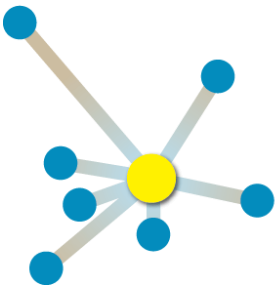
European Language Resource Coordination (ELRC) is a service contract operating under the EU's Connecting Europe Facility SMART 2019/1083 programme.



European Language Resource Coordination

Connecting Europe Facility

Deliverable D3.2.7
Task 3

ELRC Workshop Report for Slovakia

Author(s): Miroslav Zumrík

Dissemination Level: Public

Version No.: V1.0

Date: 2021-07-05

Contents

1 Executive Summary

The third Slovak ELRC workshop took place on 25 May 2021 from 9.00 to 12.30 CET as a virtual event. It was organized by the Department of Slovak National Corpus at the Ľ. Štúr Institute of Linguistics, Slovak Academy of Sciences (further referred to as SNK/JULS/SAV). The (Slovak) organisational unit for the workshop were Miroslav Zumrík (Slovak T-NAP, co-host and moderator), Kristína Bobeková (co-host) and Katarína Rausová (back-end moderator), in close cooperation with ELRC representatives Maria Giagkou, Stefania Racioppa and Eileen Schnur.

The main objective of the workshop has been to raise awareness on the state-of-the-art, possibilities and challenges for the Slovak AI and NLP research communities and LT producers, as well as the contemporary data and LT policies within the Slovak public administration and state authorities.

In line with the extended scope for the third series of ELRC workshops, the content was also aimed at representatives of Slovak SMEs and their actual or possible LT-related needs. In this context, the agenda provided a presentation of the automated CEF eTranslation service, offered by the EC to a now larger scale of users.

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, including public authorities and their policies.

The general purpose of the event can thus be characterized as an initiative for creating a more functional Slovak research and industrial network in this area. In other words, the workshop should help to create a snowball effect with the aim of supporting development towards a more digitalized and knowledge-based society in Slovakia.

2 Workshop Agenda

09:00 – 09:10	**Welcome and introduction** ***Miroslav Zumrík**, Ľ. Štúr Institute of Linguistics, SAS*
09:10 – 09:30	**The potential of Artificial Intelligence and Language Technology – where we are, where we should be heading** ***Mária Bieliková**, Kempelen Institute of Intelligent Technologies*
9:30 – 10:30	**Language Technologies in Slovakia / for Slovak - Panel session** ***Miroslav Zumrík*** (Moderator) ***Marián Šimko**, Kempelen Institute of Intelligent Technologies* ***Marek Košta*** *and* ***Andrej Greguš**, Nettle.ai* ***Libor Bešenyi*** *and* ***Peter Kostelník**, Xolution.sk* ***Marek Šuppa**, FMFI UK/Slido*
10:30 – 10:50	***Coffee Break***
10:50 – 11:15	**The CEF AT Platform** ***Francois Thunus**, DGT, European Commission*
11:15 – 12:00	**Language technologies by/for the public sector - Panel session** ***Miroslav Zumrík*** (Moderator) ***Milan Andrejkovič**, Ministry of Investments, Regional Development and Informatization of the Slovak Republic* ***Lukáš Palaj**, Ministry of Health of the Slovak Republic* ***Tatiana Hlušková**, Ministry of Economy of the Slovak Republic* ***Iveta Zraková**, Migration Office/Ministry of Interior of the Slovak Republic*
12.00 – 12:20	**The Value of Text Data** ***Radovan Garabík**, Ľ. Štúr Institute of Linguistics, SAS*
12:20 – 12:30	**Conclusions** ***Miroslav Zumrík**, Ľ. Štúr Institute of Linguistics, SAS*

3 Summary of Content of Sessions

3.1 Welcome and introduction

The event was opened by the moderator and T-NAP for Slovakia, Miroslav Zumrík, who welcomed the participants on behalf the department, institute and consortium. As he said, the third edition of the workshop was special because of the virtual environment, but on the other hand the limitations we are facing because of the pandemic also arguably provide us with new possibilities. After all, the use of language technologies that help overcome limitations and challenges in everyday life, including the language barriers in our surrounding is the very idea behind the workshops as such.

