ELRC Data Reports

Dissemination Level	Internal
Validation Guidelines version No.	V6.2
Date	12/10/2018
Name of LR	Spanish-English website parallel corpus (Processed)
Resource ID	863
Resource Version No.	V2.0
Contact person	Valérie Mapelli (mapelli@elda.org) Núria Bel (nuria.bel@upf.edu)
Validator	ELDA
Validation Manager	Victoria Arranz (ELDA)
Validation status	☐ Changes required ☒ Validated ☐ 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)	☒	
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)	☒	
5) Content validation (see section 5 for details)	☒	
6) Declaration on the list of pre-existing rights (see section 6 for details)	☒	

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)	☒	
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?	☐	

[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

Has any conversion been performed on this resource so as 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	Spanish-English website parallel corpus (Processed)	☒	☐	☐	
Resource type	Corpus	☒	☐	☐	
PSI - Public Sector Information	Yes / Ticked	☒	☐	☐	
License	Open Under-PSI	☒	☐	☐	
Contact person – surname	Mapelli Bel	☒	☐	☐	
Contact person - email	mapelli@elda.org nuria.bel@upf.edu	☒	☐	☐	
Linguality type	Bilingual	☒	☐	☐	
Lexical conceptual resource or Language description type (n/a for corpora)	n/a	☒	☐	☐	
Language(s) name	Spanish	☒	☐	☐	
Encoding level (n/a for corpora)	n/a	☒	☐	☐	
Character encoding (applicable for corpora only)	UTF-8	☒	☐	☐	
Size	21,007	☒	☐	☐	
Size unit	TU	☒	☐	☐	
Mime type	TMX	☒	☐	☐	

Other metadata field names (to be listed if completed by submitter)	Current value	Correct	Wrong	Comments
Domain		☒	☐	Various domains (many different sources)
Conformance to classification scheme		☐	☐	
Multilinguality type	Parallel	☒	☐	
Attribution text	Attribution Details: See COPYRIGHT file which contains Source owners	☒	☐	
Allows Uses Besides DGT	Yes	☒	☐	
IPR Holder		☒	☐	
Relation type and ID of related resource	Is Processed Version of #339	☒	☐	

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”)	☒	
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
“License 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. Content Validation

AUTOMATIC VALIDATION		
Has spell checking-based TU filtering been done?	Yes ☐	No ☒
Has alignment score outlier detection-based TU filtering been done?	Yes ☐	No ☒
Has TU length ratio-based filtering been done?	Yes ☐	No ☒
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 ☐	No ☒

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

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 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:	☒

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
Language resource ID 863 (processed version of #339)	Website www.mzv.sk	Copyright and/or sui generis database right	Ministry of Foreign and European Affairs of the Slovak Republic
Language resource ID 863 (processed version of #339)	Website www.dgojuego.minhap.gob.es	Copyright and/or sui generis database right	Ministerio de Hacienda y Administraciones Públicas
Language resource ID 863 (processed version of #339)	Website www.culturanorte.pt	Copyright and/or sui generis database right	Direção Regional de Cultura do Norte
Language resource ID 863 (processed version of #339)	Website www.bl.uk	Copyright and/or sui generis database right	The British Library
Language resource ID 863 (processed version of #339)	Website www.seap.minhap.gob.es	Copyright and/or sui generis database right	Secretaría de Estado de Administraciones Públicas , Ministerio de Hacienda y

			Administraciones Públicas
Language resource ID 863 (processed version of #339)	Website www.madrid.embaixadaportugal.mne.pt	Copyright and/or sui generis database right	Ministério dos Negócios Estrangeiros
Language resource ID 863 (processed version of #339)	Website www.poliziastato.it	Copyright and/or sui generis database right	Polizia di Stato, Ministero dell'Interno
Language resource ID 863 (processed version of #339)	Website www.dipgra.es	Copyright and/or sui generis database right	Diputación de Granada
Language resource ID 863 (processed version of #339)	Website www.larioja.org	Copyright and/or sui generis database right	Gobierno de la Rioja
Language resource ID 863 (processed version of #339)	Website www.bizkaia.eus	Copyright and/or sui generis database right	Diputación Foral de Bizkaia
Language resource ID 863 (processed version of #339)	Website www.dipcas.es	Copyright and/or sui generis database right	Diputació de Castelló
Language resource ID 863 (processed version of #339)	Website www.conselldemallorca.net	Copyright and/or sui generis database right	Consell de Mallorca.net
Language resource ID 863 (processed version of #339)	Website www.ine.es	Copyright and/or sui generis database right	Instituto Nacional de Estadística
Language resource ID 863	Website www.quirinale.it	Copyright and/or sui generis	Presidenza della Repubblica

