The Multilingual Corpus of Survey Questionnaires MCSQ

Dr. Lidun Hareide
Møreforsking Research institute, Norway

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

Compilation team

- Diana Zavala-Rojas – Universitat Pompeu Fabra, Spain
  - Specialist in international surveys and statistics
- Danielly Sorrato – Universitat Pompeu Fabra, Spain
  - Computer scientist, specialist in natural language processing
- Knut Hofland – formerly University of Bergen, Norway
  - Expert in corpus compilation, creator of ENPC, Norwegian Newspaper corpus
- Lidun Hareide – Møreforsking Research Institute, Norway
  - Corpus based translation studies, corpus linguistics and comparative linguistics, compiled NSPC (Hareide &Hofland 2012)

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: 48
(23 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

The MCSQ: the first multilingual corpus of international survey texts

- compiled from:

- the European Social Survey (ESS),
- the European Values Study (EVS)
- and the Survey of Health, Ageing and Retirement in Europe (SHARE)

- Open, searchable, aligned and annotated.

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:

Source language: British English

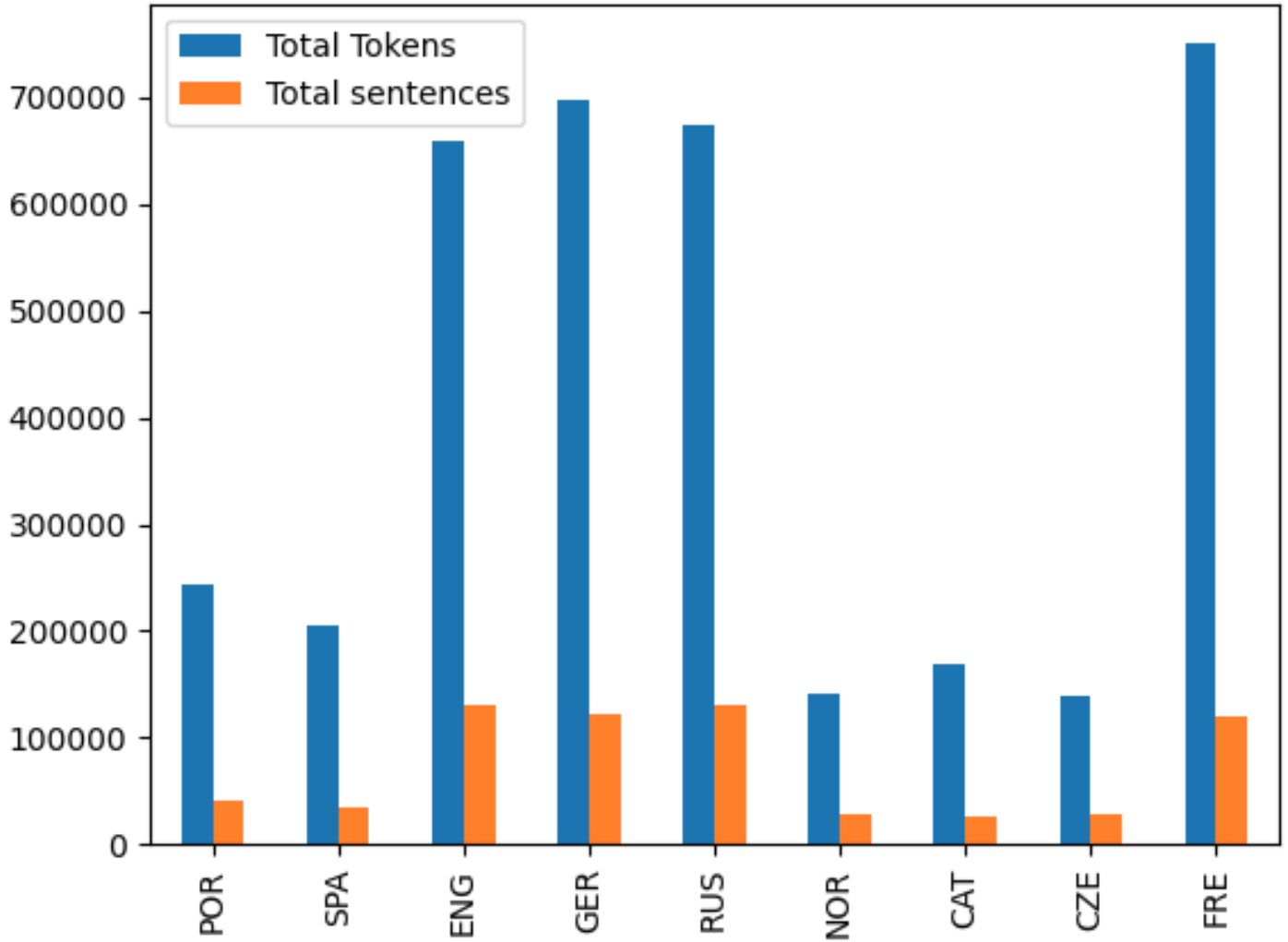
- 8 target languages: 29 target language varieties:

