Preparing and sharing data with the ELRC 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

Second ELRC Workshop, Warsaw, 18.12.2017
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)
The notion of language data

Data

• any piece of electronically stored content

(Textual) Language Data

• any piece of electronically stored text
The notion of data in the context of eTranslation

Central Statistical Office Dataset

Two Polish-English publications of the Polish Central Statistical Office available in the XLIFF format:
1. "Statistical Yearbook of the Republic of Poland 2015" is the main summary publication of the Central Statistical Office, including a comprehensive set of statistical data describing the condition of the... Read More

Appropriateness for DSI: Open Data Portal

Distribution
- Availability: Available
- Licence: CC-BY 4.0
- Attribution Details: Central Statistical Office Dataset was created for the European Language Resources Coordination Action (ELRC) (http://elrc-coordination.eu) by Oggrodniczek Maciej, Institute of Computer Science, Polish Academy of Sciences, with primary data copyrighted by the Central Statistical Office of Poland (http://stat.gov.pl/en) and its licensed under "CC-BY 4.0" (https://creativecommons.org/licenses/by/4.0).
- Allows Uses Besides DGT ✓
- Contact Person: Maciej Oggrodniczek

Bilingual text corpus
- Languages: Polish (pl), English (en)
- Liguality: Bilingual
- Text Format: XML
- Size: 1,532 Translation Units
- Character encoding: UTF-8
- Domains:
  - SOCIAL QUESTIONS (Demography And Population)
  - ENVIRONMENT (Natural Environment)
  - SOCIAL QUESTIONS (Social Framework)

Metadata
- Created: Sept. 18, 2015
- Last Updated: Dec. 15, 2016
- Metadata Language: English (en)
- Metadata Creator: Kanella Pouli, Maciej Oggrodniczek
The notion of data in the context of eTranslation

File01_pl.txt
File01_en.txt
File02_pl.txt
File02_en.txt
File03_pl.txt
File03_en.txt
...

Zapewnienie równości szans dla kobiet i mężczyzn oraz pełnoprawnego uczestnictwa w życiu społeczeństwa jest jednym z podstawowych praw człowieka.

Ensuring equality of chances for men and women, as well as full participation in the social life is one of the basic human rights.
Data used by eTranslation

DGT translation memory

European Commission
CVRIA
European Parliament
European Investment Bank
European Central Bank
EUROSYSTEM
Council of the European Union
CURIA RATIONUM
European Committee of the Regions

Second ELRC Workshop, Warsaw, 18.12.2017
Such data are already available
BUT
they are not enough...
What does eTranslation need?

• Language data residing in local public organisations, produced in-house or outsourced, e.g.
  – Reports
  – Communication
  – News
  – Web Content that is managed for several languages
  – Policies
  – Terminologies
  – Archives
  – Forms
  – FAQs
• In principle, any **electronically stored text** in any of the EU languages, plus Norwegian and Icelandic (i.e. the CEF languages)

• Ideally, **texts and their translations** in one or more of the CEF languages (i.e. parallel bilingual or multilingual)

---

**Polish text**

Sprawozdanie Prezesa Urzędu Zamówień Publicznych o funkcjonowaniu systemu zamówień publicznych w 2010 roku.

Sprawozdanie z funkcjonowania systemu zamówień publicznych obejmuje okres od 1 stycznia do 31 grudnia 2011 r. i zostało opracowane na podstawie informacji zawartych w oficjalnych dokumentach i publikacjach urzędowych jak również innych dokumentach, raportach i analizach z zakresu zamówień publicznych, będących w dyspozycji Urzędu Zamówień Publicznych.

---

**Translation in English**

Report of the President of Public Procurement Office on functioning of public procurement system in 2010.

The report on the functioning of the public procurement system covers the period from 1 January to 31 December 2011 and it has been prepared on the basis of information obtained from official documents and publications as well as other documents, reports and analyses regarding public procurement which were available to the Public Procurement Office.
Can also be a list of terms and their translations, i.e. a terminology

<table>
<thead>
<tr>
<th>English</th>
<th>Polish</th>
</tr>
</thead>
<tbody>
<tr>
<td>aircraft commander</td>
<td>dowódca statku powietrznego</td>
</tr>
<tr>
<td>aircraft configuration</td>
<td>konfiguracja statku powietrznego</td>
</tr>
<tr>
<td>aircraft flight manual</td>
<td>instrukcja użytkowania w locie statku powietrznego</td>
</tr>
<tr>
<td>aircraft hangar</td>
<td>hangar dla statków powietrznych</td>
</tr>
<tr>
<td>airframe</td>
<td>płatowiec</td>
</tr>
<tr>
<td>airworthiness certificate</td>
<td>świadectwo zdatności do lotu</td>
</tr>
<tr>
<td>automatic flight control system</td>
<td>automatyczny układ sterowania lotem</td>
</tr>
<tr>
<td>European Technical Standard Order</td>
<td>Europejska Norma Techniczna</td>
</tr>
<tr>
<td>...</td>
<td></td>
</tr>
</tbody>
</table>
What data are useful for eTranslation as per format

