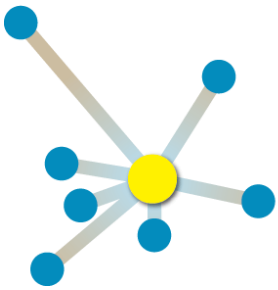
European Language Resource Coordination (ELRC) is a service contract operating under the EU's Connecting Europe Facility SMART 2014/1074 programme.



European Language Resource Coordination

Connecting Europe Facility

Deliverable Task 6

ELRC Workshop Report for Poland

Author(s): Maciej Ogrodniczuk (Institute of Computer Science, Polish Academy of Sciences)

Dissemination Level: Public

Version No.: <V1.0>

Date: 2016-03-22

Contents

1 Executive Summary

This document reports on the ELRC Workshop in Poland, which took place in Warsaw on the 9th of March 2016 at the premises of The European Commission Representation in Poland (Jasna 14/16a). 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 63 participants spanning a wide range of ministries and public organisations. The dedicated event webpage together with all presentations can be found at http://lr-coordination.eu/pl/poland.

2 Workshop Agenda

08:00 – 09:00 Registration

09:00 – 09:10 Opening and welcome
(Maciej Ogrodniczuk – Institute of Computer Science, Polish Academy of Sciences, the ELRC National Anchor Point, Witold Naturski – European Commission)

09:10 – 09:30 Aims and Objectives
(Stelios Piperidis – ELRC/ILSP)

09:30 – 10:00 Europe and Multilingualism
(Jacek Wasik – European Commission, Polish DGT Officer)

10:00 – 10:30 Languages and Language Technologies in Poland
(Marcin Miłkowski – Institute of Philosophy and Sociology, Polish Academy of Sciences)

10:30 – 11:00 Panel 1: Multilingual Public Services in Poland
(moderator: Marcin Miłkowski, participants: Małgorzata Alberti – Civil Aviation Office, Ewa Andrzejuk – National Bank of Poland, Jarosław Deminet – Government Legislation Centre, Monika Popiołek – Head of Technical Committee 256: Principles and Methods of Terminology Work at the Polish Committee for Standardization)

11:00 – 11:30 Coffee break

11:30 – 12:00 Automated Translation: How does it work?
(Krzysztof Łoboda – Jagiellonian University)

12:00 – 12:30 How can Public Institutions benefit from the CEF.AT Platform?
(Szymon Klocek – European Commission)

12:30 – 13:30 Lunch break

13.30 – 14.00 What Data is needed? Why?
(Piotr Pęzik – University of Łódź)

14:00 – 14:30 Legal framework for Contributing Data
(Krzysztof Izdebski – e-Państwo/Fundament)

14:30 – 15:00 Panel 2: Data and Language Resources in Poland
(moderator: Krzysztof Izdebski; participants: Agenor Hofmann-Delbor – localize.pl, Piotr Pęzik, Peter Reynolds – TM-Global)

15:00 – 15:30 Coffee break

15:30 – 16:00 Data and Language Resources: Technical and Practical Aspects
(Stelios Piperidis)

16:00 – 16:30 Discussion: How can we engage?
(moderators: Maciej Ogrodniczuk, Jacek Wasik)

16:30 – 16:45 Wrap-up, On site Conclusions and Commitments
(Maciej Ogrodniczuk, Stelios Piperidis)

3 Summary of Content of Sessions

3.1 Opening and welcome

Maciej Ogrodniczuk from the Institute of Computer Science, Polish Academy of Sciences, the local ELRC representative, opened the event by welcoming the audience and introducing the key persons in conceiving and organizing the event, the ELRC consortium, the EC/DGT representatives, speakers and panelists. Witold Naturski from European Commission welcomed the audience on behalf of the Commission.

3.2 Aims and Objectives

Stelios Piperidis (ELRC/ILSP) presented the context of the workshop: multilinguality of Europe, strongly supported by EU, confronted with translation challenges. The workshop objectives have been set as finding right data for CEF.AT to develop automated translation platform for EU citizens providing better support for each EU language. ELRC was presented briefly together with workshop agenda.

3.3 Europe and Multilingualism

Jacek Wasik (European Commission, Polish DGT Field Officer) presented the landscape of European official languages, policies of the European Union towards languages and several statistics concerning multilinguality in Europe (languages other than English, languages most needed by SMEs etc.) He also reported on the successes of EC in Poland in implementation of European Language Label, promoting European Language Day and Polish participation in translation competitions such as Juvenes Translatores.

3.4 Languages and Language Technologies in Poland

Marcin Miłkowski (Institute of Philosophy and Sociology, Polish Academy of Sciences) concentrated on distinctive features of Polish, the most frequently used West Slavic language in the world, featuring complex morphology and free word order. The author also presented the newest tendencies in Polish and latest developments in linguistic processing of Polish.

3.5 Automated Translation: How does it work?

Krzysztof Łoboda (Jagiellonian University) explained the principles of Statistical Machine Translation which informed the audience that data quality is vital for CEF.AT platform and involvement of public institution in the process of its implementation would facilitate development of services for this particular institution and all EC citizens.