  - 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

MCSQ in numbers: sentences and tokens



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

The MCSQ: Vital data

Version 2 (Mileva Marić-Einstein):

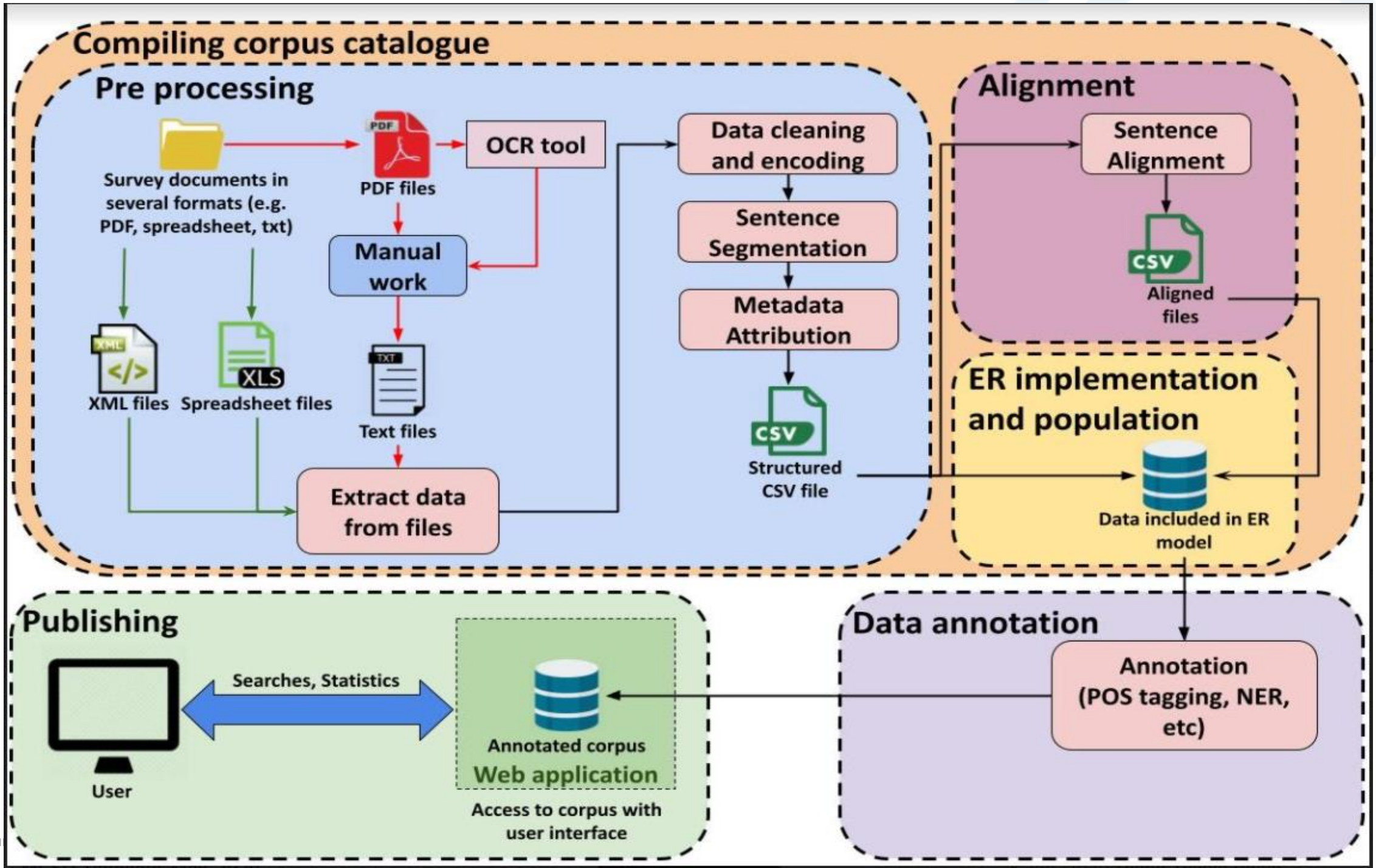
- 263 distinct questionnaires from the ESS, EVS, and SHARE
- 3.5 million words
- approximately 657 000 sentences.
- 80% is aligned – alignment is algorithm-based
- Annotated (POS – Universal dependencies tagset)
- Follows FAIR (Findable Accessible Interoperable Reproducible) principles
- Open access – open source
- TMX files: compatible with CAT 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

