European Language Resource Coordination (ELRC) is a service contract operating under the EU's Connecting Europe Facility SMART 2014/1074 programme.

European Language Resource Coordination

Connecting Europe Facility

ELRC Workshop Report for Hungary

Authors:	Tamás, Váradi (RILMTA) Réka, Kovács (RILMTA)
Dissemination Level:	Public
Version No.:	<V1.1>
Date:	2015-12-17

Contents

1 Executive Summary

This document reports on the ELRC Workshop in Hungary, which took place in Budapest, on the 23rd of November 2015 at the Hungarian Representation of the European Commission, popularly known as Európa Pont. It includes the agenda of the event (Section 2) and briefly informs about the content of each individual, interactive and panel workshop session (Sections 3 & 4). The event was attended by 35 participants, spanning a wide range of stakeholders, including ministries, the language industry, libraries, public organisations and language technology partners.

2 Workshop Agenda

08:00 – 09:00	Registration
09:00 – 09:05	Opening of the workshop (Dr. Péter Princzinger – Ministry of Human Capacities)
09:05 – 09:15	Welcome by the ELRC (Stelios Piperidis – ILSP/ELRC)
09:15 – 09:25	Welcome by the European Commission (Miklós Mátyássy – EU - DGT Translation)
09:25 – 09:35	Welcome by the Local Anchor Points in Hungary
	Goal of the Workshop and the Agenda (Váradi Tamás – ELRC / RILMTA)
09:35 – 09:50	Europe and Multilingualism (Miklós Mátyássy – EU - DGT Translation)
09:50 – 10:20	Language and Language Technologies in Hungary (Gábor Prószéky – MorphoLogic)
10:20 – 11:00	Panel: Multilingualism in Hungarian Public Services (Gábor Prószéky – MorphoLogic)
11:00 – 11:30	***Coffee Break and Networking***
11:30 – 12:00	How does Machine Translation work? (Márton Makrai – RILMTA)
12:00 – 12:30	How can Public Institutions benefit from the CEF.AT Platform? (Spyridon Pilos – EC, DGT)
12:30 – 13:30	***Lunch Break***
13:30 – 14:00	What Data is needed? (András Kornai – MTA-SZTAKI / RILMTA)
14:00 – 14:30	Legal framework for Contributing Data (Nóra Emese Takács – Hungarian Intellectual Property Office)
14:30 – 15:10	Panel: The data and Language Resources in Hungary: Where to look for what? (András Kornai– MTA-SZTAKI / RILMTA)
15:10 – 15:30	***Coffee Break and Networking***
15:30 – 16:00	Sharing Data and Language Resources: Technical and Practical aspects (Csaba Oravecz – RILMTA)
16:00 – 16:45	How can We Engage? (Váradi Tamás – ELRC / RILMTA)
16:45 – 17:00	Open Discussion: Wrap-up, On site Conclusions and Commitments (Váradi Tamás – ELRC / RILMTA, Stelios Piperidis – ILSP/ELRC)

3 Summary of Content of Sessions

3.1 Session 1: “Opening of the workshop and welcomes”

The workshop was addressed by Dr. Péter Princzinger, Head of Cabinet, Ministry of Human Resources, who welcomed the audience and the ELRC initiative and emphasized that the technological support of the Hungarian language is crucial for the digital presence and vitality of the Hungarian economy in the European Digital Single Market.

The ELRC was represented by Stelios Piperidis. He introduced the key persons in conceiving and organizing the event, namely the ELRC consortium and the EC/DGT representatives.

The European Commission was represented by Miklós Mátyássy, Head of the Hungarian Unit at the DGT. He discussed the challenges that multilingual Europe was facing including the, so far, predominantly monolingual nature of digital services.

3.2 Session 2: “Welcome by the Local Anchor Points in Hungary and Goal of the Workshop and the Agenda”

Tamás Váradi, local organizer of the workshop and the Hungarian coordinator of the ELRC in his welcome message greeted the audience and thanked particularly the speakers, the panelists and the potendial data providers. He emphasized that although the attendance of the LSPs and Language Technology Partners was very important, the target audience and the key persons of the workshop were the data providers. He stressed the importance of breaking down the barriers of linguistic isolation and he pointed out that this workshop was a great initiative for fruitful cooperation in the future and thanked all the participants for coming.