The moderator spoke about three points: firstly, he introduced the idea or concept behind the ELRC mission, secondly, he went through the event's programme and thirdly, he mentioned some of the organisational technicalities for the event.

As for the idea of the workshops, the moderator pointed to the fact that information is being generated at an exponential rate in contemporary society. The flood of data must be processed so that they exhibit their hidden potential for knowledge extraction. Thus, it is important to seek methods for effective information processing. One should also consider the existence of institutional settings, which are one of the most prolific producers of information, but where the potential value remains often unused. The information and data produced has at least three potential values: the original communication value for concerned parties, the research value for scientists, and technological value, serving as material and means for developing new processing technologies. This is also the case with the use of actual human translations in various language combinations for the purpose of training the EC's machine translation system, eTranslation. Apart from the use of authentic and large data amounts, it is also needed to create vivid human and knowledge transfer and networks, which is another rationale for the series of ELRC workshops. The participants represented a wider spectrum of branches and domains (public sector, research and technological development, academia, SMEs, as specified in section 6), which could enable exchange of experience and expertise.

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

The moderator gave the floor to the first presenter, Prof. Mária Bieliková from the Kempelen Institute of Intelligent Technologies, a rather new and independent institute for research within various branches of AI, including NLP.

Prof. Bieliková focused on the definition of AI and the challenges of the field. She stressed that the context of European AI research is important for Slovakia, and there are at least two aspects of AI being reflected by the research community: the aim for excellence and a risk-based approach. She mentioned the recently published Act on AI (April 2021) which regulates which AI applications should be banned, or restricted because of serious ethical risk. She identified a major challenge with respect to AI development, the lack of transparency in many projects, and at the same time stated that for further development, more data is needed. Machine learning has in any case changed the research paradigm, although the insiders now know how much effort and time this learning process takes. In this respect, Slovakia needs collaboration and connections.

At the European level, there are some relevant well established initiatives and projects, such as the EC's White Paper on AI (2020)[1], the AI4EU and Taylor projects (where also Slovakia participates) that started the effort for building knowledge centers. The possibilities for further development are vast and concern not only language, although it is language that defines humans as human. Among others, the digital economy and social indexing (as a high-risk enterprise) are some of the fields that are considered to be highly affected by the advent of AI.

With regard to AI development, Slovakia, however, still seems to struggle with a large scale use of AI. According to the Digital Economy and Society Index (DESI 2020), Slovakia remains in the last ten countries with regard to the use of connectivity, human capital, use of internet services, integration of digital technology and digital public services[2]. Two of the main factors that could help Slovakia advance in AI are, according to prof. Bieliková, human capital and connectivity.

There is no clear definition of AI, or, there are too many. They all, however, seem to share the statement that AI is software that enables the machines to behave intelligently, using human defined objectives that solve a problem. Various techniques and approaches (logic, knowledge-oriented, statistically-oriented) are defined in the AI Act. Statistical machine learning has therefore been introduced as part of AI, while in recent years the deep learning paradigm has emerged, which requires considerable computational power.

In the field of language centric AI, technology developers aim at systems that understand and generate human language, as is the case for instance in speech recognition and dialogue systems. The types of data required to train such systems are unstructured text or speech. Three basic approaches within AI are employed: a) rule-based, b) statistic – developed thanks to the availability of large corpora, and c) neural networks consisting of multilayered networks that yield interesting results in image, but also in language, processing. Natural language contains a large degree of variability, which is hard for machines to process, but neural networks and knowledge representation techniques, aka language models, attempt to address this issue.

A language model represents human language with millions of features. Building language models is a demanding task, in terms of not only the man- and computational power required but also in terms of its carbon footprint.

