Research and Innovation Action

Social Sciences & Humanities Open Cloud

Project Number: 823782 Start Date of Project: 01/01/2019 Duration: 40 months

Deliverable 4.3 Survey specific parallel corpora

THE [MCSQ]: MULTILINGUAL CORPUS OF SURVEY QUESTIONNAIRES

Dissemination Level	PU
Due Date of Deliverable	30/06/20 (M18)
Actual Submission Date	30/06/20
Work Package	WP4 - Innovations in Data Production
Task	T4.2 Preparing tools for the use of Computer Assisted Translation
Type	Other (database)
Approval Status	Waiting EC approval
Version	V1.0
Number of Pages	p.1 – p.14

Abstract:
This is the accompanying document to the [MCSQ]: Multilingual Corpus of Survey Questionnaires (MCSQ), a database of survey questionnaires' texts, submitted as Deliverable 4.3 of the SSHOC project. It summarizes technical information about Version 1.0 (Ada Lovelace) of the MCSQ, dated in June 2020. It links to the repository to access the code and files generating the database. The MCSQ is temporarily stored in a virtual machine provided by Universitat Pompeu Fabra.

The information in this document reflects only the author's views and the European Community is not liable for any use that may be made of the information contained therein. The information in this document is provided "as is" without guarantee or warranty of any kind, express or implied, including but not limited to the fitness of the information for a particular purpose. The user thereof uses the information at his/ her sole risk and liability.

History

SSHOC, "Social Sciences and Humanities Open Cloud", has received funding from the EU Horizon 2020 Research and Innovation Programme (2014-2020); H2020-INFRAEOSC-04-2018, under the agreement No. 823782

Version	Date	Reason	Revised by
1.0	25/06/2020	SSHOC project management review by Martina Drascic	Diana Zavala-Rojas
1.1	25/06/2020	Address SSHOC project management comments	Diana Zavala-Rojas

Author List

Organisation	Name	Contact Information
ESS ERIC (UPF)	Diana Zavala-Rojas	diana.zavala@upf.edu
UPF	Danielly Sorato	danielly.sorato@upf.edu

Executive Summary

This document describes the [MCSQ]: Multilingual Corpus of Survey Questionnaires (MCSQ), a database of survey questionnaires' texts. The report summarizes technical information about Version 1.0 (Ada Lovelace) of the MCSQ, dated in June 2020. It links to the repository to access the code and files generating the database. The corpus is compiled from questionnaires from the of European Social Survey (ESS) and the European Values Study (EVS) in the English source language and their translations into Catalan, Czech, French (produced for France, Switzerland, Belgium and Luxembourg), German (produced for Austria, Germany, Switzerland and Luxembourg), Norwegian, Portuguese, Spanish and Russian (produced for Israel, Latvia, Lithuania, Russian Confederation, Ukraine, Estonia). The MCSQ is temporarily stored in a virtual machine provided by Universitat Pompeu Fabra.

This document summarizes the research output of Task 4.2: Preparing tools for the use of Computer Assisted Translation, of the Social Science and Humanities Open Cloud (SSHOC) project. This document is closely related to D4.4, "Guidelines for building survey-specific corpora: the compilation of the [MCSQ]: Multilingual Corpus of Survey Questionnaires." D4.4 provides a detailed description about the design and implementation of the corpus, whereas D4.3 includes the database itself.

Abbreviations and Acronyms

CAT	Catalan language
CAPI	Computer Assisted Personal Interview
CSV	comma-separated values
CZE	Czech language
ENG	English language
ESS	European Social Survey
ER	Entity-Relationship
EVS	European Values Study
FRE	French language
GER	German language
GGP	Generations and Gender Programme
MCSQ	Multilingual Corpus of Survey Questionnaires
NER	Named Entity Recognition
NLP	Natural Language Processing
NLTK	Natural Language Toolkit
NOR	Norwegian language
OCR	Optical character recognition or optical character reader
PDF	Portable Document Format
POR	Portuguese language
POS, PoS	Part-of-speech
RUS	Russian language
SHARE	Survey of Health, Ageing and Retirement in Europe
SPA	Spanish language
SQL	Structured Query Language
SSHOC	Social Science and Humanities Open Cloud
TMT	Translation management tool
TRAPD	Translation, Review, Adjudication, Pretesting and Documentation
XLS	Microsoft Excel Spreadsheets
XML	Extensible Markup Language

Table of Contents

1.Introduction

Large-scale comparative survey projects such as the European Social Survey (ESS), the European Values Study (EVS) and the Survey of Health, Ageing and Retirement in Europe (SHARE) provide cross-national and cross-cultural data to the Social Sciences and Humanities (SSH). Empirical social research is often based on data gathered by administering survey questionnaires to representative samples of across countries.

