

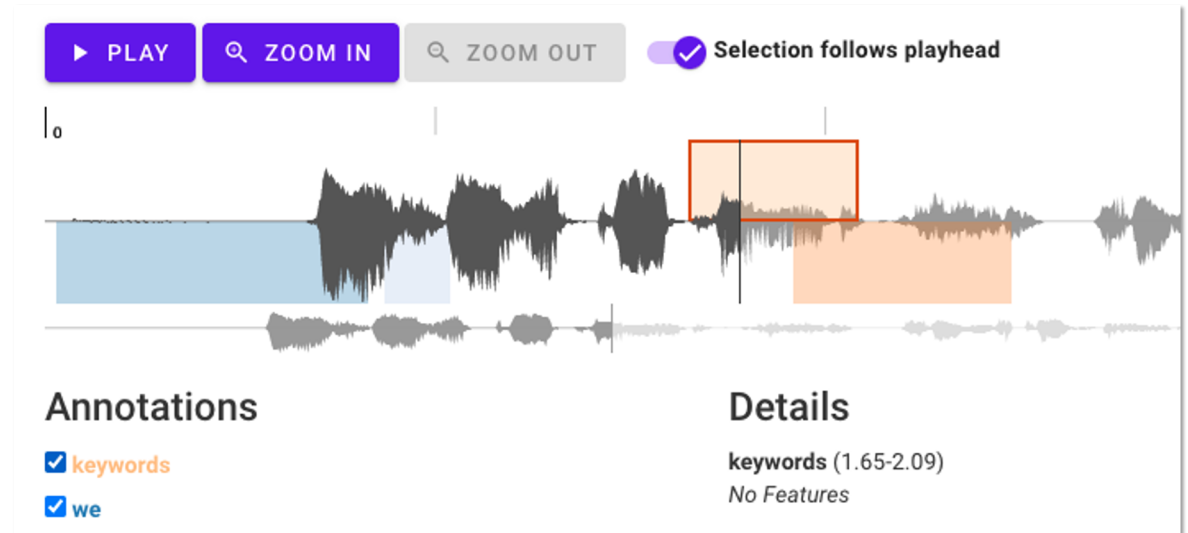
European Language Grid – META-FORUM 2022

Kalina Bontcheva – European Language Grid: LT Services and Resources

- ELG Release 3

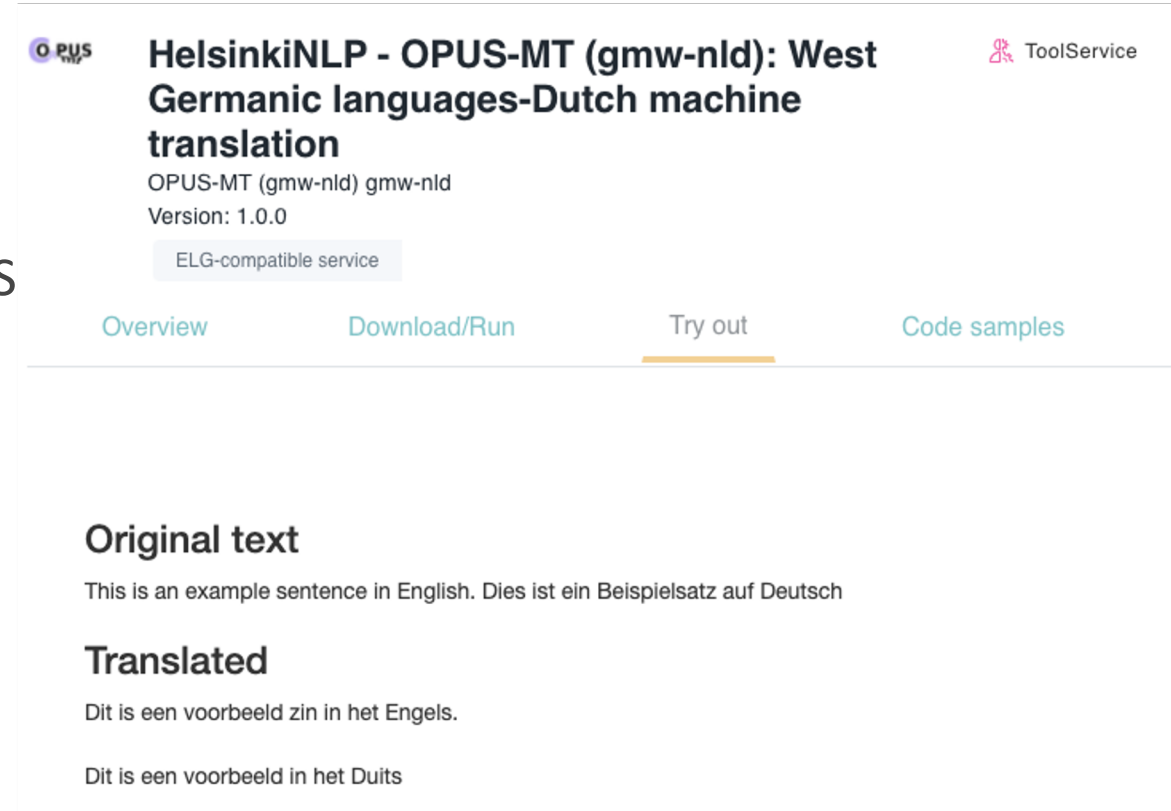
LT Services on the European Language Grid

