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 D3.2.19

Task 3

ELRC Workshop Report for Hungary

Authors: Kinga Jelencsik-Mátyus, Noémi Vadász

Dissemination Level: Public

Version No.: v1

Date: 2022-02-07

Contents

1 Executive Summary

The 3rd ELRC Workshop in Hungary took place on 7th February 2022. Due to the pandemic circumstances we decided to organise the event via Zoom. Although we were sorry not to meet the presenters and the audience personally, the Zoom platform definitely made the event far more accessible, and it could be seen in the high number of attendees. The main topics of the event were language-centric AI, language data and automated translation. Beside the keynote presentations on these fields we also had three panel discussions with distinguished representatives of the Hungarian LT industry, the local public sector, and important national infrastructures relevant to LT and language data. The whole event as well as the panel discussions were moderated by the ELRC National Anchor Points: Tamás Váradi, and Zoltán Bódi.

After a short welcome and introduction given by Tamás Váradi, the workshop started with the keynote presentation of Philippe Gelin, head of sector Multilingualism at DG CNECT, about the new Digital Europe Programme and the Language Data Space. After that, Gábor Prószéky, director general of the Hungarian Research Centre for Linguistics talked about the potential of Language Technology and Artificial Intelligence. Right after this presentation the representatives of LT providers gathered to discuss the actual questions of LT in Hungary. After a short coffee break Ágnes Farkas introduced the CEF AT platform, then we turned to LT in the public sector. In the session after the lunch break we concentrated on language data, which is at the core of today's NLP. First, Ádám Feldmann from the University of Pécs talked about the value of data for the development of top quality LT, then, after a short presentation of the changes in the legislation pertaining to linguistic data by Gergely Csősz from the Hungarian Intellectual Property Office, the representatives of national infrastructures discussed their views about language data. The event ended with a demo session of five outstanding projects. Finally, Tamás Váradi concluded the workshop.

Probably the most important message the event conveyed about AI in Hungary is that the amount and quality of language data is crucial for the development of LT, especially in the era of language models. The LT community should take prompt action to collect language data. It was also emphasised that local LT suppliers should work together with international big players to develop effective digital LT solutions for the Hungarian language.

The keynote presentations are available at the event website https://lr-coordination.eu/hungary3rd, and the video of the whole workshop is also available upon request.

2 Workshop Agenda

Time	Session
09:00 – 09:15	**Welcome and introduction** *Tamás Váradi, NYTK*
09:15 – 09:40	**The new Digital Europe Programme and the Language Data Space** *Philippe Gelin, DG-CONNECT*
09:40 – 10:05	**The potential of Language Technology and AI – where we are, where we should be heading** *Gábor Prószéky, NYTK*
10:05 – 11:05	**Language Technologies in Hungary - Panel session** *Invited speakers:* *Gábor Bessenyei - MorphoLogic Lokalizáció Ltd.* *János Horváth-Varga - T-Systems* *György Körmendi - Clementine* *Péter Szekeres - Neticle* *Pál Vadász - Montana* *Gábor Varga - Microsoft Hungary* *Moderator: Tamás Váradi*
11:05 – 11:25	***Coffee Break***
11:25 – 11:50	**The CEF AT Platform** *Ágnes Farkas, DG-Translation*
11:50– 12:35	**Language technologies by/for the public sector - Panel session** *Invited speakers:* *László Boa - Artificial Intelligence Coalition* *László Jobbágy - Digitális Jólét NP Kft.* *István Szviridov - Idomsoft* *Ádám Tarcsi - NATUK (ITM-DJP)* *Moderator: Zoltán Bódi*
12:35 – 13:35	***Lunch Break***
13:35 – 13:50	**The value of data for the development of top quality LT** *Ádám Feldmann, University of Pécs*
13:50 – 14:50	**Language data creation, management and sharing: existing practices and challenges - Panel session** *Invited speakers:* *Gergely Csősz - Council of Copyright Experts* *Richárd Farkas - Artificial Intelligence National Laboratory* *Ádám Feldmann - University of Pécs* *Lotár Csaba Schin - OTP Bank* *Miklós Sebők - Centre for Social Sciences* *Moderators: Tamás Váradi and Zoltán Bódi*

14:50– 15:50	**Language technologies demonstrations and networking** *ELE project - and the roadmap of Hungarian NLP resources* *Clementine* *Montana* *MorphoLogic Lokalizáció Ltd.* *Neticle*
15:50 – 16:00	**Conclusions** *Tamás Váradi, NYTK*

