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 Germany

Author(s):	Christian Dugast (DFKI)
Dissemination Level:	Public
Version No.:	<V1.1>
Date:	2015-11-03

Contents

1 Executive Summary 3

2 Agenda 4

3 Summary of Content of Sessions 5

3.1 "Opening by ELRC" 5

3.2 Session S1: "Welcome by the Local Anchor Points in Germany" 5

3.3 Session S2: "Welcome by the EC" 5

3.4 Session S3: "Europe and Multilingualism" 5

3.5 Session S4: "Zielsetzung des Workshops / Goal of the Workshop" 5

3.6 Session S5: "Sprachen und Sprachtechnologien in Deutschland / Language and Language Technologies in Europe“ 6

3.7 Session S7: "Wie funktioniert automatisierte Übersetzung? / How machine translation works?“ 6

3.8 Session S8: "How can Public Institutions benefit from the CEF.AT Platform?" 6

3.9 Session S9: "What Data is needed?" 7

3.10 Session S10: "Rechtliche Rahmenbedingungen zur Bereitstellung von Sprachdaten /Legal aspects“ 7

3.11 Session S12: "Data and Language Resources: Technical and Practical Aspects" 7

3.12 Session S13: "Wie können wir uns beteiligen? / How can we participate?“ 7

3.13 Session S14: "Offener Diskurs: Empfehlungen für die Zukunft – Mehrwerte für Ihre Daten / Recommendation for the future - More value for your data“ 7

3.14 Session S15: "Zusammenfassung und Ausblick / Summary and next steps“ 8

4 Synthesis of Workshop Discussions 9

4.1 Panel 1 - Session S6: "Mehrsprachigkeit im öffentlichen Dienst in Deutschland / multilingualism in German public services“ 9

4.2 Panel 2 - Session S11: "Offener Diskurs "Sprachdaten in Deutschland" / Language Data Landscape in Germany“ 9

5 Workshop Presentation Materials 10

1 Executive Summary

This document reports on the ELRC Workshop in Germany, which took place in Berlin, on the 29th of September 2015 at the Bundeshaus. 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 34 participants spanning a wide range of ministries and public organisations. The dedicated event webpage can be found at http://lr-coordination.eu/germany.

2 Agenda

08:00 – 09:00 Registration

09:00 – 09:05 Opening of the workshop (Prof Josef van Genabith, Deutsches Forschungszentrum für Künstliche Intelligenz)

09:05 – 09:15 Welcome by local representatives (Karola Peters, Bundesministerium des Inneren), Sabine Scheidemann, DGT Local Field Office Berlin)

09:15 – 09:25 Welcome by the European Commission (Kimmo Rossi, EC)

09:25 – 09:35 Goal of the Workshop (Prof Josef van Genabith, Deutsches Forschungszentrum für Künstliche Intelligenz)

09:35 – 09:50 Europe and Multilingualism (Kimmo Rossi, European Commission)

09:50 – 10:20 Language and Language Technologies in Germany (Prof Alex Waibel, Karlsruhe Institute of Technology)

10:20 – 11:00 Panel: Multilingualism in German Public Services (Dr Georg Rehm, Deutsches Forschungszentrum für Künstliche Intelligenz)

11:00 – 11:30 Coffee Break and Networking

11:30 – 12:00 How does Machine Translation work? (Prof Josef van Genabith, Deutsches Forschungszentrum für Künstliche Intelligenz)

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

12:30 – 13:30 Lunch Break

13.30 – 14.00 What Data is needed? (Dr Khalid Choukri, European Language Resources Association)

14:00 – 14:30 Legal Aspects (Dr Yuri Glickman, Fraunhofer FOKUS - Open Data Portal; John H. Weitzmann, iRights)

14:30 – 15:00 Open Discussion: Language Data in Germany (Prof Andreas Witt, Institut für Deutsche Sprache)

15:00 – 15:30 Coffee Break and Networking

15:30 – 16:00 Data and Language Resources: Technical and Practical Aspects (Dr Khalid Choukri, European Language Resources Association)

16:00 – 16:30 How can we participate? (Prof Josef van Genabith, Deutsches Forschungszentrum für Künstliche Intelligenz)

16:30 – 17:00 Open Discussion: Recommendation for the – More value for your data (Prof Felix Sasaki, Deutsches Forschungszentrum für Künstliche Intelligenz)

17:00 – 17:15 Summary and next steps (Prof Josef van Genabith, Deutsches Forschungszentrum für Künstliche Intelligenz)

3 Summary of Content of Sessions

3.1 "Opening by ELRC"

Josef van Genabith, the local ELRC representative, opened the event by welcoming the audience and introducing the key persons in conceiving and organizing the event, namely the ELRC consortium and the EC/DGT representatives.

3.2 Session S1: "Welcome by the Local Anchor Points in Germany"

Karola Peters the national anchor point in Germany welcomed the participant in this lovely old building in Berlin, a building that took part in the internal resistance during World War II.

