EUROPEAN LANGUAGE RESOURCE COORDINATION (ELRC) AND BEYOND

ANDREA LÖSCH (DFKI), STEFANIA RACIOPPA (DFKI), EILEEN MARRA (DFKI), THIERRY DECLERCK (DFKI)

1

ELRC CONTEXT AND OBJECTIVES

European Language Resource Coordination
Connecting Europe Facility

European Commission

WHO WE ARE

THE ELRC CONSORTIUM

DFKI · CROSSLANG TRANSLATION AUTOMATION · el Dà · IEA LSP · TILDE

STARTED
in April 2015

CURRENT CONTRACT
until end of 2022

GOAL
To provide language resources to improve coverage and performance of automated translation solutions in the context of current and future CEF digital services

3

European Language Resource Coordination
Connecting Europe Facility

European Commission

ELRC OBJECTIVES

Language data are key for machine translation / language technology!

the
publi
SME
dom

language data

and SMEs

related to
language data
sharing

tory for
language data in
Europe

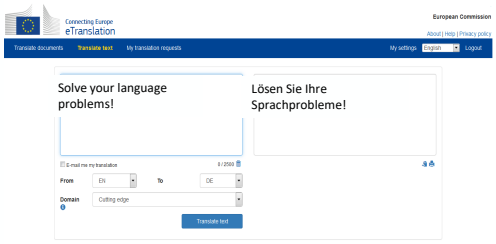
4

European Language Resource Coordination
Connecting Europe Facility

European Commission

ELRC: Language Data for eTranslation

- CEF eTranslation helps public services being multilingual:
  - Neural machine translation tool provided by the European Commission to all EU bodies but also public services and public administrations across Europe
  - for 24 official languages of the EU (24 <-> 24), Russian, Chinese, Turkish, …
  - Secure & scalable
  - Facts and Figures:
    - 250+ applications connected to eTranslation
    - More than 18.000 registered users end of 2021
    - More than 240 Mio. pages translated
- Goal: ELRC provides language resources to make eTranslation work for different languages and domains!



5

European Language Resource Coordination
Connecting Europe Facility

European Commission

ELRC: Language Data for Inclusion

- Additional dimension of ELRC: To provide language resources to overcome language barriers in public services and the European Digital Single Market in general
- Example: Digital Service Infrastructures (DSIs)

BRIS
Citizens and business partners need legal certainty when doing business cross-border

ODR
Citizens need to solve disputes online across borders

eHealth
Citizens need to have access to medical information and their patient data when abroad

EESSI
Citizens enjoy seamless exchange of social security information when abroad

It also includes
e-Justice
Open Data
eProcurement

6

European Language Resource Coordination
Connecting Europe Facility

European Commission

LANGUAGE RESOURCES (LR) COLLECTED WITHIN ELRC

ELRC collects the following types of language data:

Corpora

Set of documents or text in one or more languages

- official documents / translations in the official administration (decisions, legal acts etc.);
- articles, reports, magazines, newsletters, etc.;
- Translation Memories (i.e. aligned text segments in the source and target language)

Models

Language or translation models

Resources

Lexical & Conceptual Resources

- Terminologies
- Glossaries
- Thesauri
- Wordnets
- Ontologies

7

European Language Resource Coordination
Connecting Europe Facility

European Commission

ELRC NETWORK & OUTREACH ACTIVITIES

European Language Resource Coordination
Connecting Europe Facility

European Commission

ELRC Network

National Anchor Points (NAP) in each country

www.lr-coordination.eu/anchor-points

9

European Language Resource Coordination
Connecting Europe Facility

European Commission

ELRC Progress and Growth

PHASE 1:
- Initial Roadshow (Country Workshops)
- Set-up of ELRC Network (T-NAP plus P-NAPs)
- Set-up of data collection & validation workflow
- Technical and Legal Helpdesk

PHASE 2:
- Sustainability, local ownership
- Extension of ELRC Network on national level (600 data donors)
- Extended data collection (crawling) & validation workflow (Generic Services Projects)
- Addition: On-site assistance

PHASE 3:
- Sustainability, re-usability, opening (SME)
- Extension of ELRC Network (> 6.000 data donors and users)
- Extended data collection (speech, tools and services, …)
- New languages (RU, CN, …)