3.3 Session 3: “Europe and Multilingualism”

Mr. Mátyássy (EU, DG Translation) stressed that Europe is committed to multilingualism. A key pillar of the Digital Agenda is the Digital Single Market. However, multilingualism, though a highly valued asset, is also a major obstacle to building the Digital Single Market. Human translation effort is simply not adequate to the rise of the challenges that modern day digital communication presents. The Solution is automated translation. CEF Digital provides pan-European Digital Service infrastructure for eJustice, eProcurement, eHealth, Europeana and Open Data Portal. It is an enabler for multilingual e-services for citizens, businesses and public administrations. CEF AT (Automated Translation platform) is part of CEF Digital to provide automatic translation services with the goal of making digital services accessible to anyone everywhere from whatever language into the user’s language.

3.4 Session 4: “Language and Language Technologies in Hungary”

Gábor Prószéky started by presenting the language situation inside and outside Hungary. As a result of the Trianon Treaty, Hungary, once a multi-ethnic and multilingual country became practically monolingual, therefore multilingualism is mostly regional, confined to areas just across the border and a solid region in Transylvania. Unfortunately, Hungarians lag behind all other nations in Europe in terms of competence of foreign languages among the population aged 25-64. The presentation gave an overview of the language technology scene in Hungary as well as a survey of key resources and tools developed.

3.5 Session 5: “How does Machine Translation work?”

In his presentation Márton Makrai first dealt with the questions why AT is required and why it is difficult. The need for AT is overwhelming and it can be partly met by current technology as long as expectations are realistic. It is important to develop a sense of why is AT difficult, particularly for the Hungarian language, given its extremely rich morphology and, more importantly, its free constituent order. Current systems, predominantly the open-source Moses system produce acceptable results between similar language pairs and Mr. Makrai presented in detail how they work. He then went on to describe a new model, based on neural networks that promises to be more suitable to the likes of Hungarian, Finnish and Estonian. Both systems rely on machine learning and therefore depend on huge amounts of (preferably) parallel corpora. Therefore, the objectives of the present workshop apply whatever the actual implementation involved.

3.6 Session 6: "How can Public Institutions benefit from the CEF.AT Platform?"

This session was handled by Spyridon Pilos, Head of Language Applications of DGT. After a historical overview of how the language problem became more and more complex with the addition of member states, Mr. Pilos elaborated the reasons why MT was the only viable solution to tackle the multilingualism challenge in Europe. He proceeded to give an account of MT@EC and how MC@EC was now on the verge of opened up to public administration in member states and ultimately to all citizens, free of charge until 2020. He described the digital services that are already supported by MT@EC their number will increase in the future. Mr Pilos concluded by describing the drive from MT@EC to the CEF.AT platform, emphasizing that the focus will be the coverage of more domains (not just typical EU texts) and increasingly in-domain texts are needed to implement this move at sufficiently high quality. Mr. Pilos kindly agreed to take questions including the one asking about steps taken to enhance the quality of Hungarian MT, which was generally perceived as one of the underperforming language pairs.

3.7 Session 7: "What Data is needed?"

András Kornai started the presentation by introducing parallel corpora and explaining briefly why they are important and how they are prepared. At the same time he also emphasized that it was not only corpora but lexical databases, terminologies and ontologies that are required as well. Mr Kornai discussed how the various formats are useful to varying degrees and stressed the importance of metadata in producing valuable resources. He described how production of language resources can be automated.

3.8 Session 8: "Legal framework for Contributing Data?"

Dr. Nóra Takács, a specialist on IPR from the National Intellectual Property Office gave an overview of the effective Hungarian legislation on intellectual property rights. She listed in details the exemptions from IPR legislations and described Hungarian legislation on open data and the recent changes effective as of 1st January on rules of access to public open data. She also stated that the Creative Commons licenses have become widespread in Hungary so far. The main problem is that under Hungarian law a contract must be signed in written form between that parties. She also raised doubt about the authentication of the Creative Commons licences. Nevertheless, we can conclude from her talk that despite the above legal difficulties, there remains enough ground for gathering public open data from Hungarian public bodies.