3 Summary of Content of Sessions

3.1 Welcome and introduction

In the opening session of the event the host, Tamás Váradi after welcoming the participants introduced the Connecting Europe Facility (CEF) program and the European Language Resource Coordination (ELRC). The Connecting Europe Facility is a large-scale infrastructure-financing programme focusing on transportation, energy and telecommunication. The CEF Telecom branch, besides developing 5G networks, is aiming at demolishing language barriers in multilingual Europe in order to build a single digital market. The programme supports the vision that the citizens of Europe can access the virtual space without language barriers. Cultural, touristic, health-related or legal services should be available to everyone, regardless of the language. To achieve all this, the European Commission created the CEF automated translation system. Automated translation systems today need a lot of language data to be trained on, and this is where the European Language Resource Coordination (ELRC) comes in. ELRC coordinates the collection and dissemination of language resources (all kinds of language datasets and LT tools) in the EU countries. However, nowadays ELRC pays special attention to multilingual NLP, thus parallel, and also monolingual corpora so as to provide sufficient training material for language models, especially multilingual language models. There has been a huge paradigm shift in NLP in recent years. Neural networks made it possible to create large language models, and multilingual language models. This latter enables transfer learning, providing a solution for the challenges of multilingualism, which is one of the main fields of language-centric Artificial Intelligence (AI). Therefore the motto of the 3rd ELRC workshop is: Reshaping multilingual Europe: Language-centric AI.

3.2 The new Digital Europe Programme and the Language Data Space

The first keynote presenter of the 3rd ELRC workshop in Hungary was Philippe Gelin, Head of Sector Multilingualism at DG CNECT. Philippe Gelin introduced the new Digital Europe Programme and the Language Data Space. To start from a wider context, a definition of LT was given: it means all the LT tools dealing with language, from automatic translation, through anonymisation, up to speech transcription, to name just a few. There has been a large progress in the field of LT due to the emergence of language models, which also accelerated the development of several applications that help to overcome language barriers. With the rapid growth regarding both the technical background and the language models solving the challenges of multilingualism, discovering new cultures is already possible for everyone merely by the touch of a mobile phone.

The motto of the EU is “united in diversity”, referring to the endeavours to equally support the 24 official languages of the EU (and the over 60 non-official languages) as sources of diversity, and to foster the development of language technology, most importantly translation applications and

everything that is needed in the background to build these applications. A strong motive for doing so is the predominance of the English language on the Internet, which may lead to the digital extinction of the less-resourced languages (those that simply do not have language data/datasets large enough to support the building and training of large language models). Digital presence, backed with massive language technology is also an essential component of the economic competitiveness for each language.

Therefore the new Digital Europe Programme and the Language Data Space provides solutions to handle not only all European languages but also other large languages on the international market like Chinese or Japanese. The European Commission offers support in 3 areas: legislative help, coordination across member states and a funding scheme. The presenter emphasised that there is a plethora of European Programmes. Good examples are the CEF2, the InvestEU, the Cohesion Fund and the Creative Europe Fund, the well-known Horizon Europe Programme. And last but not least a newcomer: the Digital Europe Programme.

LT is already the focal point of a number of ongoing initiatives, such as the European Language Equality project which is supported by the European Parliament. This project has 52 partners including all EU countries, research centres as well as representatives from the industry, along with pan-European infrastructures. The aim is to develop a strategic research, innovation and implementation agenda and a roadmap to achieve digital language equality in Europe by 2030. The Horizon program concentrates on the research side, targeting endangered languages in Europe in the cultural heritage strand, and some IT driven initiatives dealing with language-centric AI, for example. The new programme, Digital Europe (DIGITAL), supports the deployment of technologies mentioned above and will complement the funding provided by the programmes already mentioned. DIGITAL affects the common service platform and the industry, but most importantly SMEs, arranging a wide use of digital innovations in several fields, through initiatives like the European Digital Innovation Hubs (EDIH), the local shops where companies "test before investing" in new digital applications. The Testing & Experimentation Facilities (TEF), on the other hand, will be large scale reference sites (specialised in given domains) available to all European LT providers to test and experiment SOTA AI solutions in real-world environments. The AI on Demand Platform collects all cutting-edge AI related initiatives from applications to research. The platform contains the European Language Grid (ELG) where most of the European multilingual tools (among all kinds of LT tools and datasets) can be found.

