[MCSQ] The Multilingual Corpus of Survey Questionnaires

Lidun Hareide, Møreforsking Research Institute, Norway
Danielly Sorato, Universitat Pompeu Fabra, Spain
Diana Zavala-Rojas, Universitat Pompeu Fabra, Spain, ESS ERIC

European Survey Research Association (ESRA)
July 09, 2021
This project is funded from the EU Horizon 2020 Research and Innovation Programme (2014-2020) under Grant Agreement No. 823782.

Duration: 40 months (January 2019 – 30 April 2022)

Partners: 45 (20 beneficiaries + 25 LTPs)

SSH ESFRI Landmarks and Projects & international SSH data infrastructures

Project budget: €14,455,594.08

Type of action & funding:
Research and Innovation action (INFRAEOSC-04-2018)

Project website: www.SSHopencloud.eu

Objectives:

• creating the social sciences and humanities (SSH) part of European Open Science Cloud (EOSC)
• maximising re-use through Open Science and FAIR principles (standards, common catalogue, access control, semantic techniques, training)
• interconnecting existing and new infrastructures (clustered cloud infrastructure)
• establishing appropriate governance model for SSH-EOSC
[MCSQ]: The Multilingual Corpus of Survey Questionnaires

• The MCSQ is the first publicly available corpus of survey questionnaires

• Version 2 (Mileva Marić-Einstein): 263 distinct questionnaires from the ESS, EVS, and SHARE
  – More than 3.5 million words
  – ≈ 657,000 sentences

• Open access, searchable, aligned and annotated database
• FAIR (Findable, Accessible, Reproducible and Interoperable) by design

• A powerful instrument for the further development of best practice in design of source questionnaire and questionnaire translation methodologies

• Accessible at https://www.upf.edu/web/mcsq
What is a corpus?

• A collection of
  1. Machine readable
  2. Authentic texts
  3. Sampled to be
  4. Representative of a particular language/language variety/domain (e.g. Literary works, medical texts, etc.)

• «A significant advantage of the corpus linguistic method is that it allows for the analyst to approach the study of language from the context for the scientific method» - Geoffrey Leech
Languages included in the MCSQ

**Source language:** English localized for Great Britain

- **8 target languages adding to 30 language varieties:**
  - Catalan
  - Czech
  - French localized for France, Switzerland, Belgium and Luxembourg
  - German localized for Austrian, German, Swiss and Luxembourg
  - Norwegian localized for Bokmål
  - Portuguese localized for Portugal and Luxembourg
  - Spanish localized for Spain
  - Russian localized for Azerbaijan, Belarus, Estonia, Georgia, Israel, Latvia, Lithuania, Moldavia, Russia and Ukraine

- The MCSQ is representative of the *specialized language of surveys in the 8 languages, but not of the 8 languages in general*
A corpus of highly specialized text

- Questionnaires in the MCSQ follow the Ask the same question (ASQ method) and the translation teams should minimize adaptation.
- Any translation is expected to produce texts that are *functionally equivalent* for the purpose of statistical analysis.
- Concepts to be measured must be kept the same across languages
  - Keep the same psychometric properties and capture the same psychological variables (e.g. opinions and attitudes) across linguistic contexts (Harkness et al., 2010; Mohler & Johnson, 2010, Zavala-Rojas et al., 2018)
  - Low quality translations hamper data comparability and increase errors of measurement (Davidov & De Beuckelaer, 2010; Oberski et al., 2007).
MCSQ in numbers: sentences and tokens (words)

More than 3.5 million words in total
Visualizing the structure of survey items

A survey item can be decomposed into the following types:

And now a few questions about you and your life.

SHOW CARD 18

Taking all things together, how happy would you say you are?

Please use this card.

0 Extremely unhappy

10 Extremely happy

INTRODUCTION

REQUEST

INSTRUCTION

INSTRUCTION

RESPONSE

RESPONSE
Visualizing the alignment

- Sentence alignment in the MCSQ was conducted as a computational task that establishes the correspondence between a given sentence in a source language and its translation in the target languages.
- The sentence alignment allows the creation of translation memories (TMX format).

And now a few questions about you and your life.

SHOW CARD 18

INSTRUCTION

MOstrar targeta 18

Taking all things together, how happy would you say you are?