3.6 How can Public Institutions benefit from the CEF.AT Platform?

Szymon Klocek (European Commission) stressed the importance of MT as the key solution for multilingual Europe and the only technical means of providing quick and cheap access to foreign language information. He also gave information about the MT@EC service (opened in 2013) and its usage. Finally he commented on the process of transforming MT@EC into CEF.AT.

3.7 What Data is needed? Why?

Piotr Pęzik (University of Łódź) again stressed the need of large amounts of good-quality data, balanced and most recent, for development of good-quality MT systems. He also discussed the accuracy vs. fluency issue in translation and showed numerous examples of idiomatic translations.

3.8 Legal framework for Contributing Data

Krzysztof Izdebski (e-Państwo/Fundament) presented the legal perspective of the use and reuse of public sector information and stressed the need of putting some input to get better results in the future.

3.9 Data and Language Resources: Technical and Practical Aspects

Stelios Piperidis (ELRC/ILSP) presented in detail the value chain activity consisting of identification and selection of language data, its documentation, cleaning, validation, processing and sharing. The talk addressed several issues related to this process, concerning both legal and technical issues (such as data formatting, anonymization etc.) The session ended with the presentation of the ELRC portal, technical and legal support helpdesk, forum and repository for sharing language resources.

3.10 Discussion: How can we engage?

The discussion was moderated by Maciej Ogrodniczuk and Jacek Wasik and concentrated mostly on actions ELRC could undertake to help public institutions in the process of negotiation of contracts with translation agencies to ensure legality of making the output data available.

The participants expressed their opinion on the role of EC in the process commenting that the examples coming from the top are still much better received in Polish public institutions than bottom-up initiatives. Another remark concerned the need of appropriate presentation of project goals; since the platform will be used by translators, they should be made aware that the system is supporting their work rather than being intended to replace them.

3.11 Wrap-up, on site conclusions and commitments

Stelios Piperidis summarized the day and ended the meeting with a statement that the Polish workshop was the first one to suggest an important improvement to the method of operation of ELRC: the participants expressed the strong need for implementation of the ‘code of conduct’ book gathering the most important best practices for CEF.AT data provision. Such document could act as a lighthouse for the general European public sector.

4 Synthesis of Workshop Discussions

4.1 Panel 1: Multilingual Public Services in Poland

Moderator: **Marcin Miłkowski** – Institute of Philosophy and Sociology, Polish Academy of Sciences
Participants: **Małgorzata Alberti** – Civil Aviation Office
Ewa Andrzejuk – National Bank of Poland
Jarosław Deminet – Government Legislation Centre,
Monika Popiołek – Head of TC256, Polish Committee for Standardization

All participants of the panel were representing institutions carrying out specialist translations (related to aviation, banking, legislation, standardization) and all stressed how important is quality and involvement of experts in the process of translation.

Jarosław Deminet commented on terminological problems concerning legal translation and lack of transparency among public institutions related to opening their terminological databases for further reuse (with the Public Procurement Office being an exception). He also stressed his enthusiasm towards MT related to his professional background (IT).

Ewa Andrzejuk from the banking sector commented on the role of her institution in deployment of EU legal regulations which requires development of Polish terminology. She expressed doubts that any technology could solve this problem since specialist translation involves establishing terminology which requires expert knowledge and creative processes.

Małgorzata Alberti raised the problem of frequent changes in the regulations in her field (aviation) and the fact that translation errors may result in security problems. She also complimented the EC by supporting the translation processes, making the documents available in advance and providing contact with translators.

Monika Popiołek commented the topic from two perspectives: the head of the Polish Committee for Standardization and CEO of a translation agency. She expressed doubts in MT as a general solution to the problem of translation but at the same time she confirmed that it can work well in certain fields such as getting a general idea of the text topic, product description etc.

4.2 Panel 2: Data and Language Resources in Poland

Moderator: **Krzysztof Izdebski** – e-Państwo/Fundament
Participants: **Agenor Hofmann-Delbor** – localize.pl
Piotr Pęzik – University of Lodz
Peter Reynolds – TM-Global

Krzysztof Izdebski asked participants to comment on challenges in data and resource sharing: which reasons should we use to persuade institutions to make the data open, which are the motivations and obstacles, what could translators and users do in this respect?

Peter Reynolds illustrated his speech with the translation market matrix and noted that the most important role for MT solutions implemented by EC should be encouragement of global business by translating from European to non-European languages even more than just concentrating on internal markets. He also pointed out the importance of non-disclosure agreements in contracts with translation companies.

Agenor Hofmann-Delbor noted differences between public institutions in their policies, formats and habits which hinders identification and reuse of data, especially as most documents are still not available in more than one language variant, not to mention ready-to-use translation memories. He also called for implementation of processes facilitating reuse of documents starting at their creation and persuading translation companies to share bilingual data.

Krzysztof Izdebski commented the role of early adopters in promotion of translation technologies.

Piotr Pęzik mentioned the legal problems with collection of texts for the National Corpus of Poland relating them to even more complicated situation of collecting bilingual data (due to copyrighted both source and target). He commented that translation agencies are not interested in sharing their data which makes the role of public institutions in changing this attitude even more important. He concluded that in cases when translation memories might be too difficult to obtain, frameworks for creating on-site language models from original texts without distributing the sources could be equally useful.

5 Workshop Presentations

All presentations are available online on the ELRC website: http://www.lr-coordination.eu/poland_agenda.