Task 4.2: Preparing tools for the use of Computer Assisted Translation, of the Social Science and Humanities Open Cloud (SSHOC) project consists of compiling a corpus of survey questionnaires and providing guidelines about how it was designed and implemented.

In this report, the SSHOC Task 4.2 team summarizes technical information about version *1.0 (Ada Lovelace)* of the Multilingual Corpus of Survey Questionnaires (MCSQ), dated in June 2020, to our knowledge the very first publicly available corpus of survey questionnaires. Version 1 *Ada Lovelace* of the MCSQ is made up of survey questionnaires from ESS and the EVS. The planned *version 2.0 (Mileva Marić-Einstein)* will include questionnaires from SHARE. A main, immediate objective of the MCSQ is to allow for the retrieval and preservation of past translations, and to provide textual data in survey translation activities and research. In the SSHOC project, part of this data will be used in Task 4.3: Applying Computer Assisted Translation tools in Social Surveys to conduct translation research.

2. General description of the MCSQ

2.1 Data sources and pre-processing

The available information was retrieved from the websites of the ESS, EVS and SHARE studies to compile a catalogue of survey questionnaires by round/wave[1], year, study, language of administration, and questionnaire file format. For all studies listed a 'source questionnaire' version, written in localized British English exists. Target languages are Catalan, Czech, French (produced for France, Switzerland, Belgium and Luxembourg), German (produced for Austria, Germany, Switzerland and Luxembourg), Norwegian, Portuguese, Spanish and Russian (produced for Azerbaijan, Belarus, Estonia, Israel, Latvia, Lithuania, Russia and Ukraine).

Version Ada Lovelace includes about 107 ESS questionnaires and 44 EVS questionnaires. The eight primary languages have country-localised versions which add to 52 language-country combinations. In total there are approximately 400,000 segments (sentences) in the database. Table 1 shows ESS and EVS questionnaires included in version Ada Lovelace of the MCSQ per study, edition, country and language variety.

Table 1: Summary of country-language questionnaires in the corpus, per round/wave and study

	ESS						EVS		
Language & country	**Round 1**	**Round 2**	**Round 3**	**Round 4**	**Round 5**	**Round 6**	**Wave 3**	**Wave 4**	**Wave 5**
CAT Spain	X	X	X	X	X	X			
CZE Czechia	X	X		X	X	X	X	X	X
ENG Great Britain	X	X	X	X	X	X	X	X	X
ENG Ireland	X	X	X	X	X	X	X	X	X
ENG Montenegro								X	X
ENG Source	X	X	X	X	X	X	X	X	
ENG Luxemburg		X							
FRE Belgium	X	X	X	X	X	X	X	X	X
FRE Switzerland	X	X	X	X	X	X		X	X
FRE France	X	X	X	X	X	X	X	X	X

[1] Study's editions in the ESS are numbered by Round (Round 1, Round 2, etc.), whereas in the EVS, they are numbered by Wave (Wave 1, Wave 2, etc.).

FRE Luxemburg	X	X					X	X	X
GER Austria	X	X	X	X	X		X	X	X
GER Switzerland	X	X	X	X	X	X		X	X
GER Germany	X	X	X	X	X	X	X	X	X
GER Luxemburg		X						X	X
NOR Norway	X	X	X	X	X	X			X
POR Portugal	X	X	X	X	X	X	X	X	X
POR Luxemburg							X	X	X
RUS Azerbaijan								X	X
RUS Belarus							X	X	X
RUS Estonia		X	X	X	X	X	X	X	X
RUS Israel	X			X	X	X			
RUS Lithuania				X	X				X
RUS Latvia			X	X			X	X	X
RUS Russia			X	X	X	X	X	X	X
RUS Ukraine		X	X	X	X	X	X	X	X
SPA Spain	X	X	X	X	X	X	X	X	X

Survey questionnaires are made up of *survey items*. Survey items constitute the basic unit of analysis in the MCSQ (called documents). They were divided into sentences which constitute segments in the database. The survey questionnaires included in this corpus are administered as in person oral interviews. The answers are recorded in a standardized way either on paper or in a Computer Assisted Personal Interview (CAPI) device.

