Research and Innovation Action

Social Sciences & Humanities Open Cloud

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

Deliverable 4.4

Guidelines for building survey-specific corpora

The compilation of the [MCSQ]: Multilingual Corpus of Survey Questionnaires

Dissemination Level	PU
Due Date of Deliverable	30/06/20 (M18)
Actual Submission Date	26/06/20
Work Package	WP4 - Innovations in Data Production
Task	T4.2 Preparing tools for the use of Computer Assisted Translation
Type	Report
Approval Status	Waiting EC approval
Version	V1.1
Number of Pages	p.1 – p.24

Abstract: This report describes the design and compilation of the Multilingual Corpus of Survey Questionnaires (MCSQ), a database of survey questionnaires' texts. The report is based on the compilation of version 1.0 (Ada Lovelace) dated in June 2020. 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 translations into 9 languages. Using it as an example, the deliverable provides guidelines on the creation of corpora in survey research.

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

History

Version	Date	Reason	Revised by
1.0	24/06/2020	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
Møreforsking Volda	Lidun Hareide	lidun.hareide@himolde.no
Computational Linguistics Senior Advisor at UPF	Knut Hofland	knut.hofland@gmail.com

Executive Summary

This report describes the design and implementation of the [MCSQ]: Multilingual Corpus of Survey Questionnaires (MCSQ), a database of survey questionnaires' texts. It documents 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. By using this database as an example, the deliverable aims at providing guidelines on the creation of corpora in survey research. This document is closely related to Deliverable 4.3: Survey specific parallel corpora: the [MCSQ]: Multilingual Corpus of Survey Questionnaires, which corresponds to the database itself and its source code. The report is based on the compilation of version 1.0 (Ada Lovelace) dated in June 2020.

The corpus is compiled from European Social Survey (ESS) and the European Values Study (EVS) questionnaires 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).

To prepare the social sciences for the greater adoption of gold-standards in translation procedures, such as Computer-Assisted Tools or translation memories, domain-specific corpora of survey questionnaires is needed. In line with the focus on open-source, open-access principles of the SSHOC project, this corpus is openly accessible (in a format which is compatible with CAT tools) and will represent an important resource for both corpus linguists, computational linguists, statisticians, typologists, social scientists, as well as translation scholars and localizers. The planned version 2.0 (Mileva Marić-Einstein) will expand to include the Survey of Health, Ageing and Retirement in Europe (SHARE) questionnaires. In the SSHOC project, part of the data in the MCSQ will be used in Task 4.3: Applying Computer Assisted Translation tools in Social Surveys to conduct translation research.

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
TMX	Translation Memory Exchange
TRAPD	Translation, Review, Adjudication, Pretesting and Documentation
UPF	Universitat Pompeu Fabra
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. In this deliverable, this report describes the design and the compilation of version *1.0 (Ada Lovelace)* of the Multilingual Corpus of Survey Questionnaires (MCSQ), to the team's knowledge the very first publicly available corpus of survey questionnaires. Version *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.

Rigorous multilingual translation of survey questionnaires has become an important area of methodology for survey design, as evidence suggests that low quality translations hamper data comparability and increase errors of measurement[1],[2]. A general principle in the translation of survey questions is that they should be made *functionally equivalent* for the purposes of statistical analysis. *Functional equivalence* implies that the indicators obtained from translated survey instruments should represent the same concepts and maintain the intended psychometric properties across multilingual contexts, they do however not need to be made identical or equivalent in the common-sense meaning[3],[4]. To achieve high-quality functional equivalent questionnaire translations, Harkness (2003) suggested the Translation, Review, Adjudication, Pretesting and Documentation (TRAPD) method, a team or committee

[1] Davidov, E., & De Beuckelaer, A. (2010). How Harmful are Survey Translations? A Test with Schwartz's Human Values Instrument. *International Journal of Public Opinion Research*, 22(4), 485–510. https://doi.org/10.1093/ijpor/edq030

[2] Oberski, D., Saris, W. E., & Hagenaars, J. A. P. (2007). Why are there differences in measurement quality across countries? In G. Loosveldt & M. Swyngedouw (Eds.), *Measuring Meaningful Data in Social Research*. Retrieved from http://daob.nl/wp-content/uploads/2013/03/Oberski-Saris-Why-are-there-differences-in-measurement-quality-across-countries.pdf

[3] Harkness, J. A., Villar, A., & Edwards, B. (2010). Translation, Adaptation, and Design. In J. A. Harkness, M. Braun, B. Edwards, T. P. Johnson, L. E. Lyberg, P. P. Mohler, … T. W. Smith (Eds.), *Survey Methods in Multinational, Multiregional, and Multicultural Contexts* (pp. 115–140). https://doi.org/10.1002/9780470609927.ch7