3.9 Session 9: "Sharing Data and Language Resources: Technical and Practical aspects"

Csaba Oravecz (Research Institute for Linguistics, Hungarian Academy of Sciences) provided a thorough but accessible overview of the practical issues of tracking down data and collecting and annotating them. In addition to the detailed technical issues he also discussed the legal aspects of the access and reuse of relevant datasets. He ended the presentation describing the services that the ELRC consortium can provide to facilitate the work of the national consortia i.e. the Help-desk, the ELRC repository, the legal advisory service and the website.

3.10 Session 10: "How can We Engage?"

In this session Tamás Váradi encouraged participants, particularly data owners from the public sector to come forward and identify datasets under their control that they could make available for the purposes of the ELRC projects. He also solicited suggestions for other stakeholders to involve in the process. He also stressed that the present workshop is the opening of a long-term effort where the EC can be expected to help less resourced and difficult languages like Hungarian but at the same time this also relies on national efforts.

3.11 Session 11: "Open Discussion: Wrap-up, On site Conclusions and Commitments"

This last session was opened by Mr. Stelios Piperidis who provided an impromptu summary of the workshop, which he considered very successful and worthy of previous similar workshops. The issues discussed in Budapest were quite similar to those raised at almost every workshop without exception. He was impressed by the wide range of stakeholders present, from the language sector to decision makers. One very encouraging moment in the final session was when participants were asked to declare their commitments to the objectives of the ELRC effort: there were data owners who made some clear commitments that they were ready to cooperate with the local ELRC coordinators to make relevant datasets available. Finally, Mr. Piperidis thanked the local organizers, the local DGT officer and the host institution and wished a successful cooperation between stakeholders in the future.

4 Synthesis of Workshop Discussions

4.1 Panel 1: Multilingualism in Hungarian Public Services

The moderator of the first panel was Gábor Prószéky. The participants consisted of representatives from the governmental sector (Ministry of National Development, Hungarian Intellectual Property Office), the translation sector (Hungarian Association of Professional Language Service Providers, Hungarian Office for Translation and Attestation Ltd.), and the Language Technology sector (Kilgray Ltd., MorphoLogic Localisation Ltd.)

The first points of the discussion were multi-linguality, languages and language priorities in Hungary. Everyone agreed that one country's economy, the changes, the directions (import, export), and the economy's state as well as political changes can be easily seen and assessed in tendency of the translation market. Unlike the old trend when the German and the English translation was primary, recently the Chinese, the Russian and the Japanese translation has grained ground.

The main problem which translation providers have to face is the lack of experts in certain small languages such as Slovak and Romanian, not only observable among human translators but in machine translation (MT) systems as well. Such systems have only general knowledge and are not able to handle the specialized texts. Also one of the main problems is that those who need to be supported by MT do not know about it. The lack of digitization of texts was another reported problem, while there are certain types of documents requiring human translation. Participants also discussed text quality and quantity, stressing that quantity is very important in statistical MT tasks. However much noise from too much data can overlie the small but important phrases or terms. Hence the proper categorisation (labeling) and the quality assurance should be considered as an important aspect of data collection.

4.2 Panel 2: The data and Language Resources in Hungary: Where to look for what?

The moderator of the second panel was András Kornai and the participants came from the Terminology Council of the Hungarian Language, the Hungarian Office for Translation and Attestation Ltd., the national libraries (National Széchényi Library, Parliamentary Library), and the Hungarian Standards Institution.