Finally, the importance of transparency was once again stressed. In this respect the research aim of the Kempelen Institute is to deepen the knowledge in the AI area, with search for new methods and answers of how to design AI systems ethically and transparently.

Prof. Bieliková concluded her talk by highlighting the issues that are critical for Slovakia in order to advance in AI: to identify the strategic sectors where AI systems should be developed; to build research capacities and communities (including academia, industry and other stakeholders); capacity building; participation in European and global research communities and infrastructures, such as CLARIN or DARIAH.

The moderator remarked that the rapid development of AI and its terminology could be reflected and researched by linguists as well, who could focus on this branch of specialized discourse. In fact, the research of newly developed terminologies and neologisms has been ongoing at the Ľ. Štúr Institute for quite some time now. The moderator then made the transition to the first panel section, as he

[1] https://ec.europa.eu/info/sites/default/files/commission-white-paper-artificial-intelligence-feb2020_en.pdf
[2] (https://ec.europa.eu/digital-single-market/en/desi)

stated that after seeing the vast possibilities and forms of AI domain, it is now time to look at whether/how these possibilities are used in Slovakia/for Slovak language.

3.3 Language Technologies in Slovakia / for Slovak (Panel session)

The first round of talks was devoted to presentations by participants:

Marián Šimko is the leader of NLP team at the Kempelen Institute of Intelligent Technologies (kinit.sk). The aim of this team is to push forward research in Slovak, especially with regard to development of linguistic methods, models, as well as transparency and interpretability of language models.

The representatives from the company nettle.ai, Marek Košta (chatbot development) and Andrej Greguš (business part) deal with the more practical or applied side of NLP (software engineering, deployment), but also theoretical sides are covered (language models).

The second delegation of the applied NLP team consisted of Libor Bešenyi (technical director) and Peter Kostelník (linguist) from the company Xolution.sk. The company offers Slovak digital assistants (chatbots) but it has still to overcome the scepticism in public/customers, as the first generation of chatbots was not as functional as expected by the broad public.

The last panelist, Marek Šuppa, represented Slido, a platform for managing audience interaction during larger-scale events. He is also a lecturer at the Faculty of Mathematics and Physics (applied informatics).

After the presentation round, the moderator posed the first question. Given that the possibilities are vast as presented by M. Bieliková, and that the situation within AI and LT development leaves something to be desired in Slovakia, as some of the panelist already hinted, the question was whether the panelists can identify AI areas where Slovakia excels.

The answers from all panelists were neither unequivocally positive nor negative, but rather mixed. A recurrent theme has been the comparison with the neighboring Czech Republic. Marián Šimko commented that the Czech market is much more developed (with regard to speech recognition, text to speech, text mining, dialogue systems), but at the same time, Czech language and Czech research also can and help with solutions for Slovak (e.g. the company Genea). It has to be stressed, however, he added, that the Czechs are also ahead with respect to commercialization of NLP products and technologies.

The representative from nettle.ai noted that Slovakia still has a long way to catch up with global developments in AI, although research at Slovak research centers and academia, such as the JULS Institute, is rather established and widely known in the research and development community. Slovak belongs to the less resourced languages, which poses a number of challenges when it comes to processing it. For instance, the development of Slovak chatbots needs much work and time and the gap between us and the surrounding world might even never be fully diminished.

The representatives from Xolution pointed at parallels between Czech and Slovak languages and LT situations, which can also be seen as motivating. In Slovakia, however, the resources and institutions are wide-spread, but this applies also partly to Czech academic and research centers (such as UFAL Institute at Charles University in Prague, or the University of Brno).

The representative from Nettle reported with that, although we are not the top-notch actor in this area, we can make use of niches that can evoke interest also abroad.

The other representatives agreed that the sketched situation in Slovakia can even be seen as an advantage, and that this might be good for future development after all, as the Slavic languages in general show less ambiguity and a renaissance of interest for the structural approach in linguistics (now somehow omitted) could be a welcome change.