[4] Zavala-Rojas, D., Saris, W. E., & Gallhofer, I. N. (2018). Preventing Differences in Translated Survey Items using the Survey Quality Predictor. In T. P. Johnson, B.-E. Pennell, I. A. L. Stoop, & B. Dorer (Eds.), *Advances in Comparative Survey Methods: Multinational, Multiregional and Multicultural Contexts* (3MC) (pp. 357–384). https://doi.org/https://doi.org/10.1002/9781118884997.ch17

approach in a multistep process[5]. Variants of the TRAPD method are used to translate questionnaires in the cross-national surveys included in the MCSQ. The TRAPD method is the gold-standard approach in survey translation. This approach results in rich local variations within the same language and within written language varieties, as well as within groups of related languages. An example of the scope of variation that can be found in the corpus is shown in the three segments in French language (FRE) from Belgium (BE), France (FR) or Switzerland (CH) (back translated to English language):

```
FRE_BE: La plupart des personnes sont dignes de confiance (Most people are trustworthy)

FRE_FR: On peut faire confiance aux gens (We can trust people)

FRE_CH: On peut faire confiance à la plupart des personnes (We can trust most people)
```

The TRAPD method, however, does not include any guidelines for the use of corpora or translation technologies. With the MCSQ, the SSHOC Task 4.2 team aims at contributing to the consolidation and the improvement of translation procedures in multilingual survey projects by providing an openly-accessible text database of multilingual survey questionnaires.

Linguistic corpora are compelling tools for linguistic/sociolinguistic research, both from theoretical and application-orientated perspectives[6]. Corpora are powerful sources of information, able to hold vast amounts of machine-readable data, either written or spoken. Corpus driven approaches allow us, for instance, to quickly retrieve and investigate massive quantities of text data, much more than a human being could ever manage to manually collect or analyse in a lifetime. Furthermore, the construction of corpora of multilingual/multicultural data, such as the MCSQ, provides valuable resources for minority as well as majority languages and cultures, tools for cross linguistic comparisons, and last but not least: a valuable tool for quantitative translations studies both across languages and language varieties.

2. Building a corpus of survey questionnaires

The framework used to create parallel corpora is suitable for a multilingual corpus with numerous languages and language varieties. It is depicted in Figure 1 and follows general best practices for the treatment and processing of multilingual text data[7]. This framework specifies three main stages (each including several steps), namely: (i) compiling corpus catalogue; (ii) corpus alignment and annotation and;

[5] Harkness, J. A. (2003). Questionnaire translation. In J. A. Harkness, F. J. R. van de Vijver, & P. P. Mohler (Eds.), *Cross-cultural survey methods* (pp. 35–56). Hoboken: Wiley & Sons. ISBN: 978-0-471-38526-4

[6] Marlén Izquierdo, Knut Hofland, and Øystein Reigem. The actres parallel corpus: an english–spanish translation corpus. Corpora, 3(1):31–41, 2008. DOI: 10.3366/E1749503208000051

[7] González, H. S. (2017). Creación de un Framework para el tratamiento de corpus lingüísticos. Universidad de León. Doctoral dissertation. Retrieved from: https://dialnet.unirioja.es/servlet/libro?codigo=727065

(iii) corpus publishing and optimisation. The corpus compilation stage comprises the pre-processing and Entity-Relationship (ER) model implementation and population steps, where the survey texts to be entered into the corpus are cleaned and transformed into plain text (all coding and figures are removed) separated into sentences, encoded into the new comma-separated values (CSV) format and the data is included in an ER model.

In the corpus annotation step, the data alignment and annotation (e.g., PoS-Tagging, Named Entity Recognition) takes place. The final stage, corpus publishing, corresponds to the optimization and publishing steps. Based on this framework, below a description of each of the stages and specific steps needed to create version 1.0 (Ada Lovelace) of the MCSQ. Sections also point out the next steps to be implemented in version 2.0 (Mileva Marić-Einstein).