Depending on the file format and study of the source files, distinct steps of pre-processing were applied[2]. As of June 2020, the ESS has published nine editions (called *Rounds*) and EVS, five (called *Waves*). The format of the data sources compiled to create the corpus varies depending on the study and wave/round year. For ESS Round 01 to Round 06, we retrieved questionnaires in Portable Document Format (PDF) from the ESS website[3].

[2] The scripts to build the MCSQ are hosted in https://github.com/dsorato/MCSQ_compiling

[3] "European Social Survey https://www.europeansocialsurvey.org/"; [June 2020]

Rounds 08 and 09 of the ESS will be exported from the Translation Management Tool[4] (TMT) - a CAT tool in which the translation process of such rounds has been documented- in spreadsheet or XML formats. Those rounds will be available in the second version of the corpus (Mileva Marić-Einstein). For EVS, the source files were obtained from the GESIS/EVS data repository either in spreadsheet (wave 05) or XML (wave 03 and wave 04). EVS wave 01 and wave 02 were not included due to only being available in scanned images with low quality resolution. Therefore, they would have to be retyped before being pre-processed. SHARE's questionnaires will be retrieved from the TMT and included in the second version of the corpus.

Source files available only in PDF format were first converted into plain text format using a combination of both manual work and Optical Character Reader (OCR) tools. After transforming PDFs to plain text, the text files were converted to spreadsheet format. Questionnaires that were already in XLS or XML formats did not have to pass through format conversion. Spreadsheet and XML formats are both machine readable, and this means the files have clear data structures that can be easily interpreted by a computer. We were therefore able to extract the desired data from these files developing specific scripts for this purpose.

The texts were normalized converting them into a more convenient, standard form. All text passed through the following pre-processing procedure:

UTF-8 encoding; removal of unnecessary elements (e.g., trailing spaces, markup tags such as bold and italic , dots sequences); tokenization (segmentation) of the words; sentence segmentation; standardised label attribution to metadata; regex-based language specific recognition of instructions.

All data file conversion, extraction and pre-processing steps were performed algorithmically using the Python 3.6 programming language[5],[6] and NLP libraries, such as the Natural Language Toolkit (NLTK)[7]. To make sure the questionnaires were represented accordingly and eliminate human errors; we also validated these files. The validation was performed manually both by survey experts and linguists. During the conversion/validation step we used gitflow[8], a series of guidelines for Git[9], which is an open-source version control system. The usage of gitflow facilitates parallel work in teams and makes the versioning of files substantially easier.

2.2 Data nomenclature

As distinct survey projects compose the raw data for MCSQ, we established a common nomenclature to distinguish them in the corpus. This nomenclature also facilitates the process of checking metadata, as it carries

[4] Martens, M. (2017) Uploaded and modularized TMT. Deliverable 3.12 of the SERISS project funded under the European Union's Horizon 2020 research and innovation programme GA No: 654221. Retrieved from https://seriss.eu/wp-content/uploads/2017/07/SERISS-Deliverable-3.12_TMT_final.pdf."

[5] "Python https://www.python.org/"; [June 2020]

[6] Python scripts and other code used for developing the MCSQ can be accessed at the repository: "https://github.com/dsorato/MCSQ_compiling"

[7] "NLTK: https://www.nltk.org/"; [June 2020]

[8] "Gitflow: https://www.atlassian.com/git/tutorials/comparing-workflows/gitflow-workflow"; [June 2020]

[9]"Git: https://git-scm.com/"; [June 2020]

certain information. Namely, the information contained in this nomenclature is the study, round or wave, year, language and country, with the following digits: SSS_RRR_YYYY_LLL_CC.

This nomenclature is used both for identifying questionnaire files in the repository and for identifying documents, or survey items, in the corpus. For instance, the questionnaire file for ESS round 1, performed in the year 2002, written in the French of France would be named (as indicated in the example below *(survey)*:

```
survey = ESS_R01_2002_FRE_FR
survey_item_id = ESS_R01_2002_FRE_FR_i
```

Following the nomenclature rule, the survey items are named as in the example (*survey_item_id*), where *i* is the sequential number of each document as it is displayed in the questionnaire.

2.3 Entity-Relationship (ER) Model

In order to represent and store MCSQ data, we designed an *Entity-Relationship* (ER) model. An ER model is a conceptual representation of interrelated objects of interest inside a given domain. It is composed of entities (objects of interest) and the relationships between them. An entity is an abstraction of some aspect of the real world that can be uniquely identified, whereas a relationship between two certain entities specify how they relate to each other.