10

European Language Resource Coordination
Connecting Europe Facility

European Commission

Traditional ELRC Outreach and Networking Actions

European Language Resource Coordination
Connecting Europe Facility

European Commission

Since COVID-19 Pandemic: ELRC on Social Media

European Language Resource Coordination
Connecting Europe Facility

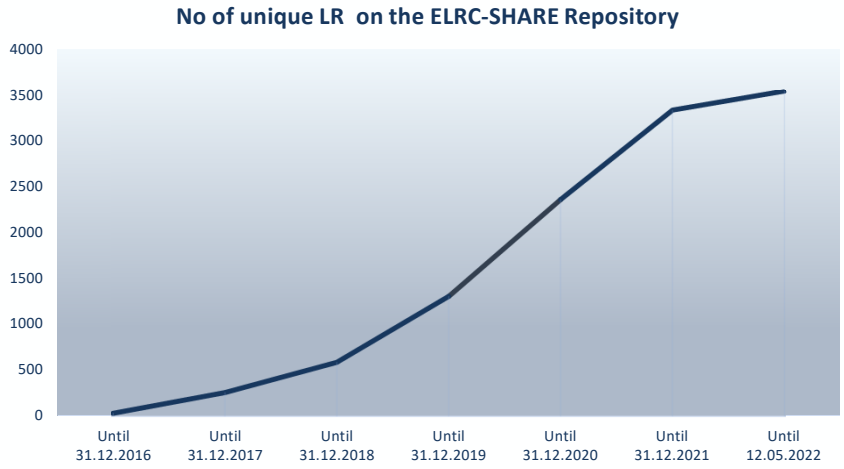
European Commission

ELRC: ACHIEVEMENTS

European Language Resource Coordination
Connecting Europe Facility

European Commission

ELRC ACHIEVEMENTS: COLLECTION OF RESOURCES

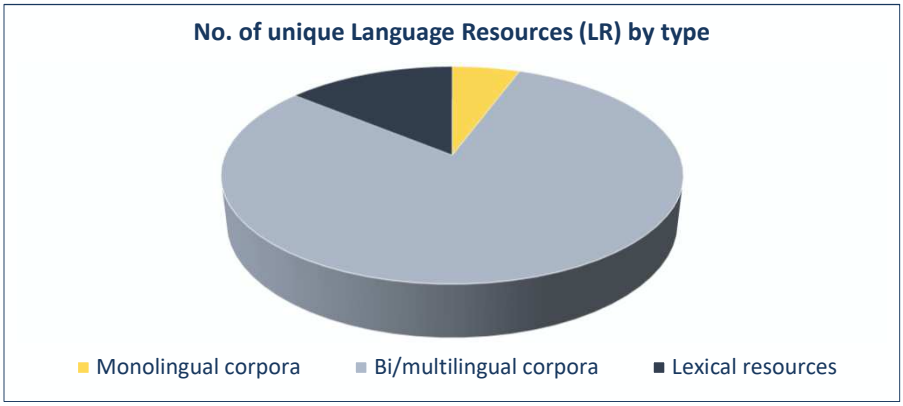


14

European Language Resource Coordination
Connecting Europe Facility

European Commission

ELRC Achievements: Number of Resources

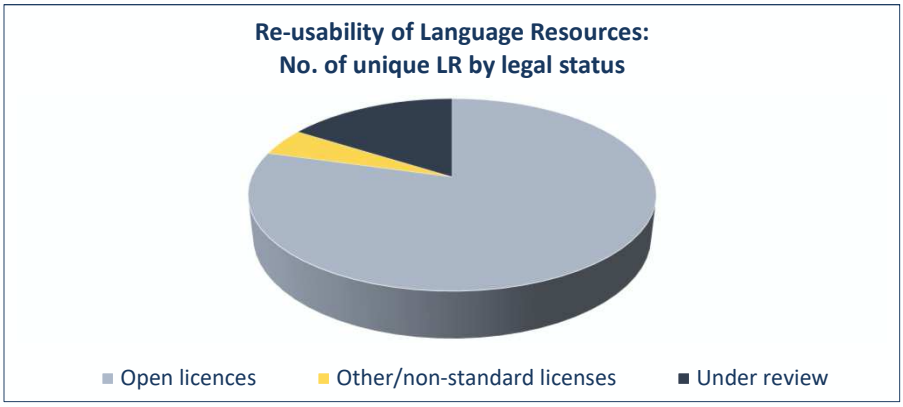


15

European Language Resource Coordination
Connecting Europe Facility

European Commission

ELRC Achievements: Reusability of Resources

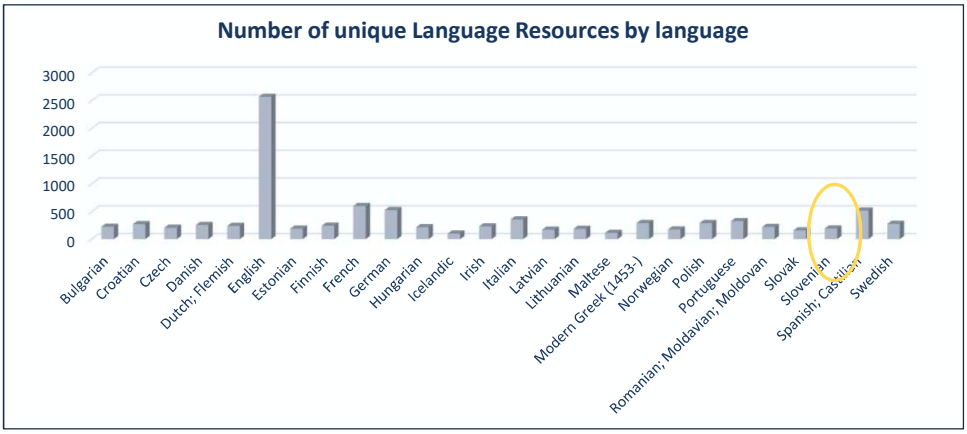


16