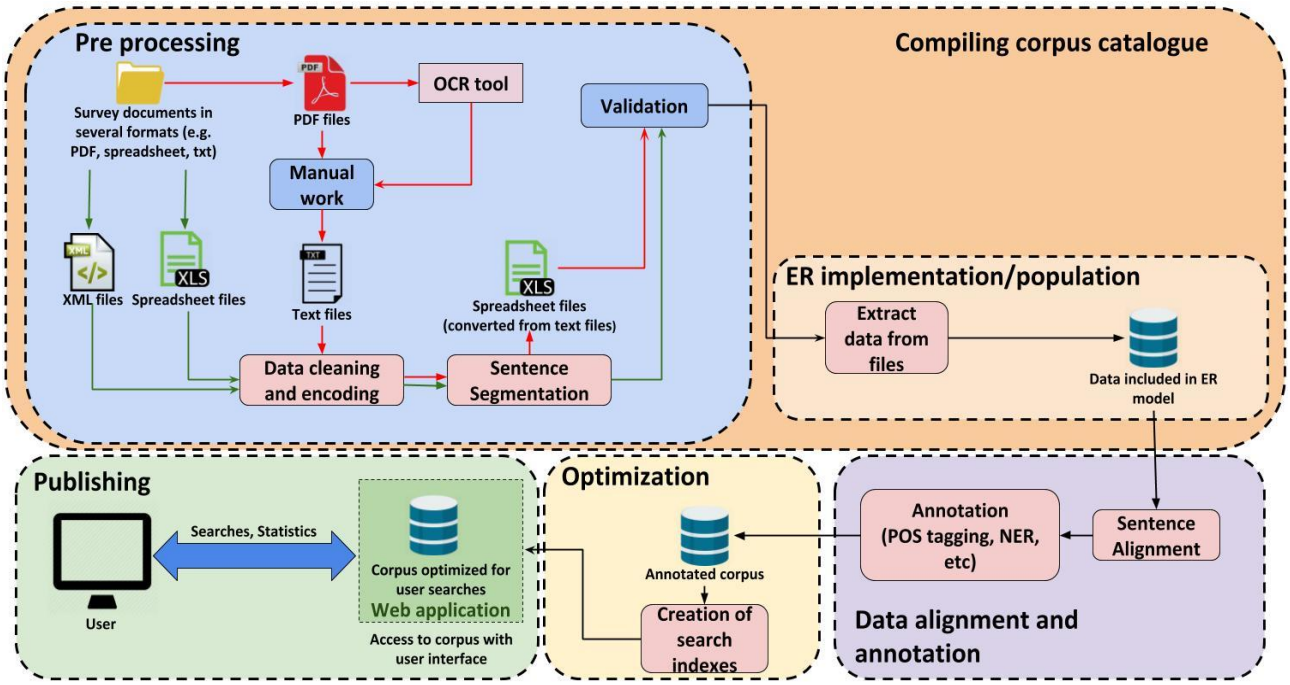

FIGURE 1: FLOWCHART OF A FRAMEWORK FOR THE CREATION OF PARALLEL CORPORA

Note: processes with red background were executed using computational algorithms, processes with blue background are dependent on manual work.

3. Compiling the corpus catalogue

3.1. Contents of the corpus

Task team retrieved available information from the websites of the ESS, EVS and SHARE studies to compile a catalogue of survey questionnaires by round/wave[8], year, study, language and questionnaire file format. For all studies listed a 'source questionnaire' version, written in localized British English exists. Questionnaires are made up of survey items. Commonly, survey items are a 'request for an answer' with its 'answer options', they may include additional textual elements guiding interviewers and clarifying the information that should be understood and provided by respondents[9]. Figure 2 shows Saris and Gallhofer's model of the decomposition of a survey item. Survey items constitute the basic unit of analysis in the MCSQ (called documents). They were divided into sentences which constitute segments in the database.

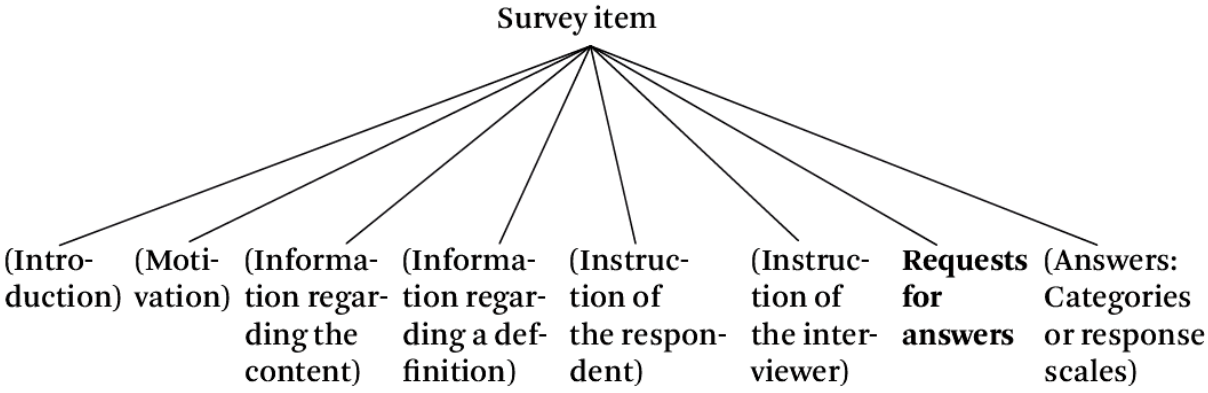

FIGURE 2: INDICATION OF STRUCTURAL ELEMENTS OF SURVEY ITEMS. SOURCE SARIS & GALLHOFER (2014)

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. A survey questionnaire performs a dual role by being both a guide to a communicative event between two persons, and at the same time an instrument for transforming that communicative event into data. These highly formatted texts are therefore complex, normally featuring scales, ticking boxes, columns, as well as routing guidelines for the interviewer. Across the survey projects, there is no industry standard for the creation of the questionnaire' documents. Some files are produced in Microsoft

[8] 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.).