We present the MCSQ ER model in Figure 7. Eight distinct entities (or tables) compose this model, namely *Survey, Module, Survey Item, Introduction, Request, Instruction, Response* and *Alignment*. One survey is composed of several instances of survey items, which are our unit of analysis. Therefore, the relationship between *Survey* and *Survey Item* indicates that one entity type *Survey* can have many entity types *Survey* Item associated with it. The tables Introduction, *Request, Instruction* and *Response* are elements that may compose a survey item. The survey item elements have a zero-to-many relationship with survey items because they may not be present, i.e. not all survey items necessarily have all four substructures. The *Alignment* entity indicates the relationship between *Survey Item* entity types in English language (source) and their translations in other languages (target). Namely, this table holds the information of what is the *Survey Item* translation segment that corresponds to a given *Survey Item* in source language (English). The segments have correspondence at sentence level. To avoid the repetition of sentences in the database, we manipulated the data to identify unique segments throughout the questionnaires. Only unique elements in the introduction, instruction, request and response tables are included in the corpus and repeated elements are referenced by their IDs.

We developed the ER model of the MCSQ to represent in a structured manner how a survey questionnaire, survey items and its elements relate to each other. This design enables the inclusion of new data in MCSQ, as the database architecture is simple and easy to extend.

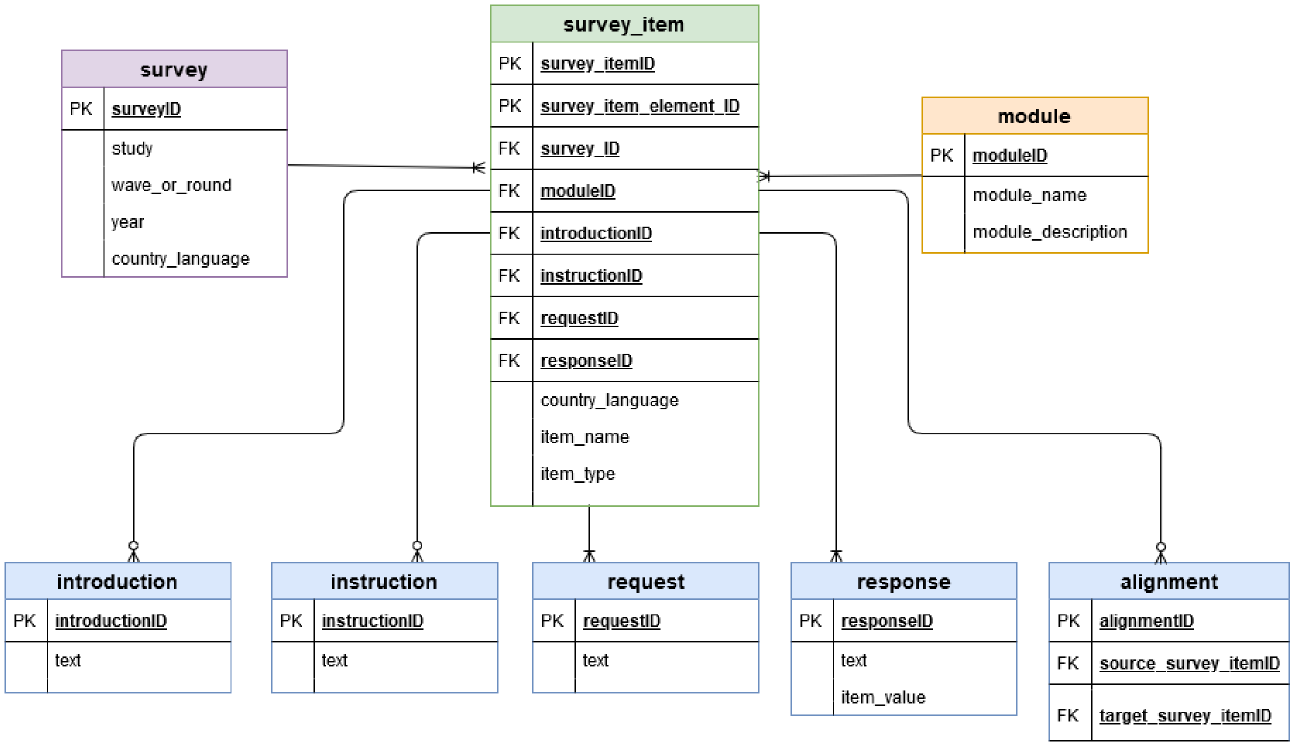

Figure 1: MCSQ Entity-Relationship (ER) Model diagram.