European Language Resource Coordination
Connecting Europe Facility

European Commission

ELRC Achievements: Resources by Language

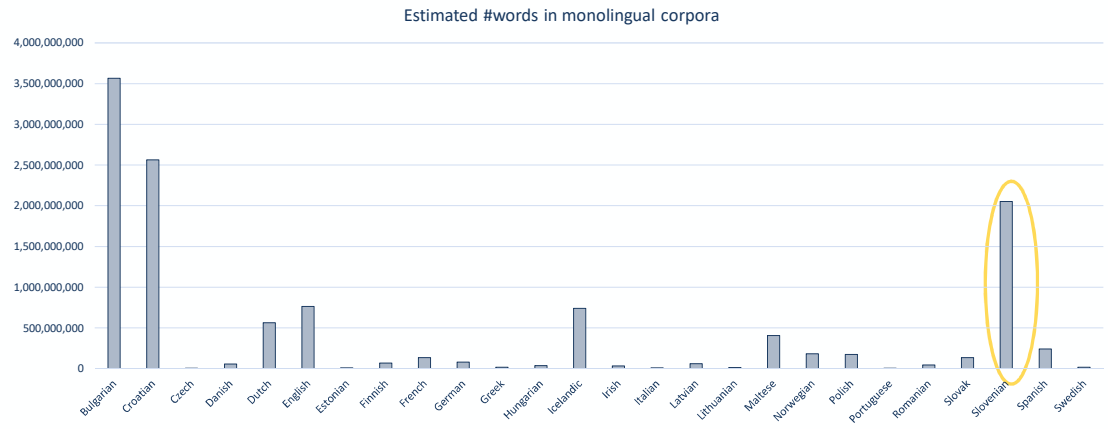


17

European Language Resource Coordination
Connecting Europe Facility

European Commission

ELRC Achievements: Monolingual Corpora



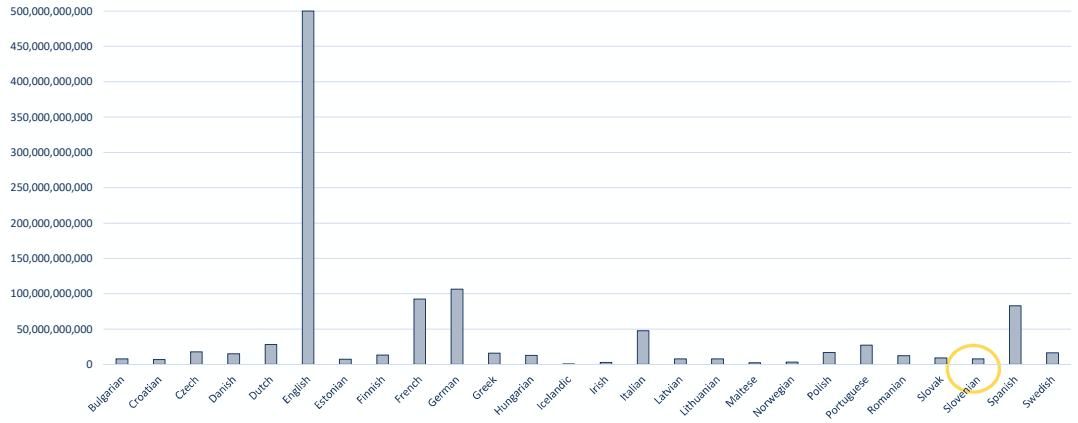
18

European Language Resource Coordination
Connecting Europe Facility

European Commission

ELRC Achievements: Bilingual Corpora

Estimated #words in bilingual corpora



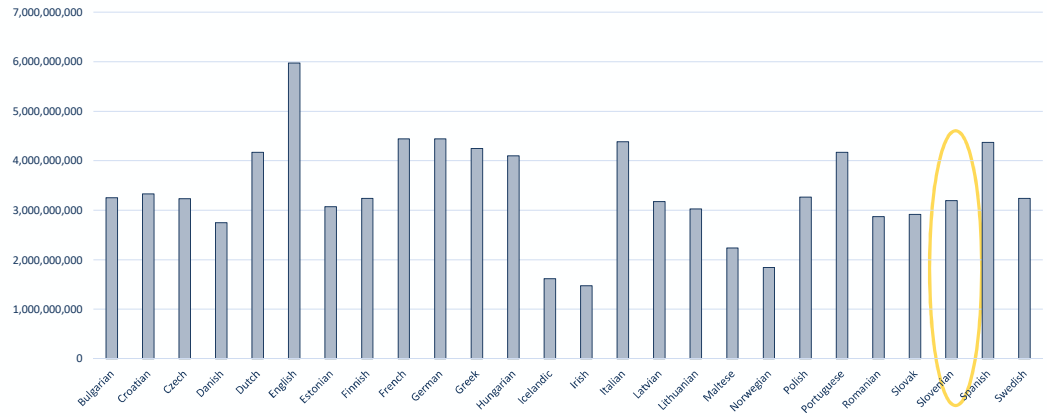
19

European Language Resource Coordination
Connecting Europe Facility

European Commission

ELRC Achievements: Multilingual Corpora

Estimated #words in multilingual corpora

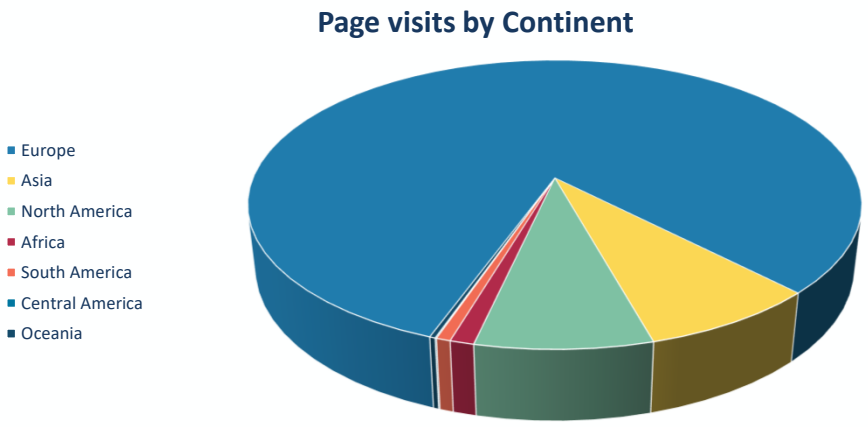


20

European Language Resource Coordination
Connecting Europe Facility

European Commission

ELRC Achievements: Page visits worldwide

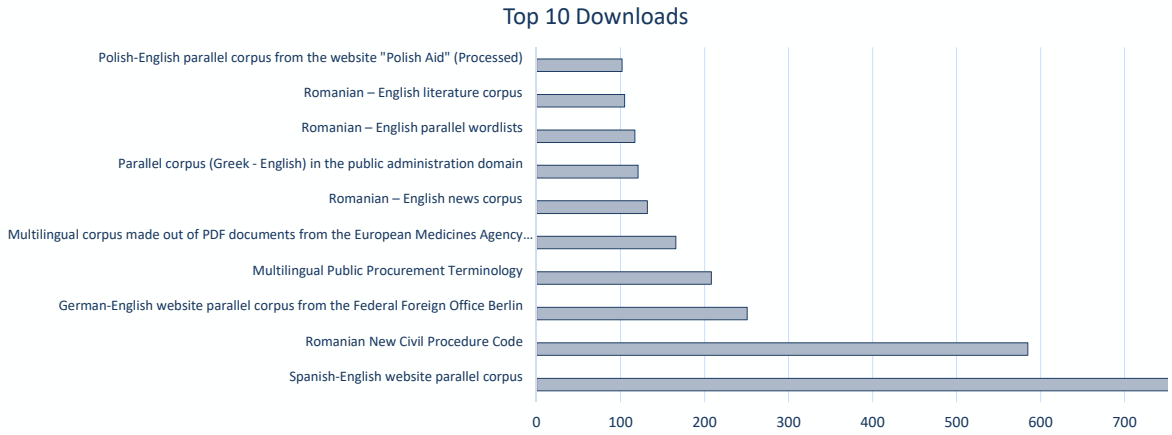


21