MCSQ framework

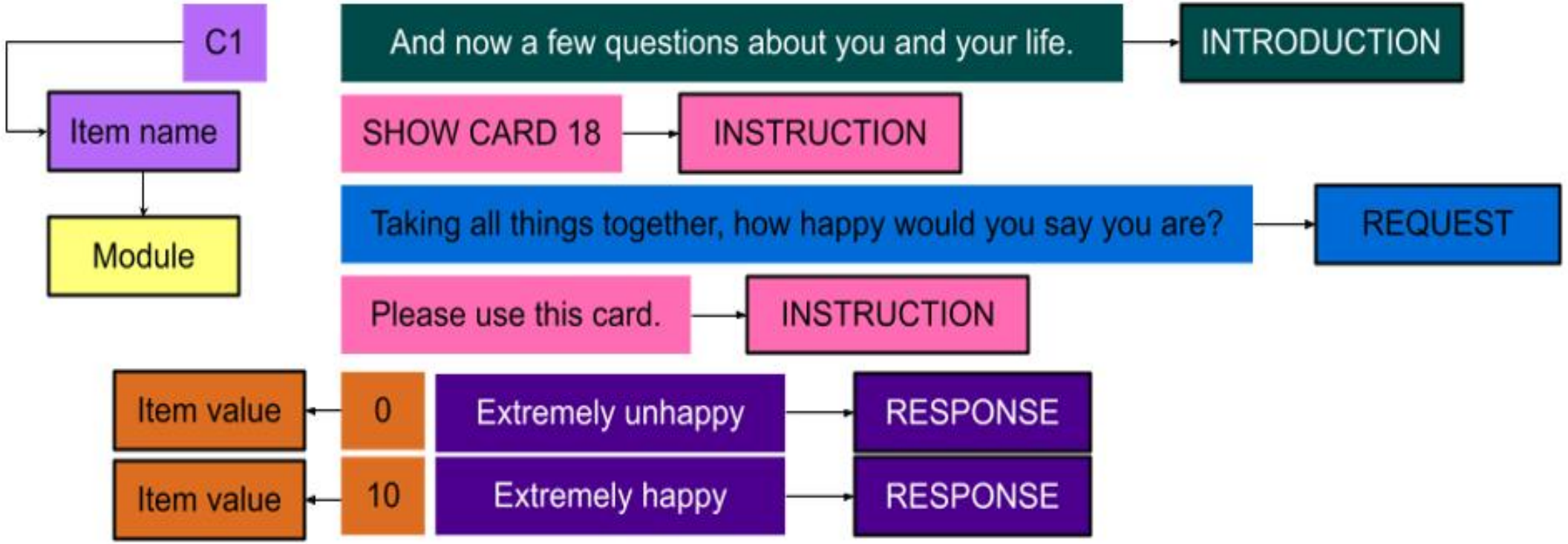
Adapted from González (2017)



