



**European Language
Resource Coordination**
Connecting Europe Facility

ELRC Data Validation Report

Dissemination Level	Internal
Validation Guidelines version No.	2.0
Date	30 June 2017
Name of LR	Parallel Corpus from the Web Site of the the MFA of Latvia
Contact person	Jānis Ziediņš, KISC
Validator	Roberts Rozis, TILDE
Validation Manager	Aivars Bērziņš, TILDE
Validation status	<input type="checkbox"/> Changes required <input checked="" type="checkbox"/> Validated <input type="checkbox"/> 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)	<input checked="" type="checkbox"/>	
2) Quick content check (see section 2 for details)	<input checked="" type="checkbox"/>	
3) LR Metadata (see section 3 for details)	<input checked="" type="checkbox"/>	
4) Legal issues (see section 4 for details)	<input checked="" type="checkbox"/>	

ELRC Data Validation Report**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)	<input checked="" type="checkbox"/>	
Language(s) of the data content ¹ (not the documentation)	<input checked="" type="checkbox"/>	

2. Quick content check

	Validated (check box if yes)	Comments
Readability of files	<input checked="" type="checkbox"/>	
Data content acceptability (no empty files, correct alignment for parallel corpora, ...)	<input checked="" type="checkbox"/>	

3. Validation of LR Metadata**a. General information**

	Validated (check box if yes)	Comments
Language used in free text fields are CEF languages	<input checked="" type="checkbox"/>	
Does the “resource name” field contain an English version	<input checked="" type="checkbox"/>	
Does Language(s) in “description” field contain an English version	<input checked="" type="checkbox"/>	
Is there any information mentioning Pre-processing done by the provider	<input type="checkbox"/>	Checked. No such info.
Is there any information mentioning Pre-processing done through ELRC services	<input type="checkbox"/>	Checked. No such info.
Has any converting been performed on this resource to make it directly useful for training MT engines of the Automated Translation platform	<input checked="" type="checkbox"/>	

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

ELRC Data Validation Report

b. Accuracy of completed metadata with respect to provided LR

Mandatory metadata field names	Current value	Correct	Wrong	Missing	Comments
Resource name	Parallel Corpus from the Web Site of the the MFA of Latvia	<input checked="" type="checkbox"/>	<input type="checkbox"/>	<input type="checkbox"/>	
Resource type	Corpus	<input checked="" type="checkbox"/>	<input type="checkbox"/>	<input type="checkbox"/>	
PSI - Public Sector Information	YES	<input checked="" type="checkbox"/>	<input type="checkbox"/>	n/a	
Licence	CC-BY-SA	<input checked="" type="checkbox"/>	<input type="checkbox"/>	<input type="checkbox"/>	
Contact person surname	Ziediņš	<input checked="" type="checkbox"/>	<input type="checkbox"/>	<input type="checkbox"/>	
Contact person email	janis.ziedins@kis.gov.lv	<input checked="" type="checkbox"/>	<input type="checkbox"/>	<input type="checkbox"/>	
Linguality type	Bilingual	<input checked="" type="checkbox"/>	<input type="checkbox"/>	<input type="checkbox"/>	
Lexical conceptual resource or Language description type (n/a for corpora)	-	<input type="checkbox"/>	<input type="checkbox"/>	<input type="checkbox"/>	
Language(s) name	EN LV	<input checked="" type="checkbox"/>	<input type="checkbox"/>	<input type="checkbox"/>	
Encoding level (n/a for corpora)		<input type="checkbox"/>	<input type="checkbox"/>	<input type="checkbox"/>	
Character encoding (applicable for corpora only)	UTF-8	<input checked="" type="checkbox"/>	<input type="checkbox"/>	<input type="checkbox"/>	
Size	7194	<input checked="" type="checkbox"/>	<input type="checkbox"/>	<input type="checkbox"/>	
Size unit	Translation units	<input checked="" type="checkbox"/>	<input type="checkbox"/>	<input type="checkbox"/>	
Mime type	TMX	<input checked="" type="checkbox"/>	<input type="checkbox"/>	<input type="checkbox"/>	

Other metadata field names (to be listed if completed by submitter)	Current value	Correct	Wrong	Comments
Domain	INTERNATIONAL RELATIONS	<input checked="" type="checkbox"/>	<input type="checkbox"/>	
Conformance to classification scheme	EUROVOC	<input checked="" type="checkbox"/>	<input type="checkbox"/>	
Multilinguality type	Parallel	<input checked="" type="checkbox"/>	<input type="checkbox"/>	
Attribution text	The Parallel Corpus from the Web Site of the the MFA of Latvia has been created and donated to the European Language Resources Coordination Action (ELRC) (http://lr-coordination.eu/), and its primary data is licensed under "CC-BY-SA 4.0" (https://creativecommons.org/licenses/by-sa/4.0/).	<input checked="" type="checkbox"/>	<input type="checkbox"/>	
Allows Uses Besides DGT	Ticked "Yes"	<input checked="" type="checkbox"/>	<input type="checkbox"/>	

European Language Resource Coordination

ELRC Data Validation Report

IPR Holder	Culture Information Systems Centre	<input checked="" type="checkbox"/>	<input type="checkbox"/>	
------------	------------------------------------	-------------------------------------	--------------------------	--

ELRC Data Validation Report**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”)	<input checked="" type="checkbox"/>	
If attribution is required, IPR Holder(s) is identified in the “IPR holder” field	<input checked="" type="checkbox"/>	
Privacy/Confidentiality (if the resource is identified as private or confidential, is “Personal Data Included” or “Sensitive Data Included” box ticked?)	<input type="checkbox"/>	No Privacy / Confidentiality

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”)	<input type="checkbox"/>	
If attribution is required, IPR Holder(s) is identified in the “IPR holder” field	<input type="checkbox"/>	
Privacy/Confidentiality (if the resource is identified as private or confidential, is “Personal Data Included” or “Sensitive Data Included” box ticked?)	<input type="checkbox"/>	

5. Further comments

During conversion, parallel Plain text files were converted into a parallel corpus in TMX format.