[9] Saris, W. E., & Gallhofer, I. (2014). *Design, evaluation, and analysis of questionnaires for survey research*. In Wiley Series in Survey Methodology (Second, Vol. 548). https://doi.org/10.1111/j.1751-5823.2008.00054_20.x

Word or a similar word processor, whereas some are created as technical documents for programming the interview in a CAPI-device. The latter contain extensible visible coding and do therefore not exist in printable versions. The interfaces for retrieving and downloading the questionnaires have different formats, as the different survey projects' teams have different archiving systems. Some require granted data access, meaning that files cannot be downloaded automatically from their websites.

3.2. Data nomenclature

As distinct survey projects compose the raw data for MCSQ, a common nomenclature was established in order to distinguish them in the corpus. This nomenclature also facilitates the process of checking metadata, as it carries certain information. Namely, the information contained in this nomenclature is the study, round or wave, year, language and country, with the following number of 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.

3.3. Data sources and pre-processing

Preprocessing is one of the crucial tasks in the initial phase of corpus building. It is necessary in order to clean, standardize, and in some cases harmonize data inconsistencies before inclusion in the corpus. In particular, when mixing data from several sources, such as the case of MCSQ, special attention is required in this step. Each survey study produces data for its own intents. Therefore, survey items from different studies may diverge structurally. Depending on the file format and study of the source files, distinct steps of preprocessing were applied[10].

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, questionnaires were retrieved in Portable Document Format (PDF)

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

from the ESS website[11]. Rounds 08 and 09 of the ESS will be exported from the Translation Management Tool[12] (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 questionnaires require dedicated pre-processing, those questionnaires will be retrieved from the TMT and included in the second version of the corpus. 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 QUESTIONNAIRES IN THE MCSQ BY STUDY, WAVE/ROUND, COUNTRY AND LANGUAGE

	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							

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

[12] 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."

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
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

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. OCR tools are able to transform PDFs and images to plain text and typically achieve good performance when extracting simple structures (e.g., paragraphs of books, newspaper articles). However, in the corpus data source files there are some complex structures that OCR tools were not able to extract correctly. To exemplify this, Figure 3 depicts a complex structure from the EVS.

Q125 **Nous désirons analyser les résultats de cette étude en fonction des revenus familiaux des personnes que nous avons interrogées. Voici une échelle de revenus annuels NETS. Pouvez-vous me dire à quel niveau vous vous situez, en comptant toutes les rendtrées d'argent de votre foyer : salaires, allocations familiales, pensions, autres revenus... Quel est la lettre qui correspond à la situation de votre foyer ?**

☐☐ (v353)

		Somme approximative par SEMAINE	Somme approximative par MOIS	Somme approximative par AN
1	A	Moins de 125 euros	Moins de 500 euros	Moins de 6000 euros
2	B	De 125 euros à moins de 250 euros	De 500 euros à moins de 1 000 euros	De 6000 euros à moins de 12 000 euros
3	C	De 250 euros à moins de 321.50 euros	De 1 000 euros à moins de 1 250 euros	De 12 000 euros à moins de 15 000 euros
4	D	De 321.50 euros à moins de 375 euros	De 1 250 euros à moins de 1 500 euros	De 15 000 euros à moins de 18 000 euros
5	E	De 375 euros à moins de 437.50 euros	De 1 500 euros à moins de 1 750 euros	De 18 000 euros à moins de 21 000 euros
6	F	De 437.50 euros à moins de 500 euros	De 1 750 euros à moins de 2 000 euros	De 21 000 euros à moins de 24 000 euros
7	G	De 500 euros à moins de 562.50 euros	De 2 000 euros à moins 2 250 euros	De 24 000 euros à moins de 27 000 euros
8	H	De 562.50 euros à moins de 625 euros	De 2 250 euros à moins de 2 500 euros	De 27 000 euros à moins de 30 000 euros
9	I	De 625 euros à moins de 750 euros	De 625 euros à moins de 750 euros	De 30 000 euros à moins de 36 000 euros
10	J	De 750 euros à moins de 937.50 euros	De 3 000 euros à moins de 3 750 euros	De 36 000 euros à moins de 45 000 euros
11	K	De 937.50 euros à moins de 1 250 euros	De 3 750 euros à moins de 5 000 euros	De 45 000 euros à moins de 60 000 euros
12	L	De 1 250 euros à moins de 1 875 euros	De 5 000 euros à moins de 7 500 euros	De 60 000 euros à moins de 90 000 euros
13	M	De 1 875 euros à moins de 2 500 euros	De 7 500 euros à moins de 10 000 euros	De 90 000 euros à moins de 120 000 euros
14	N	2 500 euros et plus	10 000 euros et plus	120 000 euros et plus

FIGURE 3: EXAMPLE OF COMPLEX STRUCTURE IN QUESTIONNAIRE EVS 2008

For a human, it would be trivial to interpret this image as a table. However, a computer needs clear indications of what composes a table structure in order to interpret it, such as what is a row, a column, etcetera. These indications are not internally represented in a PDF file, and this means it is not structured data and a computer program would achieve poor results trying to extract it automatically. Also, the layout of the table itself does not favour the transformation into plain text. Due to this reason, manual work had to be carried out to transform these structures into machine readable formats.

