The European Language Resources Coordination (ELRC)

Prof. Josef van Genabith, Dr. Andrea Lösch

DFKI GmbH, Germany
josef.van_genabith@dfki.de, andrea.loesch@dfki.de

META-FORUM 2016
Lisbon, Portugal – July 04/05, 2016
Why ELRC …?

- Facts & figures - Invisible borders in the Digital Single Market (DSM)
  - While 99% of all European businesses are small and medium-sized enterprises (SMEs), only 7% of them sell cross-border and cross-language.
  - Similarly, 90% of all European customers prefer to browse websites in their own language.

- Automated Translation platform of the Connecting Europe Facility (CEF AT) to facilitate multilingual communication and exchange of documents in key public administration scenarios:
  - Consumer rights, health, public procurement, social security, culture, justice.
  - Public online services: Open Data Portal, Europeana, Online Dispute Resolution, eJustice etc.
What is ELRC about?

- ELRC is about
  - Collecting data to adapt CEF AT to the day-to-day needs of public services in all CEF-affiliated countries
  - Raising awareness on the value and importance of data held by public services in overcoming language barriers
  - Engaging the public sector in the identification and sharing of data for CEF AT
  - Helping with legal and technical issues associated with the collection and/or provision of data by public service administrations

- So we do no research, we do not work scientifically, we do not develop technologies – we help to collect the language resources needed to adapt CEF AT to the needs of European public administrations
The ELRC Consortium

- Deutsches Forschungszentrum für Künstliche Intelligenz (DFKI) – Josef van Genabith
- Evaluations and Language Resources Distribution Agency (ELDA) – Khalid Choukri
- TILDE – Andrejs Vasiljevs
- ILSP (Institute for Language and Speech Processing) – Stelios Piperidis

PLUS:
- 30 ELRC Technological NAPs (one per CEF affiliated country)
- 30 ELRC Public Services NAPs (one per CEF affiliated country)
- Legal advisors (e.g. iRights)
The ELRC Workshops

- Localized workshops conducted in each of the 30 participating countries
- Target audience: National public service administrations
- Goals:
  - To raise awareness about the importance of language data held by public administrations for public administrations
  - To understand the needs of national public service administrations with regard to automated translation
  - To jointly identify relevant sources of multi-lingual language resources
  - To discuss any technical and legal issues involved in the use of data for automated translation
The ELRC Workshops l2

- Past workshops:
  - 24.09.15 Greece
  - 29.09.15 Germany
  - 05.08.15 Latvia
  - 23.11.15 Hungary
  - 01.12.15 Cyprus
  - 08.12.15 Slovenia
  - 15.12.15 Czech Republic
  - 26.01.16 Spain
  - 28.01.16 Ireland
  - 11.02.16 Estonia
  - 19.02.16 Finland
  - 24.02.16 Lithuania
  - 26.02.16 Malta
  - 01.03.16 Portugal

- 07.03.16 Copenhagen
- 09.03.16 Poland
- 10.03.16 Sweden
- 15.03.16 Italy
- 18.03.16 Bulgaria
- 23.03.16 Romania
- 14.04.16 Slovakia
- 15.04.16 Austria
- 19.04.16 Netherlands
- 21.04.16 Croatia
- 11.05.16 France
- 08.06.16 Norway
- 14.06.16 Luxemburg

http://www.meta-net.eu
The ELRC Helpdesk

- Continuous support for data contributors
- Accessible through
  - Phone, Skype, Email
- Services:
  - Answering any technical questions related to the use, production, collection, processing, and sharing of language resources.
  - Answering any legal questions related to the use, production, collection, processing, and sharing of language resources.
- Response times:
  - 24 hours (simple query)
  - Up to 5 days (complex query)
Got a question? We're here to help!

We are happy to answer any questions on the technical or legal aspects related to the use, production, collection, processing, and sharing of language resources.

