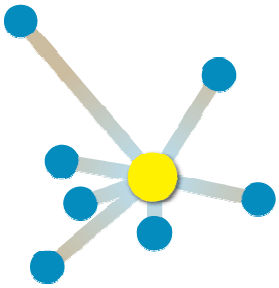

European Language Resource Coordination

Connecting Europe Facility

ELRC Data Validation Report

DisseminationLevel	Internal
Validation Guidelines version No.	3.2
Date	22.12.2017
Name of LR	Bilingual Bulgarian-English corpus from the National Revenue Agency (BG) (Processed)
Resource ID	471
Resource Version No.	2.0
Contact person	Rusinova Annie
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 (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	☐	
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	☒	

b. Accuracy of completed metadata with respect to provided LR

Mandatory metadata field names	Current value	Correct	Wrong	Missing	Comments
Resource name	Bilingual Bulgarian-English corpus from the National Revenue Agency (BG) (Processed)	☒	☐	☐	
Resource type	Corpus	☒	☐	☐	
PSI - Public Sector Information	Yes	☒	☐	n/a	
Licence	Open Under-PSI	☒	☐	☐	
Contact personsurname	Rusinova	☒	☐	☐	
Contact person email	a.rusinova@nra.bg	☒	☐	☐	
Linguality type	Bilingual	☒	☐	☐	
Lexical conceptual resource or Language description type (n/a for corpora)	n/a	☒	☐	☐	
Language(s) name	Bulgarian, English	☒	☐	☐	
Encoding level (n/a for corpora)	n/a	☒	☐	☐	
Character encoding (applicable for corpora only)	UTF-8	☒	☐	☐	
Size	1292	☒	☐	☐	
Size unit	Translation Units	☒	☐	☐	
Mime type	TMX	☒	☐	☐	

Other metadata field names (to be listed if completed by submitter)	Current value	Correct	Wrong	Comments
Domain	FINANCE(Taxation)	☒	☐	
Conformance to classification scheme	Eurovoc	☐	☐	
Multilinguality type	Parallel	☒	☐	
Allows Uses Besides DGT	Yes	☒	☐	
Relation type and ID of related resource	Is Aligned Version of #447	☒	☐	

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 not available	☒	
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

Bilingual Bulgarian-English corpus of administrative documents on the Refund of Value Added Tax from the Bulgarian National Revenue Agency. It was offered as a collection of documents by the Bulgarian National Revenue Agency. Modules of the ILSP Focused Crawler was used for the normalization, cleaning, (near) de-duplication and identification of parallel documents. The Maligna sentence aligner was used for extracting segment alignments from crawled parallel documents. As a post-processing step, alignments were merged into one TMX file and several filters were applied to discard/annotate alignments that might be incorrect or of limited for training MT engines.

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