The European Language Resource Coordination (ELRC): why and how you can engage and benefit

Stelios Piperidisis
Institute for Language and Speech Processing / Athena R.C.
ELRC
The ELRC context
Heads & organisations behind the ELRC

The ELRC Consortium:

- DFKI (German Research Center for Artificial Intelligence)
- ELDA
- IEA ILSP
- TILDE
Heads & organisations behind the ELRC

30 ELRC Technological National Anchor Points

30 ELRC Public Services National Anchor Points
The ELRC anchor points in Hungary

Tamás Váradi
Research Institute for Linguistics, Hungarian Academy of Sciences

Zoltán Bódi
Institute for Hungarian Language Strategy

Janos Szenthe
Department of EU Law at the Ministry of Justice
The ELRC anchor points in Hungary

Tamás Váradi
Research Institute for Linguistics, Hungarian Academy of Sciences

Zoltán Bódi
Institute for Hungarian Language Strategy

Janos Szenthe
Department of EU Law at the Ministry of Justice
Why ELRC...?

• To facilitate multilingual communication and exchange of information in key public service scenarios
• How? With the right language resources!
• Examples of public services scenarios to which eTranslation may be applied:
  – Consumer Rights (ODR),
  – Social Security (EESSI),
  – Justice (eJustice),
  – Commerce (BRIS),
  – Open Data (EDP),
  – ...

THE EUROPEAN PARLIAMENT VOTED IN FAVOR OF THE RESOLUTION ON "LANGUAGE EQUALITY IN THE DIGITAL AGE" ON 11. SEPTEMBER
“18. Calls on the Commission and the Member States to develop strategies and policy action to facilitate multilingualism in the digital market; requests, in this context, that the Commission and the Member States define the minimum language resources that all European languages should possess, such as data sets, lexicons, speech records, translation memories, annotated corpora and encyclopaedic content, in order to prevent digital extinction;”
“21. Calls on the Commission to make as a priority of language technology those Member States which are small in size and have their own language, in order to pay heed to the linguistic challenges that they face;”
Committee on Industry, Research and Energy (11.9.18): stresses "the role of the EU, the Member States, universities and other public institutions in contributing to the preservation of their languages in the digital world and in developing databases and translation technologies for all EU languages, including languages that are less widely spoken; calls for coordination between research and industry with a common objective of enhancing the digital possibilities for language translation and with open access to the data required for technological advancement".
• Cracking the Language Barrier Federation (2016): “We recommend that Europe actively makes an effort to compete in the global landscape for research and development in language technology since we cannot expect third parties from other continents to solve our translation and knowledge management problems in a way that suits our specific communicative, societal and cultural needs. Language Technology made for Europe in Europe is the key. It will contribute to future European crossborder and cross-language communication, economic growth and social stability.”
## What does ELRC do?

<table>
<thead>
<tr>
<th>Action</th>
<th>Description</th>
</tr>
</thead>
<tbody>
<tr>
<td>Collect</td>
<td>Language resources</td>
</tr>
<tr>
<td>Identify</td>
<td>Needs of public services</td>
</tr>
<tr>
<td>Engage</td>
<td>With the public sector in the identification of LR</td>
</tr>
<tr>
<td>Help</td>
<td>With any technical or legal issues</td>
</tr>
<tr>
<td>Act</td>
<td>Observatory for language resources across Europe</td>
</tr>
</tbody>
</table>
How to make it work…?

With the right language resources!

- Add in-domain text
- In-domain data
- Good translation
How to make it work…?

In-domain data

Out-of-domain data

➔ Your language resources make a difference!
What has been achieved so far...?

The ELRC network

More than 400 identified data holders

60 National Anchor Points

ELRC consortium

Today

April 2016

April 2015
What has been achieved so far...?

ELRC Data in April 2017

A total of 225 LRs provided by ELRC

- Open data / cc licenses: 78%
- Non-standard/other licenses: 17%
- Under review: 5%
What has been achieved so far…?

ELRC Data in October 2018

More than 500 LRs contributed through ELRC

- cc / open licences: 73%
- non-standard licences: 12%
- under review: 15%
What has been achieved so far...?
ELRC Data by language

Development of LR collection by language

BG CS DA DE EL ES ET FI FR GA HR HU IS IT LT LV MT NL NO PL PT RO SK SL SV

Apr 17  Oct 18
What has been achieved so far...?