She gave the word to Mrs Scheidemann, DGT Local Field Office Berlin, who mentioned the European Day of Languages a few days before as a very good opportunity to talk about communicating within European Countries: "who speaks other languages can understand the other". For her we have a great opportunity by combining artificial intelligence with human intelligence to make things happen. She recalled that MT@EC can be used by Public Services for free and should be used to build the Digital Single Market Europe is looking at.

3.3 Session S2: "Welcome by the EC"

Kimmo Rossi, Head of Sector, DG CONNECT, European Commission, expressed his welcome message to the audience by setting the Connecting Europe Facility (CEF) scene from the perspective of the multilingual Digital Single Market (DSM) strategy. He identified the market potential of the DSM that can bring a significant additional growth to the European economy while bringing multilingual challenges to the European public services. He insisted on the support of the EC to digital multilingualism. He reported on the nature and the objectives of the CEF Digital and explained the rationale behind the CEF Automated Translation platform (CEF AT) and the expected benefits for the German public services.

3.4 Session S3: "Europe and Multilingualism"

Kimmo Rossi, Head of Sector at the EC, set the European scene with its 24 official languages and 60 major regional/minority pointing out that all languages are equal in Europe and that Europe is committed to multilingualism.

Kimmo Rossi followed with a brief overview with regard to the translation services in the EU, i.e. the volume of translated documents, the number of appointed and freelance translators, and the tools they use in their everyday translation process. He also reported on the challenges that the German language encounters in terms of machine translation and in view of the multilingual Digital Single Market, stressing the need for multilingual support and securing digital inclusion.

3.5 Session S4: "Zielsetzung des Workshops / Goal of the Workshop"

Josef van Genabith (DFKI) presented the context in which Europe is remembering that the 24 European official languages are the base of the culture and the identity of Europe. Europe is not only working against discrimination, it is working on enabling opportunities, helping Europeans to communicate and make business in their own language.

The only way to support this diversity is to ease the access to and give the means for translation. As a consequence, translation demand and volume will increase exponentially what can be covered only with the help of Automated Translation (AT) supporting Human Translation (HT).

Josef van Genabith finished his talk with a strong statement: Supporting our languages is supporting Europe and supporting Europe is supporting our languages.

3.6 Session S5: "Sprachen und Sprachtechnologien in Deutschland / Language and Language Technologies in Europe"

Presentation by Alex Waibel (KIT) who showed that multilingualism in Germany is important, either through minority populations, local languages, regional border languages or international business. Further, Germany drives language technology development in the world.

Alex Waibel demonstrated a voice to text translation system of general purpose in which the whole presentation has been made in English and translated automatically into Spanish, all this happening in real time with no human intervention.

3.7 Session S7: "Wie funktioniert automatisierte Übersetzung? / How machine translation works?"

Josef van Genabith (DFKI) illustrates the difficulty of translation by showing nice ambiguous examples based for example on context.

He presents further the intuition behind Statistical Machine Translation technology in two steps first by showing how it works at word level and then showing how it works with a larger context of a few words (so called phrases). This intuition illustrates clearly why data is very important for developing MT technology.

He continues by explaining how context is important: a system trained on Harry Potter cannot translate well a lawyer's text. And vice versa.

CEF AT is looking for the good data, the data behind the public services it wants to translate.

3.8 Session S8: "How can Public Institutions benefit from the CEF.AT Platform?"

The presentation begins by illustrating the typical interactions between administrations, citizens, and businesses as much within a Member State as between Member States or between Member States and the European Union and it asks the question of accessing any service or business in any Member State using its own language.

Alone translating the Europa.eu website in all 24 European Languages with all 2.000 translators the EU has would take 100 years. It is clear based on this illustration that MT is the only viable solution to provide cross lingual access to information in all Member States.

Further it shows how MT can be used with respect to Human Translation, for example by giving a rough idea of the content of a document written in a foreign language what gives the opportunity to decide whether this document is relevant or not and if it need be translated accurately (by a Human Translator) or not.

MT@EC is than presented as existing already since 26th of June 2013. It has a web user interface in 24 languages for a human-to-machine use case, or can be used as a web service in a machine-to-machine scenario.

It is known that the quality for German translation is not high enough for a heavy usage, but this is why we are here to make it improve with data.

CET AT will augment the service provided by MT@EC by allowing for everyday language, by being able to be more knowledgeable of domains and by being a multilingual enabler not only based on MT.

3.9 Session S9: "What Data is needed?"

Khalid Choukri (ELDA) insisted on the fact that ELRC is looking at bags of words.

ELRC is not only looking at parallel data but at dictionaries, terminologies, ontologies. So called Comparable Collections where both text are more or less telling the same story are also of interest to ELRC.

A list of particularly data formats is presented (*tmx, *.xliff, *.txt, *.doc, *.docx, *.odt, *.ppt).

He mentioned that PDF format is not really useful.

Data identifiers or meta-data sets are important (e.g. based on Dublin core) and should belong to the data.