The European Commission presented its data strategy in 2020[1] with special emphasis on language data, to make it available for commercial purposes, especially for SMEs, for the benefit of European customers. 20 sectoral areas were defined where data-connected actions will be taken - all of which has its corresponding Data Space, language is one of them. The Language Data Space will help the exchange of language data between the various actors, with the aim of building efficient language tools and services - this is becoming ever more important since the rapid growth of AI-related solutions in LT.

Philippe Gelin ended his presentation with calling the attention of the audience to three websites of practical information:

- the Connecting Europe Language Tools[2]
- the European Language Grid[3]
- CEF Automated Translation Catalogue of Services[4]

[1] https://digital-strategy.ec.europa.eu/en/policies/strategy-data
[2] https://language-tools.ec.europa.eu/
[3] https://www.european-language-grid.eu/
[4] https://cef-at-service-catalogue.eu/

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

Gábor Prószéky, Director General of the Hungarian Research Centre for Linguistics started his presentation by explaining the relation between AI, machine learning and deep learning. Historically, when AI was invented in the 1950-60s its primary focus was to truly model human intelligence. In the 1980s machine learning emerged, launching very complex algorithms that, in a specific area, are capable of learning through experience, provided with a sufficient (usually huge) amount and quality of data - making even unsupervised learning possible. In the core of machine learning a new approach emerged: deep learning (mostly called AI by the media), algorithms that develop themselves to successfully solve different tasks. The special features of deep learning are neural networks, and the need for unimaginably huge sets of data.

In our everyday lives AI is ubiquitous today, from digital communication, through agriculture to intelligent cars. Concerning linguistics, there are numerous applications already achieving or even exceeding human performance in very well specified tasks. Although in linguistics non-compositional meanings, metaphors, and the role of context impose complex challenges on researchers, deep learning is present at several fields of LT: automatic speech recognition, text-to-speech, chatbots, sentiment analysis, named-entity recognition, and automated translation, to name just a few.

Deep learning in LT began about 10 years ago. Back then, first, there were some efficient solutions for example for spam filtering, or named entity recognition. Second, there were some promising research directions concerning opinion mining or information extraction, while, third, question answering or dialogue systems seemed rather distant problems to solve. In 2022 we have very good solutions for the first two categories, and really promising experiments for the third. The biggest change in the past 10 years was in the role of data: AI solutions need an incredible amount of data for training language models. Researchers need to collect or generate data for their research and development project, and should also assure the safe storage, management, sharing and reuse of these data. The presenter drew the attention of the audience to the fact that although data is the new fuel for AI, particular attention should be paid to the features of data. For example, to develop a language model of standard written Hungarian, data cannot be collected from Internet forums, only from curated texts.

Gábor Prószéky also highlighted the contribution of the Hungarian Research Centre for Linguistics to the utilisation of language-centric AI, the most prominent being the HilBERT, the BERT-large model for Hungarian, and numerous other experimental language models also developed in the HILANCO project.[5] Some demo versions of these models are already available,[6] and the best is yet to come!

3.4 Language Technologies in Hungary / for Hungarian (Panel session)

The first panel session of the event introduced the LT supply side at the national level. The participants of the session were six representatives of distinguished Hungarian LT providers. The session started with a short introduction on the status of LT in Hungary, given by the moderator, Tamás Váradi. Ten years ago, the White Paper entitled "The Hungarian Language in the Digital Age"[7] depicted the level of LT support for European languages. According to that survey there were three languages (Spanish, English and French) that met at least the moderate level of LT support, a vast majority of the other languages appeared in the language technology danger zone due to lack of LT resources. Nowadays, LT is ubiquitous in everyday life thanks to smart devices. At least - as the moderator emphasises - for

[5] https://hilanco.github.io/
[6] https://juniper.nytud.hu/demo/nlp
[7] http://www.meta-net.eu/whitepapers/e-book/hungarian.pdf

English, but not yet for Hungarian. The session discussed the situation of LT in Hungary concerning the language models paradigm.

