Preparing and sharing data with the ELRC-SHARE repository and what happens next

Maria Giagkou
Institute for Language and Speech Processing / Athena R.C.
ELRC
The notion of data

- Weather / climate
- Geospatial
- Educational
- Public spending
- Public health and insurance
- Demographic
The notion of data

Data: the oil of the 21st century
The notion of data
The notion of language data

Data
• any piece of electronically stored content

(Textual) Language Data
• any piece of electronically stored text
Legea penală prevede pedepsele aplicabile și măsurile educative ce se pot lua față de persoanele care au săvârșit infracțiuni, precum și măsurile de siguranță ce se pot lua față de persoanele care au comis fapte prevăzute de legea penală.

Criminal law establishes applicable penalties and educational measures that can be ruled against persons who committed offenses, as well as security measures that can be ruled against persons who committed actions covered by criminal law.
Data used by eTranslation

DGT translation memory

Institutions and organizations supplying data to eTranslation:

- European Parliament
- European Commission
- CVRIA
- Council of the European Union
- European Central Bank
- European Committee of the Regions
- Connecting Europe Facility (CEF)
Such data are already available
BUT
they are not enough...
What data are useful for eTranslation as per type |

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

### Romanian text

Legalitatea sancțiunilor de drept penal

1. Legea penală prevede pedepsele aplicabile și măsurile educative ce se pot lua față de persoanele care au săvârșit infracțiuni, precum și măsurile de siguranță ce se pot lua față de persoanele care au comis fapte prevăzute de legea penală.

2. Nu se poate aplica o pedeapsă ori nu se poate lua o măsură educativă sau o măsură de siguranță dacă aceasta nu era prevăzută de legea penală la data când fapta a fost săvârșită.


### Translation in English

Lawfulness of criminal penalty

1. Criminal law establishes applicable penalties and educational measures that can be ruled against persons who committed offenses, as well as security measures that can be ruled against persons who committed actions covered by criminal law.

2. No penalty, educational or security measure can be ruled that was not stipulated in criminal law at the date when the violation was committed.

3. No penalty can be ruled and enforced outside the law’s general limits.
What data are useful for eTranslation as per type |2

- List of terms and their translations, i.e. a **terminology**

<table>
<thead>
<tr>
<th>Romanian</th>
<th>English</th>
</tr>
</thead>
<tbody>
<tr>
<td>Risc</td>
<td>risk</td>
</tr>
<tr>
<td>acceptarea riscului</td>
<td>risk acceptance</td>
</tr>
<tr>
<td>risc acceptabil</td>
<td>acceptable risk</td>
</tr>
<tr>
<td>risc de incendiu</td>
<td>fire risk</td>
</tr>
<tr>
<td>risc de inundație</td>
<td>flood risk</td>
</tr>
<tr>
<td>risc rezidual</td>
<td>residual risk</td>
</tr>
<tr>
<td>analiza riscului</td>
<td>risk analysis</td>
</tr>
<tr>
<td>modificarea climei</td>
<td>climate change</td>
</tr>
<tr>
<td>eroziune</td>
<td>erosion</td>
</tr>
<tr>
<td>degradarea mediului</td>
<td>environmental degradation</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
File formats for parallel texts
• 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
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.

Aceasta este traducerea în română a paragrafului de mai sus. Aceasta este traducerea în română a paragrafului de mai sus. Aceasta este traducerea în română a paragrafului de mai sus. Aceasta este traducerea în română a paragrafului de mai sus. Aceasta este traducerea în română a paragrafului de mai sus.


Aceasta este traducerea în română a paragrafului de mai sus. Aceasta este traducerea în română a paragrafului de mai sus. Aceasta este traducerea în română a paragrafului de mai sus.
Preparing your data

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. This is a paragraph in English.

A sentence in English.


### Don’ts

<table>
<thead>
<tr>
<th>English</th>
<th>Română</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>Aceasta este traducerea în română a paragrafului din stânga. Aceasta este traducerea în română a paragrafului din stânga. Aceasta este traducerea în română a paragrafului din stânga. Aceasta este traducerea în română a paragrafului din stânga.</td>
</tr>
</tbody>
</table>
Use **identical filenames** for each document pair (source – translation).
Preparing your data | 5

Include language identifiers in the filename

Do's

filename01_EN.txt
filename01_RO.txt
filename02_EN.txt
filename02_RO.txt
filename03_EN.txt
filename03_RO.txt
filename04_EN.txt
filename04_RO.txt
filename05_EN.txt
filename05_RO.txt
filename06_EN.txt
filename06_RO.txt
filename07_EN.txt
filename07_RO.txt
filename08_EN.txt
filename08_RO.txt
filename09_EN.txt
filename09_RO.txt
filename10_EN.txt
filename10_RO.txt
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 – Romanian*
  - *Documents in English – Romanian - French*

- **Domain:** each collection ideally belongs to a single domain, e.g.
  - *Collection of texts in English – Romanian in the culture domain*
  - *Social security documents in English – Romanian - 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
To ELRC-SHARE

- At the ELRC portal click on the “Language resource submission” button
- Or
- Type in the url address: elrc-share.eu

What are Language Resources?

The term language resources refers to sets of language data and descriptions in machine readable form, including written and spoken corpora, grammars, and terminology databases. Language resources can be used to build, improve, or evaluate natural language systems such as machine translation engines.