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.

A special case of data cleaning is carried out in the SHARE files. This was necessary due to the existence of non-natural language items in these files, which will refer to as *fills* throughout this document. The sentences below are examples of survey items containing fills:

*Can I just confirm? You were born in **{FLMonthFill} {FLYearFill}**?*
*And about how many hours of help did **{FL_XT025_1}** receive during a typical day?*
*Enter an amount in **{FLCurr}***

In order to replace such fills, first it was necessary to evaluate and classify its types. Different actions were taken depending on the fill type. Namely, the fills were either (i) substituted by proposed static values or; (ii) deleted from the sentence. For instance, for texts such as the one in the example sentence number 2, substitution of the fill `{FL_XT025_1}` into a proper noun; `Tom` took place. The resulting text in this case is

And about how many hours of help did Tom receive during a typical day?

The texts were normalized[13]. In this context, normalization refers to a computational task that comprises a series of steps to preprocess the text converting it into a more convenient, standard form. Regardless of file formats conversion, all text passed through the following pre-processing procedure:

1. UTF-8 encoding;
2. Removal of unnecessary elements (e.g., trailing spaces, markup tags such as bold and italic, dots sequences);
3. Tokenization (segmentation) of the words;
4. Sentence segmentation;
5. Standardised label attribution to metadata;
6. Regex-based language specific recognition of instructions.

Dedicated scripts were created to implement all the aforementioned file format conversion and data extraction using the Python 3.6 programming language[14],[15]. Pre-processing steps were performed algorithmically with Python and its NLP libraries, such as the Natural Language Toolkit (NLTK)[16]. Step 5 (standardised label attribution to metadata) is a step that concerns the harmonization of the distinct

[13]Jurafsky, D. and Martin, J. H. (2000 Speech and language processing. Computational Linguistics, and Speech Recognition, UK: Prentice-Hall Inc, pages 22–105.

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

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

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

survey item types found in studies. For instance, some of the data sources subdivided requests item types into introduction, request and sometimes even sub requests, whereas other sources did not. As the aim of the research activity is to create a concise unique model for these sources and minimize manual annotation, the team simplified and standardised such labels in the first iteration of the corpus. For the example given above, all subdivisions e.g. introduction, question, request received the label 'request'. In following iterations of the corpus, manual annotate of the corpus can help to automatically unfold the subdivisions of the labels.

Step 6 refers to metadata attribution to include indications of the structural elements of the survey items for data sources where this information is absent, i.e., questionnaires converted to plain text. The project's team aim at implementing a rich structural set of elements, similar to the model by Saris & Gallhofer (2014) presented in Figure 2. Although such a level of details would not be feasible due to the necessity of time-consuming manual annotations in the corpus, the team was able to decompose a survey item into *introduction, instruction, request* and *response (answer)*.

For *request*, *introduction* and *response* elements, a file specification was developed containing textual tags. Later on, these tags were interpreted by a script which then attributes the correct metadata for the segments. As for the *instructions*, language specific regex patterns were developed to automatically identify them. Such regexes are capable of recognizing various types of instructions based on the tokens of a sentence. Examples of segments identified by the aforementioned regexes are: *'Please use this card to answer'*, *'Read out'*, *'Show card'*, *'Choose the answer that is closest to your opinion'* and its translations in Catalan, Czech, French, German, Norwegian, Portuguese, Spanish, and Russian.

In order to make sure the questionnaires were represented accordingly and eliminate human errors; these files underwent a validation process. The validation was performed manually both by survey experts and linguists. During the conversion/validation step gitflow[17] was used, a series of guidelines for Git[18], 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. Gitflow will be used as an users' friendly way to share the data with other partners in the SSHOC project, especially with the team of SSHOC Task 4.3: Applying Computer Assisted Translation tools in Social Surveys.

3.4. Entity-Relationship (ER) Model

Designing a database is a challenging task. Besides correctness and readability, scalability, performance and maintainability factors need to be taken into account. Due to this reason, designing a database is frequently a process of iterative changes and adaptations. Once the texts have been selected and preprocessed for inclusion in a corpus, a decision has to be made regarding how they should be

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

represented in electronic form[19]. In order to represent and store MCSQ data, the team 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.

There are no specific rules to design an ER model. Its conception depends crucially on the specific domain and intended usage. The designed model is an abstraction that serves as a guideline to implement a real ER database. To illustrate this abstract concept, entity-relationship examples are shown in Figure 4. First, suppose one wants to represent the fact that one parent can have one or more children. In an ER model, an entity *Parent* has a one-to-many relationship with an entity *Child*. This relationship specifies that a given entity *Parent* can have one or more entities type *Child* associated with it, but not the other way around. The second example in Figure 4 is a zero-to-many relationship that defines that a child can have zero or more toys. Finally, the last relationship example says one person can have exactly one national identity document, therefore the entity *Person* has a one-to-one relationship with the entity *National Identity Document*. Other types of relationships can be established such as many-to-many or zero-to-one. The symbols in the diagram indicate types of cardinality that an entity may have in a relationship, where the ring represents zero, the dash represents one and the crow's foot represents many. The first case has a dash and a crow's foot, representing a one-to-many relationship.

