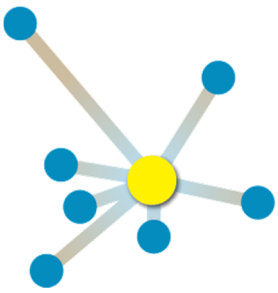
European Language Resource Coordination (ELRC) is a service contract operating under the EU's Connecting Europe Facility SMART 2014/1074 programme.



European Language Resource Coordination
Connecting Europe Facility

ELRC Data Validation Report

Dissemination Level	Internal
Validation Guidelines version No.	2.0
Date	30 June 2017
Name of LR	Parallel corpus from Estonian Ministry of Foreign Affairs
Contact person	Martin Luts martin.luts@tilde.ee
Validator	Roberts Rozis, TILDE
Validation Manager	Aivars Bērziņš, TILDE
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 (see section 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 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	☐	Checked. No such info.
Is there any information mentioning Pre-processing done through ELRC services	☐	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	☒	

[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	Parallel corpus from Estonian Cabinet of Ministers	☒	☐	☐	
Resource type	Corpus	☒	☐	☐	
PSI - Public Sector Information	Yes	☒	☐	n/a	
Licence	CC-BY	☒	☐	☐	
Contact person surname	Luts	☒	☐	☐	
Contact person email	martin.luts@tilde.ee	☒	☐	☐	
Linguality type	Bilingual	☒	☐	☐	
Lexical conceptual resource or Language description type (n/a for corpora)	-	☐	☐	☐	
Language(s) name	EN ET	☒	☐	☐	
Encoding level (n/a for corpora)		☐	☐	☐	
Character encoding (applicable for corpora only)	UTF-8	☒	☐	☐	
Size	17296	☒	☐	☐	
Size unit	Translation units	☒	☐	☐	
Mime type	TMX	☒	☐	☐	

Other metadata field names (to be listed if completed by submitter)	Current value	Correct	Wrong	Comments
Domain	POLITICS	☒	☐	
Conformance to classification scheme	EUROVOC	☒	☐	
Multilinguality type	Parallel	☒	☐	
Attribution text	The corpus of the Government Office was collected for the European Language Resources Coordination Action (ELRC) (http://lr-coordination.eu/) and primary data copyrighted by Government Office of Estonia and is licensed under "CC-BY 4.0" (https://creativecommons.org/licenses/by/4.0/).	☒	☐	
Allows Uses Besides DGT	Ticked “yes”	☒	☐	
IPR Holder	Government Office	☒	☐	

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")	☒	
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?)	☐	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")	☐	n/a
If attribution is required, IPR Holder(s) is identified in the "IPR holder" field	☐	n/a
Privacy/Confidentiality (if the resource is identified as private or confidential, is "Personal Data Included" or "Sensitive Data Included" box ticked?)	☐	n/a

5. Further comments

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