The introduction of the panellists started with Gábor Bessenyei, CEO of MorphoLogic Lokalizáció Ltd. Morphologic offers a machine translation solution, called Globalese, which can be trained with the customers' own data. The next speaker was János Horváth-Varga, deputy branch director at T-Systems Hungary Ltd. who was the head of the group developing Vanda, the digital customer service assistant of Hungarian Telecom. He talked about his hopes that Vanda, as a first complex solution, will be followed by a lot of other digital assistants. Péter Szekeres, CEO and co-founder of Neticle, introduced his company as the representative of start-ups and scale-ups at this panel. As a company dealing with NLP and text analysis for 10 years, they have their own NLP engine focusing on sentiment analysis, topic recognition, entity analysis etc. for 30 European languages. Therefore they have experience on how to make NLP solutions fit to be solved and this is the aspect they could represent in the session. György Körmendi, managing director at Clementine Consulting, reported on providing NLP solutions predominantly for bigger companies. Their new product, Hanga, presented in detail in the demo session, will be a new solution for voice assistants for Hungarian. Then, Pál Vadász, managing director at Montana, introduced his company that has spent more than 30 year in LT, specialised in legaltech. They have provided LT solutions, for example a search facility for the national court network that enables the 3000 judges to find relevant regulations and precedent cases. Finally, Gábor Varga, national technology officer at Microsoft Hungary, emphasised the role and accomplishments of Microsoft in AI, and talked about how important it is to present these results to the audience and partners. So the role of their company in the present session is to highlight the areas a global player of LT can offer for the local sector.

The focus of the panel session was whether Hungarian language is ready for the AI era, whether there is sufficient LT industry and research in Hungary, or do we depend on international stakeholders? Gábor Varga started the discussion by drawing the attention of the audience to the layered nature of LT with the long-long way from research to the end user. A possible solution is to build a structure where the different layers represent different complexity, and each layer should be "operated" by those who have the most efficient application/knowledge/etc. for that given layer. There are LT providers on the market who offer services operated by them from the lowest levels to the end users, but there are also other examples, for instance Microsoft also offers an AI, deep neural platform, on which other participants of the LT ecosystem can build language functions with a great value for the end users. To sum up, according to Gábor Varga, the roles of international and local players should be reconsidered. Péter Szekeres added that AI is a rocketing field. There are some outstanding sectors where AI and NLP can offer excellent solutions, like bankruptcy prediction, or marketing segmentation, and it is a very good trend. However, there are only a few companies who have sufficient research or budget for the development of such applications, and although Hungarian is a good starting point, regional coverage (i.e. covering several languages in the regions) should be heavily considered. János Horváth-Varga continued by introducing the architecture of Vanda. It was built on Nuance company's leading modules in speech recognition and a lot more, and it was extended by the T-Systems. He also highlighted that based on their experience of promoting Vanda in neighbouring countries, it is extremely important to have professionals who are native speakers of the target language if a company aims at expanding.

Pál Vadász, as a representative of SMEs emphasised that gaining a sufficient amount and quality of data is a big problem for companies similar to Montana. Although there are principles about providing data for companies who process data, it is still very hard. Gábor Bessenyei shared these thoughts: technology is ready but data is scarce. György Körmendi depicted the picture of complex services, like

voice assistants, as building blocks, and the problem is that expertise is needed to build these blocks together.

At the end of this panel session the participants had the opportunity to share their experience with language technologies through a Zoom poll. 75 participants filled the poll which popped up after the panel session. The first question sought to answer which language technologies are used the most. Since a respondent could select more responses, we got a fairly detailed picture. The most used technology is machine translation applications, followed by grammar and spell checkers and information search and retrieval applications. Speech recognition technologies and virtual assistants, chatbots were in the fourth place, while the least used technology was text-to-speech synthesis which received one-tenth of the votes. It also turned out that these technologies are used equally in English and in Hungarian, a quarter of respondents use language technologies in these languages. As far as satisfaction of these language technologies for Hungarian is concerned, it can be seen that one third of the respondents are very satisfied with the quality and reliability of these technologies, however, there is still room for improvement.

Which of the following language technologies do you use more? (up to 3 choices)		
Automated translation	61	30.8%
Speech recognition (e.g. you dictate messages on your mobile phone)	10	5.2%
Text to Speech synthesis (e.g. you mobile phone assistant reads out your messages)	5	2.6%
Virtual assistants/chatbots (e.g. on your mobile phones or call centers)	10	5.2%
Information search and retrieval (e.g. web search)	50	26.0%
Grammar and spell checker	56	29.2%
When using any of the technologies above, which language do you usually use them with?		
Hungarian	68	42.5%
English	66	41.3%
Other	26	16.3%
As a simple user, how satisfied are you with the quality and reliability of such technologies for Hungarian?		
Completely satisfied	1	1.3%
Very satisfied	25	33.3%
Neutral	17	22.7%
Somewhat satisfied	29	38.7%
Not at all satisfied	2	2.7%
I don’t know/No answer	1	1.3%