ELRC Data by language

Development of LR collection by language

- Apr 17
- Oct 18

Languages: BG, CS, DA, DE, EL, EN, ES, ET, FI, FR, GA, HR, HU, IS, IT, LT, LV, MT, NL, NO, PL, PT, RO, SK, SL, SV
The notion of data

- Weather / climate
- Geospatial
- Educational
- Public spending
- Public health and insurance
- Demographic

European Language Resource Coordination
Connecting Europe Facility
The notion of data

Data: the oil of the 21st century
The notion of data
Basic concepts:

- **Data**: any piece of electronically stored content
- **Dataset (or resource)**: the collection of one or many data files grouped according to certain criteria
- **Metadata**: *data about the data*, i.e. description of a dataset with properties (e.g. title, publisher, description of the content and URL)
Language data (resources) for eTranslation through ELRC
The notion of language data

Data
• any piece of electronically stored content

(TEXTUAL) LANGUAGE DATA
• any piece of electronically stored text
International human rights law is grounded on the premise that all individuals are entitled to the same rights and freedoms as set out in the Universal Declaration of Human Rights. Human rights shall apply equally to every human being, regardless of sexual orientation or gender identity.
Data used by eTranslation
Such data are already available
BUT
they are not enough...
What data are useful for eTranslation as per type | 1

- Any **electronically stored text** in an EU language plus NO and IS
- **Texts and their translations** (i.e. parallel bilingual or multilingual)

**Hungarian text**

A feladatok harmadik csoportját az Országgyűlés saját felkészülése képezi, amelynek során a Parlament felkészül az elnökség alatt megrendezendő parlamenti bizottsági találkozókra. Ennek irányítására létrejött egy Koordinációs csoport, amelynek vezetője az Országgyűlés Külgügyi Hivatalának vezetője, tagjai az EU Főoszály vezetője, a Protokoll Főosztály vezetője, valamint a Főtitkárság, a Sajtóiroda, a Gazdasági Főigazgatóság és az Országgyűlés Kutatószolgálatának képviselői, illetve az Országgyűlés brüsszeli állandó irodájának vezetője.

**Translation in English**

The third group of the preparatory tasks is the National Assembly's own preparation, where the Parliament prepares for the meetings of the parliamentary committees to be held during its presidency. To coordinate these meetings, a Coordination Group was established that is led by the head of the Office for Foreign Relations of the National Assembly, its members are the head of the EU Department, the head of the Protocol Department and the representatives of the General Secretariat, the Press Office, the Economic Directorate and the Research Service of the National Assembly, as well as the Permanent Representative of the National Assembly to the EU in Brussels.
List of terms and their translations, i.e. a **terminology**

<table>
<thead>
<tr>
<th>English</th>
<th>Hungarian</th>
</tr>
</thead>
<tbody>
<tr>
<td>accident insurance</td>
<td>balesetbiztosítás</td>
</tr>
<tr>
<td>health insurance</td>
<td>betegbiztosítás, egészségbiztosítás</td>
</tr>
<tr>
<td>household insurance</td>
<td>háztartási biztosítás</td>
</tr>
<tr>
<td>insurance broker</td>
<td>biztosítási ügynök</td>
</tr>
<tr>
<td>medical insurance</td>
<td>betegbiztosítás</td>
</tr>
<tr>
<td>sickness insurance</td>
<td>betegbiztosítás</td>
</tr>
<tr>
<td>unemployment insurance</td>
<td>munkanélküliségi biztosítás</td>
</tr>
<tr>
<td>health insurance card</td>
<td>egészségbiztosítási kártya</td>
</tr>
<tr>
<td>nursing care insurance</td>
<td>gondozási biztosítás</td>
</tr>
<tr>
<td>...</td>
<td>...</td>
</tr>
</tbody>
</table>
In principle, any text in machine readable format

But, some formats are more “MT-ready” than others, i.e. they require less manual or automatic processing

More processing introduces more errors in the final output, making it less useful for eTranslation
The following formats are particularly useful (in descending order):

- For bilingual/multilingual parallel texts
  1. Translation memories (.tmx)
  2. XML translation files (.xliff)
  3. Plain text (.txt, .csv)
  4. Spreadsheets (e.g. xlsx)

- For terminologies
  1. TermBase eXchange (.tbx)
  2. Plain text (.txt, .csv)
  3. Spreadsheets (e.g. xlsx)

- For monolingual texts
  1. Plain text (.txt, .csv)
File formats of parallel texts and their manipulation
Preparing your data | 1

This is a paragraph in English. This is a paragraph in English. This is a paragraph in English. This is a paragraph in English. This is a paragraph in English. This is a paragraph in English. This is a paragraph in English. This is a paragraph in English.

Don’ts


Ez a második bekezdés magyar fordítása. Ez a második bekezdés magyar fordítása. Ez a második bekezdés magyar fordítása. Ez a második bekezdés magyar fordítása. Ez a második bekezdés magyar fordítása. Ez a második bekezdés magyar fordítása. Ez a második bekezdés magyar fordítása.
Don’ts

- This is a paragraph in English. This is a paragraph in English. This is a paragraph in English. This is a paragraph in English. This is a paragraph in English. This is a paragraph in English. This is a paragraph in English. This is a paragraph in English. This is a paragraph in English. This is a paragraph in English. This is a paragraph in English. This is a paragraph in English. This is a paragraph in English. This is a paragraph in English.