- ELG Release 3 (June 2022) has close to 2,000 ELG-compatible services (1959)
 - Majority are language dependent
 - 220 are language independent
- APIs for various types of services, including ASR, IE, MT, TTS, OCR
- All 24 official EU languages covered and many more (24 other EU; 69 other languages)
- 1349 services for the official EU languages
 - 765 MT; 472 TA; rest - other services
- 198 services for other languages in EU
 - 98 MT; 88 TA; rest - other services
- 412 services for other languages
 - 228 TA; 86 MT; rest - other services



LT Services: From the Pilots

- Many services have been contributed by the ELG-funded pilot projects
- Services include:
 - Over 100 MT language pairs from OPUS-MT
 - Clinical NER in 5 languages from E3C
 - Terminology extraction service from Text2TCS
 - Italian classification services from EVALITA
 - Multilingual WSD, SRL and AMR annotation services (around 100 languages)
 - Sign language explanations



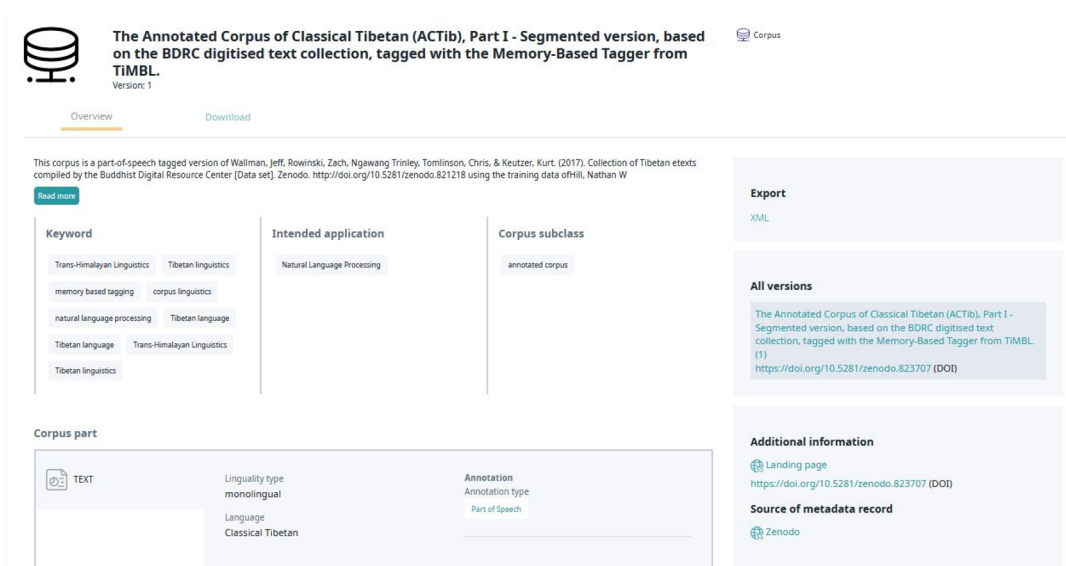
The screenshot shows the interface for the HelsinkiNLP - OPUS-MT (gmw-nld) tool. The title is "HelsinkiNLP - OPUS-MT (gmw-nld): West Germanic languages-Dutch machine translation". Below the title, it says "OPUS-MT (gmw-nld) gmw-nld" and "Version: 1.0.0". There is a button that says "ELG-compatible service". The interface has four tabs: "Overview", "Download/Run", "Try out", and "Code samples". The "Try out" tab is currently selected. Under the "Try out" tab, there are two sections: "Original text" and "Translated". The "Original text" section shows an example sentence in English: "This is an example sentence in English." and its translation in German: "Dies ist ein Beispielsatz auf Deutsch". The "Translated" section shows the same sentence translated into Dutch: "Dit is een voorbeeld zin in het Engels." and "Dit is een voorbeeld in het Duits".

