



**European Language
Resource Coordination**
Connecting Europe Facility

ELRC Data Reports

Dissemination Level	Internal
Validation Guidelines version No.	6.2
Date	05.09.2018
Name of LR	English-Estonian corpus from Finnish Information Bank (Processed)
Resource ID	717
Resource Version No.	2.0
Contact person	Bartis Imre imre.bartis@helsinki.fi
Validator	Rinalds Vīksna, Tilde
Validation Manager	Roberts Rozis, Tilde
Validation status	<input type="checkbox"/> Changes required <input checked="" type="checkbox"/> Validated <input type="checkbox"/> 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)	<input checked="" type="checkbox"/>	
2) Quick content check (see section 2 for details)	<input 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"/>	
5) Content validation (see section 5 for details)	<input checked="" type="checkbox"/>	
6) Declaration on the list of pre-existing rights (see section 6 for details)	<input checked="" type="checkbox"/>	

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)	<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 checked="" type="checkbox"/>	
Is there any information mentioning Pre-processing done through ELRC services?	<input type="checkbox"/>	

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

Data Quality Indicators

Has any conversion been performed on this resource so as to make it directly useful for training MT engines of the Automated Translation platform?	<input type="checkbox"/>	
--	--------------------------	--

b. Accuracy of completed metadata with respect to provided LR

Mandatory metadata field names	Current value	Correct	Wrong	Missing	Comments
Resource name	English-Estonian corpus from Finnish Information Bank (Processed)	<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	
License	CC-BY-4.0	<input checked="" type="checkbox"/>	<input type="checkbox"/>	<input type="checkbox"/>	
Contact person – surname	Bartis Imre	<input checked="" type="checkbox"/>	<input type="checkbox"/>	<input type="checkbox"/>	
Contact person - email	imre.bartis@helsinki.fi	<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	English, Estonian	<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	8321	<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"/>	

Data Quality Indicators

Other metadata field names (to be listed if completed by submitter)	Current value	Correct	Wrong	Comments
Domain	SOCIAL QUESTIONS	<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	"English-Estonian corpus from Finnish Information Bank" has been compiled from primary data of http://www.infopankki.fi/ by Tilde and provided for the European Language Resources Coordination Action (ELRC) (http://lr-coordination.eu/), licensed under "CC-BY 4.0" (https://creativecommons.org/licenses/by/4.0/).	<input checked="" type="checkbox"/>	<input type="checkbox"/>	
Allows Uses Besides DGT	Yes	<input checked="" type="checkbox"/>	<input type="checkbox"/>	
IPR Holder	The City of Helsinki	<input checked="" type="checkbox"/>	<input type="checkbox"/>	
Relation type and ID of related resource	Is version of #304	<input type="checkbox"/>	<input type="checkbox"/>	

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")	<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"/>	

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

	Validated (check box if yes)	Comments

Data Quality Indicators

"License 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. Content Validation

AUTOMATIC VALIDATION		
Has spell checking-based TU filtering been done?	Yes <input type="checkbox"/>	No <input checked="" type="checkbox"/>
Has alignment score outlier detection-based TU filtering been done?	Yes <input type="checkbox"/>	No <input checked="" type="checkbox"/>
Has TU length ratio-based filtering been done?	Yes <input checked="" type="checkbox"/>	No <input type="checkbox"/>
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 <input checked="" type="checkbox"/>	No <input type="checkbox"/>
MPFilter tool was used to remove segments with low similarity		
Corpus cleaner tool was used to remove segments with duplicate content		

MANUAL VALIDATION				
Has manual TU validation been done?	Yes <input checked="" type="checkbox"/>	No <input type="checkbox"/>		
If yes, indicate manually-annotated sample percentage (in terms of the number of TUs)	< 1 %	<input type="checkbox"/>		
	1-3 %	<input checked="" type="checkbox"/>		
	3-5 %	<input type="checkbox"/>		
	5-10 %	<input type="checkbox"/>		
	> 10 %	<input type="checkbox"/>		
Has fined-grained error annotation been done?	Yes <input type="checkbox"/>	No <input checked="" type="checkbox"/>		
If yes, indicate error type likelihoods (if available)	Unlikely (< 10 %)	Likely (10 – 60 %)	Very likely (> 60 %)	Undetermined (untreated)
Language identification error	<input checked="" type="checkbox"/>	<input type="checkbox"/>	<input type="checkbox"/>	<input type="checkbox"/>
Tokenisation error	<input checked="" type="checkbox"/>	<input type="checkbox"/>	<input type="checkbox"/>	<input type="checkbox"/>
Translation error	<input checked="" type="checkbox"/>	<input type="checkbox"/>	<input type="checkbox"/>	<input type="checkbox"/>
Machine-translated text	<input checked="" type="checkbox"/>	<input type="checkbox"/>	<input type="checkbox"/>	<input type="checkbox"/>
Free translation	<input checked="" type="checkbox"/>	<input type="checkbox"/>	<input type="checkbox"/>	<input type="checkbox"/>
Character formatting error	<input checked="" type="checkbox"/>	<input type="checkbox"/>	<input type="checkbox"/>	<input type="checkbox"/>
Alignment error	<input checked="" type="checkbox"/>	<input type="checkbox"/>	<input type="checkbox"/>	<input type="checkbox"/>

Data Quality Indicators

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 <input type="checkbox"/>	No <input checked="" type="checkbox"/>

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.	<input checked="" type="checkbox"/>
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:	<input type="checkbox"/>

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?	<input checked="" type="checkbox"/>	
Has automatic text extraction from scanned documents (via Optical Character Recognition – OCR) been performed?	<input type="checkbox"/>	
Has automatic text extraction from PDF or DOC(X) documents been performed?	<input type="checkbox"/>	
Has automatic document pair detection been performed?	<input type="checkbox"/>	

Data Quality Indicators

Has automatic sentence-level alignment been performed?	<input checked="" type="checkbox"/>	
Has TMX cleaning been performed?	<input checked="" type="checkbox"/>	
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)	<input type="checkbox"/>	
	<input type="checkbox"/>	
	<input type="checkbox"/>	