[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

SSHOC social sciences & humanities open cloud

Project: SSHOC
social sciences & humanities open cloud

Horizon 2020
European Union Funding
for Research & Innovation

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

Partners: 45
(20 beneficiaries + 25 LTPs)

SSH ESFRI Landmarks and Projects
& international SSH data infrastructures

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

Project budget:
€ 14,455,594.08

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**

This project is funded from the EU Horizon 2020 Research and Innovation Programme (2014-2020) under Grant Agreement No. 823782

SSHOC social sciences & humanities open cloud

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

This project is funded from the EU Horizon 2020 Research and Innovation Programme (2014-2020) under Grant Agreement No. 823782

SSHOC
social sciences & humanities open cloud

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**

This project is funded from the EU Horizon 2020 Research and Innovation Programme (2014-2020) under Grant Agreement No. 823782

SSHOC
social sciences & humanities open cloud

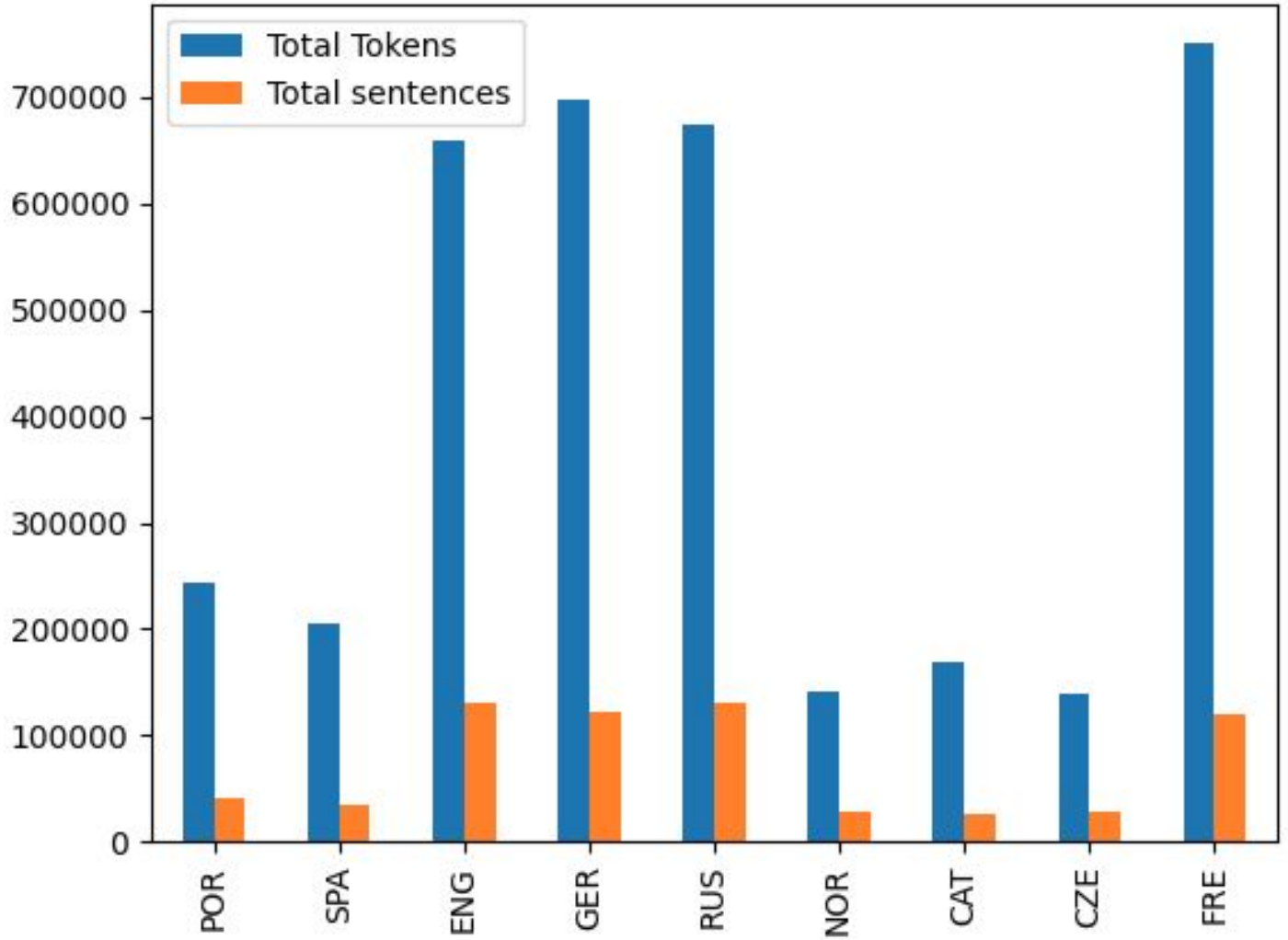
A corpus of highly specialized text

- 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).

This project is funded from the EU Horizon 2020 Research and Innovation Programme (2014-2020) under Grant Agreement No. 823782

SSHOC
social sciences & humanities open cloud

MCSQ in numbers: sentences and tokens (words)

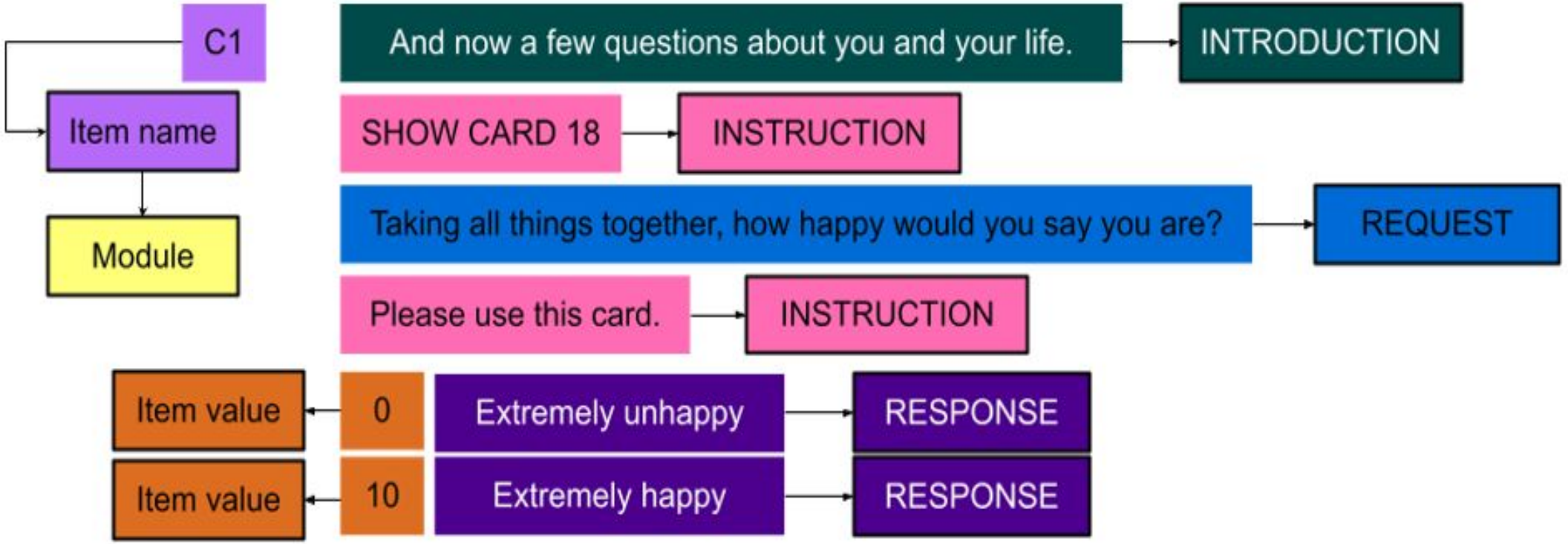


- More than 3.5 million words in total

This project is funded from the EU Horizon 2020 Research and Innovation Programme (2014-2020) under Grant Agreement No. 823782