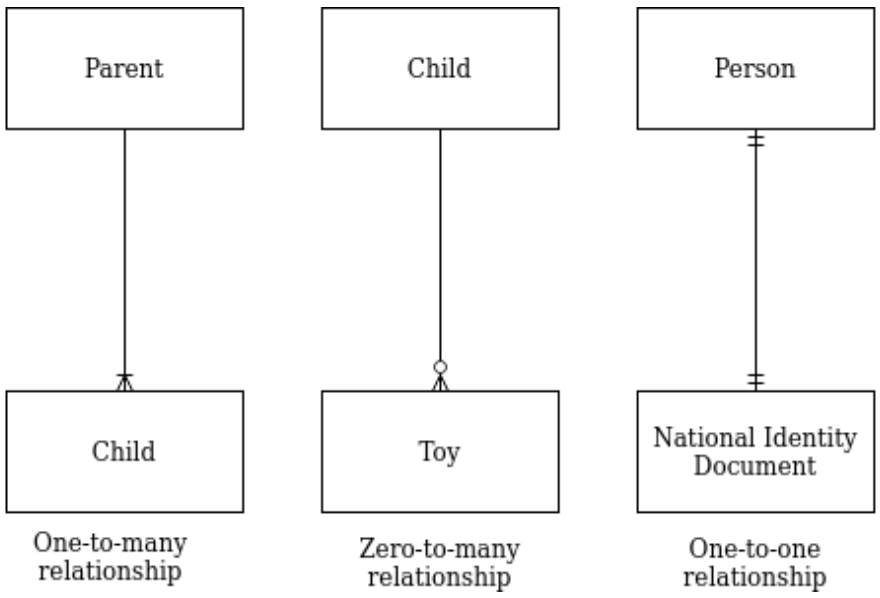

FIGURE 4: EXAMPLE OF ENTITIES AND THEIR RESPECTIVE RELATIONSHIPS

An entity, or table, also has attributes, also known as *fields*, *metadata* or *paradata*. Metadata meaning literally data about data, in other words, attributes are relevant characteristics of the entities. Each entity

[19] Kenny, D. (1998) Corpora in translation studies. Routledge encyclopedia of translation studies, pages 50–53.

corresponds to a database table and each attribute within an entity represents a column in such a table. The metadata is what defines what kind of entries are stored in the database and also what type of operations can be executed on them.

Having presented the theoretical basis about entity-relationship (ER) models to facilitate the readers' understanding, authors now 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 the corpus 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. The information about correspondence between source and target sentences can be used, for instance, in a translation memory.

The field *SurveyID* in the table *Survey* is marked with the acronym **PK**, because it is a **P**rimary **K**ey. That means that this attribute is responsible for identifying uniquely the entity *Survey*. While there may be two or more distinct records within the same *wave_or_round*, *year* and *language* in the table *Survey*, the survey identification number, *SurveyID,* is a unique number. In the entity *SurveyItem* the field *SurveyID* is marked with the acronym **FK**. The reason is that this field is a **F**oreign **K**ey in this table. A foreign key is a primary key from another table, in this case, the primary key of *Survey*. This explanation holds for all fields marked as **PK** or **FK.**

In conclusion, the model depicted in Figure 5 was developed 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.