This project is fu

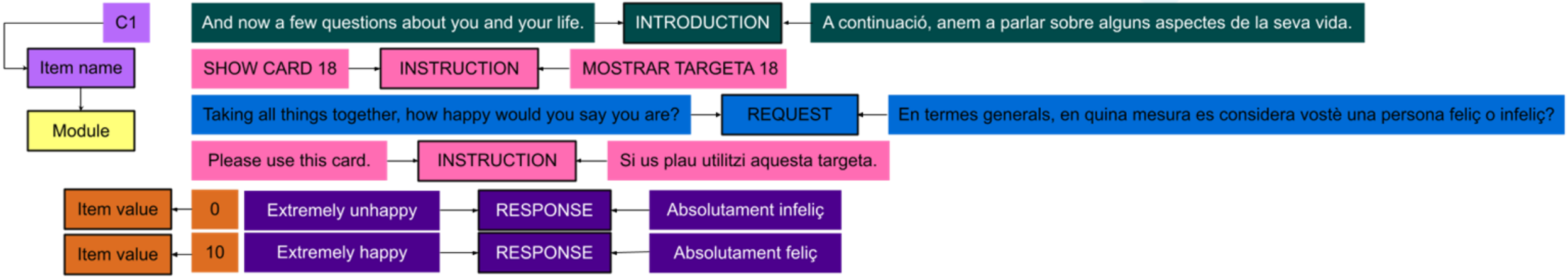
A corpus of highly specialized text

A survey item concists of the following segments



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

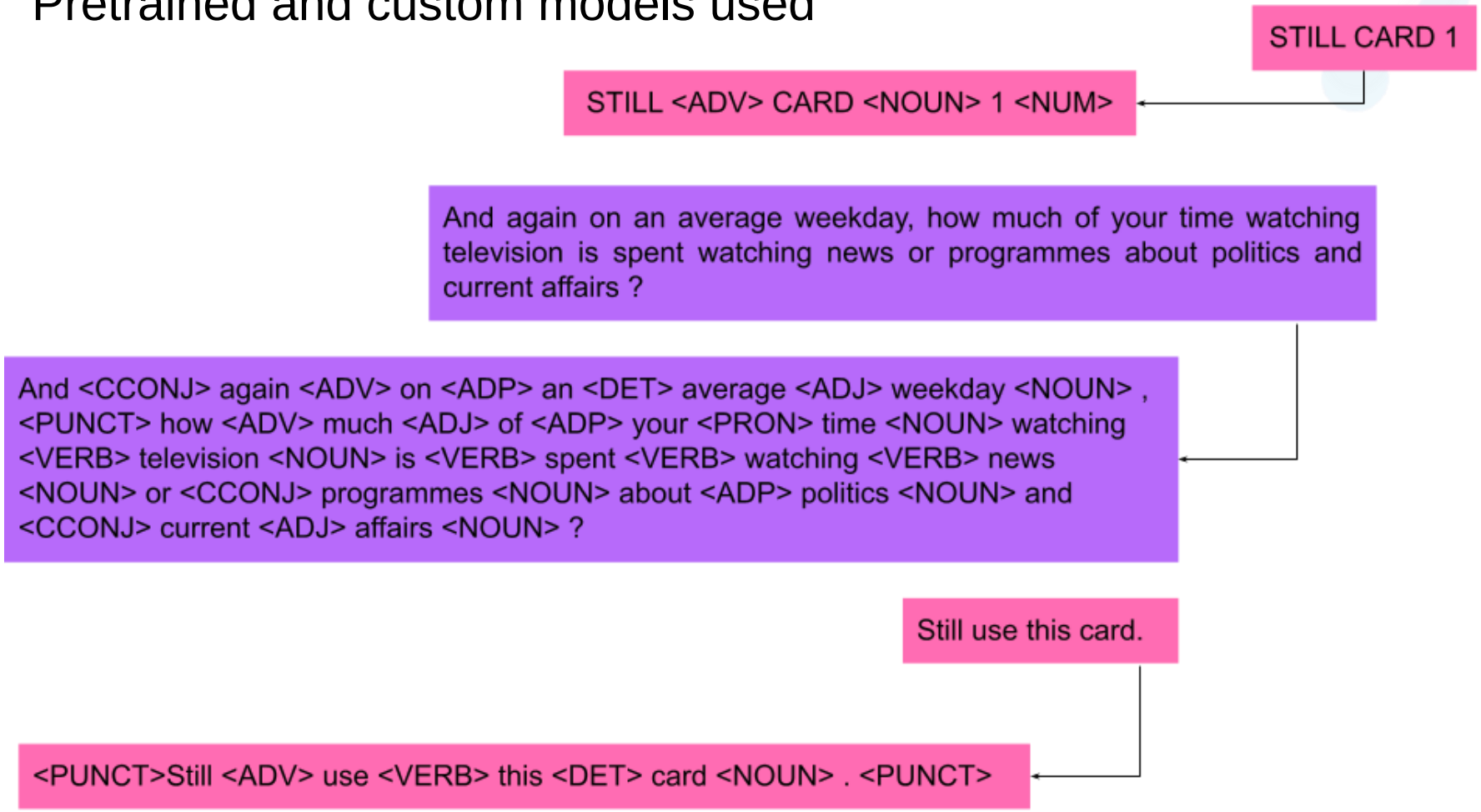
Visualization of the alignment



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

Annotation

- Part-of-speech tags
  - Universal Dependencies tagset
  - Pretrained and custom models used



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

Why is the MCSQ needed?

- Searchable database
  - Tool for systematic analysis of previous errors and successes in surveys
  - Tool for checking the translation of concepts across languages and surveys
- Repository for previous rounds/waves of surveys
  - Allows for retrieval and preservation of source and translated questionnaires
  - Provides textual data for for survey translation activities and research
  - Facilitates visualization and statistical analysis of previous translation decisions across languages
  - Allows for the integration of translation analysis into the design of the source questionnaire
- Valuable database for training new survey designers and translators

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

Ask the same question (ASQ method)

- Assumed to produce texts that are functionally equivalent for the purpose of statistical analysis.
- Concepts to be measured must be kept the same across languages
  - Need to be functionally equivalent for the purpose of statistical analysis
  - 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).
- The MCSQ allows for comparison across language varieties and surveys

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

Our case studies on the MCSQ show inconcistencies in translation that may hamper data comparability

Example: segment (ESS_R06_2012_ENG_Source_31):

- Most people can be trusted.
- (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

A valuable corpus resource also for

- “Smaller” languages and language varieties (i.e. Catalan and Norwegian Bokmål)
- Data to feed translation engines (machine translation)
- The study of the specialized language of surveys
- Analyzing linguistic patterns of survey items
- Analyzing survey translation
- Multilingual dictionary of survey terms
- Cross-linguistic comparison of specialized use of language

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: a powerful instrument for

- The further devellopment of best practice in design of source questionnaire and questionnaire translation methodologies.

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

- Official website https://www.upf.edu/web/mcsq/

- Open source
  - Github repository containing developed code https://github.com/dsorato/MCSQ_compiling
  - Technical documentation in Read the Docs https://mcsq-compiling.readthedocs.io/en/latest/

To cite the corpus:

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

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, 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

Thank you for your attention!

Join our community

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

SSHOC
social sciences & humanities open cloud