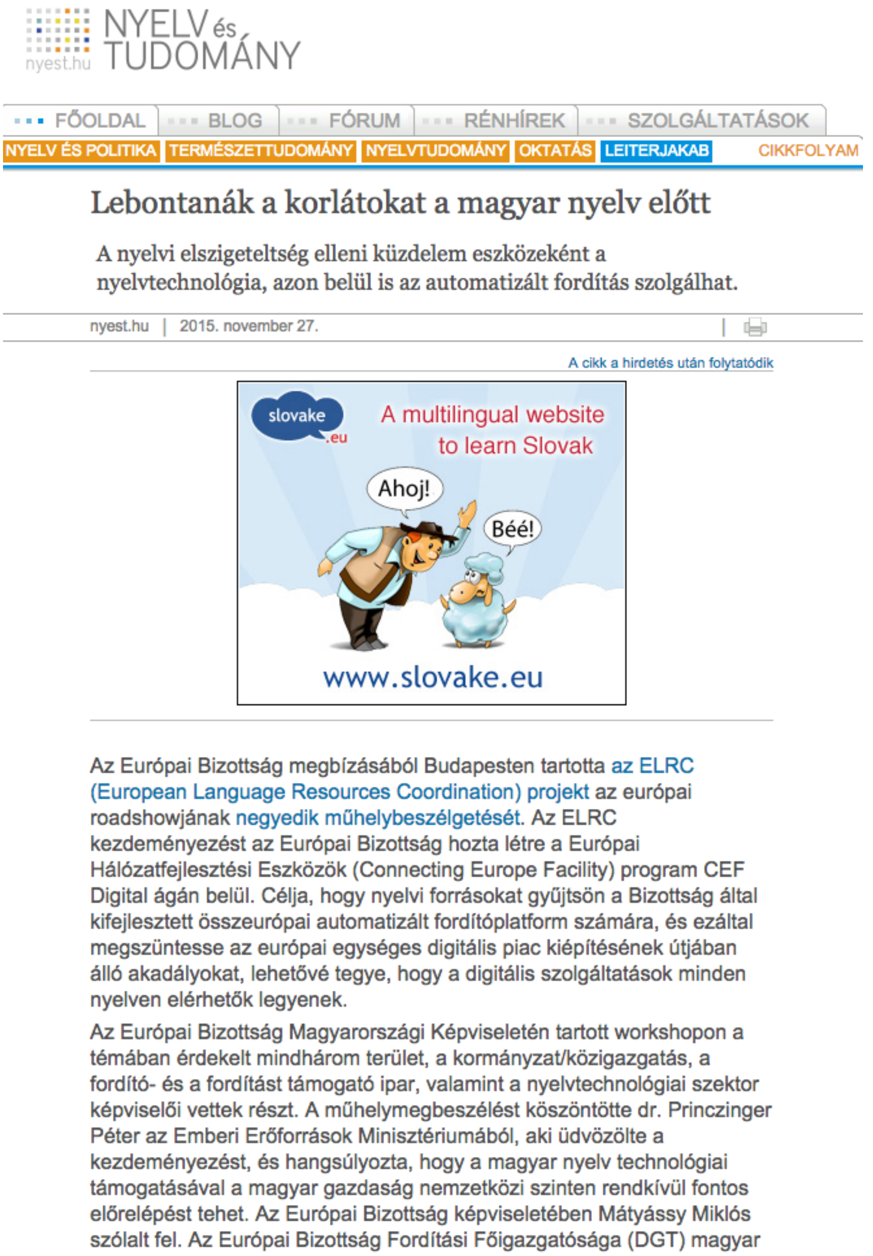
Mr. Kornai started the panel discussion by pointing out that not all data can be used for the purpose of increasing efficiency in machine translation system. On the other hand many legally protected texts can be used with some preparation of anonymization on the corpus. Mr. Kornai suggested a technique called Roaming (Randomize and mix) to overcome legal concerns over copyright problems. He also mentioned further potential organisations to be contacted and collaborated with such as the Ministry of Foreign Affairs and Trade or National Archives.

The Hungarian Office for Translation and Attestation Ltd. has huge amount of parallel texts with such document as administrative documents or official documents, and it started a terminological project a year ago aiming at setting up a legal and administrative terminology database in 20 languages. The Hungarian Standards Institution has roughly 4-5 thousand standards coming into force every year, nonetheless the Hungarian economy doesn't seem to demand the standars. Mr. Váradi in his answer made the ELRC's purpose clear again that the ELRC is not intrested in the content of the standard, but the text of the standard itself. The standars, especially the international standards with Hungarian translation would meet the ELRC's needs. The National Széchényi Library cannot provide parallel corpora but it can offer monolingual texts requested for the design and development of the language model in CEF.AT. The Library of the Parliament confirmed their intention to provide relevant data as the Library contains a lot of parliamentary documents, also parallel texts.

