[MCSQ] The Multilingual Corpus of Survey Questionnaires

Danielly Sorato, Universitat Pompeu Fabra, Spain

Digital Approaches to Multilingual Text Analysis

January 27, 2022

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

[MCSQ]: The Multilingual Corpus of Survey Questionnaires

- **The MCSQ is the first publicly available corpus of survey questionnaires**

  - Version 3 (Rosalind Franklin): 306 distinct questionnaires from the European Social Survey (ESS), the European Values Study (EVS), Survey of Health Ageing and Retirement in Europe (SHARE), and the WageIndicator Survey (WIS)
    - More than 4 million words
    - ≅ 766.000 sentences
  - Questionnaires comprise more than 40 years of survey research from large-scale comparative survey projects that provide cross-national and cross-cultural data to the Social Sciences and Humanities (SSH)
  - **Open access, searchable, sentence aligned, and annotated (Pos-tagging e NER)**

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

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

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

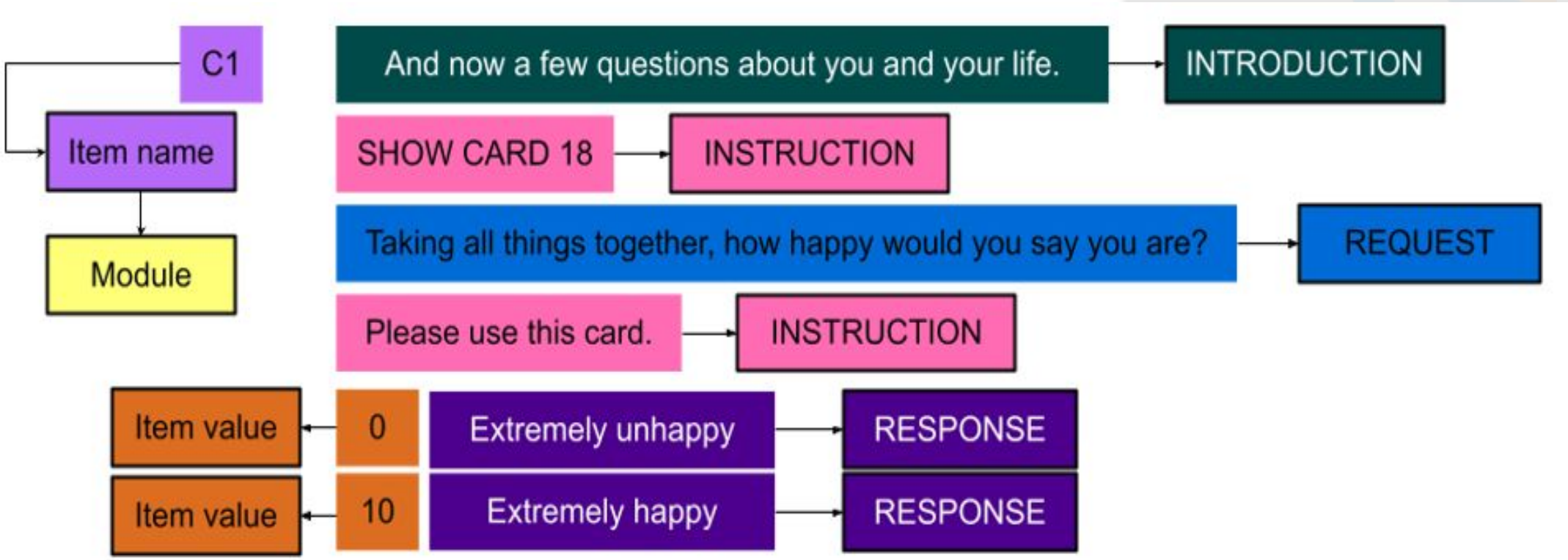
A corpus of highly specialized text

- Questionnaires in the MCSQ follow the Ask the same question (ASQ method)
  - 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 in order to capture the same psychological variables (e.g. opinions and attitudes)
- Low quality translations hamper data comparability and increase errors of measurement

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

Visualizing the structure of survey items



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

To sum up: the [MCSQ] as a resource

- MCSQ is open source and open access that follows **FAIR** (Findable, Accessible, Interoperable Reproducible) principles
- The MCSQ data is freely available in the CLARINO repository
  - https://repo.clarino.uib.no/xmlui/handle/11509/142
- The corpus and all its metadata are available for visualization and download through the MCSQ interface: http://easy.mcsq.upf.edu
  - Registration free of charge
  - Many interesting functionalities: word search with metadata filters, customization of translation memories, frequencies and collocations, etc
- Code
  - MCSQ compiling: https://github.com/dsorato/MCSQ_compiling
  - MCSQ interface: https://github.com/dsorato/MCSQ_interface

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

Thank you for your attention!

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

danielly.sorato@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