LT Services: from elsewhere

- Attend Session 3 to see details of selected LT services provided by others:
 - Machine Translation:
 - NTEU – Pangeanic
 - OPUS-MT – University of Helsinki
 - Text-to-speech:
 - Elhuyar Basque ASR/TTS – Elhuyar
 - Automatic speech recognition:
 - Welsh ASR Service – Bangor University
 - Microservices at your Service – Lingsoft
 - A collection of information extraction services, basic linguistic pre-processing, MT

Language Resources: metadata description and harvesting

- ELG has imported metadata from many other repositories:
 - ELRA Catalogue, ELRC-SHARE, ELRA-SHARE-LRs, LINDAT/CLARIAH-CZ, CLARIN Poland repository, CLARIN Slovenia repository, META-SHARE-DFKI, META-SHARE-ELDA and META-SHARE-ILSP, Hugging Face, Quantum Stat, Zenodo
 - Some metadata can be imported automatically, some need manual correction



The Annotated Corpus of Classical Tibetan (ACTib), Part I - Segmented version, based on the BDRC digitised text collection, tagged with the Memory-Based Tagger from TiMBL.
Version: 1

Overview Download

This corpus is a part-of-speech tagged version of Wallman, Jeff, Rowinski, Zach, Ngawang Trinley, Tomlinson, Chris, & Keutzer, Kurt. (2017). Collection of Tibetan etexts compiled by the Buddhist Digital Resource Center [Data set]. Zenodo. <http://doi.org/10.5281/zenodo.821218> using the training data of Hill, Nathan W

Keyword

- Trans-Himalayan Linguistics
- Tibetan linguistics
- memory based tagging
- corpus linguistics
- natural language processing
- Tibetan language
- Tibetan language
- Trans-Himalayan Linguistics
- Tibetan linguistics

Intended application

- Natural Language Processing

Corpus subclass

- annotated corpus

Export

- XML

All versions

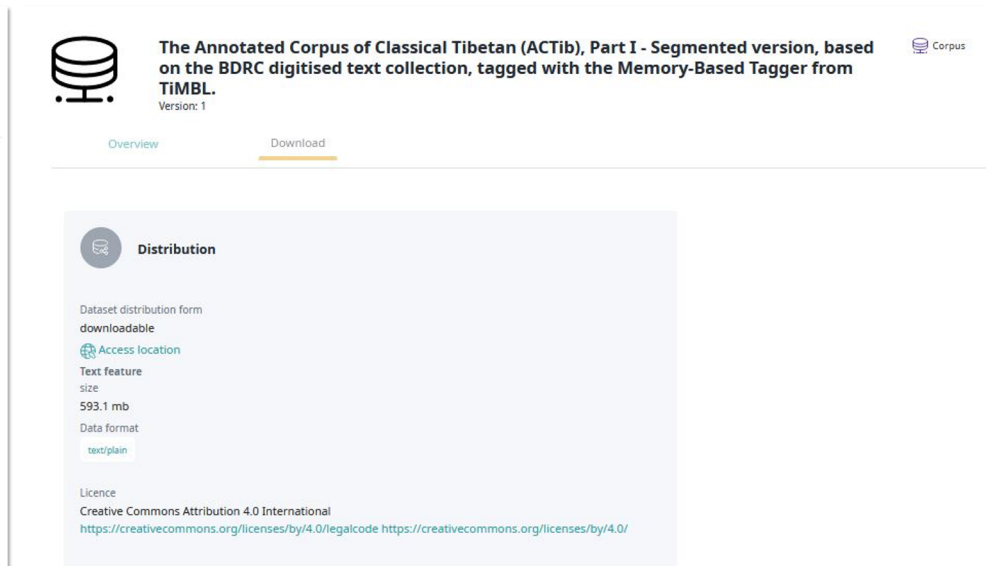
- The Annotated Corpus of Classical Tibetan (ACTib), Part I - Segmented version, based on the BDRC digitised text collection, tagged with the Memory-Based Tagger from TiMBL (1) <https://doi.org/10.5281/zenodo.823707> (DOI)