Visualizing the structure of survey items

A survey item can be decomposed into the following types:



This project is funded from the EU Horizon 2020 Research and Innovation Programme (2014-2020) under Grant Agreement No. 823782

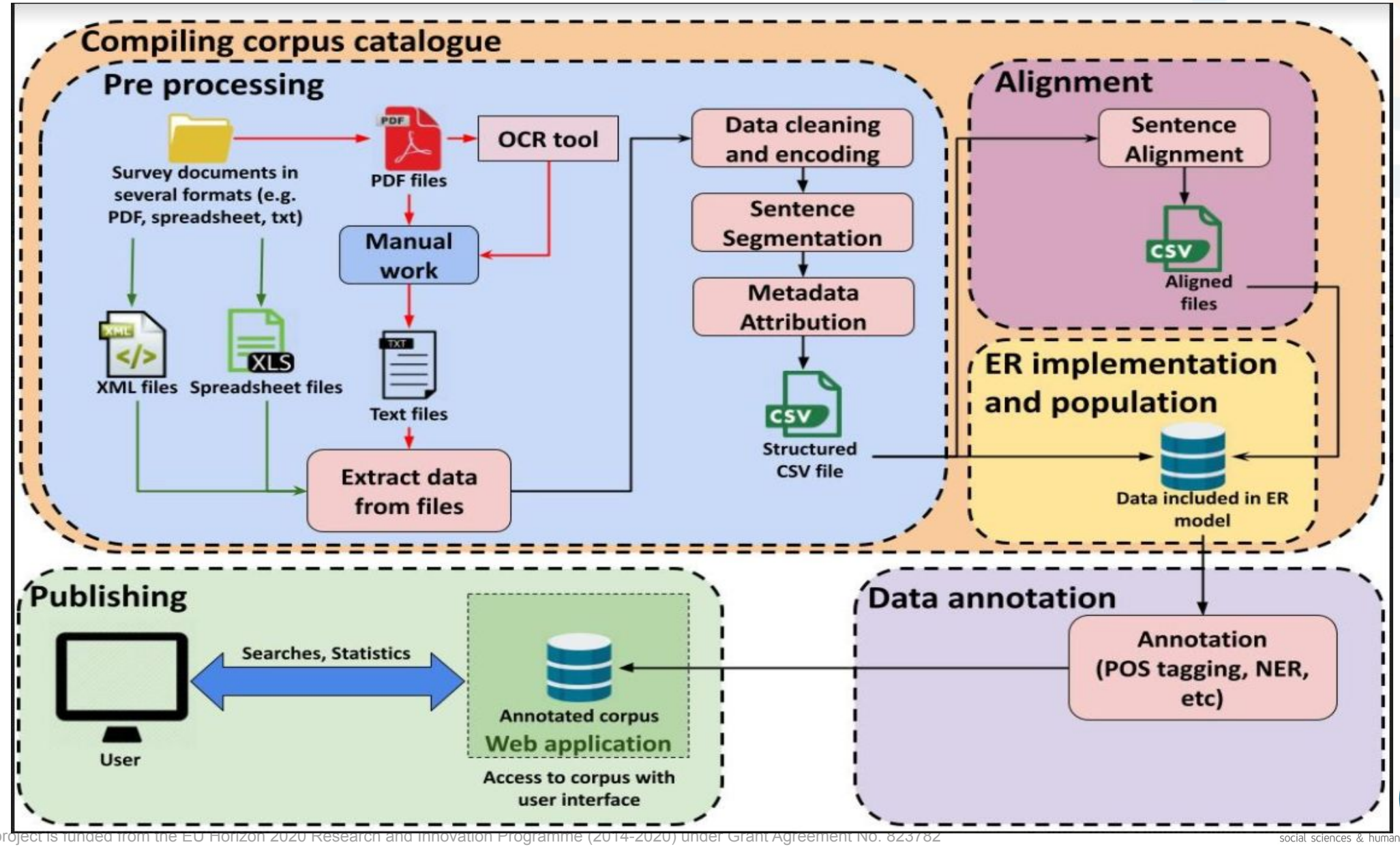
Visualizing the alignment

- 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)