3.5 The CEF AT platform

Ágnes Farkas from the DG-Translation delivered a presentation about the Connecting Europe Facility (CEF) Automated Translation (AT) platform. She started with a historic overview: the machine translation system used before the current one was MT@EC, a statistical machine translation (MT) system focusing on the legal domain. The present system, eTranslation, has been used by European institutions since 2017. It covers more domains, and it is a neural MT system, also available for online public services. The latest initiative here is the DIGITAL (replacing CEF), with the aim of making digital technology available for businesses, citizens and public administration.

eTranslation addresses two types of audiences: at first it was developed to support translator and other employees of EU organisations, and it is integrated in the digital services and home pages of EU organisations. The other target contains everyone supported by the CEF programme: Pan-European digital public services, public administration in EU member states (plus Iceland and Norway), universities, free-lancer translators working for the EU, and SMEs. The use-cases of eTranslation mostly include their web interface for human users, and a machine-to-machine service (API) which can be integrated in web pages (this is currently available for public services only), for example. eTranslation is a free service where registration is necessary, with a high level of security, as it is protected by the Commission's firewalls. All data is deleted after 24 hours (or can be deleted right after using the service). The domains covered are: EU formal language, general text, Court of Justice Case Law, finance (ECB), public health etc., available for all official EU languages, plus six other big languages. All the uploaded documents/texts can be translated to any, or even all of these languages in one go, and the result can be requested in email or at the storage of the account.

The most important factor, however, is quality. eTranslation is best for EU texts and documents, obviously, as the EU general (legal) translation was trained with the databases of the EU, containing all EU translations so far, with all the parallel corpora. For smaller languages, like Hungarian, unfortunately, there is less data. It is less efficient in case of new words, creative texts, or words/expressions without context. As for additional services and the future, the web page offers other NLP services (like speech recognition, or multilingual tweet), and anonymisation and language recognition are also parts of the plan.

At the end of the presentation the participants were asked to express their views on machine translation via a live Zoom poll. All the 58 respondents use machine translation applications. The vast majority use machine translation systems outside eTranslation (CEF AT), including paid or free service like Google translate or Bing. One-tenth of the respondents use eTranslation and other services, and only less than 8% use solely eTranslation. However, none of the respondents are completely satisfied with the quality of automated translation to or from Hungarian, a third of them are very satisfied. The fact that a quarter of them feels neutral and approximately 40% are only somewhat satisfied shows that there is still ground to achieve higher user satisfaction.

Have you ever used an Automated Translation system?		
Yes	58	100.0%
No	0	0.0%
If yes, which one?		
eTranslation (CEF AT)	5	7.9%
Another free or proprietary system (e.g. Google translate, Bing etc.)	52	82.5%

All of the above	6	9.5%
I don't know/No answer	0	0.0%
As a simple user, how satisfied are you with the quality of automated translation to/from Hungarian?		
Completely satisfied	0	0.0%
Very satisfied	19	32.8%
Neutral	14	24.1%
Somewhat satisfied	24	41.4%
Not at all satisfied	1	1.7%
I don't know/No answer	0	0.0%

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

Zoltán Bódi, the moderator of this panel session, started with an introduction of application areas of NLP (and possibly language-centric AI) in eGovernment, like case routing, automatic message answering, phone call summation or opinion mining with the help of sentiment analysis. All of these examples process a huge amount of digital data, and as NLP research is data-driven, eGovernment should also be data-driven. After the short lead-in, the participants of the session introduced themselves, and shared their thoughts about the topic. László Boa, professional lead at the Hungarian Artificial Intelligence Coalition presented his organisation: the Coalition, with its more than 360 members, works on supporting the LT activities in the government and other sectors in using AI. They contributed to the compilation of the AI strategy of Hungary, with two prioritised areas: the digitalisation of customer service and the Hungarian language. Then, László Jobbágy, managing director at Digitális Jólét Nonprofit Ltd., focused on development of e-administration, working together with the Ministry of Interior, and creation of the digitalisation strategy of the public collections through the Digital Education Strategy. István Szviridov, head of System Integration Division at the IdomSoft Ltd., introduced the company he represented as dealing with law enforcement and administration in the Ministry of Interior and connected sectors. His task is to integrate AI in these services to ease the administrative procedures of the citizens. They have several solutions for the processing of texts, pictures, and biometric data. As it entered into force in December 2021, there are 3 AI services provided by the company: the Mia chat robot, a speech-to-text and a text-to-speech service. Finally, Ádám Tarcsi, division leader at the Nemzeti Adatgazdasági Tudásközpont (NATUK), said that the responsibilities of the organisation he represented were the execution of the AI strategy actions, especially those connected to data-driven economy. It is a really new organisation, launched in September 2021, and their aim is to foster data-intensive operation of SMEs. He emphasised that prompt actions need to be taken for the Hungarian language to be able to keep pace with the digital era.