Additional information

- Landing page <https://doi.org/10.5281/zenodo.823707> (DOI)
- Source of metadata record [Zenodo](https://zenodo.org)

Corpus part

TEXT	Linguality type monolingual	Annotation Annotation type Part of Speech
	Language Classical Tibetan	



The Annotated Corpus of Classical Tibetan (ACTib), Part I - Segmented version, based on the BDRC digitised text collection, tagged with the Memory-Based Tagger from TiMBL.
Version: 1

Overview Download

Distribution

Dataset distribution form
downloadable

Access location

Text feature

- size
593.1 mb
- Data format
text/plain

Licence


Creative Commons Attribution 4.0 International
<https://creativecommons.org/licenses/by/4.0/> <https://creativecommons.org/licenses/by/4.0/>

Statistics on ingested LRs that are publicly visible (June 2022)

Repository	Corpora	Lexical/Conceptual Resources	Models & Computational grammars	Total
ELRA	635	545	–	1180
ELRC-SHARE	1249	50	–	1299
META-SHARE	52	12	7	71
ELRA-SHARE-LRs	105	37	2	144
LINDAT/CLARIAH-CZ	290	89	–	400
CLARIN.SI	140	78	–	218
CLARIN.PL	225	26	31	282
Quantum Stat	255	6	–	261
Zenodo	342	134	37	513
Tudatalib	1	–	–	1
HuggingFace	385	–	–	385
ELG / ELE	2580	1277	367	4224
TOTAL	6259	2254	444	8978

Language Resources: hosted datasets


- Data resources can also be hosted by ELG for direct download, e.g., from pilot projects



European Clinical Case Corpus

E3C-Corpus
Version: 2.0.0 (09/08/2021)

Overview **Download** Related LRTs



Distribution


Dataset distribution form downloadable


Text feature size
151384 token

Data format
[XMI](#)

Character encoding
[UTF-8](#)

Licence
Creative Commons Attribution Non Commercial 4.0 International
<https://creativecommons.org/licenses/by-nc/4.0/legalcode>
<https://creativecommons.org/licenses/by-nc/4.0/>

Download 



Turku Paraphrase Corpus

TurkuParaC
Version: 1.0.0 (automatically assigned) (30/06/2021)

Overview **Download** Corpus

The Turku Paraphrase corpus consists of over 100,000+ manually selected Finnish paraphrases, most of which are in their document context. The paraphrases are manually classified using a scheme capturing the degree of contextual dependence, as well as a possible subsumption relation and other flags such as style and min

[Read more](#)

Keyword

paraphrases Finnish

Swedish

meaning representation

representation learning

Intended application

Paraphrasing

Training of language models

Conversational systems building

Information Retrieval

Natural Language Generation

Corpus subclass

annotated corpus



Export

[XML](#)


All versions

[Turku Paraphrase Corpus \(1.0.0 \(automatically assigned\)\)](#)


Resource provider

 TurkuNLP
 [Website](#)


Additional information

 [Landing page](#)

Contact




Corpus part

 TEXT Linguality type multilingual


Language Resources: other LR types

- Besides downloadable LRs, ELG also supports LRs accessed via a query interface
- Coreon pilot project integrated SPARQL endpoints queryable through ELG



Coreon SPARQL endpoint: Eurovoc combi

EuroVoc MKS SPARQL endpoint
Version: 1.0.0

 LexicalConceptualResource

Overview
Download
Try out

Eurovoc

Eurovoc is the European Union's multilingual thesaurus. It has been converted from SKOS/rdf. This data set is highly multilingual, covering 20+ languages. The data is characterised by a very structured, levelled approach. Upper levels contain classificatory numbers (/ 36 science / 3611 humanities ...). Through Coreon's multilingual concept map approach, the data is explorable in any of the available languages. For instance, switch to Spanish as source language and see immediately how the concept map is rendered in Spanish.