This project is funded from the EU Horizon 2020 Research and Innovation Programme (2014-2020) under Grant Agreement No. 823782

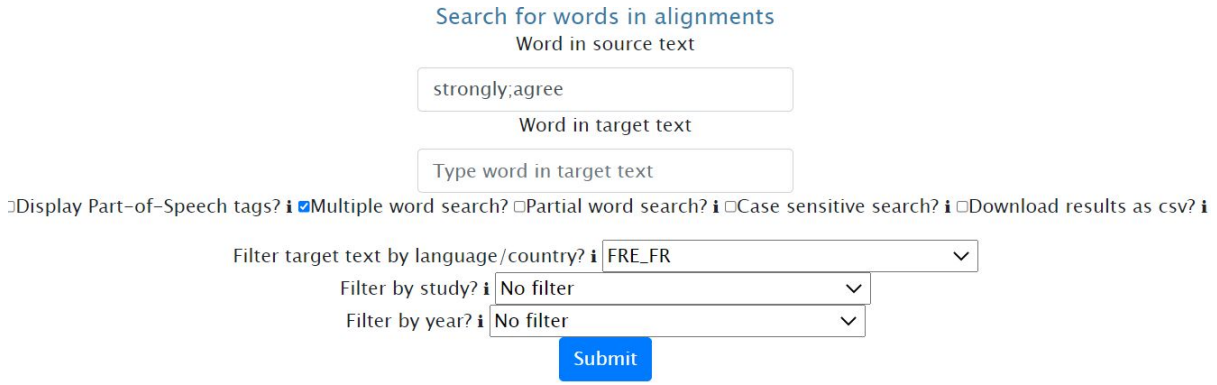
SSHOC
social sciences & humanities open cloud

Compilation framework



This project is funded from the EU Horizon 2020 Research and Innovation Programme (2014-2020) under Grant Agreement No. 823782

A snapshot of the MCSQ interface



Search results retrieved from MCSQ Alignment Collection

	source_survey_itemid	target_survey_itemid	Source Text	Target Text
0	SHA_R08_2019_ENG_SOURCE_999	SHA_R08_2019_FRE_FR_1058	Would you say you strongly agree, agree, disagree or strongly disagree?	(Diriez-vous que vous êtes tout à fait d'accord, d'accord, pas d'accord, pas du tout d'accord?)
1	ESS_R08_2016_ENG_SOURCE_791	ESS_R08_2016_FRE_FR_146	Agree strongly	Tout à fait d'accord
2	ESS_R02_2004_ENG_SOURCE_332	ESS_R02_2004_FRE_FR_349	Agree strongly	Tout à fait d'accord

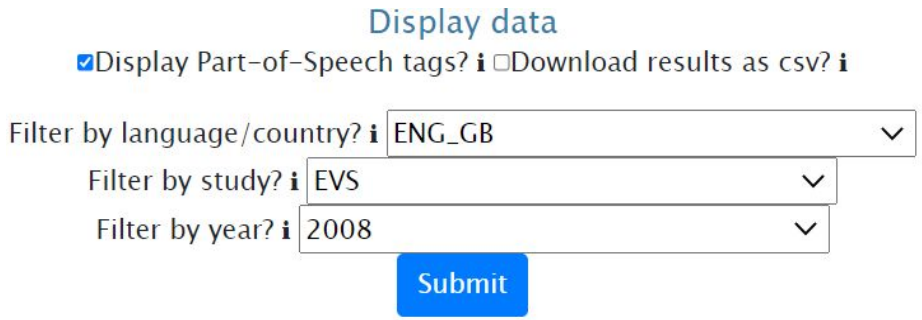
	A	B	C	D	E	F	G	H	I	J
1	source_survey_itemid	target_survey_itemid	Source Text	Target Text						
2	SHA_R08_2019_ENG_SOURCE_999	SHA_R08_2019_FRE_FR_1058	Would you say you strongly agree,	(Diriez-vous que vous êtes tout à fait d'accord, d'accord, pas d'accord, pas du tout d'accord?)						
3	ESS_R02_2004_ENG_SOURCE_348	ESS_R02_2004_FRE_FR_365	Agree strongly	Tout à fait d'accord						
4	ESS_R02_2004_ENG_SOURCE_356	ESS_R02_2004_FRE_FR_373	Agree strongly	Tout à fait d'accord						
5	ESS_R02_2004_ENG_SOURCE_898	ESS_R02_2004_FRE_FR_833	Agree strongly	Tout à fait d'accord						