REQUEST

En termes generals, en quina mesura es considera vostè una persona feliç o infeliç?

Please use this card.

INSTRUCTION

Si us plau utilitzi aquesta targeta.

0

Extremely unhappy

RESPONSE

Absolutament infeliç

10

Extremely happy

RESPONSE

Absolutament feliç
Compilation framework

Pre processing
- Survey documents in several formats (e.g. PDF, spreadsheet, txt)
- Extract data from files
- OCR tool
- Manual work

Alignment
- Sentence Alignment
- Sentence Segmentation
- Metadata Attribution
- Structured CSV file

ER implementation and population
- Data included in ER model

Data annotation
- Annotation (POS tagging, NER, etc)

Publishing
- Searches, Statistics
- Annotated corpus

Compilation corpus catalogue
A snapshot of the MCSQ interface

- Looking for translations of “strongly agree” in French-France across all survey projects

- MCSQ allows for immediate identification of the texts, in this case, examples are from SHARE wave 8 and ESS Round 8 and Round 2

- Users can customize and download data
A snapshot of the MCSQ interface

Display data

- Asking MCSQ to display EVS 2008 questionnaire in British English showing Part of Speech tags
- MCSQ metadata allows for the identification of the type of text, name in the questionnaire
- Users can customize and download data
The TRAPD method

• The Translation, Review, Adjudication, Pretesting and Documentation (TRAPD, Harkness 2003) is an approach to translate questionnaires under the ASQ framework
  • Questionnaires in the MCSQ were translated using the TRAPD method.
  • Gold standard approach to survey translation.

• Human work intensive

• Translations are not necessarily harmonized across languages
  • Variations may reflect the teams choices and not necessarily linguistic differences
    • May hamper data comparability
    • Translation options multiply – hindering replicability
    • Managing, storing, analysing and reusing translation documentation is challenging
How can the MCSQ contribute to the TRAPD?

- **Searchable database**
  - Facilitates visualization and statistical analysis of previous translation decisions across languages
  - Tool for checking the translation of concepts across languages and surveys
- **Repository for previous rounds/waves of surveys**
  - Allows for the retrieval and preservation of source and translated questionnaires
  - Provides textual data for survey translation activities and research
  - Allows for the integration of translation analysis into the design of the source questionnaire
- **Valuable database for training new survey designers and translators**
- **Can be downloaded as a translation memory and used in a Computer Assisted Translation Tool**
The [MCSQ] shows some inconsistencies in translation that may hamper data comparability

Example:

- **Most people can be trusted.** (ESS R06)
- **(BE)** *La plupart des personnes sont dignes de confiance.* [Lit] (Most people are trustworthy.)
- **(CH)** *On peut faire confiance à la plupart des personnes.* [Lit] (One can trust most people.)
- **(FR)** *On peut faire confiance aux gens.* [Lit] (One can trust people.)

A more standardized approach to translation across countries and languages is needed to enhance comparability.

- The MCSQ was created to this end
To sum up: the [MCSQ] as a resource

- MCSQ is open source and open access
  - Follows FAIR (Findable, Accessible, Interoperable Reproducible) principles
  - Corpus data can be accessed and downloaded through the interface,
  - Website: https://www.upf.edu/web/mcsq

How to cite the MCSQ:

MCSQ is a valuable resource for

- Regional languages and language varieties (i.e. Catalan, Norwegian Bokmål, Swiss German)
- Facilitates visualization and statistical analysis of previous translation decisions across languages
- Cross-linguistic comparison of specialized use of survey language
- The creation of translation memories (TMX format)
  - Can be built and downloaded directly from the interface
  - Compatible with Computer Assisted Translation tools

- Website: https://www.upf.edu/web/mcsq
Thank you for your attention!

https://www.upf.edu/web/mcsq

lidun.hareide@moreforsking.no
danielly.sorato@upf.edu
diana.zavala@upf.edu

https://www.sshopencloud.eu

@SSHOpenCloud

info@sshopencloud.eu

/in/sshopencloud
Works cited


Hareide, L. (2013). *The Norwegian-Spanish Parallel Corpus*. [http://hdl.handle.net/11509/73](http://hdl.handle.net/11509/73)