Please feel free to contact us through one of the following channels:

<table>
<thead>
<tr>
<th>Channel</th>
<th>Contact Information</th>
</tr>
</thead>
<tbody>
<tr>
<td>a Web forum</td>
<td>Web forum</td>
</tr>
<tr>
<td>Telephone*</td>
<td>+33 970 440 522</td>
</tr>
<tr>
<td>reach the Secretariat Support at</td>
<td>+49 681-8575 5285</td>
</tr>
<tr>
<td>Skype</td>
<td>CEF-AT-Helpdesk</td>
</tr>
<tr>
<td>E-mail</td>
<td><a href="mailto:help@cef-at-helpdesk.org">help@cef-at-helpdesk.org</a></td>
</tr>
</tbody>
</table>
The ELRC Conferences

- 2nd Conference: October 2016 in Brussels together with the Translating Europe Forum
  - **Target audience:**
    - National public service administrations in all CEF-affiliated countries
    - 120 participants, at least two representatives per country
  - **Focus and contents:**
    - Demonstrate the benefits and possibilities of MT for public service administrations and raise awareness on the importance of corresponding data and language resources
    - Illustrate what kind of data is needed and how public service administrations can provide such data
    - Give an insight into the services of the CEF Digital Service Infrastructure that should in the future be available to public service administrations in a multi-lingual manner

http://www.meta-net.eu
ELRC’s Language Resource Board

- **Major tasks:**
  - Local support to ELRC activities – executive arm (e.g. workshop organisation)
  - Guidance, feedback and local expertise
  - Setting up continuous data supply mechanism

- **Composition:** 2 representatives from each country
  - 30 Technological National Anchor Points
    - highly regarded language or language technology expert.
    - Often with distinguished academic or research background, or representing a national language institution
    - Support when local technical expertise is needed

http://www.meta-net.eu
Composition: 2 representatives from each country (cont.)

- 30 Public Services National Anchor Points
  - Representative of national public services, public administration or a ministry
  - Officially appointed by the EC
  - Acts as a liaison contact person to the national, regional and local administrations; can effectively mobilizes and spreads the word about the importance of language resources and the ELRC effort among the respective public authorities/ministries

Operations:

- Bi-annual face-to-face meetings
- Monthly Online Q&A Sessions per region
- Individual collaboration for data collection strategy in each country
Within the ELRC contract, the consortium has committed itself to supply at least 1 new data set of the following types for each CEF language:

- parallel corpus (at least two languages),
- large monolingual corpus (here the best monolingual corpora for each language should be considered, e.g. where possible official corpus of a language or other)
- term collection relevant to CEF DSIs (including Online Dispute Resolution (ODR), Europeana, Open data Portal, eJustice and Electronic Exchange of Social Security Information (EESSI))

The aforementioned list will constitute approx. 3x28=84 data sets already. Within ELRC, a total of at least 200 new data sets will be supplied.
A non-exhaustive list of typical datasets that will be sought by ELRC includes:

- aligned parallel corpora,
- translation memories,
- translation/language models,
- comparable corpora,
- monolingual corpora,
- terminologies,
- grammars, etc.

Key: Close collaboration with the National Anchor Points in each country to define and implement the relevant data collection strategy.
Data collection l3

The ELRC data collection process at a glance

- ELRC Workshop conducted
- Analyze participant's feedback (ELRC engagement forms)
- Identify potential data contributors
- Get in touch with potential data contributors
- Forward potential data contributors or data to ELRC

Timeline:
- 2 weeks after workshop
- 3-4 weeks after workshop
- 4 weeks after workshop
- Anytime
Submission of language resources any time possible, either via email (data@lr-coordination.eu) or through upload to the ELRC-SHARE Repository.

The ELRC-SHARE Repository
- is used for documenting, storing, browsing and accessing LR that are collected through ELRC
- is available online via http://www.lr-coordination.eu/resources

It currently holds 25 language resources from various contributors.
Data collection |4

Welcome to the ELRC-SHARE repository!

The ELRC-SHARE repository is used for documenting, storing, browsing and accessing Language Resources that are collected through the European Language Resource Coordination and considered useful for feeding the CEF AT platform.

If you want to contribute resources, all you have to do is register here and go on to describe and upload your data with a simple form.

All data resources gathered by this initiative will be provided exclusively to the European Commission for use in the CEF Automated Translation platform.

http://www.meta-net.eu
Thank you for your attention!

elrc-secretariat@dfki.de

http://www.lr-coordination.eu