In the ensuing discussion, another comparison with the Czech Republic was made, this time regarding business networking (with actors like Vocals, Newton, Phonexia). The Czech actors tend to create consortia (similarly to Nuance from USA) and aim at creating synergies, while networking is largely absent in Slovakia, despite a few attempts like the one by Peter Bednár from Košice, or by founders of the Slovak.AI initiative. In contrast to the Czech consortia, synergies and infrastructures, the Slovak actors need to rely more on themselves.

The participants pointed out that the Czech market itself is different and that the conditions there can hardly, if ever, be recreated in Slovakia. So, the general conclusion might be that we need to be realistic, while at the same time not lose the future vision and perspective. The situation is not ideal, there is room for improvement. Better networking is definitely needed, so we can work on better language models for Slovak.

The moderator affirmed that the synergies do not emerge easily, and pointed out at a similar online workshop (NLP in Slovakia) last June, where participants expressed their willingness to cooperate.

The next question was how fast such a development of a single solution proceeds, or how many solutions can companies like Nettle and Xolution provide per year. To this, it was replied that the customers have high expectations which must be adjusted to the real and harsh conditions of technology development. This might have caused the aforementioned scepticism after the first generation of chatbots was introduced. Now, in order to achieve a solution with high accuracy, some 10 months' work is needed, which is practically not feasible for a company to survive. The onboarding of customers must thus be faster and we need to change our approach and goals (set at approximately 10 chatbots/5 months), we need to consider the human capital and to create incentives for talent retention. The representative of Nettle also pointed out that the type of the customer matters (banks as big vs other, small enterprises) and that the future could eventually bring solutions retailed in packages, which would be an easy and fast solution for customers.

The representative for Slido and FMFI, Marek Šuppa, made the point that the reality of product development is more complicated than the general public may think. A so-called feedback loop is very important and time consuming. In order to be ready for market demands, we need labeled data (metadata) and investments. Slido is said to manage 1-2 projects in a year.

Marián Šimko added that the focus at the KINIT institute is mainly on technologies and applications with the aim to customize them, and perhaps focus on efficient sentiment analysis.

As for the input from the audience, Dr. Garabík reminded participants that within the ongoing ELE (European Language Equality) project[3], stakeholders and engaged parties (including the JULS Institute) will be contacted in a few months with regard to survey conduct and future cooperation in the area. The required information can concern all kinds of Slovak corpora, as the aim is to cover most languages and domain, including the commercial domain.

Towards the end of the discussion, Mária Bieliková remarked that, as the focus had been on comparison with the situation in Czech Republic, one aspect of it is the role of the Czech and Slovak

[3] https://european-language-equality.eu/

states. The Czech Republic invests annually approximately 1 million euro into the research infrastructures CLARIN and DARIAH, which needs to also be done for Slovakia. The good news is that the roadmap for the foundation of a national Slovak research infrastructure has been approved recently and, the active membership of Slovakia in these infrastructures could be possible in the foreseeable future. At the same time, a large part of the Pandemics Renovation Plan is devoted to digitalization and language development, so the prospects are indeed good.

In the section's very end, one question was addressed to panelists from the audience, namely, how to engage interest for AI issues and technologies in schools. To this, one panelist replied with his experience from an industrial high school in Prešov, Eastern Slovakia, where robotization of processes is taught. The students' response was overwhelmingly enthusiastic and the panelist stressed that they are very open for and supportive of educational initiatives for future AI experts.

3.4 The CEF AT platform

Francois Thunus (DGT, EC) presented the EC's work on automated translation, the eTranslation platform. Statistical MT was developed by the EC more than 20 years ago, followed in recent years by neural machine translation systems. Francois Thunus stressed one of the EC's principles and objectives, namely, to offer solution for ideally all European languages, not just some of them.