5 Workshop Presentation Materials

The workshop presentations are provided at the event webpage, at http://lr-coordination.eu/budapest_agenda.

6 Appendix



Lebontanák a korlátokat a magyar nyelv előtt

A nyelvi elszigeteltség elleni küzdelem eszközeként a nyelvtechnológia, azon belül is az automatizált fordítás szolgálhat.

nyest.hu | 2015. november 27.

A cikk a hirdetés után folytatódik

Az Európai Bizottság megbízásából Budapesten tartotta az ELRC (European Language Resources Coordination) projekt az európai roadshowjának negyedik műhelybeszélgetését. Az ELRC kezdeményezést az Európai Bizottság hozta létre a Európai Hálózatfejlesztési Eszközök (Connecting Europe Facility) program CEF Digital ágán belül. Célja, hogy nyelvi forrásokat gyűjtsön a Bizottság által kifejlesztett összeurópai automatizált fordítóplatform számára, és ezáltal megszüntesse az európai egységes digitális piac kiépítésének útjában álló akadályokat, lehetővé tegye, hogy a digitális szolgáltatások minden nyelven elérhetők legyenek.

Az Európai Bizottság Magyarországi Képviseletén tartott workshopon a témában érdekelt mindhárom terület, a kormányzat/közigazgatás, a fordító- és a fordítást támogató ipar, valamint a nyelvtechnológiai szektor képviselői vettek részt. A műhelymegbeszélést köszöntötte dr. Princzinger Péter az Emberi Erőforrások Minisztériumából, aki üdvözölte a kezdeményezést, és hangsúlyozta, hogy a magyar nyelv technológiai támogatásával a magyar gazdaság nemzetközi szinten rendkívül fontos előrelépést tehet. Az Európai Bizottság képviseletében Mátyássy Miklós szólalt fel. Az Európai Bizottság Fordítási Főigazgatósága (DGT) magyar