Then the moderator went on with a question about whether the development and prompting of NLP should be driven by the market or should it be connected more to public administration. The participants of the panel agreed that big international LT players already have excellent solutions for several tasks. The question is whether Hungarian NLP research and development is ready and able to take part in the rapid progress of the field or whether international bigtech companies will provide solutions for Hungarian. For the Hungarian LT companies to be successful, a cooperation of research, government and market sectors is necessary- as some examples of it are already visible in the media. István Szviridov added that the digitisation of public administration, and the development of LT for

the public administration should be done by Hungarian companies mostly due to legal reasons, concerning law protecting personal data, and emphasised that research results and mature technological solutions are already available. Even national legislation should be adapted to digital public services, to enable the use of the new portals. And the data collected through digital public services would also be very useful for training AI applications - an aspect that is still not resolved in legislation.

After the panel session, the participants were asked for their opinion as citizens on the digital readiness of the Hungarian public services. 58 participants took part in this poll. Only one participant was completely satisfied, only 9% of them were very satisfied, 28% felt neutral about it, more than 50% were only somewhat satisfied, one-tenth were not at all satisfied. These results show that the digitization of the Hungarian public administration still has room for improvement. The second question explored the opinions about the role of government in supporting the development of language-centric AI. Respondents could choose more than one option, therefore we got a very detailed picture of their opinion on the topic. The results show that, according to voters, the government should be involved in the development of language-centric AI in the following ways (in descending order of preference): financier or direct investor, regulator, user and service provider, convener and standards-setter, data steward and finally, smart buyer and co-developer of technology.

As a citizen, how satisfied are you with the digital readiness of the Hungarian public services?		
Completely satisfied	1	1.7%
Very satisfied	5	8.6%
Neutral	16	27.6%
Somewhat satisfied	30	51.7%
Not at all satisfied	6	10.3%
I don’t know/No answer	0	0.0%
Which of the following do you believe is the Administration’s most important role in supporting the development of Language-centric Artificial Intelligence? (up to 2 choices)		
financier or direct investor	32	24.6%
regulator	30	23.1%
convener and standards-setter	18	13.8%
data steward	15	11.5%
smart buyer and co-developer of technology	12	9.2%
user and service provider	23	17.7%

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

Ádám Feldmann, Head of Applied Data Science and Artificial Intelligence Group at the University of Pécs, started his presentation underlining that the development of LT triggers the progress of the whole area of AI, especially since the emergence of transformer-based models. It seems that the language models are the engine of LT. These SOTA models need a lot of data, several times larger than before, and once we manage to gather this incredibly huge amount, it enables the rapid growth of the

language models. With growth in the number of parameters of a model, new, unexpected features arise, for example much less data is sufficient for document classification tasks, named entity recognition or sentiment analysis. This way few-shot (only some training data) or even zero-shot (no training data at all) learning will be possible. The promise of these models is that domain specific solutions need only a fragment of this huge data. So, if AI helps to reduce the number of steps in an LT task, and makes the whole process a lot faster, while providing the same quality (compared to previous models), it means it is worth collecting the amount of data needed. Ádám Feldmann presented a calculation showing that model size, the training dataset size and the amount of compute resources used for training must be scaled up in tandem for optimal performance. Close attention should be paid to the quality of data as well.

In his summary the presenter emphasised that we should start collecting the data to be able to build large language models. Hungarian data in itself may not be sufficient in size, but transfer learning enables us to use the solutions that are developed for the English language in another (smaller) language without the need of training data for the smaller language. It is very similar to human learning: we can use in Hungarian what we learn in English. The next step may be the collection of multimodal data, where also prompt actions need to be taken for the Hungarian language to keep pace with the rapidly growing AI era.

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

The most expected panel of the event was about the language data. The topic was present all along the workshop, but there was a whole panel session dedicated to it. The reason why data is such a hot topic in Hungary is the novelty, and at the same time uncertainty of the legal background. Therefore the session started with a short presentation of Gergely Csősz, an expert of copyright protection from the Hungarian Intellectual Property Office. He introduced the regulations of text and data mining (TDM) from the perspective of copyright. Any original, genuine work from the fields of art or science falls under the regulations of copyright - anything from a master thesis to a Facebook comment. There are two groups of the relevant rights: individual rights (e.g. indicating the name of the author) and the rights presenting assets (regulating the industrial application of protected data). Free application of data is also possible in well-specified cases, like quotation, research, education, private use, institutional use (typically internal use within the author institution, most dominantly in the field of cultural heritage). If free application is not possible, data users need to apply for a permission.