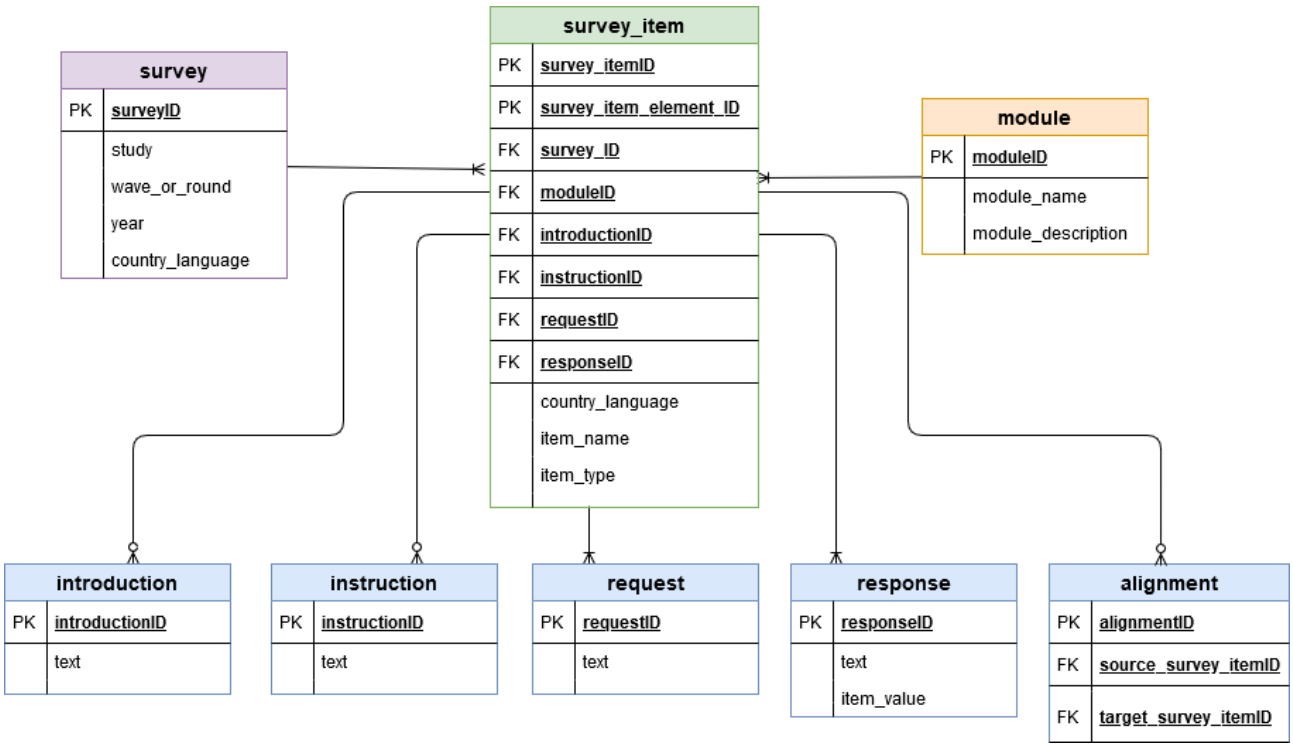

FIGURE 5: MCSQ ENTITY-RELATIONSHIP (ER) MODEL DIAGRAM

3.5. Implementation and Population

The MCSQ ER model was implemented using PostgreSQL[20], a database management system, and SQLAlchemy[21], an open-source Structured Query Language (SQL) toolkit and object-relational mapper for Python programming language. The SQLAlchemy toolkit facilitates the implementation of the database, eliminating the necessity of writing SQL code for implementing the model. It enables the creation and manipulation of database objects, i.e. entities and relationships, through high-level programming language, such as Python. Due to this, it promotes an easier way of making changes in the ER model structure as well as populating the database.

To populate the database, scripts were developed in order to extract information from the distinct source files used in this project. Then, the necessary information is extracted from the source files in spreadsheet or XML format by exploring the file structure, as they are already structured and machine readable. PDF files converted to plain text were transformed into spreadsheet format and subsequently the desired information was extracted from the spreadsheets.

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

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

To avoid the repetition of sentences in the database, unique segments were identified 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.

4. Data Alignment and Annotation

Due to the large amount of data available and the opportunity of leveraging structural information in the alignment phase, the Task team applied a strategy of pre alignment based on metadata. This strategy allows for the inclusion of more data, as it does not rely on more time-consuming alignment strategies. One drawback is that errors in automatic metadata attribution are further propagated to the alignment phase. The sentence segmentation and automatic metadata attribution in the pre-processing phase are crucial in order to achieve good quality in the pre alignment. The team developed an algorithm[22], 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. The large amount of data included in the database hinders the process of manual revision of the alignments. For the second iteration, manual review of a sample of at least 50% of the files in the corpus will be conducted. 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 questions that have country-localized response categories are those about affiliation to religious denominations, preference for political parties, or formal education levels.

5. Optimization and publishing the corpus

The MCSQ database is stored in a virtual machine provided by Universitat Pompeu Fabra (UPF), Barcelona which runs a Debian Operating System Linux distribution, with 70GB of disc capacity. This is a temporary solution available during the SSHOC project timeline, for the long-term preservation of the data, the Task team will apply to a CLARIN repository. The database was designed to be compatible with the standards of CLARIN for preservation. For easy access and search of the data during the SSHOC project, a public domain will be made available by UPF: mcsq.upf.edu

[22] In Python 3.6 programming language

6. Conclusion

This report documents 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 report provides guidelines on the creation of corpora in survey research. The Task team designed and implemented the [MCSQ]: Multilingual Corpus of Survey Questionnaires (MCSQ), a database of survey questionnaires' texts. The report is based on the compilation of version 1.0 (Ada Lovelace) dated in June 2020. The corpus is compiled from European Social Survey (ESS) and the European Values Study (EVS) questionnaires 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).

To prepare the social sciences for the greater adoption of gold-standards in translation procedures, such as Computer-Assisted Tools or translation memories, domain-specific corpora of survey questionnaires is needed. In line with the focus on open-source, open-access principles of the SSHOC project, this corpus is openly accessible (in a format which is compatible with CAT tools) and will represent an important resource for corpus linguists, computational linguists, statisticians, social scientists, as well as translation scholars and localizers. In the SSHOC project part of this corpus will feed into the activities of Task 4.3: Applying Computer Assisted Translation tools in Social Surveys to conduct translation research.