European Language Resource Coordination
Connecting Europe Facility

European Commission

ELRC Achievements: Top 10 Resources

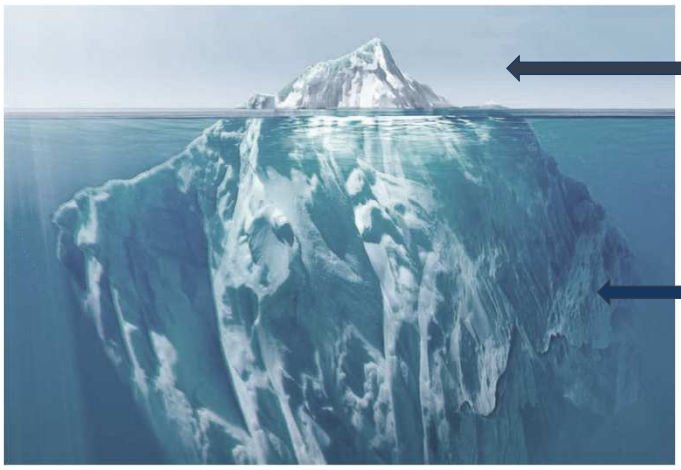


22

European Language Resource Coordination
Connecting Europe Facility

European Commission

STATUS QUO: THE TIP OF THE ICEBERG



What we have…

… and what there is!

23

European Language Resource Coordination
Connecting Europe Facility

European Commission

DONATING LANGUAGE DATA TO ELRC

ELRC-SHARE

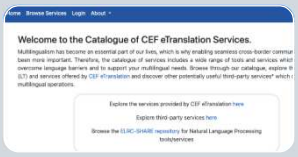
ELRC Language Resource Repository

www.elrc-share.eu

ELRC Helpdesk

for any questions on technical or legal aspects related to the use, production, collection, processing and sharing of LR.

www.lr-coordination.eu /helpdesk



Catalogue of CEF eTranslation Services

676 language tools/services offered by CEF eTranslation and third-party services

https://cef-at-service-catalogue.eu

24

European Language Resource Coordination
Connecting Europe Facility

European Commission

ELRC: CHANCES, CHALLENGES AND OUTLOOK

European Language Resource Coordination
Connecting Europe Facility

European Commission

THE ELRC WHITE PAPER

- Stakeholders
- Infrastructure for LR sharing
- Challenges for LR sharing
- Recommendations & Actions
- Annex: Country Profiles

- Please visit: www.lr-coordination.eu
  ➔ Helpdesk ➔ Infopoint

European Language Resource Coordination
Connecting Europe Facility

European Commission

MAJOR CHANCES

- Great number of potential data holders!
- More than 4.000 individuals from several hundred organisations
- But again: This is only the tip of the iceberg!

European Language Resource Coordination
Connecting Europe Facility

European Commission

MAJOR CHALLENGES

- Lack of appreciation of the value of language data
- Structural chal
- language data management practices
- Limited access to outsources translations
- Legal concerns

Updated ELRC White Paper in October 2022!

Scan for more info!