- 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
United Nations Convention against Corruption

Preamble

The States Parties to this Convention,

Concerned about the seriousness of problems and threats posed by corruption to the stability and security of societies, undermining the institutions and values of democracy, ethical values and justice and jeopardizing sustainable development and the rule of law,

Concerned also about the links between corruption and other forms of crime, in particular organized crime and economic crime, including money-laundering,

Concerned further about cases of corruption that involve vast quantities of assets, which may constitute a substantial proportion of the resources of States, and that threaten the political stability and sustainable development of those States,

Convinced that corruption is no longer a local matter but a transnational phenomenon that affects all societies and economies, making international cooperation to prevent and control it essential,

Convinced also that a comprehensive and multidisciplinary approach is required to prevent and combat corruption effectively,
The following formats are particularly useful (in descending order):

- **For 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

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.

To jest polskie tłumaczenie poprzedniego akapitu. To jest polskie tłumaczenie poprzedniego akapitu. To jest polskie tłumaczenie poprzedniego akapitu. To jest polskie tłumaczenie poprzedniego akapitu. To jest polskie tłumaczenie poprzedniego akapitu. To jest polskie tłumaczenie poprzedniego akapitu.

A second paragraph in English. A second paragraph in English. A second paragraph in English. A second paragraph in English.

To jest polskie tłumaczenie drugiego akapitu. To jest polskie tłumaczenie drugiego akapitu. To jest polskie tłumaczenie drugiego akapitu. To jest polskie tłumaczenie drugiego akapitu.
Preparing your data | 2

**Don’ts**

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

A second paragraph in English. A second paragraph in Polish.

---

This is a paragraph in English. This is a paragraph in Polish. To jest polskie tłumaczenie poprzedniego akapitu. To jest polskie tłumaczenie drugiego akapitu.
<table>
<thead>
<tr>
<th>English</th>
<th>Polskie</th>
</tr>
</thead>
<tbody>
<tr>
<td>This is a paragraph in English. This is a paragraph in English.</td>
<td>To jest polskie tłumaczenie poprzedniego akapitu.</td>
</tr>
<tr>
<td>This is a paragraph in English. This is a paragraph in English.</td>
<td>To jest polskie tłumaczenie poprzedniego akapitu.</td>
</tr>
<tr>
<td>This is a paragraph in English. This is a paragraph in English.</td>
<td>To jest polskie tłumaczenie poprzedniego akapitu.</td>
</tr>
<tr>
<td>This is a paragraph in English. This is a paragraph in English.</td>
<td>To jest polskie tłumaczenie poprzedniego akapitu.</td>
</tr>
<tr>
<td>A second paragraph in English. A second paragraph in English.</td>
<td>To jest polskie tłumaczenie drugiego akapitu.</td>
</tr>
<tr>
<td>A second paragraph in English. A second paragraph in English.</td>
<td>To jest polskie tłumaczenie drugiego akapitu.</td>
</tr>
<tr>
<td>A second paragraph in English. A second paragraph in English.</td>
<td>To jest polskie tłumaczenie drugiego akapitu.</td>
</tr>
</tbody>
</table>
Do's

Use **identical filenames** for each document pair (source – translation)
Include **language identifiers** in the filename.

- filename01_EN.txt
- filename01_PL.txt
- filename02_EN.txt
- filename02_PL.txt
- filename03_EN.txt
- filename03_PL.txt
- filename04_EN.txt
- filename04_PL.txt
Criteria for grouping your data

• Remember: 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 – German*
    • *Documents in English – Norwegian - Finnish*
  
  – **Domain**: each collection ideally belongs to a single domain, e.g.
    
    • *Collection of texts in English – German in the culture domain*
    • *Social security documents in English – Norwegian - Finnish*
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</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

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
- Register
- Activate account

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

Welcome to the ELRC-SHARE repository!
How to Register (2/2)

- 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)

• Two modes for contributing your data

Contribution Mode
- Upload ZIP archive
- Provide URL of resources

Upload Resource
- Choose File
  - No file chosen

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

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 Contribute Data (4/6)

1. Click on Choose file
2. Locate your resource in your hard disk
3. Click on Submit
• Alternatively indicate a url (directory listing)
• 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
• All datasets are processed to result in tmx/tbx/txt files
• Data will indicatively undergo the following processing:
  – cleaning
  – format conversion
  – sentence alignment
  – metadata completion
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
Assistance will be provided in close cooperation with a broad network of language experts
On-site assistance

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!
Language processing services on-site

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
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 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
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]
Helpdesk for Language Resources

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:

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

lr-coordination.eu/helpdesk
ELRC consortium – come talk to us!
Dziękuję za uwagę!