Figure 1 Screenshot of published news of Budapest ELRC

Figure 2 Notebook and pen with ELRC logo

Figure 3 Photo 1

Figure 4 Photo of the Panel 1

Figure 5 Photo of the Panel 2

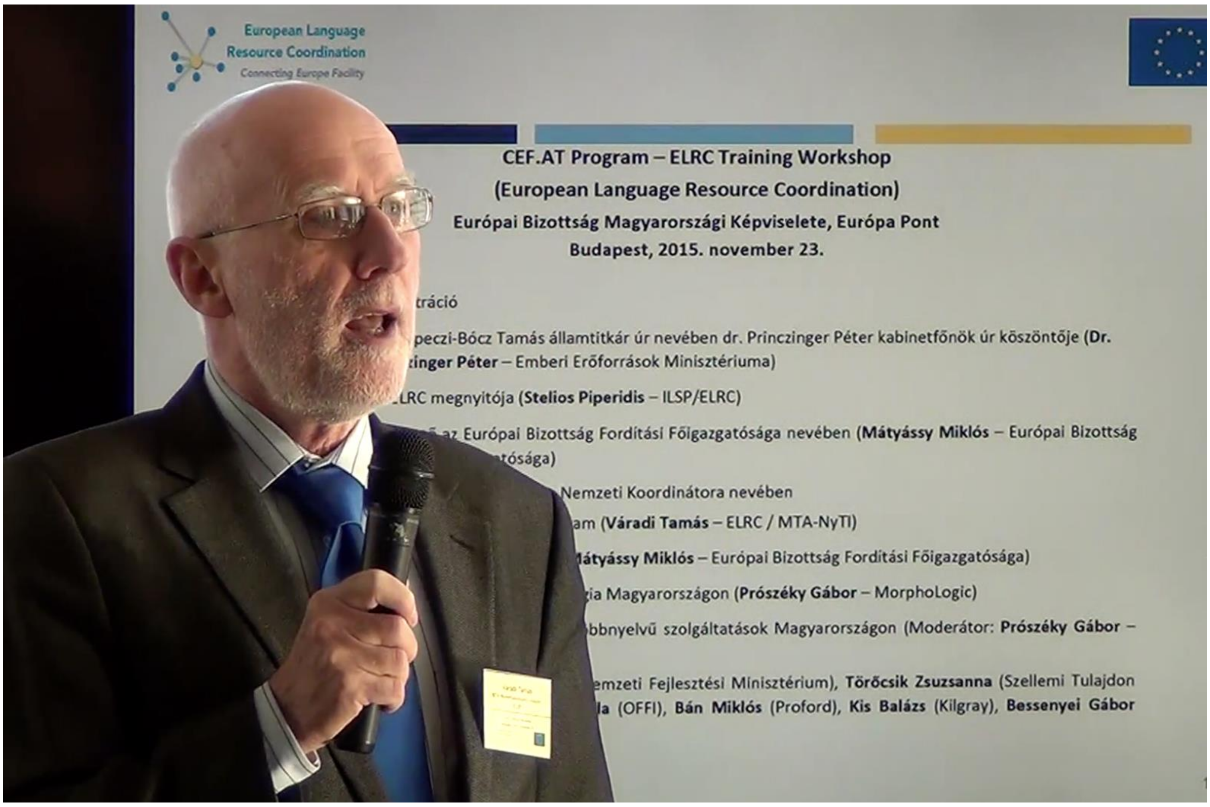

European Language
Resource Coordination
Connecting Europe Facility
CEF.AT Program – ELRC Training Workshop
(European Language Resource Coordination)
Európai Bizottság Magyarországi Képviselete, Európa Pont
Budapest, 2015. november 23.