(processed version of #339)		database right	
Language resource ID 863 (processed version of #339)	Website www.sligococo.ie	Copyright and/or sui generis database right	Sligo county council
Language resource ID 863 (processed version of #339)	Website www.agenciatributaria.es	Copyright and/or sui generis database right	Agencia Tributaria, Gobierno de España
Language resource ID 863 (processed version of #339)	Website www.ivdp.pt	Copyright and/or sui generis database right	Instituto dos Vinhos do Douro e do Porto, I. P.
Language resource ID 863 (processed version of #339)	Website www.agenciaidea.es	Copyright and/or sui generis database right	Consejería de Empleo, Empresa y Comercio, Junta de Andalucía
Language resource ID 863 (processed version of #339)	Website www.aragon.es	Copyright and/or sui generis database right	Gobierno de Aragón
Language resource ID 863 (processed version of #339)	Website www.dgsfp.mineco.es	Copyright and/or sui generis database right	Dirección General de Seguros y Fondos de Pensiones
Language resource ID 863 (processed version of #339)	Website www.mineco.gob.es	Copyright and/or sui generis database right	Ministerio de Economía y Competitividad
Language resource ID 863 (processed version of #339)	Website www.asturias.es	Copyright and/or sui generis database right	Gobierno del Principado de Asturias
Language resource ID	Website www.dival.es	Copyright and/or sui	Diputació de València

863 (processed version of #339)		generis database right	
Language resource ID 863 (processed version of #339)	Website www.msssi.gob.es	Copyright and/or sui generis database right	Ministerio de Sanidad, Servicios Sociales e Igualdad
Language resource ID 863 (processed version of #339)	Website www.jgpa.es	Copyright and/or sui generis database right	Junta General del Principado de Asturias
Language resource ID 863 (processed version of #339)	Website www.aecosan.msssi.gob.es	Copyright and/or sui generis database right	Agencia española de consumo, seguridad alimentaria y nutrición
Language resource ID 863 (processed version of #339)	Website www.minetur.gob.es	Copyright and/or sui generis database right	Ministerio de Industria, Energía y Turismo
Language resource ID 863 (processed version of #339)	Website www.spanien.diplo.de	Copyright and/or sui generis database right	German Foreign Ministry
Language resource ID 863 (processed version of #339)	Website www.csd.gob.es	Copyright and/or sui generis database right	Consejo Superior de Deportes, Ministerio de Educación, Cultura y Deporte
Language resource ID 863 (processed version of #339)	Website www.turismodeportugal.pt	Copyright and/or sui generis database right	Turismo de Portugal, I.P.
Language resource ID 863 (processed version of #339)	Website www.ine.es	Copyright and/or sui generis database right	Instituto Nacional de Estadística

Language resource ID 863 (processed version of #339)	Website www.asambleamadrid.es	Copyright and/or sui generis database right	Asamblea de Madrid
Language resource ID 863 (processed version of #339)	Website www.catastro.meh.es	Copyright and/or sui generis database right	Dirección General del Catastro
Language resource ID 863 (processed version of #339)	Website www.carabinieri.it	Copyright and/or sui generis database right	Ministero della Difesa
Language resource ID 863 (processed version of #339)	Website www.aeval.es	Copyright and/or sui generis database right	Agencia Estatal de Evaluación de las Políticas Públicas y la Calidad de los Servicios
Language resource ID 863 (processed version of #339)	Website www.dppireland.ie	Copyright and/or sui generis database right	Director of Public Prosecutions
Language resource ID 863 (processed version of #339)	Website www.mecd.gob.es	Copyright and/or sui generis database right	Ministerio de Educación, Cultura y Deporte
Language resource ID 863 (processed version of #339)	Website www.aecid.es	Copyright and/or sui generis database right	Agencia Española de Cooperación Internacional para el Desarrollo, Ministerio de Asuntos Exteriores y Cooperación

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?	☒	
Has automatic text extraction from scanned documents (via Optical Character Recognition – OCR) been performed?	□	
Has automatic text extraction from PDF or DOC(X) documents been performed?	□	
Has automatic document pair detection been performed?	□	
Has automatic sentence-level alignment been performed?	□	
Has TMX cleaning been performed?	☒	
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)	□	
	□	
	□	