[MCSQ]: The Multilingual Corpus of Survey Questionnaires

Danielly Sorato

MSc in Computer Science

Researcher at RECSM

PhD candidate in Language and Translation Sciences

RECSM webinar - April 6, 2020

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

Outline

- General information and structure
- Alignment and Annotation
- Applications
- Design
- Access and Examples
- Limitations
- Next steps
- Conclusions

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

The Multilingual Corpus of Survey Questionnaires

- The Multilingual Corpus of Survey Questionnaires (MCSQ) is the first publicly available **corpus** of survey questionnaires
  - **Corpus** = large and structured language resource (text, audio). In this context, **survey items texts**

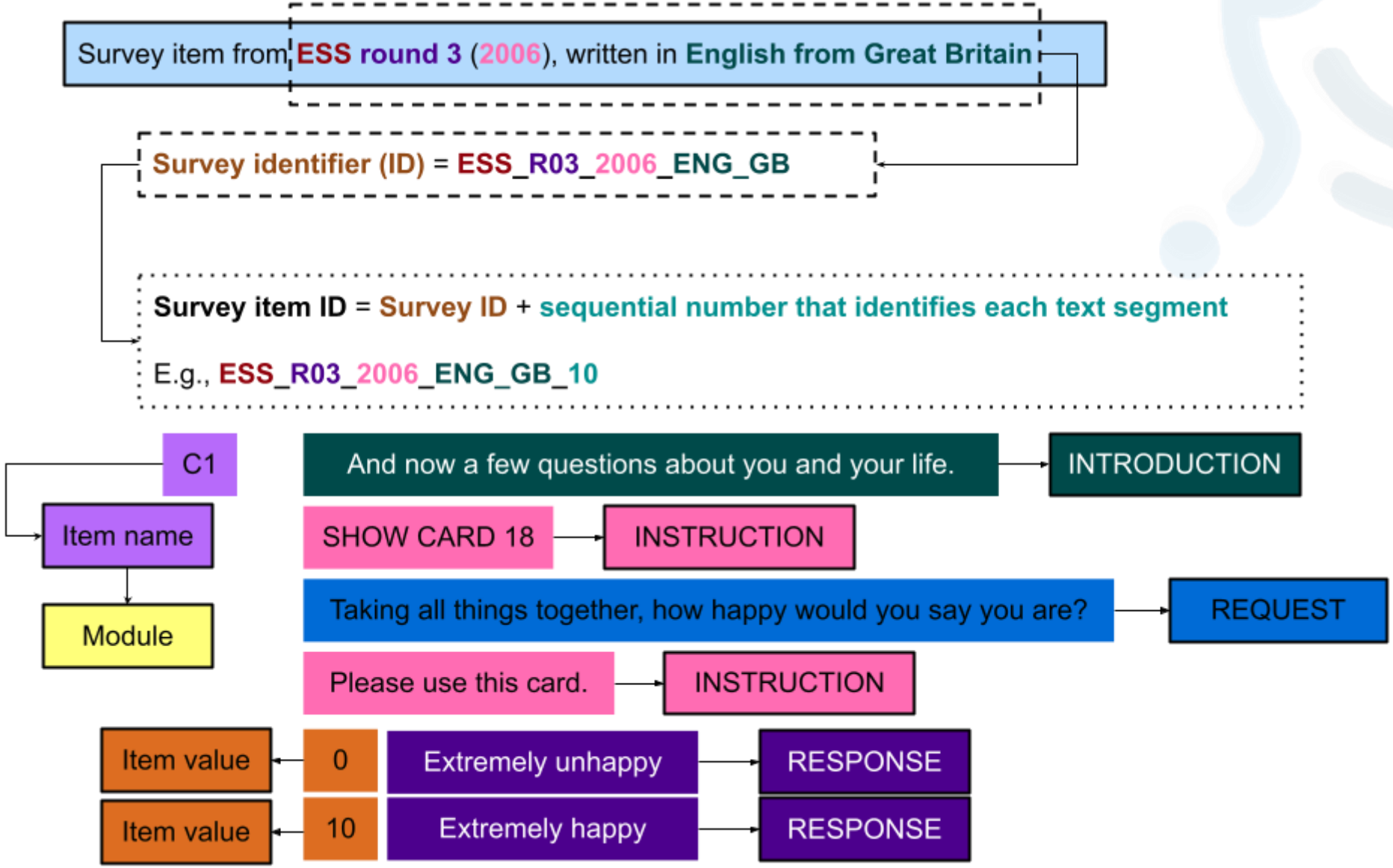
- Version 2 (Mileva Marić-Einstein): 263 distinct questionnaires from the ESS, EVS, and SHARE
  - More than 3.5 million words
  - ≅ 657.000 sentences

- English source and their translations into Catalan, Czech, French, German, Norwegian, Portuguese, Spanish and Russian, adding to 29 language varieties (e.g. French-Switzerland)

- Nearly 80% of the corpus is aligned
  - Source sentences (in English) are linked to their translations

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

And what is in there?