European Language Resource Coordination
Connecting Europe Facility

European Commission

DEVELOPMENTS AND OUTLOOK

Extending Country Profiles and White Paper through Country Workshops	• 26 ELRC Country Workshops conducted under the current contract since end of 2020, another 3 are in the planning for 2022 • Dedicated session on "Language Data creation, management and sharing" • Define ways of overcoming obstacles in each country and across Europe
Increasing language coverage	• Initially only EU official languages plus Norwegian and Icelandic (as CEF affiliated countries) • Recently also Russian (end of 2019), Chinese (mid 2020), and other languages added based on increasing demand for new languages
Opening of eTranslation for SMEs	• eTranslation is available to SMEs as well since March 2020

29

European Language Resource Coordination
Connecting Europe Facility

European Commission

DEVELOPMENTS AND OUTLOOK

Extending collection of LR	• Relevant language tools for all EU official languages, Icelandic and Norwegian are available in the ELRC-SHARE (https://elrc-share.eu/) • Speech and multimodal data are being considered too for inclusion
Adjusting outreach activities to COVID-19 pandemic limitations	• ELRC Social Media Campaign since summer 2020 with help of the LRB
Defining and developing Language Technology Specifications	• The purpose is to support the creation of technical specifications or best technical practices for the language technology applications in CEF eTranslation. • Anonymisation API specification ready for implementation