Initially, MT@EC, eTranslation's predecessor, was developed to address internal needs at DGT, but it subsequently opened up to European public authorities, universities, and SMEs, although this automated translation system is still not offered completely publicly. This is because the systems should not be overload, so the quota system is applied. It is always possible and advisable to get in touch with eTranslation service desk for requesting access to eTranslation.

eTranslation is available in the 24 official EU languages plus RUS, TR, JP, CN, IS, NO and Arabic. It provides not only a general language engine, but also domain-adapted engines, such as the EU formal language engine, health, culture etc. Depending on the availability of training data, tailored engines can be trained for specific domains.

Francois Thunus then accentuated that the demand for (big) data remains the crucial problem. For instance, in order to create a so-called "toy engine", one needs to collect at least 100k aligned sentences, whereas DGT uses some 200 million sentence pairs. Another valuable source for future development are translation memories, which are used by default and work well, given that these are official document translations. The aspect of context dependency is a strong one: Francois Thunus showed an example of homonymic expression (chair as chairman and a piece of furniture) that can yield misunderstanding, when the engine is not fed by appropriate type/genre/style of texts.

The EC is working on extending the domain coverage (e.g. scientific texts); on supporting additional non-EU languages of social & economic importance, and regional languages; on developing more language technologies, such as speech recognition, anonymization, named-entity recognition and a basic Computer-Aided Translation tool. Some of these tools have already been made publically available at https://language-tools.ec.europa.eu/.

Francois Thunus then provided an overview of the eligible users (Public Administrations, Universities, CEF-funded projects, SMEs) and described the steps and links to self-register and use eTranslation, which are as follows:

1 Self-registration via https://webgate.ec.europa.eu/etranslation/public/welcome.html

2 Web service (API) Technical documentation: https://ec.europa.eu/cefdigital/wiki/display/CEFDIGITAL/How+to+submit+a+translation+request+via+the+CEF+eTranslation+webservice

3 eTranslation Service Desk: help@cefat-tools-services.eu

4 Access to eTranslation web user interface: https://webgate.ec.europa.eu/ETRANSLATION

As for the audience questions, Mária Bieliková firstly thanked for the presentation which made her glad to be European, and asked for confirmation, whether Francois Thunus believes KINIT, as a research legal body, would in his estimation be considered a legit user of eTranslation.

Francois Thunus firstly remarked that a lot of important points have already been mentioned in Bieliková's presentation, and confirmed that gaining access for the institute should not be a major problem. He also elaborated on the problem of infrastructures, that is, why eTranslation is not freely open to the general public. Everything namely, must flow through and be processed at DIGIT (the Commission's Directorate General for IT), where they work on secure premises. Although the data is not going anywhere, there is a constraint for servers, which have their limits. In the case of interest shown by a research institute, granting access should not be a problem for the EC. Francois Thunus also mentioned the volume of texts being translated every year at European institutions, i.e. more than 70 million documents.

A representative from the National Bank of Slovakia posed a question whether the EC considers a feature allowing for uploading reviewed translated documents back, as well as the possibility of rating the translation output. Francois Thunus replied this type of feedback is not possible.

Conclusively, the moderator wanted to know (also with regard to the attendance of representatives from the Slovak Migration Office, whether the expansion towards more "exotic", non-European languages would be continued by the EC. Francois Thunus confirmed that adding RUS, AR, TUR reflects the growing need for such expansion, however, the problem remains the availability of data.

A final question from the audience concerned the possibility to select the domain in future versions of eTranslation. Francois Thunus answered that this is not a likely scenario, given the limitations of infrastructure and the sheer number of language combinations/potentially required engines.

3.5 Language technologies by/for the public sector (Panel session)

The second panel consisted of representatives of Slovak public/state authorities, where the question was to compare data and language policy and projects at the respective ministries and the first round was focused on individual presentation of their agendas.