- 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

This project is funded from the EU Horizon 2020 Research and Innovation Programme (2014-2020) under Grant Agreement No. 823782

SSHOC social sciences & humanities open cloud

A snapshot of the MCSQ interface



	survey_itemid	item_type	Text	POS Tagged Text	item_name	country_language
0	EVS_R04_2008_ENG_GB_0	INTRODUCTION	WE START WITH SOME QUESTIONS ABOUT LIFE IN GENERAL, LEISURE TIME ACTIVITIES AND WORK.	WE <PRON> START <VERB> WITH <ADP> SOME <DET> QUESTIONS <NOUN> ABOUT <ADP> LIFE <NOUN> IN <ADP> GENERAL <PROPN> , <PUNCT> LEISURE <NOUN> TIME <NOUN> ACTIVITIES <NOUN> AND <CCONJ> WORK <NOUN> . <PUNCT>	Q1	ENG_GB
1	EVS_R04_2008_ENG_GB_1	INSTRUCTION	Show card 1	Show <NOUN> card <NOUN> 1 <NUM>	Q1	ENG_GB
2	EVS_R04_2008_ENG_GB_2	REQUEST	Please say, for each of the following, how important it is in your life.	Please <INTJ> say <VERB> , <PUNCT> for <ADP> each <DET> of <ADP> the <DET> following <VERB> , <PUNCT> how <ADV> important <ADJ> it <PRON> is <VERB> in <ADP> your <PRON> life <NOUN> . <PUNCT>	Q1	ENG_GB
3	EVS_R04_2008_ENG_GB_3	REQUEST	Work	Work <NOUN>	Q1a	ENG_GB
4	EVS_R04_2008_ENG_GB_4	RESPONSE	very important	very <ADV> important <ADJ>	Q1a	ENG_GB
5	EVS_R04_2008_ENG_GB_5	RESPONSE	quite important	quite <ADV> important <ADJ>	Q1a	ENG_GB
6	EVS_R04_2008_ENG_GB_6	RESPONSE	not important	not <ADV> important <ADJ>	Q1a	ENG_GB
7	EVS_R04_2008_ENG_GB_7	RESPONSE	not at all important	not <ADV> at <ADV> all <ADV> important <ADJ>	Q1a	ENG_GB
8	EVS_R04_2008_ENG_GB_8	RESPONSE	Don't know	Do <VERB> n't <ADV> know <VERB>	Q1a	ENG_GB