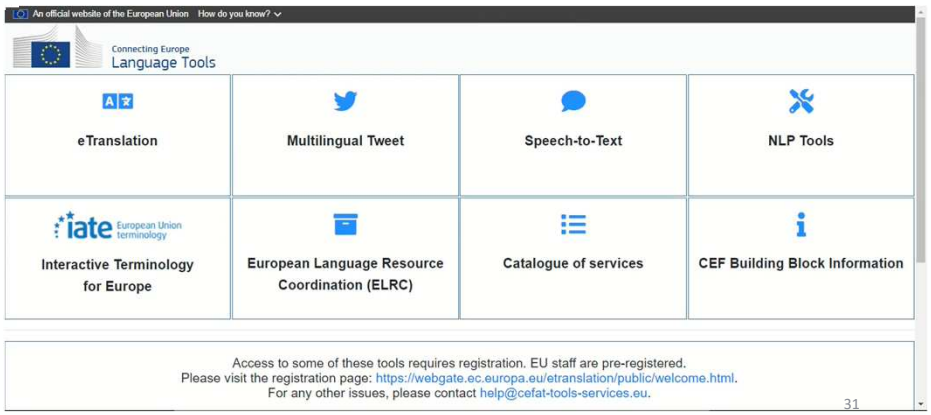
30

European Language Resource Coordination
Connecting Europe Facility

European Commission

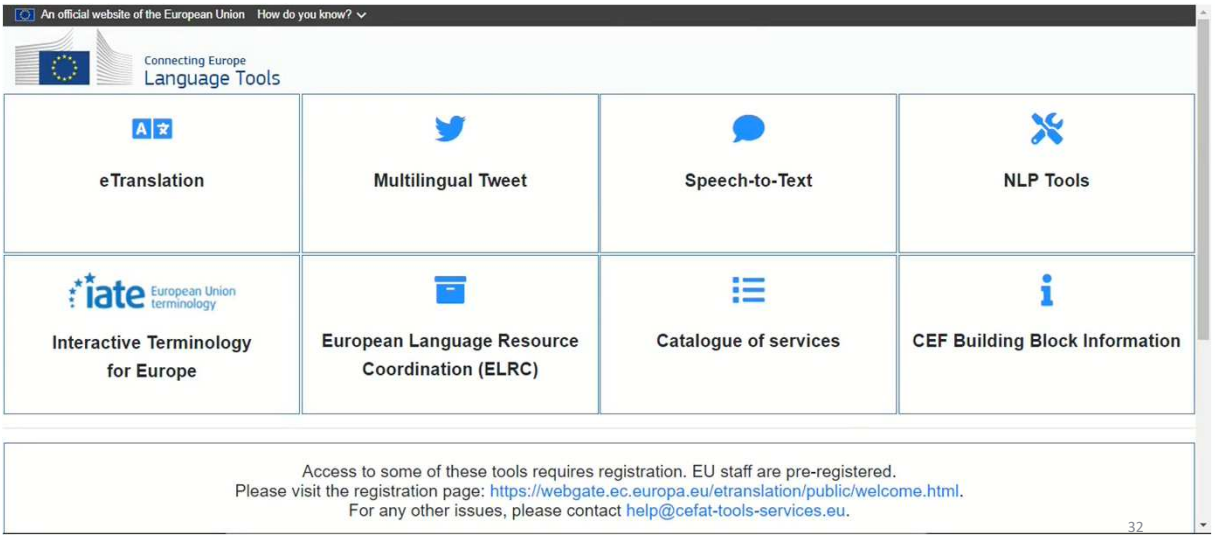
SPEECH-TO-TEXT TRANSCRIPTION

https://language-tools.ec.europa.eu/SpeechServices/Transcription



European Language Resource Coordination
Connecting Europe Facility

European Commission



European Language Resource Coordination
Connecting Europe Facility

European Commission

CATALOGUE OF SERVICES

https://cef-at-service-catalogue.eu/

Choose from 700 Language Tools and Services available in and for Europe!

33

European Language Resource Coordination
Connecting Europe Facility

European Commission

STAY TUNED!

Follow us on Social Media

EuropeanLanguage ResourceCoordination

LRCoordination

LR_Coordination

34

THANK YOU FOR YOUR ATTENTION!

Website: www.lr-coordination.eu
Twitter: @LR_Coordination
Email: info@lr-coordination.eu

DATA PROCESSING & VALIDATION WITHIN ELRC

Back-up slides

European Language Resource Coordination
Connecting Europe Facility

European Commission

DATA PROCESSING WITHIN ELRC

- Each LR is analysed and processed by ELRC experts to ensure compliance with the **Language Resources Data Formats Specification** agreed with the EC.
- According to this specification, data should be provided in the following form:
  - **Parallel data** should be provided in the TMX format in UTF-8 encoding, without optional data fields (e.g. translator id, adjacent segments) without non-printable control characters.
  - **Monolingual corpora** are to be delivered in plain text format without any additional annotation, in UTF-8 encoding, single file by language and resource, segmented into paragraphs.
  - **Terminology resources** should be provided in the TBX format.
- **Explanatory notes:**
  - **TMX** stands for „Translation Memory eXchange". TMX is an XML specification for the exchange of translation memories.
  - **TBX** stands for "TermBase eXchange" TBX is an XML-based format for the representation and exchange of terminology data

37

European Language Resource Coordination
Connecting Europe Facility

European Commission

DATA VALIDATION WITHIN ELRC

- **Validation** = quality control of a language resource against a list of relevant criteria
- Because of the different processes of gathering the data and their varying quality level, validation may be conducted in **two different ways**:
  - **Quick Content Check (QCC)**: for high quality data (e.g. human translations)
  - **Extended Content Validation**: e.g. for data derived from automatic processing
- Note: Both ways include first the **technical validation** of the resource and then the **legal validation** (IPR Clearance)

38

European Language Resource Coordination
Connecting Europe Facility

European Commission

DATA VALIDATION WITHIN ELRC: QCC

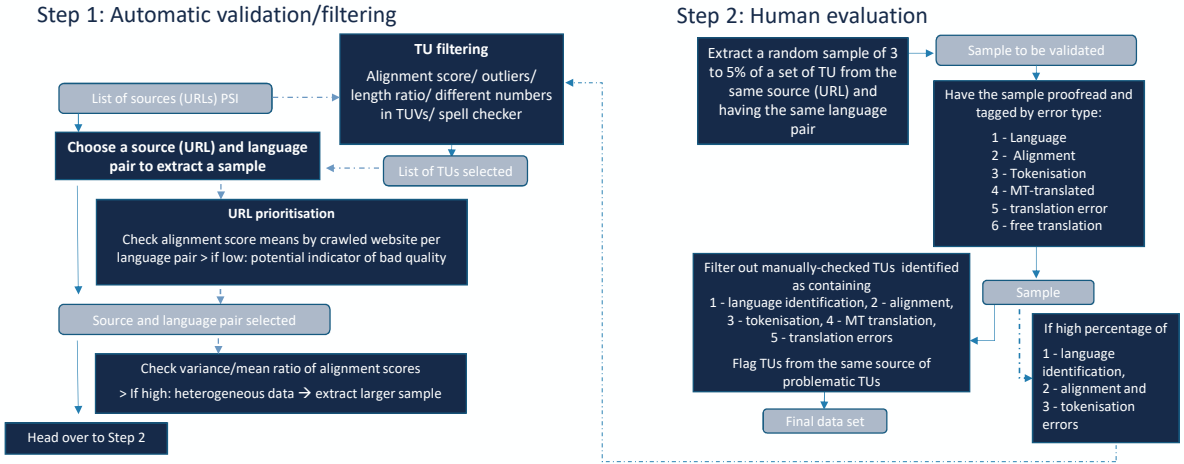
- **Quick Content Check (QCC)**: for high quality data (e.g. human translations)
  - check compliance of data with the ELRC objectives and **scope**,
  - check the **format** of provided data,
  - check that the **metadata** fields have been correctly filled in and are compliant with the data **content**, and
  - check whether the legal information provided is compliant with the ELRC scope (➔ **legal validation**).

