# ELRC Data Reports

| Dissemination Level | Internal |
|---|---|
| **Validation Guidelines version No.** | 6.2 |
| **Date** | 30.07.2018 |
| **Name of LR** | Expression of interest (Processed) |
| **Resource ID** | 662 |
| **Resource Version No.** | 2.0 |
| **Contact person** | Dimitris Kapopoulos |
| **Validator** | Vassilis Papavassiliou - ILSP |
| **Validation Manager** | Kanella Pouli – ILSP |
| **Validation status** | ☐ Changes required<br>☒ Validated<br>☐ Rejected |

# 1. Validation Report

## Summary sheet

The validation results for this resource are as follows (please refer to the Validation Guidelines for the meaning of the various items):

| Validation steps | Validated (check box if yes) | Comments |
|---|---|---|
| 1) ELRC scope (see section 1 for details) | ☒ | |
| 2) Quick content check (see section 2 for details) | ☒ | |
| 3) LR Metadata (see section 3 for details) | ☒ | |
| 4) Legal issues (see section 4 for details) | ☒ | |
| 5) Content validation (see section 5 for details) | ☒ | |
| 6) Declaration on the list of pre-existing rights (see section 6 for details) | ☒ | |

If relevant, for details about the processing of the LR, see section 2 (Processing Report) at the end of this document.

## 1. Compliance with ELRC scope

| | Validated (check box if yes) | Comments |
|---|---|---|
| Data origin (comes from public institutions or relevant to the general administrative/regulatory domain and does not come from the European Commission) | ☒ | |
| Language(s) of the data content[1] (not the documentation) | ☒ | |

## 2. Quick content check

| | Validated (check box if yes) | Comments |
|---|---|---|
| Readability of files | ☒ | |
| Data content acceptability (no empty files, correct alignment for parallel corpora, …) | ☒ | |

## 3. Validation of LR Metadata

### a. General information

| | Validated (check box if yes) | Comments |
|---|---|---|
| Language used in free text fields are CEF languages | ☒ | |
| Does the "resource name" field contain an English version? | ☒ | |
| Does Language(s) in "description" field contain an English version? | ☒ | |
| Is there any information mentioning Pre-processing done by the provider? | ☐ | |
| Is there any information mentioning Pre-processing done through ELRC services? | ☒ | |
| Has any conversion been performed on this resource so as to make it directly useful for training MT engines of the Automated Translation platform? | ☒ | |

---

[1] Parallel / multilingual corpora LRs should contain English and, at least, one of the following languages: Bulgarian, Croatian, Czech, Danish, Dutch, Estonian, Finnish, French, German, Greek, Hungarian, Icelandic, Irish, Italian, Latvian, Lithuanian, Maltese, Norwegian, Polish, Portuguese, Romanian, Slovakian, Slovenian, Spanish or Swedish. Monolingual corpora and terminology LRs should contain, at least, one of the following languages: Bulgarian, Croatian, Czech, Danish, Dutch, English, Estonian, Finnish, French, German, Greek, Hungarian, Icelandic, Irish, Italian, Latvian, Lithuanian, Maltese, Norwegian, Polish, Portuguese, Romanian, Slovakian, Slovenian, Spanish, Swedish

## b. Accuracy of completed metadata with respect to provided LR

| Mandatory metadata field names | Current value | Correct | Wrong | Missing | Comments |
|---|---|---|---|---|---|
| Resource name | Expression of interest (Processed) | ☒ | ☐ | ☐ | |
| Resource type | Corpus | ☒ | ☐ | ☐ | |
| PSI - Public Sector Information | Yes | ☒ | ☐ | n/a | |
| License | Public Domain | ☒ | ☐ | ☐ | |
| Contact person – surname | Kapopoulos | ☒ | ☐ | ☐ | |
| Contact person - email | d.kapopoulos@gsis.gr | ☒ | ☐ | ☐ | |
| Linguality type | Bilingual | ☒ | ☐ | ☐ | |
| Lexical conceptual resource or Language description type (n/a for corpora) | n/a | ☒ | ☐ | ☐ | |
| Language(s) name | English, Modern Greek (1453-) | ☒ | ☐ | ☐ | |
| Encoding level (n/a for corpora) | n/a | ☒ | ☐ | ☐ | |
| Character encoding (applicable for corpora only) | UTF-8 | ☒ | ☐ | ☐ | |
| Size | 84 | ☒ | ☐ | ☐ | |
| Size unit | Translation Units | ☒ | ☐ | ☐ | |
| Mime type | TMX | ☒ | ☐ | ☐ | |

| Other metadata field names (to be listed if completed by submitter) | Current value | Correct | Wrong | Comments |
|---|---|---|---|---|
| Domain | EMPLOYMENT & WORKING CONDITIONS (personnel management and staff remuneration); ECONOMICS | ☒ | ☐ | |
| Conformance to classification scheme | Eurovoc | ☒ | ☐ | |
| Multilinguality type | Parallel | ☒ | ☐ | |
| Attribution text | | ☐ | ☐ | |
| Allows Uses Besides DGT | Yes | ☒ | ☐ | |
| IPR Holder | Hellenic Statistical Authority (EL.STAT.) | ☒ | ☐ | |
| Relation type and ID of related resource | Is Aligned Version of #199 | ☒ | ☐ | |