To develop the automated translation systems for the CEF Automated Translation platform, the ELRC initiative aims to gather language resources in all official languages of EU. The initiative seeks large general-domain corpora, whether monolingual (e.g. official corpora of national languages) or multilingual, as well as domain-specific language resources in the fields of consumer rights, culture, legal domain, social security, health, public procurement, etc.

Read more about what language resources are needed

How to contribute?

Any contributor may submit Language Resources to us at any exploitation stage: simple internet links to websites (Sources), raw data, or fully-packaged data (Language Resources).

Click below if you can indicate a potential source for relevant data

Click below if you are a language resource owner and are willing to share it for the purposes of CEF.AT
Welcome to the ELRC-SHARE repository!
How to Contribute Data

- Register
- Activate account

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

Welcome to the ELRC-SHARE repository!

Type in your keywords, please...
• 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
How to Contribute Data (1/6)

Data Contribution

New Resource

**Resource Title**

The name by which the resource is already known or by which you would like it to be known; e.g. “The GSRT bilingual corpus of Greek-English bulletins”
How to Contribute Data (2/6)

• Fill in the details of the dataset
**How to Contribute Data (3/6)**

- 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
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*
www
Please provide a URL containing the files you wish to contribute

Submit Reset
• 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

CEF Digital
Connecting Europe
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!
Language processing services

Data conversion
If your data isn’t converted to the proper formats, we can help convert it

Tag removal
Does your data contain unneeded tags? We can assist in removing them!

Alignment
Translations aren’t aligned? We’ll do it for you with our tools!

Metadata
Metadata are crucial! We can organise and validate metadata for your team
What has happened to your data?

After processing:

- File01_ro.txt
- File01_en.doc
- File02_ro.pdf
- File02_en.txt
- File03_ro.doc
- File03_en.doc

XML:

```xml
<tu tuid="269">
  <tuv xml:lang="en">
    <seg>Ancillary penalties, additional penalties, and security measures in case of multiple offenses</seg>
  </tuv>
</tu>

<tu tuid="270">
  <tuv xml:lang="ro">
    <seg>Pedepsele complementare, pedepsele accesorii și măsurile de siguranță în caz de pluralitate de infracțiuni</seg>
  </tuv>
</tu>

<tu tuid="269">
  <tuv xml:lang="en">
    <seg>(1) If one of the committed violations carries a ancillary penalty, such penalty shall be awarded alongside the main penalty.</seg>
  </tuv>
</tu>

<tu tuid="270">
  <tuv xml:lang="ro">
    <seg>(1) Dacă pentru una dintre infracțiunile săvârșite s-a stabilit și o pedeapsă complementară, aceasta se aplică alături de pedeapsa principală.</seg>
  </tuv>
</tu>
```
General Romanian-English bilingual corpus

Attribution details: General Romanian-English bilingual corpus was created for the European Language Resources Coordination Action (ELRC) (http://lr-coordination.eu) by Tufis Dan, Institutul de Cercuri pentru Intelectual Artificila "Mihai Draganscu", Academia Romana (www.racal.ro) with primary data copyrighted by Wikipedia (https://en.wikipedia.org) and is licensed under "CC-BY-SA 3.0" (https://creativecommons.org/licenses/by-sa/3.0/).

Romanian – English corpus built from a Wikipedia dump.

DSI Relevance: Europeana

Distribution
Availability: Available
Licences
CC-BY-SA-3.0
Conditions: Attribution, Share Alike
Distribution Details
Attribution Details: General Romanian-English bilingual corpus was created for the European Language Resources Coordination Action (ELRC) (http://lr-coordination.eu) by Tufis Dan, Institutul de Cercuri pentru Intelectual Artificila "Mihai Draganscu", Academia Romana (www.racal.ro) with primary data copyrighted by Wikipedia (https://en.wikipedia.org) and is licensed under "CC-BY-SA 3.0" (https://creativecommons.org/licenses/by-sa/3.0/).

IPR Holders
Wikipedia, the free encyclopedia

Contact Person
Dan Tufis

Bilingual text corpus
Languages
English (en) (2,671,991 Words)
Romanian, Moldavian, Moldovan (ro) (2,729,213 Words)

Linguality
Linguality type: Bilingual
Multi-linguality type: Parallel

Text Format
Plain Text

Size
5,622,367 Words
Character encoding
UTF-8

Resource Creation
Funding Project
Connecting Europe Facility - European Language Resource Coordination (CEF-ELRC - LANGUAGE RESOURCE COORDINATION - SMART 2014/1074 - 30-CF-009678/00-64)
URL: http://www.lr-coordination.eu

Funding Type: Service Contract
Funder: European Commission
Funding Country: European Union (EU)
Project duration: 29/03/2015 - 16/04/2017

Metadata
Created: 18/11/2016
Last Updated: 16/12/2016
Metadata Language: English (en)

Metadata Creator
Kenele Poulk
Dan Tufis

Version
Version: 1.0
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
ELRC onsite assistance

www.lr-coordination.eu/request-onsite-assistance
ELRC Helpdesk

www.lr-coordination.eu/helpdesk

Helpdesk for Language Resources

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

<table>
<thead>
<tr>
<th>Phone</th>
<th>+33 970 440 522</th>
</tr>
</thead>
<tbody>
<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>
Mulțumesc!
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