The planned version 2.0 (Mileva Marić-Einstein) will expand to include the Survey of Health, Ageing and Retirement in Europe (SHARE) questionnaires.

This document is closely related to Deliverable 4.3: Survey specific parallel corpora: the [MCSQ]: Multilingual Corpus of Survey Questionnaires, which corresponds to the database itself and its source code.

The interfaces for retrieving and downloading the questionnaires have different formats, as the different survey projects' teams have different archiving systems. Some require granted data access, meaning that files cannot be downloaded automatically from their websites. Survey questionnaires are complex documents, they are highly formatted texts, normally featuring scales, ticking boxes, columns, as well as routing guidelines for the interviewer, some questionnaires are created as technical documents for programming the interview in a CAPI-device. The latter contain extensible visible coding and do therefore not exist in printable versions.

Compiling corpus is a complex multidisciplinary activity. The creation of this database required the collaboration of survey experts, statisticians, computational linguists, and corpus linguists. The database was designed using an ER model. It aims 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. Compiling a corpus requires a combination of intensive manual and computational tasks. To create the MCSQ, the team used gold-standard framework and procedures summarized in this concluding section. They can serve as guidelines for the creation of corpora in survey research.

(i) compiling corpus catalogue:

This is the most important step. It aims at transforming source files with the texts of interest into data. It encompasses several steps of plain text generation, file format conversion and data extraction. It also includes modelling the representation of the database, and populating it by transforming the data into a digital object.

(ii) corpus alignment and annotation;

Once data sources have been pre-processed and integrated into a database, the alignment matches English source segments (sentences) with their translations. This step requires the selection of a strategy that will balance the amount of manual work and computational tasks.

(iii) optimisation and corpus publishing

Finally, the corpus should be published and optimized. The management, utilization, and preservation of databases require a highly skilled team, but the main objective is that the data is used by the SSH research community. Therefore, the data should be made compatible with common data formats used in survey research and translation, such as CVS or TMX. SSHOC Task 4.2 team designed the database in a way it is compatible with the requirements for being permanently hosted in a CLARIN repository, and at the time of submission has started the application process to host it in CLARIN. This is important for long-term preservation of the data.

7. References

Davidov, E., & De Beuckelaer, A. (2010). How Harmful are Survey Translations? A Test with Schwartz's Human Values Instrument. *International Journal of Public Opinion Research*, 22(4), 485–510. https://doi.org/10.1093/ijpor/edq030

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

González, H. S. (2017). Creación de un Framework para el tratamiento de corpus lingüísticos. Universidad de León. Doctoral dissertation. Retrieved from: https://dialnet.unirioja.es/servlet/libro?codigo=727065

Harkness, J. A. (2003). Questionnaire translation. In J. A. Harkness, F. J. R. van de Vijver, & P. P. Mohler (Eds.), *Cross-cultural survey methods* (pp. 35–56). Hoboken: Wiley & Sons. ISBN: 978-0-471-38526-4

Harkness, J. A., Villar, A., & Edwards, B. (2010). Translation, Adaptation, and Design. In J. A. Harkness, M. Braun, B. Edwards, T. P. Johnson, L. E. Lyberg, P. P. Mohler, … T. W. Smith (Eds.), *Survey Methods in Multinational, Multiregional, and Multicultural Contexts* (pp. 115–140). https://doi.org/10.1002/9780470609927.ch7

Jurafsky, D. and Martin, J. H. (2000 Speech and language processing. Computational Linguistics, and Speech Recognition, UK: Prentice-Hall Inc, pages 22–105.

Kenny, D. (1998) Corpora in translation studies. Routledge encyclopedia of translation studies, pages 50–53.

Marlén Izquierdo, Knut Hofland, and Øystein Reigem. The actres parallel corpus: an english–spanish translation corpus. *Corpora*, 3(1):31–41, 2008.

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

Oberski, D., Saris, W. E., & Hagenaars, J. A. P. (2007). Why are there differences in measurement quality across countries? In G. Loosveldt & M. Swyngedouw (Eds.), *Measuring Meaningful Data in Social Research*. Retrieved from http://daob.nl/wp-content/uploads/2013/03/Oberski-Saris-Why-are-there-differences-in-measurement-quality-across-countries.pdf

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

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

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

Zavala-Rojas, D., Saris, W. E., & Gallhofer, I. N. (2018). Preventing Differences in Translated Survey Items using the Survey Quality Predictor. In T. P. Johnson, B.-E. Pennell, I. A. L. Stoop, & B. Dorer (Eds.), *Advances in Comparative Survey Methods: Multinational, Multiregional and Multicultural Contexts* (3MC) (pp. 357–384). https://doi.org/https://doi.org/10.1002/9781118884997.ch17

List of Figures