There was a change in the relevant regulation as of 1 June 2021 (SZJT 35/A).[8] Before that, a written official permission form was needed, but now an online declaration is sufficient. The Hungarian legislation is slightly more detailed than the EU legislation: it has one additional point saying that apart from collecting and copying data as stated in the previous lines, it is also possible to use the source in the framework of a research cooperation, or for the peer review of an article.

Then a lively discussion started about the new legislation. Several questions were asked about the details, like 'legal access' of sources. It means that if no restrictions are implemented on the data (in the form of licences, for example), it can be used freely. Lotár Schin from OTP Bank emphasised the growing ratio of legislation-related tasks in all kinds of projects - especially the results of big language models. It is still not clear how the output of language models is regulated. As Gergely Csősz explained, the whole field is so new, that more experience is needed to be able to build an effective legislative framework.

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

3.9 Demos session

The demo session was also very welcome both by the presenters and by the audience. Apart from one, all companies accepted our invitation to give a short overview of one of their outstanding projects. The session started with the introduction of the European Language Equality (ELE) project (as already mentioned by the first keynote presenter, Philippe Gelin), and the Hungarian NLP resource roadmap. ELE is a large-pilot project with 52 partners, representatives of all EU member countries, with LT research centres and industrial stakeholders, and also distinguished pan-European organisations like CLARIN or EFNIL. The aim of the project is to develop a strategic agenda and roadmap for achieving language equality in Europe by 2030. All the partners actively participated in the project in different tasks.

The Research Centre for Linguistics was the Hungarian coordinator of the large data collection process cataloguing all possible resources built for Hungarian. Therefore in the framework of the ELE project all corpora, lexical datasets, tools, grammars and language models were collected that have a landing page, i.e. reliable information about where the given resource is available. As a result, now we have a NLP resource roadmap for Hungarian, as for all other EU official languages investigated by the ELE project. The Hungarian collection included in January 2022 353 datasets and 179 tools, language models and LT services. It is important to emphasise that it is but a snapshot of a rapidly growing field. The results of the metadata collection process is available at the Catalogue of the European Language Grid.[9] The data collection not only revealed the fields with numerous resources, like multilingual corpora, text analysing tools and toolchains, but also highlighted the areas where the lack of datasets or tools hinder development. The emergence of language models has reshaped how language data is used in most subfields of NLP. In several areas state-of-the-art results can only be achieved through an incredibly large amount of data compared to previous methodologies. This implies that researchers not only need to develop technological solutions, but also to find and create their own datasets.

The second presentation was held by György Körmendi, the managing director of Clementine. He introduced Hanga, an Interactive Voice Response (IVR) system, the new customer support solution of the company. Their aim is to increase satisfaction with chatbots from both the customers' and the companies' sides. The next presentation was delivered by Pál Vadász, managing director at Montana, about the AI4Lawyers project. He talked about the NLP related technologies used in legaltech, mostly used in English-speaking countries. He ended his presentations emphasising that the EU has a large role in supporting the following field: small languages, technology transfer, and relevant local education and technology centres. After that, Gábor Bessenyei, CEO of MorphoLogic Lokalizáció Ltd., introduced the technical background of Globalese. This neural machine translation framework provides SOTA solutions for translation, using the customer's own data. With this approach Globalese achieved better results than large, global translation applications. The last demo was the Neticle text analysis API, introduced by Péter Szekeres, co-founder and CEO of Neticle. This solution is available for 30 languages. Within this API the demo focused on the Zurvey platform, providing a detailed transparent analysis all along the application, like sentiment analysis, topic analysis, entity recognition, etc.

3.10 Take-home message and conclusions

Even in the last session of the workshop there were around 80 participants, which is in itself a sign of success. The workshop was an ideal platform for the stakeholders to gather: public administration, LT providers, research and development and the translators and interpreters were all involved. Tamás

[9] https://live.european-language-grid.eu/catalogue/?&language__term=Hungarian

Váradi ended the event with the last poll, which was a feedback form of the workshop. There were good results in all categories, showing that the participants were satisfied.