Figure 6 Photo of Mr. Tamás Váradi

Figure 7 Photo of Mr. Márton Makrai

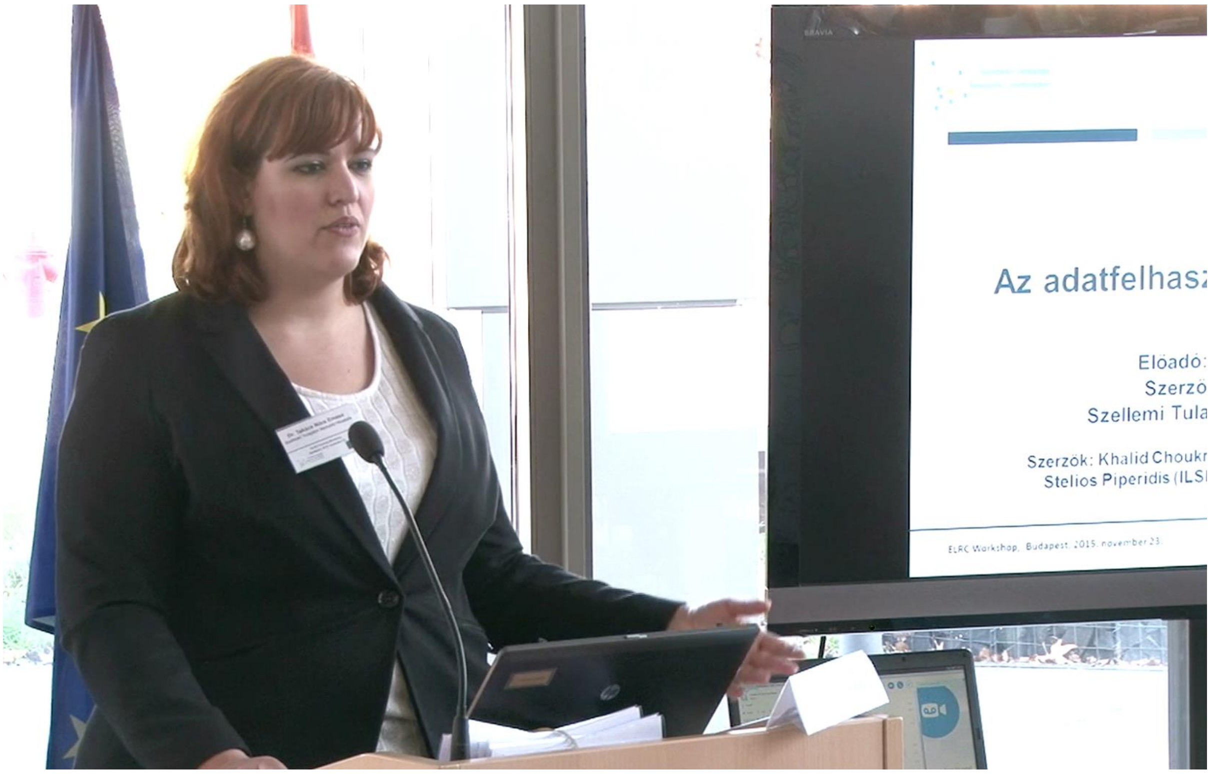


Figure 8 Photo of Mrs. Nóra Emese Takács

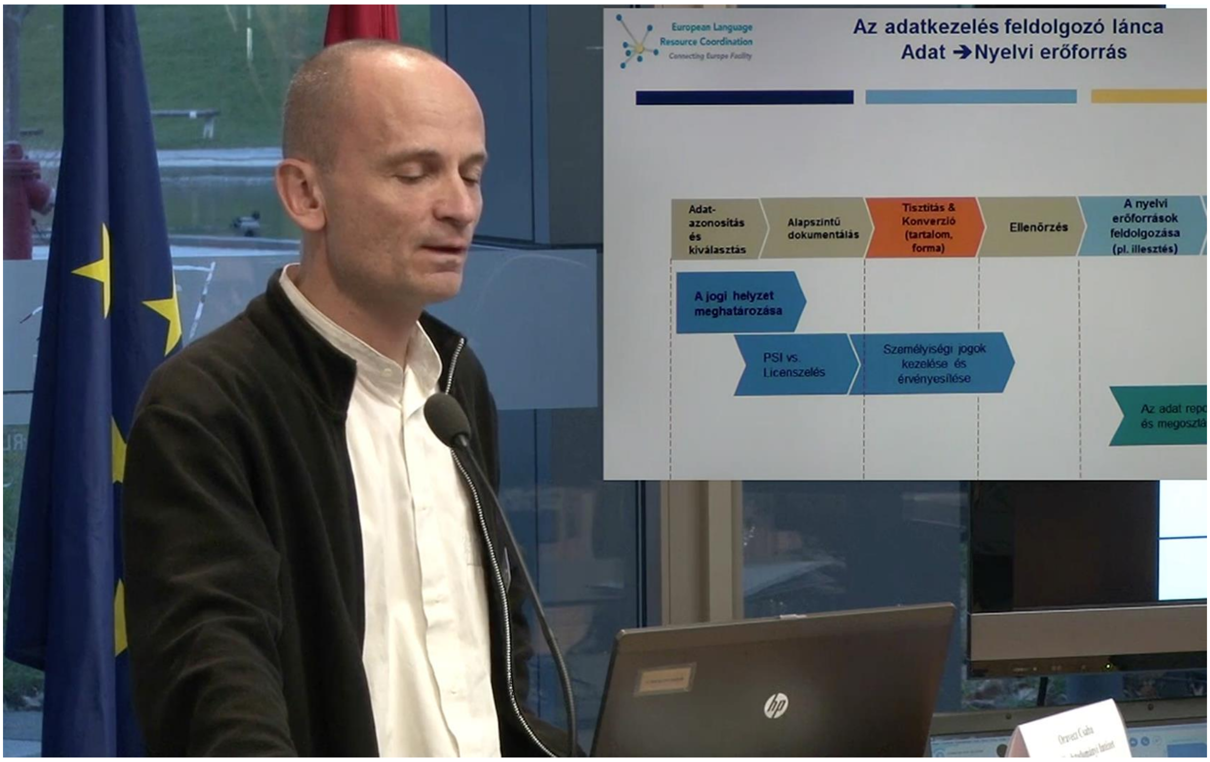

European Language
Resource Coordination
Connecting Europe Facility
Az adatkezelés feldolgozó lánca
Adat → Nyelvi erőforrás
Adat-azonosítás és kiválasztás
Alapszintű dokumentálás
Tisztítás & Konverzió (tartalom, forma)
Ellenőrzés
A nyelvi erőforrások feldolgozása (pl. illesztés)
A jogi helyzet meghatározása
PSI vs. Licenszelés
Személyiségi jogok kezelése és érvényesítése
Az adat repo és megosztá