- A sentence in English.


<table>
<thead>
<tr>
<th>English</th>
<th>Magyar</th>
</tr>
</thead>
<tbody>
<tr>
<td>This is a paragraph in English. This is a paragraph in English. This is a paragraph in English. This is a paragraph in English. This is a paragraph in English. This is a paragraph in English. This is a paragraph in English.</td>
<td>Ez a bal oldali bekezdés magyar fordítása. Ez a bal oldali bekezdés magyar fordítása. Ez a bal oldali bekezdés magyar fordítása. Ez a bal oldali bekezdés magyar fordítása. Ez a bal oldali bekezdés magyar fordítása.</td>
</tr>
</tbody>
</table>
Use identical filenames for each document pair (source – translation)
Preparing your data | 5

Do’s

Include **language identifiers** in the filename
Preparing your data

Do's

<table>
<thead>
<tr>
<th>Language</th>
<th>Files</th>
</tr>
</thead>
<tbody>
<tr>
<td>English</td>
<td>text.txt, report.txt</td>
</tr>
<tr>
<td>Bulgarian</td>
<td>tekst.txt, report in Bulgarian.txt</td>
</tr>
<tr>
<td>French</td>
<td>texte.txt, traduction française.txt</td>
</tr>
<tr>
<td>Greek</td>
<td>keimeo.txt, Metaphors of references.txt</td>
</tr>
</tbody>
</table>

- Do's:
  - Organize your data folder
  - Use descriptive file names
  - Include a README.txt file

- Example files:
  - text.txt
  - report.txt
  - README.txt
Criteria for grouping your data into meaningful datasets

- A dataset is a collection of data **grouped according to certain criteria**
- For the purpose of enhancing and adapting CEF eTranslation, two criteria are critical:
  - **Language(s)**: each collection is defined by the language or language pairs of its data, e.g.
    - *Collection of texts in English – Hungarian*
    - *Documents in English – Hungarian - French*
  - **Domain**: each collection ideally belongs to a single domain, e.g.
    - *Collection of texts in English – Hungarian in the culture domain*
    - *Social security documents in English – Hungarian- French*
Preferred domains

- Administrative/regulatory domain and
- Topics relevant to the CEF DSIs

<table>
<thead>
<tr>
<th>CEF DSI</th>
<th>Domain</th>
</tr>
</thead>
<tbody>
<tr>
<td>Online Dispute Resolution</td>
<td>Consumers’ rights, complaints</td>
</tr>
<tr>
<td>Electronic Exchange of Social Security Information</td>
<td>Social security, insurance</td>
</tr>
<tr>
<td>eProcurement</td>
<td>Public procurement, contractual agreements</td>
</tr>
<tr>
<td>European e-Justice Portal</td>
<td>Justice, Law</td>
</tr>
<tr>
<td>eHealth</td>
<td>Health, Medicine</td>
</tr>
<tr>
<td>Business Registers Interconnection System</td>
<td>Business, market</td>
</tr>
<tr>
<td>Safer Internet</td>
<td></td>
</tr>
<tr>
<td>Cybersecurity</td>
<td></td>
</tr>
<tr>
<td>Public Open Data</td>
<td></td>
</tr>
<tr>
<td>Europeana</td>
<td>Culture</td>
</tr>
</tbody>
</table>
How to contribute your data to CEF eTranslation
A step-by-step guide
The ELRC-SHARE Repository

- Access to, sharing and contribution of language resources
- Access to tools and services catalogue (upcoming)

https://elrc-share.eu
How to Contribute Data

Register
- Register
- Activate account

Contribute Data
- Login
- Describe
- Upload
How to Register (1/2)

Welcome to the ELRC-SHARE repository!
• Fill in the required info
• Read the Terms of Service and click Accept, if you agree
• Click the Create Account button
• Activate your account according to the guidelines emailed to you
• Fill in the details of the dataset
• Three modes for contributing your data

**Contribution Mode**
- **Upload ZIP archive**
- **Provide URL of resources**
- **eDelivery (Generate XML file to attach to your eDelivery contribution)**

Please select the way you wish to contribute your data. Uploading a ZIP archive is recommended.

**Upload Resource**

Choose File: No file chosen

Please upload a **.zip file** up to 100MB.