4 Synthesis of Workshop Discussions

The workshop focused on three main topics: 1) language-centric AI 2) automated translation and 3) language data. All three topics were discussed in detail during the event. Both language-centric AI and language data appeared in almost all of the presentations and discussions throughout the event.

The presentation about LT and AI, with special focus on Hungary, showed that although we can say that AI is ubiquitous in our everyday life, it is not necessarily true for the Hungarian language. In 2022 we have excellent solutions for numerous tasks like named-entity-recognition, opinion mining or information extraction, but there are still areas, like complex chatbots, where there is plenty of space for improvement. Several presenters and also participants from the audience emphasised that data is a crucial factor in the development of AI. Large-scale cooperation should be set up in order to facilitate collection of data in appropriate size for building large language models. At the same time, due to the number of speakers of Hungarian, the size of data is limited, and therefore transfer learning methods should be paid special attention to. This means that the method developed and working for English could be used for smaller languages, like Hungarian, as well.

As for automated translation, eTranslation (CEF-AT) provides high quality and secure translation to EU institutions, member-states public organisations, European universities and European SMEs. eTranslation is best for EU texts and documents, and it uses neural machine translation techniques. It was also emphasised by another presenter that the combination of NMT methods with good quality data yield the best results in translation. It must be underlined, however, that even these top-quality solutions need human post-editing.

The workshop was also an excellent platform to introduce the new language data regulation, focusing especially on text and data mining (TDM). There was a change in the relevant regulation as of 1 June 2021 (SZJT 35/A).[10]

The demo session started with the introduction of the NLP resource roadmap for Hungarian, as compiled in the framework of the European Language Equality project. The roadmap revealed the fields of NLP where numerous resources can be found for Hungarian, but, at the same time, identified the gaps, where prompt actions need to be taken. Deficiency occurs mostly regarding the amount of excellent quality data for basically all fields on NLP where neural methods are used.

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

5 Country Profile: Language data creation, management and sharing

In Hungary, there have been several changes regarding language data creation, management and sharing since the publication of the Country profile in the 2019 ELRC White Paper.[11] First of all, several large language resources have been compiled, and some of them have already been contributed to ELRC-SHARE.[12] One of the most prominent ones being the MARCELL Hungarian legislative subcorpus with its more than 30 million tokens. This domain-specific corpus can be used in the development of NMT solutions.

The NLP resource roadmap for the Hungarian language was created in the framework of the European Language Equality project, cataloguing more than 500 language resources. The results of this project show the fields where excellent solutions are available for Hungarian, like multilingual corpora, or tools and toolchains for text analysis. There are also language models built specifically for Hungarian: HuBERT and HILBERT, and several experimental language models developed in the HIILANCO project. The data collection process also highlighted the gaps, where the lack of datasets or tools mean a significant obstacle for development. There is a significant gap in NLP concerning language data. Although neural methods are used in almost all subfields of NLP, there are not enough datasets in terms of size, annotation and domain. This suggests that researchers and developers need to invest a lot of time and energy in collecting sometimes an incredibly large amount of data. The presenters and the audience agreed that prompt, co-operative actions should be taken, that could even be accompanied by changes in the relevant regulation.

An important development at the policy level is the creation of the Artificial intelligence Strategy of Hungary.[13] The foundation of two important umbrella organisations was also a salient step to help AI related topics in Hungary. The Artificial Intelligence National Laboratory (MILAB) aims at strengthening the position of Hungary in AI. The research plan of MILAB is built on the National Artificial Intelligence Strategy (2020-2030). It creates the necessary cooperation by connecting the major Hungarian research centres/universities with the industry, society and government. The other initiative, the Artificial Intelligence Coalition participated in the compilation of the National Artificial Intelligence Strategy, and defines its mission (among three other points) as to "make sure that the government, as a user of AI-powered solutions, should be actively engaged in developing the local AI ecosystem by systematically utilising the national data asset pool".[14]

As for the practices of the creation of translations as multilingual data in the Hungarian public sector, they are the same as indicated in the ELRC Country Profile of 2019. Namely: the translation of foreign language source documents is still the competence of the Hungarian Office for Translation and Attestation Ltd. (OFFI). However, other language service providers can also be contacted.

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

[11] https://www.lr-coordination.eu/sites/default/files/Documents/ELRCWhitePaper.pdf
[12] https://elrc-share.eu/
[13] https://ai-hungary.com/files/e8/dd/e8dd79bd380a40c9890dd2fb01dd771b.pdf
[14] https://ai-hungary.com/en/content/ai-coalition#mission