Figure 9 Photo of Mr. Csaba Oravecz

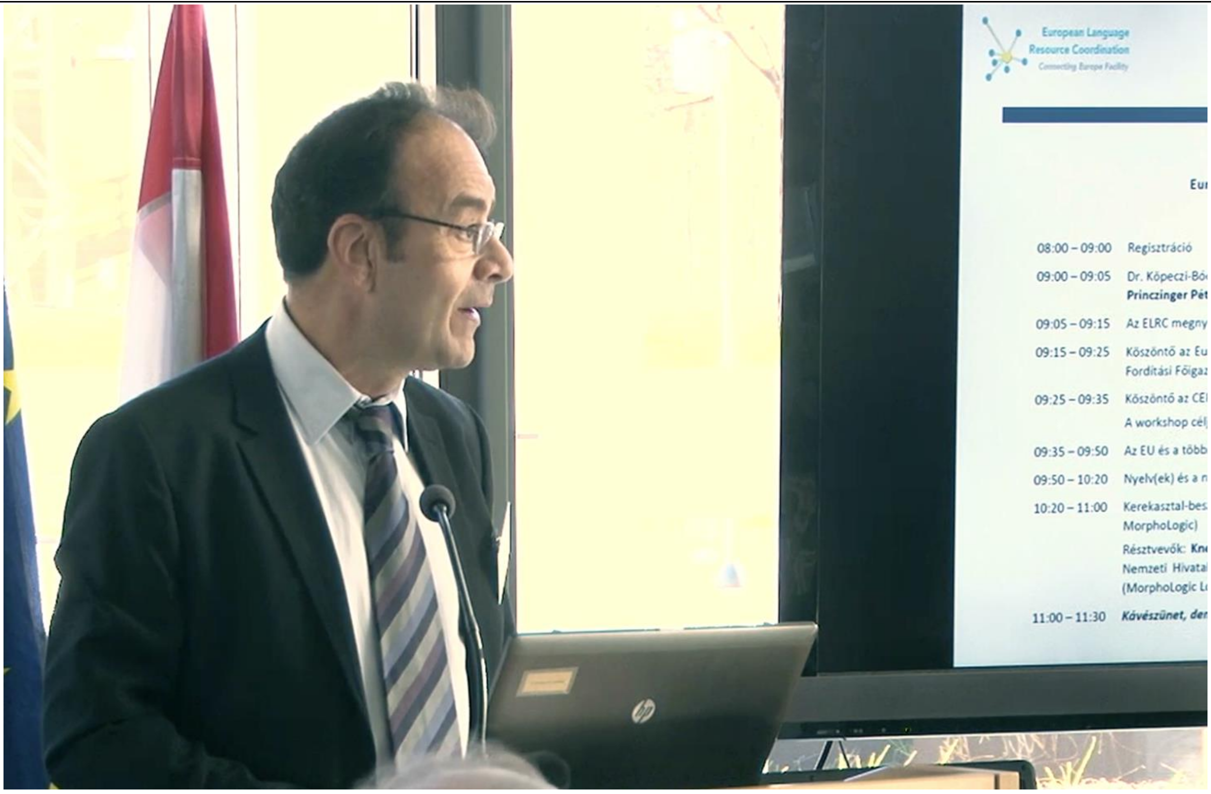

European Language
Resource Coordination
Connecting Europe Facility
Eu
08:00 – 09:00 Regisztráció
09:00 – 09:05 Dr. Köpeczi-Bó
Princzinger Pét
09:05 – 09:15 Az ELRC megny
09:15 – 09:25 Köszöntő az Eu
Fordítási Főigaz
09:25 – 09:35 Köszöntő az CE
A workshop cél
09:35 – 09:50 Az EU és a több
09:50 – 10:20 Nyelv(ek) és a n
10:20 – 11:00 Kerekasztal-bes
MorphoLogic)
Résztvevők: Kn
Nemzeti Hivata
(MorphoLogic L
11:00 – 11:30 Kávészünet, de

Figure 10 Photo of Mr. Stelios Piperidis

Figure 11 Photo of Mr. Péter Princzinger

Figure 12 Photo of Mr. Miklós Mátyássy

Figure 13 Photo of Mr. András Kornai

Figure 14 Photo of Mr. Gábor Prószéky

Figure 15 Photo of the Opening