survey_itemid	item_type	Text	POS Tagged Text	item_name	country_language
EVS_R04_2008_ENG_GB_0	INTRODUCTION	WE START WITH SOME QUESTIONS	WE <PRON> START <VERB> WIT	Q1	ENG_GB
EVS_R04_2008_ENG_GB_1	INSTRUCTION	Show card 1	Show <NOUN> card <NOUN> 1	Q1	ENG_GB
EVS_R04_2008_ENG_GB_2	REQUEST	Please say, for each of the followin	Please <INTJ> say <VERB> , <PU	Q1	ENG_GB
EVS_R04_2008_ENG_GB_3	REQUEST	Work	Work <NOUN>	Q1a	ENG_GB
EVS_R04_2008_ENG_GB_4	RESPONSE	very important	very <ADV> important <ADJ>	Q1a	ENG_GB
EVS_R04_2008_ENG_GB_5	RESPONSE	quite important	quite <ADV> important <ADJ>	Q1a	ENG_GB
EVS_R04_2008_ENG_GB_6	RESPONSE	not important	not <ADV> important <ADJ>	Q1a	ENG_GB
EVS_R04_2008_ENG_GB_7	RESPONSE	not at all important	not <ADV> at <ADV> all <ADV> i	Q1a	ENG_GB
EVS_R04_2008_ENG_GB_8	RESPONSE	Don't know	Do <VERB> n't <ADV> know <VE	Q1a	ENG_GB
EVS_R04_2008_ENG_GB_9	RESPONSE	No answer	No <DET> answer <NOUN>	Q1a	ENG_GB

- 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

ement No. 823782

SSHOC
social sciences & humanities open cloud

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

This project is funded from the EU Horizon 2020 Research and Innovation Programme (2014-2020) under Grant Agreement No. 823782

SSHOC
social sciences & humanities open cloud

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**

This project is funded from the EU Horizon 2020 Research and Innovation Programme (2014-2020) under Grant Agreement No. 823782

SSHOC
social sciences & humanities open cloud

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

This project is funded from the EU Horizon 2020 Research and Innovation Programme (2014-2020) under Grant Agreement No. 823782

SSHOC social sciences & humanities open cloud

To sum up: the [MCSQ] as a resource

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

How to cite the MCSQ:

Zavala-Rojas, D., Sorato, D., Hareide, L., & Hofland, K. (forthcoming 2021).Multilingual Corpus of Survey Questionnaires: a tool for refining survey translation. *Meta: Journal Des Traducteurs.*

This project is funded from the EU Horizon 2020 Research and Innovation Programme (2014-2020) under Grant Agreement No. 823782

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**

This project is funded from the EU Horizon 2020 Research and Innovation Programme (2014-2020) under Grant Agreement No. 823782

SSHOC
social sciences & humanities open cloud

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

This project is funded from the EU Horizon 2020 Research and Innovation Programme (2014-2020) under Grant Agreement No. 823782

Works cited

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

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

Hareide, L., & Hofland, K. (2012). Compiling a Norwegian-Spanish parallel corpus. In M. Oakes & M. Ji (Eds.), *Quantitative methods in corpus-based translation studies: A practical guide to descriptive translation research* (pp. 75–114). John Benjamins Publishing.

Harkness, Janet A. 2003. "Questionnaire Translation."In *Cross-Cultural Survey Methods*, edited by Janet A. Harkness, F. J. R. van de Vijver, and P. P. Mohler, 35–56. Hoboken: Wiley & Sons.

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, B.-E. Pennell, & T. W. Smith (Eds.), *Survey Methods in Multinational, Multiregional, and Multicultural Contexts* (pp. 115–140). John Wiley & Sons, Inc. https://doi.org/10.1002/9780470609927.ch7

Mohler, P. P., & Johnson, T. P. (2010). Equivalence, Comparability, and Methodological Progress. In *Survey Methods in Multinational, Multiregional, and Multicultural Contexts* (pp. 17–29). John Wiley & Sons, Inc. https://doi.org/10.1002/9780470609927.ch2

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*. Acco.

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). Wiley Series in Survey Methodology. https://doi.org/https://doi.org/10.1002/9781118884997.ch17

Zavala-Rojas, D., Sorato, D., Hareide, L., & Hofland, K. (forthcoming 2021). Multilingual Corpus of Survey Questionnaires: a tool for refining survey translation. Meta: Journal Des Traducteurs.

This project is funded from the EU Horizon 2020 Research and Innovation Programme (2014-2020) under Grant Agreement No. 823782

SSHOC
social sciences & humanities open cloud