## 4. Legal validation

### a. If "PSI - Public Sector Information" metadata checkbox is ticked

| | Validated (check box if yes) | Comments |
|---|---|---|
| "License field" value is identified  (any value except "Under Review") | ☒ | |
| If attribution is required, IPR Holder(s) is identified in the "IPR holder" field | ☐ | |
| Privacy/Confidentiality (if the resource is identified as private or confidential, is "Personal Data Included" or "Sensitive Data Included" box ticked?) | ☐ | |

### b. If "PSI - Public Sector Information" metadata checkbox is not ticked

| | Validated (check box if yes) | Comments |
|---|---|---|

| | | |
|---|---|---|
| "License field" value is identified  (any value except "Under Review") | ☐ | |
| If attribution is required, IPR Holder(s) is identified in the "IPR holder" field | ☐ | |
| Privacy/Confidentiality (if the resource is identified as private or confidential, is "Personal Data Included" or "Sensitive Data Included" box ticked?) | ☐ | |

## 5. Content Validation

| AUTOMATIC VALIDATION | | | | |
|---|---|---|---|---|
| Has spell checking-based TU filtering been done? | Yes | ☐ | No | ☒ |
| Has alignment score outlier detection-based TU filtering been done? | Yes | ☐ | No | ☒ |
| Has TU length ratio-based filtering been done? | Yes | ☐ | No | ☒ |
| Have any other content validation steps been applied? If yes, list them in the columns to the right, one content validation step per row (add further rows if needed) | Yes | ☒ | No | ☐ |
| | Automatic content validation has been implemented as follows: - The c-eval tool (http://www.aclweb.org/anthology/W15-4924) was used to train a parallelness classifier using the DGT-TM-release 2016 datasets (https://ec.europa.eu/jrc/en/language-technologies/dgt-translation-memory). - The classifier was applied on this ELRC language resource and **97.62 %** of its sentence pairs were classified as of good quality. | | | |
| | | | | |

| MANUAL VALIDATION | | | | |
|---|---|---|---|---|
| Has manual TU validation been done? | | Yes ☐ | No ☒ | |
| If yes, indicate manually-annotated sample percentage (in terms of the number of TUs) | | < 1 % | ☐ | |
| | | 1-3 % | ☐ | |
| | | 3-5 % | ☐ | |
| | | 5-10 % | ☐ | |
| | | > 10 % | ☐ | |
| Has fined-grained error annotation been done? | | Yes ☐ | No ☒ | |
| If yes, indicate error type likelihoods (if available) | Unlikely (< 10 %) | Likely (10 – 60 %) | Very likely (> 60 %) | Undetermined (untreated) |
| Language identification error | ☐ | ☐ | ☐ | ☐ |
| Tokenisation error | ☐ | ☐ | ☐ | ☐ |

| | | | | |
|---|---|---|---|---|
| Translation error | ☐ | ☐ | ☐ | ☐ |
| Machine-translated text | ☐ | ☐ | ☐ | ☐ |
| Free translation | ☐ | ☐ | ☐ | ☐ |
| Character formatting error | ☐ | ☐ | ☐ | ☐ |
| Alignment error | ☐ | ☐ | ☐ | ☐ |
| Have any other content validation steps been applied? If yes, list them in the columns to the right, one content validation step per row (add further rows if needed) | Yes ☐ | | No ☒ | |
| | | | | |
| | | | | |
| | | | | |

## 6. Declaration on the list of pre-existing rights

| No. | Options | Selected option |
|---|---|---|
| 1 | The results of this LR are free of rights or claims from creators or from any third parties for any use. The contracting authority may envisage and declare that the results do not contain any pre-existing rights to the results or parts of the results or to pre-existing materials as defined in the above-mentioned contract. | ☒ |
| 2 | The results of this LR and the pre-existing material incorporated in the results are free of rights or claims from creators or from any third parties for any use. The contracting authority may envisage and declare that the results contain the following pre-existing rights: | ☐ |

**For Option 2 complete the table below – one line per pre-existing right**

| Result concerned | Pre-existing material concerned | Rights to pre-existing material | Identification of rights' holder |
|---|---|---|---|
| | | | |
| | | | |

# 2. Processing Report

This report provides details on the processing steps carried out on the resource referred to above. This information is filled in by the same LR validator.

| Processing action | Check if true | Comments |
|---|---|---|
| Does the processed resource originate from ELRC sources? | ☒ | |
| Has automatic text extraction from scanned documents (via Optical Character Recognition – OCR) been performed? | ☐ | |
| Has automatic text extraction from PDF or DOC(X) documents been performed? | ☒ | |
| Has automatic document pair detection been performed? | ☒ | |
| Has automatic sentence-level alignment been performed? | ☒ | |
| Has TMX cleaning been performed? | ☒ | |
| Have any other processing steps been carried out? If yes, list them in the columns to the right, one processing step per row (add further rows if needed) | ☒ | Text extraction (semi-automatically), alignment on document and sentence level, and generation of TMX files were applied. As a post-processing task several filters were applied to discard/annotate alignments that might be incorrect or of limited use for training MT systems. |
| | ☐ | |
| | ☐ | |