European Language Resource Coordination (ELRC) is a service contract operating under the EU's Connecting Europe Facility SMART 2014/1074 programme.

European Language
Resource Coordination
Connecting Europe Facility

ELRC Data Validation Report

DisseminationLevel	Internal
Validation Guidelines version No.	3.2
Date	22.12.2017
Name of LR	Polish Ministry of Foreign Affairs Regional Dataset (Processed)
Resource ID	481
Resource Version No.	2.0
Contact person	Ogrodniczuk Maciej
Validator	Vassilis Papavassiliou – ILSP
Validation Manager	Kanella Pouli – ILSP
Validation status	☐ Changes required ☒Validated ☐Rejected

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 (seesection 2 for details)	☒	
3) LR Metadata (see section 3 for details)	☒	
4) Legal issues (see section 4 for details)	☒	

1. Compliance with ELRC scope

	Validated (check box if yes)	Comments
Data origin (comes from public institutions or relevant to the generaladministrative/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	☒	The texts of the rulings together with some metadata were acquired from the website of the Polish Ministry of Justice: http://orzeczenia.ms.gov.pl and aggregated into JSON files.The translations were manually aligned at the sentence level and encoded in the XLiFF format.
Is there any information mentioning Pre-processing done through ELRC services	☐	
Has any converting been performed on this resource to make it directly useful for training MT engines of the Automated Translation platform	☒	

[1] 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	Polish Ministry of Foreign Affairs Regional Dataset (Processed)	☒	□	□	
Resource type	Corpus	☒	□	□	
PSI - Public Sector Information	Yes	☒	□	n/a	
Licence	Open Under PSI	☒	□	□	
Contact personsurname	Ogrodniczuk	☒	□	□	
Contact person email	maciej.ogrodniczuk@gmail.com	☒	□	□	
Linguality type	Bilingual	☒	□	□	
Lexical conceptual resource or Language description type (n/a for corpora)	n/a	☒	□	□	
Language(s) name	English, Polish	☒	□	□	
Encoding level (n/a for corpora)	n/a	☒	□	□	
Character encoding (applicable for corpora only)	UTF8	☒	□	□	
Size	3653	☒	□	□	
Size unit	Translation Units	☒	□	□	
Mime type	TMX	☒	□	□	

Other metadata field names (to be listed if completed by submitter)	**Current value**	**Correct**	**Wrong**	**Comments**
Domain	INTERNATIONAL RELATIONS (International Balance)	☒	□	
Multilinguality type	Parallel	☒	□	
Allows Uses Besides DGT	Yes	☒	□	
Relation type and ID of related resource	Is Version Of #299	☒	□	

4. Legal validation

a. If “PSI - Public Sector Information” metadata checkbox is ticked

	Validated (check box if yes)	Comments
“Licence field” value is identified (any value except “under review”) or indicated as “not available” if information about the licence is notavailable	☒	
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
“Licence 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. Further comments

The dataset was provided as a collection of two xlf files. They were merged into a TMX file. 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.

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 orclaims 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