In case the **.zip file** file you wish to upload is larger than 100MB, please contact elrc-share@ilsp.gr
How to create a .zip file

• Free file compression tools (indicative):
  – 7zip
  – PeaZip
  – Hamster Free Zip Archiver
  – Universal Extractor
  – ZipItFree

• Windows embedded compression functionality
1. Click on Choose file
2. Locate your resource in your hard disk
3. Click on Submit
• Alternatively indicate a url (directory listing)

Language(s)*
Bulgarian
Czech
Croatian
Danish
Dutch; Flemish
English
Estonian
Finnish
French
German
Hungarian

The language(s) of the resource; for resources with multiple languages, hold down CTRL key to select multiple values

Contribution Mode*
- Upload ZIP archive
- Provide URL of resources

Please select the way you wish to contribute your data. Uploading a ZIP archive is recommended.

Resource URL*

Please provide a URL containing the files you wish to contribute
• Repeat the process if you want to contribute another resource, or log out
Guidelines for contributors

Documentation on the ELRC-SHARE editor

The following guidelines provide detailed information on how to use the editing facility for documenting and uploading LRs:

- Walkthrough for contributors
- Walkthrough for editors

ELRC-SHARE schema

- ELRC-SHARE schema XSD (based on the META-SHARE Schema)
- Documentation about the schema
What happens next?
What happens to your data?

Data contributor

Upload to ELRC-SHARE

ELRC processes your data

Processed data
Language processing services

Data extraction
If your data is trapped in archives and databases, we can help extract it

Anonymisation
Does your data contain private info? We can help to anonymise

Cleaning
If your data is messy (i.e., lots of noise), we will clean it up

Re-formatting
Need to re-format DOCX to XML, or PDF to WORD? Let us do it for you!
Metadata are crucial! We can organise and validate metadata for your team.
Weights: they express the typical percentages the groups of different goods or services represent within the total expenditures of households.

Weights are revised annually, based on household consumption in the year two years before the reference year.

Súlyarány: kifejezi, hogy az egyes termékek és szolgáltatások csoportjai a lakosság összes kiadásán belül milyen jellemző százalékarányt képviselnek.

A súlyok felülvizsgálata évente történik, a tárgyévet két évvel megelőző év lakossági fogyasztása alapján.
How your dataset is described

Translations of Hungarian from public websites (Processed)

A web crawl of 14 different websites covering parallel corpora of Hungarian with Polish, Czech, Swedish, Finnish, French, German, Italian, English and Slovenian (Processed)

Multilingual text corpus

- **Languages**
  - Dutch (nl)
  - Finnish (fi)
  - German (de)
  - Czech (cs)
  - Hungarian (hu)
  - English (en)
  - Polish (pl)
  - Finnish (fi)
  - Swedish (sv)
  - French (fr)

- **Linguality**
  - Linguality type: Multilingual
- **Text Format**
  - TMX

- **Size**
  - 12 files
  - 11,087 Translation Units
  - Character encoding: UTF-8

Resource Creation

- **Resource Creator**
  - Tamas Varadi

- **Funding Project**

- **URL**: http://www.l-koord...
All these services can also be offered on-site to all data contributors free of charge
Our team of experts will travel directly to assist you at your own offices
We will fix your data issues and return the processed data directly to you. We can also help to improve your data management processes. Just ask!
What happens to your data?

Data contributor

Upload to ELRC-SHARE

ELRC processes your data

Processed data

Request on-site assistance
How to request services and help
www.lr-coordination.eu/request-onsite-assistance

Submit a request for on-site assistance by filling out the form below. See a list of services here.

First name *

Last name *

Institution *

Country *

Email *

Types of assistance required *
- Legal assistance
- Data processing
- Anonymisation
- Other

Description of assistance required

Submit
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>Telephone*</td>
<td>+33 970 440 522</td>
</tr>
<tr>
<td>Secretariat Support</td>
<td>+49 681 857 7552 85</td>
</tr>
<tr>
<td>Skype</td>
<td>ELRC Helpdesk</td>
</tr>
<tr>
<td>E-mail</td>
<td><a href="mailto:help@lr-coordination.eu">help@lr-coordination.eu</a></td>
</tr>
</tbody>
</table>

*Telephone number for international calls only.
Köszönöm!

Email: info@lr-coordination.eu
Website: www.lr-coordination.eu
Icons used in this presentation

- By Michael Mellon, GB, CC-BY 3.0 US
- By Joana Pereira, BR, CC-BY 3.0 US
- By Becca O'Shea, NZ, CC-BY 3.0 US
- By Creative Stall, Basic licence www.iconfinder.com
- By Creative Stall, PK, CC-BY 3.0 US
- By Arthur Shlain, IL, CC-BY 3.0 US
- By Shmidt Sergey, US, CC-BY 3.0 US
- By Gregor Cresnar, CC-BY 3.0 US
- By anbileru adaleru, CC-BY 3.0 US
- By Vectors Market, CC-BY 3.0 US