The representative of the Migration Office, Ministry of Interior, Iveta Zraková, started with the characterisation that their crucial agenda is asylum politics and a direct contact with foreign citizens, thus multilinguality issues are the heart of the ministry's topics of interest. There are two main aspects of the language issue at the ministry: the communication with foreign citizens, and the communication with other migration offices and ministries in Europe (the latter being conducted mainly in English). On another level, it can be stated that the focus for the Slovak Migration Office lies on non-European languages. This communication has a specific nature, since the ministry/office needs to evaluate all aspects of a given asylum case. The office often works with non-professional interpreters, while there is need for certain language families and dialects (Arabic, Kurd etc.). Mrs. Zraková also stated that they are experienced with some of the presented technologies and they are continuously searching for novel ways of AI usage, especially for tools enabling recognition of dialects in order to identify the country of origin of asylum seekers. Mrs. Zraková pointed out that the situation during the asylum

process is special, often stressful, even more so in the times of pandemics, so the need for new and dynamic technologies only increases in time.

The representative of the second public authority, the Ministry of Economics, Tatiana Hlušková, started with precising that their aim is to provide better regulations for citizens, not just legislation. The regulations are perceived and processed from the semantic point of view. Mrs. Hlušková explained that they distinguish and define four elementary so-called regulations in every law norm: they are rights, obligations, definitions and sanctions. The importance for semantic processing of law regulation becomes understandable when one thinks about the amount of law regulation that in some way affect citizens in various life situations. The second part of their agenda is the creation of a list of so-called concerned persons. The aim with this processing task is to enable various subjects to get a register of all norms that concern the subject. The overall objective of such endeavour is to facilitate interaction between physical and legal personae, on the one hand, and on the other to make this interaction easier, more transparent, effective and less time consuming, for the sake of entrepreneurs and citizens.

Here, the moderator mentioned the Scandinavian, mostly Swedish, politics of klarspråk (clear language), which requires the simplification of administrative and legal texts so that they are comprehensible by all concerned parties, mainly by the wider public. From this perspective, the semantic tools for simplifying legal regulations seem to be in demand in Slovakia.

The floor was then given to the representative of the Ministry of Health, Lukáš Palaj. At the question of how they work with data at their ministry, Mr. Palaj replied that, to a large extent, they focus on speech and text processing. Medical practitioners mainly work with tools for speech to text transcription of medical records (mainly within radiology). Feedback from the radiologists who have used the tool indicated that such tools were not primarily developed for Slovak, but for Czech, so apart from Latin expressions, some Slovak expressions are problematic. The solution for this is that the doctors have to choose which part of the observation they would rather write manually and which they could dictate by using the tool. The demand for tools that have been originally developed for and trained on Slovak language data was unanimous. Another issue is how to achieve more structured records in the domain of eHealth. The medical records are numerous (65 mil. documents as of 2018), but to a certain degree unstructured. NLP may provide a very useful solution for extracting information from unstructured medical records, since unstructured records are of minimal use for further processing or even for informing future health politics. Mr. Palaj mentioned yet another tool, used in Denmark, which supports and guides medical emergencies personnel.

The last of state authorities representatives, Milan Andrejkovič, head of the Data Office (Section of Information Technologies for Public Administration, Ministry of Investments, Regional Development and Informatization), presented the Ministry's newly established Section of Digital Policies. Its aim is the digitization of Slovakia. The Data Office focuses on the collection and processing of public data, and on their transformation into structured data that can inform further political and strategic decisions. Another aim is to create semantic interoperability and to stop bureaucracy, in line with the so-called „once-only principle", i.e. a citizen should need to provide certain information to a public administration only once. Since the technological aspects are complicated, a pilot measuring data quality was required, in order to have high data quality and then be able to prepare a so-called central data model (including metadata). The ministry additionally works towards transparency: to have as much open data as possible, to make personal data of a citizen more aptly accessible for him/her and, to collect and process big data from the public administration. The Ministry's Section of Digital Policies then uses the public data to design the state's digital agenda. Mr. Andrejkovič mentioned that some first attempts for adding metadata and creating structured data were made at the Ministry of Justice, with the digitalization of case files. The Data Office, as hinted before, works for a larger scale

digitalization of all administrative processes required when life situations changes, as is for instance the birth of a child.

