[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, The Survey of Health, Aging and Retirement in Europe (**SHARE**) and the WageIndicator survey.

Since its first release (Ada Lovelace) in June 2020 the corpus grew consistently, culminating in 306 distinct questionnaires, approximately 766.000 sentences and more than 4 million words in current version 3 (Rosalind Franklin), adding questionnaires from the WageIndicator Survey and COVID-19 questionnaire 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.

Translation: the translation work is divided amongst two translators who produce independent translations.

Review: in a team meeting, the reviewer assesses and reconciles the translation versions.

Adjudication: an adjudicator is responsible for the final decisions on the different translation versions.

Pretesting: the translated questionnaire is tested before being administered to respondents.

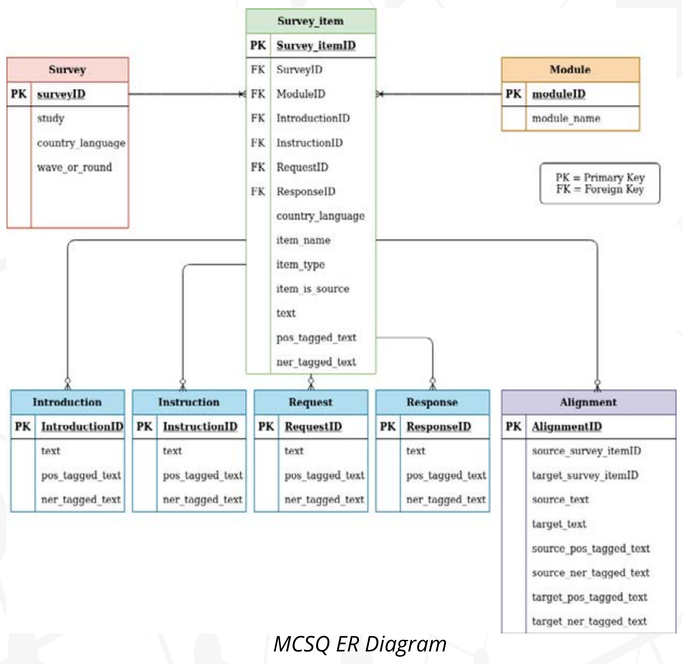
Documentation: the whole process is documented.

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 an Entity- Relationship database

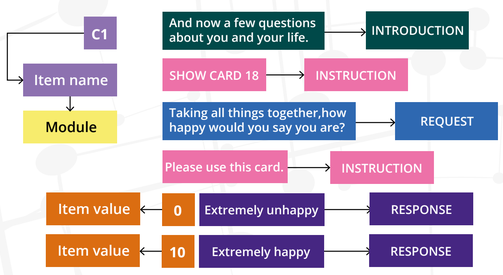
» An *Entity-Relationship* (ER) database is a representation of data as tables (entities), which have attributes (metadata) and relationships with other tables.
» ER models allow for conceptual representations of interrelated objects of interest inside a given domain.
» Eight distinct entities or tables 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

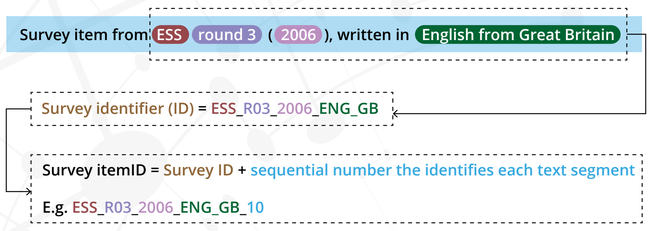
» Segment types are defined following Saris & Gallhofer 2014 model to decompose a survey item
» A survey item is a *request for an answer* with a set of *response options*, and may include additional textual information such as an *introduction* and *instructions*, among others.



» 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)



MCSQ interface

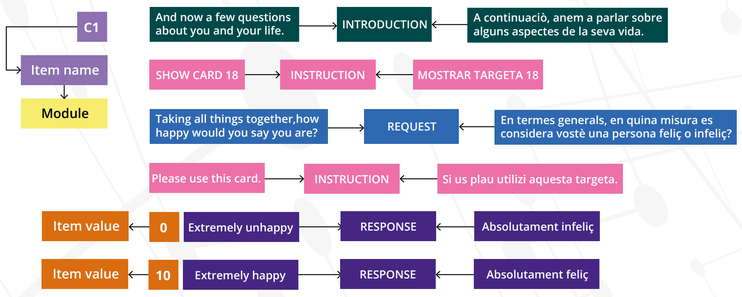
» MCSQ is hosted in a virtual machine at Universitat Pompeu Fabra, Barcelona which runs a Debian Linux OS
» The user interface of the MCSQ is a Flask application that runs on top of the ER database
» SQL alchemy library facilitates the manipulation of data and SQL objects in a high-level programming language
» Consultation with corpus linguists, survey practitioners, translators with experience in questionnaire translation, and computational linguists defined which functionalities should be implemented.
» Functionalities allow for data usage in real research contexts, such as questionnaire design, multilingual resources for domain-specific machine translation, translation verification, among others.
» The application encapsulates all queries to the database, hiding them from the users.
» Users build their queries by selecting the desired filters on a graphic interface.

Linguistic functionalities

» Word searches are possible both in aligned and non-aligned data, comprising partial, single and multiple word search. By applying the metadata filters, one can search for words or word sequences restricting results by item type (e.g., introductions, response options), language variation, year, study, etc.
» Word collocation functions allow searching for bigram or trigram collocations of a word, ranked by raw frequency. Two word collocations can be compared, for any subset of the data. Users can retrieve and compare up to 30 collocations.
» Word frequencies can be computed for single or multiple words and filtered by item type, language variation, year, and study.
» POS-tag sequence search is available, the corpus is POS annotated, which can be useful for analyzing syntactic patterns.
» Selection of any given subset of the corpus data to be visualized and downloaded fully customized datasets it's useful for expanding data analysis, or to design new questionnaires.
» Data resources can be downloaded as CSV files with tab separators
» Customized translation memories (TMX file) compatible with CAT tools. It is possible to create subsets of the data to build a translation memory by filtering the corpus according to language (and country variation) study and year.

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
» Approximately 88% aligned with respect to the source



Visualization of alignments with item type correspondence in MCSQ

Data annotation

» The MCSQ data is annotated with Part-of-speech (POS) and Named Entity Recognition (NER) tags.
» POS tags provide useful information about the syntax of the sentences, while NER tags identify entities (e.g. location, person, organization) present in text.
» We use the Universal Dependencies tagset for part-of-speech tags, which is homogeneous across languages.

FAIR Principles

Findable
» Rich metadata
» Code publicly available and findable through a persistent identifier
» Forthcoming permanent preservation in CLARIN ERIC repository

Accessible
» Data available in an open file format (CSV with tab separators)
» Data is safeguarded accessible via the interface

Interoperable
» Future formalization of the MCSQ data model in FAIRsharing
» MCSQ metadata is a simplified and adapted subset of the DDI codebook
» Adapted for CSV files instead of XML ones
» Includes linguistic metadata such as part-of-speech tags
» Universal POS tags for part-of-speech metadata tags

Reusable
» License for the MCSQ
» Remix, adapt, and build upon the work done in MCSQ
» Documentation and materials about the corpus with persistent identifiers in Zenodo

How to cite the MCSQ

The MCSQ 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*.

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

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