SPARQL query

```
SELECT * WHERE {
  ?t rdftype coreon:Term .
  ?t coreon:value ?val .
  FILTER langMatches( lang(?
val), "en" )
}
```

//

SUBMIT

Sample Queries

FETCH THE FIRST 10 TERMS

FIRST 50 ENGLISH TERMS, SORTED FROM A TO Z

Language Resources: current state

- Some large and important repositories (Zenodo, HuggingFace) are particularly complex, with sparse or variable quality metadata
 - Harvesting procedures have been implemented whenever possible to optimise ingestion (Zenodo)
- Critical issues being addressed include:
 - Duplication – the same resource in multiple repositories
 - Re-use – datasets that claim to be distinct but have data in common, e.g., WMT or RumourEval shared tasks
 - Updates – new datasets added to already imported repositories
- Legal issues
 - ELG is a complex platform with heterogeneous data flows
 - Need for monitoring for GDPR compliance and a specific data management plan
 - ELG services and datasets use over 130 different licences between them
 - The “Conditions of use” metadata field has been associated to each identified licence (to improve the search functionality)
 - All licences have been analysed and where possible – homogenised

Summary and Next Steps

- Establish ELG as the primary platform and marketplace for Language Technology in Europe.
- An initiative *from* the European LT community *for* the European LT community.
- European LT landscape is **highly fragmented**: ELG aims to provide just the right **umbrella platform**.
- Global market size by 2025 is enormous: we want the European LT community to be a **key player**.
- We want to **increase the visibility and reach of all members of the European LT landscape**.
- ELG is a long-term initiative: we will establish a **legal entity** for sustainability, which will operate and maintain the technology platform for the whole LT community as a joint marketplace.
- Contribute to **Digital Language Equality** in Europe by giving all our languages one virtual home and umbrella platform that collects **all** services and resources (**ELE**).
- **Next steps**: validation of **ELG products**; attach more **data repositories**; include ELG in **relevant infrastructures** (e.g., NFDI, GAIA-X); establish **ELG legal entity**.

ELG Tutorial Video

The video player displays several overlapping windows of JSON code. A large black text box at the bottom of the video area contains the text: "annotation responses, classification responses, text responses and audio responses." The video player interface at the bottom shows a progress bar at 11:01 / 14:02 and the title "Integrated ELG Services".

```
{
  "response":{
    "type":"texts",
    "warnings":[...], /* optional */
    "texts":[
      {
        "role":"string", /* optional */
        "content":"string of translated/transcribed text", // either
        "texts":["* same structure, recursive */], // or
        "score":number, /* optional */
        "features":{" /* arbitrary JSON, optional */ },
        "annotations":{" /* optional */
          "<annotation type>":[
            {
              "start":number,
              "end":number,
              "sourceStart":number, // optional
              "sourceEnd":number, // optional
              "features":{" /* arbitrary JSON */ }
            }
          ]
        }
      }
    ]
  }
}
```

```
{
  "response":{
    "type":"classification",
    "warnings":[...], /* optional */
    "classes":[
      {
        "class":"string",
        "score":number,
        "features":{" /* arbitrary JSON */ }
      }
    ]
  }
}
```

```
{
  "response":{
    "type":"audio",
    "warnings":[...], /* optional */
    "content":"base64 encoded audio for shorter snippets",
    "format":"string",
    "features":{" /* arbitrary JSON, optional */ },
    "annotations":{" /* optional */
      "<annotation type>":[
        {
          "start":number,
          "end":number,
          "sourceStart":number, // optional
          "sourceEnd":number, // optional
          "features":{" /* arbitrary JSON */ }
        }
      ]
    }
  }
}
```

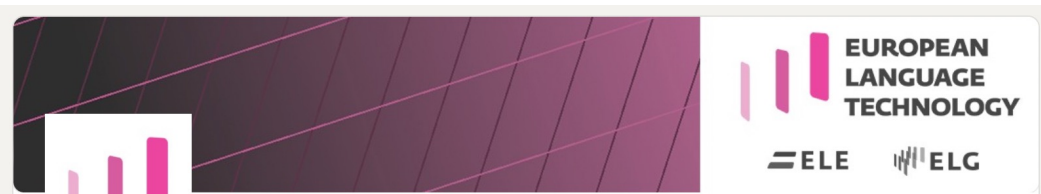
Description

- 0:00 Introduction
- 0:56 The European Language Grid
- 1:59 Browsing the ELG Catalogue
- 4:15 Running a Service
- 4:38 Organizations and Projects
- 5:09 The ELG PythonSDK
- 5:50 Becoming a User & Provider
- 7:22 Creating a Resource
- 8:47 Submitting a Resource
- 9:24 Integrated ELG Services
- 11:37 Creating an Integrated ELG Service
- 12:30 Summary

The European Language Grid (ELG) is a scalable cloud platform, providing access to hundreds of commercial and non-commercial Language Technologies for all European languages, including running tools and services as well as data sets and resources.