During the Q/A part of the panel discussion, Maria Giagkou from the ELRC consortium asked if there were any preliminary results of the Slovak Digital Transformation plan for the years 2019 – 2022 and 2030, respectively.

To this, Milan Andrejkovič mentioned several support tasks at the Digital Office, such as innovative machine learning and testing, development of the new open data portal), creation of an interoperable consolidated analytical layer, which combines data from several ministries and answers data requests from ministries and other state authorities.

3.6 The value of text data

In the last workshop presentation, Radovan Garabík raised awareness of the value of textual data and language technology tools that are already available at the JULS institute, among others. The presentation included examples of how word embedding techniques can be used to study cultural or other stereotypes.

Dr Garabík primarily focuses on textual data, that is, text corpora, and started by explaining differences between corpora of various sizes and what these different sizes enable linguists do to. While a corpus of 1 million tokens is enough for syntactic annotation, 10 million tokens are required to create a terminology, 100 million for a short dictionary, 1 billion for a big dictionary (as the ongoing publishing of the Slovak Language Dictionary), as well as for word embeddings and language models. The presenter then illustrated the immense size of such text collections by pointing out that an average person only is able to read 400 million words during a lifetime. A critical task in text processing is to divide texts into textual units, or elements, which is called tokenization. Another important issue is the accessibility of corpora, which sometimes requires registration or a fee. Dr. Garabík additionally mentioned a number of smaller, but specialized and innovative corpora and their inherent value, such as the legal corpus MARCELL with 43 million tokens of Slovak legal regulation from the years 1955 – 2020, which has been provided to CEF.AT for training the EC's machine translation engines. Corpora can also have an educational value, as is the case of web corpora ARANEA, used for teaching purposes at the Faculty of Pedagogy, Comenius University in Bratislava.

Dr. Garabík then presented a list of other specialized or general corpora and proceeded to his second theme of the talk, on how to use mathematical entities and methods to process the meaning of words. The basic idea of word embeddings is that a word is represented by a vector in n-dimensional space. Then, a one-layered neural network is used for extracting a vector, while the distance between vectors corresponds to word meaning similarities. Dr Garabík then showcased the possibilities for analysis and visualization of word embedding relations at the example of the Slovak NLP tool "Semä"[4]. Several examples have been provided, e.g. words semantically similar to "beer" in both Slovak and Czech. It appeared that the group of the Czech words related to beer are more poetic, which might serve as a first hint of how the Slovak and Czech cultures and ways of thinking differ.

The moderator then concluded that it is surely interesting to see how the data affirm some of the stereotypes in society and even reveal some new ones, and that language use can even reveal how societies in Europe perceive each other. He then thanked the participants for their attention, asked for their feedback by filling out the country survey and feedback forms, expressed hope that the series of ELRC workshops would continue and that a living network of stakeholders within LT in Slovakia

[4] https://www.juls.savba.sk/sem%C3%A4.html

would gradually emerge. Last but not least, he thanked the interpreters, who, although “invisible”, are often an indispensable precondition for communication in multilingual settings.

4 Synthesis of Workshop Discussions

One of the main issues during the workshop was the need for Slovakia to participate in international research infrastructures, such as CLARIN or DARIAH. With respect to this need, the signals from political milieu are at last positive: after several years of attempts the Road Map for Research Infrastructures (SK VI Roadmap 2020 – 2030) was published in March 2021.

The Slovak language, as stressed multiple times during the event, is a lesser spoken language, which in turn makes it inadequately supported by language technology. The nature of Slovak, or Slavic language(s) in general, can nevertheless be considered as an advantage in terms of the perspectives for future research into (systematic and comparative) linguistics and for the development of language technologies.

