[MCSQ]: The Multilingual Corpus of Survey Questionnaires

MCSQ Definition

[MCSQ]: The Multilingual Corpus of Survey Questionnaires is the first publicly available corpus of questions' texts. It includes questionnaires from European Social Survey (**ESS**), European Values Study (**EVS**) survey and The Survey of Health, Aging and Retirement in Europe (**SHARE**).

Since its first release (Ada Lovelace) in June 2020 the corpus grew consistently, culminating in 263 distinct questionnaires, approximately 657.000 sentences and more than 3.5 million words in current version 2.0 (Mileva Marić-Einstein). Version 3 adds questionnaires from the WageIndicator Survey to the database.

[MCSQ]: includes **source sentences in British English and their translations into Catalan, Czech, French, German, Norwegian, Portuguese, Spanish and Russian**, adding to **30 language varieties** (e.g. French from Switzerland). *

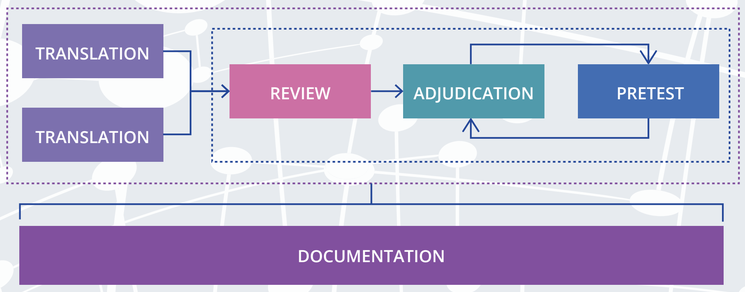
* MCSQ follows three standards: data is UTF-8 encoded, language codes follow the ISO 639- 2/B three-digit standard and country codes follow the ISO 3166 Alpha-2 two digit standard.

What can the MCSQ be used for?:

- **Analyzing translation equivalence**
- **Contrastive linguistic studies**
- **Building bilingual dictionaries of survey terms (lexicology)**
- **Building translation memories**
- **Using aligned data to train domain specific machine translation models**
- **Cross-linguistic comparison of survey terms**
- **Retrieving past items to use as reference for new translations**
- **Easily comparing survey items in multiple languages**
- **Training survey creators, translators and pollsters**

TRAPD approach

The [MCSQ]: data was produced using the **TRAPD approach** (Harkness, 2003). This is an iterative committee approach for translating questionnaires. Team members combine expertise on survey methodology, linguistics, and knowledge related to the questionnaire topic and the culture where it will be administered. The objective is to ask the same question across all cultures and countries participating in a survey project.

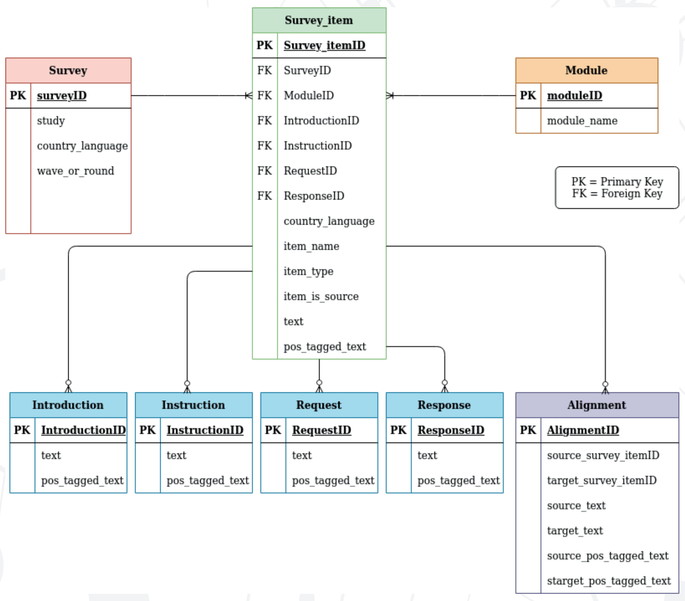


MCSQ Access

- **Official website** https://www.upf.edu/web/mcsq/
- Interact and download data from the database using the MCSQ interface: http://easy.mcsq.upf.edu/
- MCSQ is a FAIR and open-source research resource
  - Github repository containing developed code https://github.com/dsorato/MCSQ_compiling
  - Technical documentation in Read the Docs https://mcsq-compiling.readthedocs.io/en/latest/

MCSQ is an Entity- Relationship database

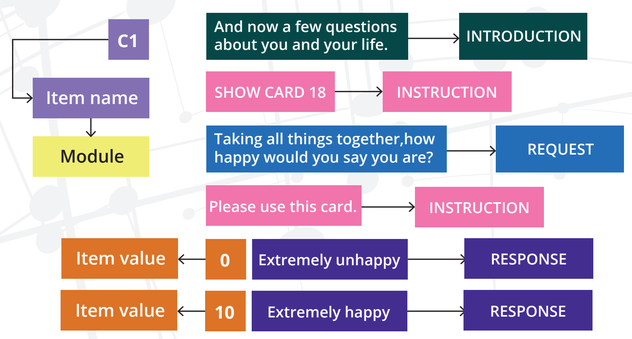
- An *Entity-Relationship* (ER) database is a representation of data as tables (entities), which have attributes (metadata) and relationships with other tables.
- ER models allow for conceptual representations of interrelated objects of interest inside a given domain.
- Eight distinct entities or tables compose the MCSQ ER model: *Survey, Module, Survey Item, Introduction, Request, Instruction, Response* and *Alignment*.
- A **PK** (**P**rimary **K**ey) uniquely identifies entities present in the database.
- A **FK** (**F**oreign **K**ey) describes relationships between entities, being an attribute in a table that references the PK of another table.



