to the recipients, as the ratings can be characterised as more or less universal acclaim. An average 4,67 and 4,76 have been obtained for both polls, respectively, based on ratings between 1 and 5, the latter being the highest.

The workshop ended with a *coffee and networking* moment, which proved to be an effective and pleasant way to establish partnerships and engage collaborations.

4 Conclusions

The main issues, questions and concerns raised by the participants would be as follows:

- The status of LT development and adoption for Slovakia and Slovak is such that while still being one of the lower-resources languages, the potential for scientific and industrial utilisation is presented and already being harvested within the projects introduced at the event. It is important that the experts and companies know about each other in a relatively small country, so that they can coordinate and use the situation on the market to the fullest. It is also very important to raise awareness about NLP and AI solutions for Slovak, but also for other languages with regard to the general public, including representatives of publishing, media houses and other bodies working with and collecting large amounts of (language) data. In other words, the awareness about the inherent, even financial value of this data needs to be strengthened. While partially slowed down by the Slovak legislative regulation, which can create obstacles for exploring new technologies and procedures, all invited participants have shown resilience in overcoming these complications in one way or another. At the same time, the technologies behind the products proved to be an interesting topic for raising general knowledge about technological solutions present in our lives as of now. In order to make the objectives and workflow of the LDS platform easily graspable, the local organisers decided to send a follow-up mail to all participants, with instructions and useful links that can help them to be engaged with ELDS afterwards. This will be done shortly, so that the event will not remain a singular, albeit positively received event. The viability of data provision through the LDS platform needs to be thoroughly discussed by local legal experts, as the domain of AI brings about not only new solutions, but also new legal issues (as shown in the presentation by Radovan Garabík, to name just one instance). Judging from the immediate and later feedback of the participants, the ELDS initiative can nevertheless count on at least several Slovak stakeholders in the future, such as coordination of mutual efforts that would help with making, for example, Slovak membership in the established research infrastructures (CLARIN, ERIC) a reality at last. Also citing one of the remarks in the feedback forms, the idea of organising workshops like the Slovak one is definitely worthwhile.