One recurrent theme in this context has been the comparison of Slovak achievements in AI and LT with the situation in the Czech Republic. It was repeated that the Czech situation is characterized by more networking and synergies between the stakeholders. It was unanimously agreed that Slovakia definitely needs a network of AI and LT stakeholders. Several participants expressed their appreciation of events like the ERLC workshop which is considered an effective means to gradually achieve a so-called snowball effect to this end. What is also important and very promising is the genuine interest showed by young Slovaks for studying, using and, eventually, developing AI and LT solutions.

The first signals that the event can be considered successful are as follows:

1. The prompt response of representatives from CVTI (the Slovak Center of Scientific and Technical Information[5]), who initiated a meeting with the JULS institute, during which they expressed interest in collecting more data for the purposes of improving their plagiarism detection system ANTIPLAG;
2. Libor Bešenyi, one of the panelists representing the company Xolution informed the T-NAP that he, together with colleagues, plans to write a more technologically oriented paper on the need for collecting big data, addressed to the linguistic community;
3. The director of KINIT Institute and opening speaker at the workshop, Prof. Bieliková, later asked for update with regard to Slovak attempts for membership in CLARIN and other infrastructures, which she was given by the Slovak T-NAP. The absence of Slovakia in these infrastructures has negatively impacted internalization of Slovak research in social sciences and humanities. Because of that, the approval of the Road Map for Research Infrastructures mentioned above is an important step towards better integration;
4. Another important point is the state's commitment to invest more in digitalization, as evident from the presentations of the representatives of the Ministry of Investments, Regional Development and Informatisaton and from the 2030 Digital Transformation Strategy for Slovakia[6].

[5] https://www.cvtisr.sk/en.html?page_id=58

[6] https://www.mirri.gov.sk/wp-content/uploads/2019/10/SDT-English-Version-FINAL.pdf

5 Country Profile: Language data creation, management and sharing

No substantial or abrupt changes in Slovakia's country profile, compared to the latest version of the ELRC Country Profile for Slovakia (2019), can be reported. Some important developments, however are the following:

- the expressed intention of the Slovak research community to request membership in the CLARIN infrastructure;
- the approval of the Road Map (SK VI Roadmap 2020 – 2030, available in Slovak by the Slovak Ministry of Education, Science, Research and Sport[7]);
- the publication (July 2019) of the Action plan for the digital transformation of Slovakia for 2019–2022[8]. This action plan contains concrete steps to build a sustainable, human-centric, and trustworthy AI ecosystem within the long-term Strategy of the digital transformation of Slovakia 2030[9]. One of the proposed projects of the Action plan is the development of a tool for natural language processing to accelerate the development of AI in the private sector and improve the quality of public services. In detail, the Action plan specifies the following: "...It will be necessary to remove barriers in the use and development of text and voice corpus of the Slovak language with specific regard to safe and practical application of such technologies in the field of public services. It will be possible to use the methods of natural language processing for monitoring of priority holistic goal, i.e. increasing transparency of the Slovak regulatory framework. Subsequently, it will be possible to use features of semantic text and voice analysis for automation and electronization of subset of services in the contact with authorities, medical facilities and schools, which will make opportunities for developing innovative packages of services and products also for commercial sector, e.g. in IT sector, in the field of data transfer security, in automobile industry as well as in other fields of the commercial sector." Additionally, the Action plan foresees the preparation of a new Act on Data to better define regulations on data protection, disclosure principles, data access and open data regulations. The proposed measure "...will result from precisely defined categorisation and classification of data based on their information value and required level of protection. It means that there will be a precise definition of rules and processes for reference data, open data and the manner how it will be possible to analytically process data (including rules for anonymising and pseudonymising data)".

[7] https://www.minedu.sk/data/files/10600_cestovna-mapa-vyskumnych-infrastruktur-sk-vi-roadmap-2020-2030.pdf
[8] https://www.mirri.gov.sk/wp-content/uploads/2019/10/AP-DT-English-Version-FINAL.pdf
[9] https://www.mirri.gov.sk/wp-content/uploads/2019/10/SDT-English-Version-FINAL.pdf