MCSQ ER Diagram

MCSQ Data structure:

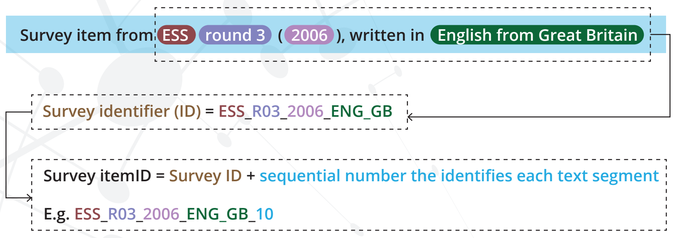
- Segment types are defined following Saris & Gallhofer 2014 model to decompose a survey item
- A survey item is a *request for an answer* with a set of *response options*, and may include additional textual information such as an *introduction* and *instruction*s, among others.



- The nomenclature to identify questionnaires in the corpus follows the following digits:

SSS_RRR_YYYY_LLL_CC
SSS - survey project or study,
RRR - edition (round or wave),
YYYY - year,
LLL - language,
CC - country,

To uniquely identify each segment in a questionnaire we add a sequential number *i* at the end of the nomenclature (SSS_RRR_YYYY_LLL_CC_*i*)



To cite the MCSQ

The [MCSQ]: Multilingual Corpus of Survey Questionnaires is an open-access and open-source research resource.
If you use part of the code, datasets, and/or findings to inspire your own scientific work, please cite the article:

Zavala-Rojas, D., Sorato, D., Hareide, L., & Hofland, K. (*forthcoming*). The Multilingual Corpus of Survey Questionnaires: a tool for refining survey translation. Meta: *Journal Des Traduceurs*.

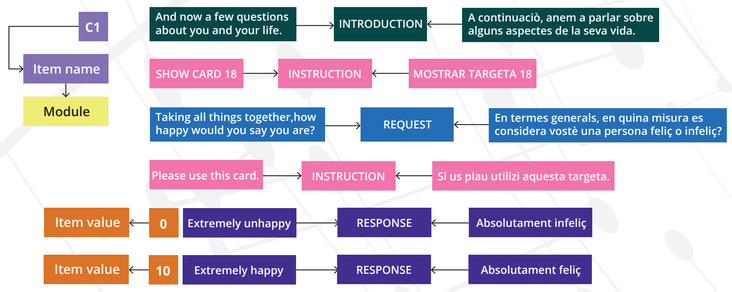
Interface functionalities

The MCSQ interface allows end-users to easily interact with and download MCSQ data from by simply applying filters in the graphical user interface. The following functionalities are currently implemented for MCSQ version 2.0:

- Building customized Translation Memories
- Word searches in both aligned and non-aligned data (by item type and regardless of item type), also including multiple and partial word search
- POS tag sequence search
- Word frequencies
- Collocations (bigrams and trigrams) and comparison of collocations
- Selection of corpus data for download independently of word search
- Comparison of questionnaires in multiple languages by item type, by word or the whole questionnaire

Data Alignment

- Sentence alignment is a computational task that finds the correspondence between a given sentence in a source language and its translation in the target languages[1].
- MCSQ aligns data based on a tailored sentence alignment algorithm that leverages metadata such as *module*, *item name*, *item type* to find the correspondences.
- This alignment strategy reduces the search space for the alignment candidates of a given source sentence.
- Approximately 80% of the corpus is aligned.
- Country-specific survey items, e.g. about religious denominations and political parties, are excluded from the alignments by design



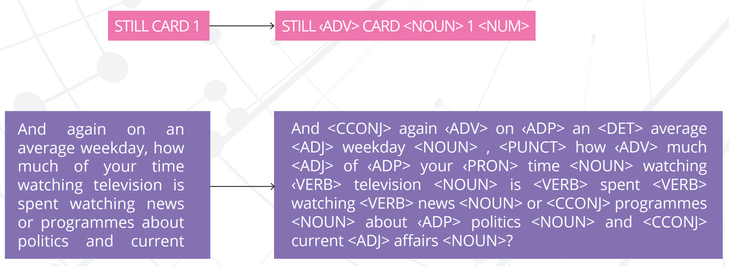
Visualization of alignments with item type correspondence in MCSQ

1 in MCSQ, response options and other short texts are considered sentences

Data annotation

Currently, the MCSQ data contains Part-of-Speech (POS) tagging annotation.
POS-tagging is a language-specific computational task that has the objective of predicting the POS tag, e.g. noun, verb, adjective, of each word in a given sentence.
MCSQ uses the Universal Dependencies tagset, which is homogeneous across languages.

In version 3, Named Entity Recognition annotation will be included.



MCSQ in numbers

Distribution of the data concerning the number of tokens and sentences across the languages (excluding punctuation).

Join the SSHOC community

info@sshopencloud.eu
sshopencloud.eu
@SSHOpenCloud
company/sshoc

SSHOC, "Social Sciences and Humanities Open Cloud", has received funding from the European Union's Horizon 2020 project call H2020-INFRAEOSC-04-2018, grant agreement #823782