ELRC can make use of a language resource factory, a set of tools that crawls websites and find the relevant data on that given website. This means that links to the deep web, the one that is not efficiently indexed and that cannot be find easily are very welcome.

3.10 Session S10: "Rechtliche Rahmenbedingungen zur Bereitstellung von Sprachdaten /Legal aspects"

John Weitzmann (iRights.Law) stressed the need for a clear and simple to follow legal framework for the public sector information and data reuse throughout the EU. He briefly presented the PSI directive and explained the structure of the data rights. Most importantly, he listed the step-by-step process to be followed to release data under the PSI Directive and presented actual case studies from various EU countries. The distinct stages in the process refer to: exclusion of confidential information; consent for, anonymization or exclusion of personal data; take care of 3rd party copyrights; follow the national PSI transposition rules; use a standard Open Government Licence, Open public licence or re-use licence, and follow the national or organisational PSI re-use policy.

Yuri Glickman from Fraunhofer presented the advantages of Open Data. He defined the Open Data framework and gave examples of open data in use.

3.11 Session S12: "Data and Language Resources: Technical and Practical Aspects"

Khalid Choukri (ELDA) presented in detail the workflow for the collection, processing and sharing of language resources. Most of the stages of this process, e.g. the identification of the data sources and datasets, the basic metadata documentation, data cleaning and privacy and ethics management are tasks in which the public sector providers will collaborate with ELRC. The presenter encouraged the audience to participate in these activities and work together with ELRC, and he showcased the mechanisms with which ELRC will fully support the providers throughout the whole process, i.e. the helpdesk and user forum mechanism, the ELRC repository and the ELRC website.

3.12 Session S13: "Wie können wir uns beteiligen? / How can we participate?"

Josef van Genabith (DFKI) presents the portal where public services can upload there data.

Organisations having given data have than access to processed data at the end.

3.13 Session S14: "Offener Diskurs: Empfehlungen für die Zukunft – Mehrwerte für Ihre Daten / Recommendation for the future - More value for your data"

Felix Sasaki (DFKI) presents the Open Linked Data concept as a recommendation for managing data in the future showing the value of Open (accessible) Linked (easy harvesting of distributed) data at nearly no additional effort.

3.14 Session S15: “Zusammenfassung und Ausblick / Summary and next steps“

Josef van Genabith (DFKI) summarises the day in presenting the next direct steps to go for:

- Finalise the requirements of the public service administrations with regard to CEF AT (when is CEF AT needed, for which situations, documents etc.; what are typical use cases with regard to languages etc.)
- Clear all aspects around availability of language data for CET AT

In return the public services receive a service from CEF AT that fits their needs.

Longer term next steps:

- Language evolves over time and brings new challenges
- We will continue working together on maintaining the multilingual data flow within Europe
- Supporting multi-linguality is supporting Europe

Josef van Genabith asked for feedback on the workshop, précising that forms are available for participants to fill.

4 Synthesis of Workshop Discussions

4.1 Panel 1 - Session S6: “Mehrsprachigkeit im öffentlichen Dienst in Deutschland / multilingualism in German public services“

Panellist:

Gisela Batzel, Language Service Managerin of the BIBB (Bundesinstitut für Berufsbildung, Institute for industrial training) supervising also the terminology database.

Thomas Santelmann, Translator within the Language Service of the German Parliament, supervising the terminology data base of the Parliament

Andrew Sims, Language Service manager of the BWM (Bundeswirtschaftsministerium, German Ministry of economy).

The role of terminology was discussed and rated as very important. But it needs time and man power. Quality is also rated as very important as Public Services do not control where translated content lands to.

In Germany, the main translation languages are: English, French, Spanish, Russian and Arabic.

It seems that bottlenecks experienced in German is the lack of English mother-tongue speakers.

The German translator community seems to be networking well.

Public services typically do not get translation memories from LSPs.

At end discussing the usage of MT as an intelligent assistant in order to optimise processes is seen positively

4.2 Panel 2 - Session S11: “Offener Diskurs “Sprachdaten in Deutschland” / Language Data Landscape in Germany“

Andreas Witt (IDS) presented IDS as an institution that manages texts in different formats from diverse sources. He gave a list of data IDS is managing.

He pointed out that like with ELRC, the data managed by IDS does not belong IDS but IDS licenses the data in different ways. At the end IDS can always work with the data it gets.

For Mr Witt, making data available to IDS is always a win for better technologies and tools.

Andreas Witt initiated a discussion around the question where which translated data is available in Germany.

A general statement made was that translation services are providing services and as a consequence they do not have ownership of the data. They are not the best contact persons to give data to ELRC. People managing the archive are in a better position to give data.

But access to archive is limited to 30 year old documents. This means the language used in the archive is not actual and 30 years old.

Some institutions are asking if it would be possible to use the data internally and train the MT engine internally and give back only the MT Model to ELRC.

Generally, it seems a consensus that access can certainly be given to brochures and all the data that is centrally stored.

5 Workshop Presentation Materials

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