39

European Language Resource Coordination
Connecting Europe Facility

European Commission

DATA VALIDATION WITHIN ELRC: EXTENDED VALIDATION



40

Fig. 2. ELRC Workflow for fine-grained content validation

European Language Resource Coordination
Connecting Europe Facility

European Commission

DATA VALIDATION WITHIN ELRC: LEGAL VALIDATION

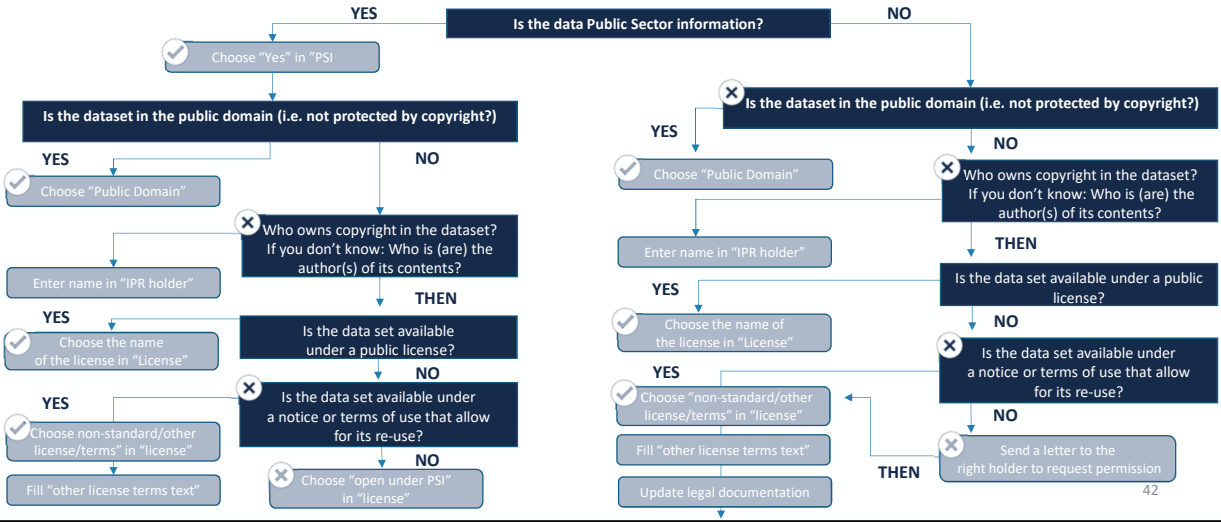
- **Objective:** To determine appropriate license
- **Key questions** to be addressed and assessed:
  - Does the data fall within the scope of the **Public Sector Information Directive (PSI)**? ➔ Public Sector Information Directive 2003/98/EC (modified in 2013 by the Directive 2013/37/UE
  - Is the data protected by **copyright**? ➔ National laws may contain rules excluding certain works from copyright protection
  - If the data is protected by copyright, can I identify the **owner of the copyright** or the **author of the work**? ➔ to get permission
  - Is the data available under a **public license**? ➔ For example, certain datasets are made available by the owner of copyright under a license that allows reuse or redistribution free of charge (e.g., cc licenses, NCGL 1.0, OGL 3.0 etc.)
  - If no public license is clearly marked on the document ➔ check the **terms of use** or if any **documentation** may help you determine the conditions of reuse of the material

41

European Language Resource Coordination
Connecting Europe Facility

European Commission

DATA VALIDATION WITHIN ELRC: LEGAL VALIDATION



42