This tutorial explains how to browse the Grid, how to access its resources, how to register as a user or provider and how to contribute services, tools and corpora to the ELG. Further information on the project and the platform can be found at

<https://www.youtube.com/watch?v=29-V2EyMn4E>

European Language Technology
 Fostering the European Language Technology community towards digital language equality in Europe by 2030!
 Research Services · Brussels · 763 followers

[Following](#) [Learn more](#) [More](#)

[Home](#) [About](#) [Posts](#) [Jobs](#) [People](#) [Videos](#)

About

This is the combined channel of the EU projects European Language Equality (ELE) and European Language Grid (ELG) – together working towards a joint network of language technology for Europe’s languages and digital language equality by 2030. Follow us for project updates, events and news from the European artifici... [see more](#)

[See all details](#)

<https://www.linkedin.com/company/european-language-technology>

<https://twitter.com/EuroLangTech>

<https://www.european-language-technology.eu>

Subscribe to our newsletter

More than 4000 subscribers already!



META-FORUM 2022




European Language Technology
 266 Tweets

European Language Technology
 @EuroLangTech

The channel of the EU projects European Language Equality and European Language Grid, fostering the LT community towards digital language equality by 2030!

📍 Europe european-language-technology.eu 📅 Joined June 2021

1,039 Following 629 Followers

[Edit profile](#)



Welcome to European Language Technology!
 | | | ELT

[Follow ELT on Twitter](#)
[Follow ELT on LinkedIn](#)

European Language Technology is the combined communication channel for the sister projects European Language Grid (ELG) and European Language Equality (ELE) – funded by the European Commission.

For further information about either project – their goals, consortium partners, and contact details etc. – please click below:



08 June 2022

09:00-10:30 **Session 1: Opening – European Language Grid**

10:30-10:45 *Coffee Break*

11:00-11:45 **Session 2: ELG Platform**

11:45-12:30 **Session 3: Selected ELG Tools, Services and Resources**

12:30-13:30 *Lunch Break + Video Expo*

13:30-14:45 **Session 4: Language-centric AI Panel**

14:45-15:15 *Coffee Break*

15:15-16:15 **Session 5: Industry Session**

16:15-16:45 **Session 6: The Future of ELG**

16:45-17:00 **Closing Session**

17:00-18:00 *Reception*

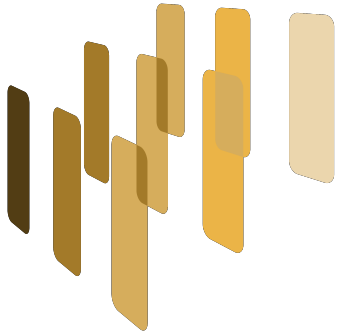


**EUROPEAN
LANGUAGE
GRID**

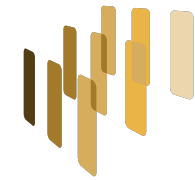
META-FORUM 2022:

Joining the European Language Grid
Together towards Digital Language Equality

Detailed Programme at <https://www.european-language-grid.eu/meta-forum-2022>



European Language Grid
European Language Equality



**EUROPEAN
LANGUAGE
GRID**



**EUROPEAN
LANGUAGE
EQUALITY**

Thank you!



The European Language Grid has received funding from the European Union's Horizon 2020 research and innovation programme under grant agreement № 825627 (ELG).

The European Language Equality project has received funding from the European Union under grant agreement № LC-01641480 – 101018166 (ELE),

Georg Rehm (DFKI), Stelios Piperidis (ILSP, R.C. "Athena"), Kalina Bontcheva (University of Sheffield),
georg.rehm@dfki.de, spip@athenarc.gr, k.bontcheva@sheffield.ac.uk

08/09-06-2022 META-FORUM 2022 – Joining the European Language Grid (hybrid conference)
<http://www.european-language-grid.eu> – <http://european-language-equality.eu>