2.4 Implementation and Population

The MCSQ ER model was implemented using PostgreSQL[10], a database management system, and SQLAlchemy[11], an open-source Structured Query Language (SQL) toolkit and object-relational mapper for Python 3.6 programming language.

2.5 Alignment

Due to the large amount of data available and the opportunity of leveraging structural information in the alignment phase, we applied a strategy of pre alignment based on metadata. We developed an algorithm[12], which aligns two given files with respect to their *item_name*, *item_type* and *item_value* (in case of response segments) metadata. The segment length was also considered to decide correspondence between segments in the source English language and their translations.

After the pre-aligned files are generated, manual revision adjusts occurrences of incorrect alignments. In version Ada Lovelace of the MCSQ, about 15% of Russian language questionnaires were manually checked. In the entity Alignment of the database, response options that correspond to country-localized categories were excluded by design because they do not have alignment correspondence with other languages. Examples of

[10] "PostgreSQL: https://www.postgresql.org/";[June 2020]

[11] "SQLAlchemy: https://www.sqlalchemy.org/";[June 2020]

[12] In Python 3.6 programming language

questions that have country-localized response categories are those about affiliation to religious denominations, preference for political parties, or formal education levels.

3.Publishing the corpus

The MCSQ database is temporarily stored in a virtual machine provided by Universitat Pompeu Fabra which runs a Debian Operating System Linux distribution, with 70GB of disc capacity. Access to the database will be linked and promoted through the public domain mcsq.upf.edu. This hosting solution will remain active for the duration of the SSHOC project. For the long-term hosting of the database, we will apply to a public CLARIN repository.

A repository with the scripts used to generate the MCSQ is hosted at https://github.com/dsorato/MCSQ_compiling. The repository has a README.md file documenting the files in the repository (see Appendix 1 to this document).

This document is accompanied by file: "mcsq.dump.version.1.0.ada.lovelace.sql" a SQL dump of the database.

4. References

European Social Survey https://www.europeansocialsurvey.org/; June 2020.

Git: https://git-scm.com/; June 2020

Gitflow: https://www.atlassian.com/git/tutorials/comparing-workflows/gitflow-workflow; June 2020

Martens, M. (2017) Uploaded and modularized TMT. Deliverable 3.12 of the SERISS project funded under the European Union's Horizon 2020 research and innovation programme GA No: 654221. Retrieved from https://seriss.eu/wp-content/uploads/2017/07/SERISS-Deliverable-3.12_TMT_final.pdf."

NLTK: https://www.nltk.org/; June 2020

PostgreSQL: https://www.postgresql.org/; June 2020

Python programming language https://www.python.org/"; June 2020

SQLAlchemy: https://www.sqlalchemy.org/;June 2020

Appendix 1

README.md file in the repository containing the code generating the MCSQ

Compiling the Multilingual Corpus of Survey Questionnaires (MCSQ)

The Multilingual Corpus of Survey Questionnaires (MCSQ) is a multilingual corpus of survey items from different studies. It is designed as an entity-relationship (ER) database. In its first version (Ada Lovelace), it comprises datasets from the European Social Survey (rounds 1 to 6) and European Values Study (rounds 3 and 4) in English source language and their translations into Catalan, Czech, French (France, Switzerland, Belgium and Luxembourg), German (Austria, Germany, Switzerland and Luxembourg), Norwegian, Portuguese, Spanish and Russian (Israel, Latvia, Lithuania, Russian Confederation, Ukraine, Estonia).

This repository contains several modules that were used in the compilation of MCSQ.

In the preprocessing directory there are scripts to preprocess EVS and ESS data.

The scripts differ concerning the format of the input source file. For EVS and ESS XML inputs please use 'evs_xml_data_extraction.py' and 'ess_xml_data_extraction.py' respectively. For EVS spreadsheet inputs please use 'evs_data_extraction.py'.

txt_to_spreadsheet is a special module to transform ESS plain text files to spreadsheet.

Many utility functions for the preprocessing step can be found in 'utils.py' (some deprecated).

In the DB directory there are the files concerning the database structure. It is a PostgreSQL Entity-Relationship (ER) database implemented using SQLAlchemy and Python. The script 'populate_tables.py' takes the preprocessed spreadsheets as inputs and populated the database tables. Data manipulation for the population of the MCSQ ER model is done in 'ess_data_inclusion.py'.

Alignment folder has the script to align two given files regarding its metadata information.