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

And how is it different from a PDF?

And now a few questions about you and your life.

C1 **CARD 18** Taking all things together, how happy would you say you are? Please use this card.

Extremely unhappy										**Extremely happy**	**(Don't know)**
00	01	02	03	04	05	06	07	08	09	10	88

REQUEST	C1		Taking all things together, how happy would you say you are?
INSTRUCTION	C1		Please use this card.
RESPONSE	C1	0	Extremely unhappy
RESPONSE	C1	10	Extremely happy
RESPONSE	C1	888	Don't know

Features	PDF file	MCSQ survey item
Machine readable?	No	Yes
Sentence segmented?	No	Yes
Item type indication?	No	Yes
Country, language, year, etc metadata?	No	Yes
Comparison	Only manual	Fully automatized

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

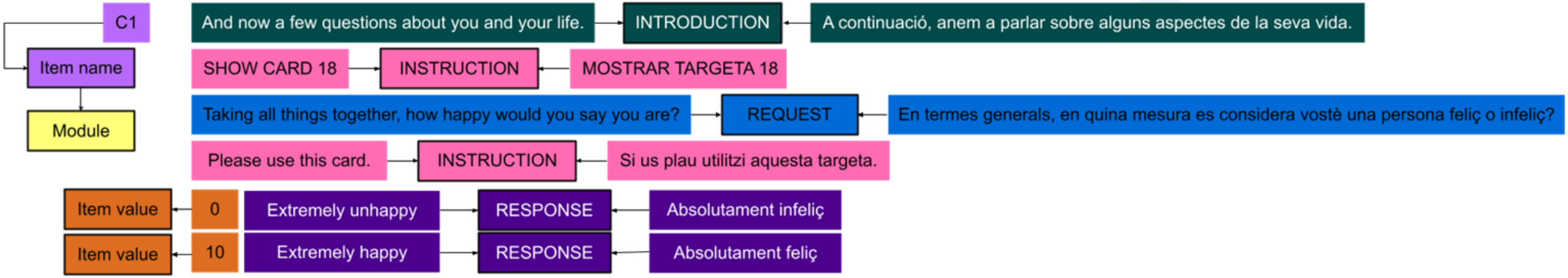
So there is the questionnaire text, what else?

module	item_type	item_name	item_value	source_text	target_text
A - Media; social trust	INSTRUCTION	A1		CARD 1	MONTREZ CARTE 1
A - Media; social trust	INSTRUCTION	A1		Please use this card to answer.	Veuillez utiliser cette carte pour répondre.
A - Media; social trust	REQUEST	A1		On an average weekday, how much time, in total,	Combien de temps passez-vous à regarder la télévision un jour de semaine habituel?
A - Media; social trust	RESPONSE	A1	0	No time at all	Pas du tout
A - Media; social trust	RESPONSE	A1	1	Less than ½ hour	Moins d'une demi heure
A - Media; social trust	RESPONSE	A1	2	½ hour to 1 hour	D'une demi heure à une heure
A - Media; social trust	RESPONSE	A1	3	More than 1 hour, up to 1½ hours	Plus d'une heure, jusqu'à une heure et demie
A - Media; social trust	RESPONSE	A1	4	More than 1½ hours, up to 2 hours	Plus d'une heure et demie, jusqu'à 2 heures
A - Media; social trust	RESPONSE	A1	5	More than 2 hours, up to 2½ hours	Plus de 2 heures, jusqu'à 2 heures et demie
A - Media; social trust	RESPONSE	A1	6	More than 2½ hours, up to 3 hours	Plus de 2 heures et demie, jusqu'à trois heures
A - Media; social trust	RESPONSE	A1	7	More than 3 hours	Plus de 3 heures
A - Media; social trust	RESPONSE	A1	888	Don't know	Ne sait pas

- Sentence alignment: finding the correspondence between a given sentence in a source language and its translation in a target language
- Country specific responses (about religion, education level, etc) excluded from alignments by design
- Approximately 80% of the corpus is aligned (concerning a total of ≅ 657.000 sentences)

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

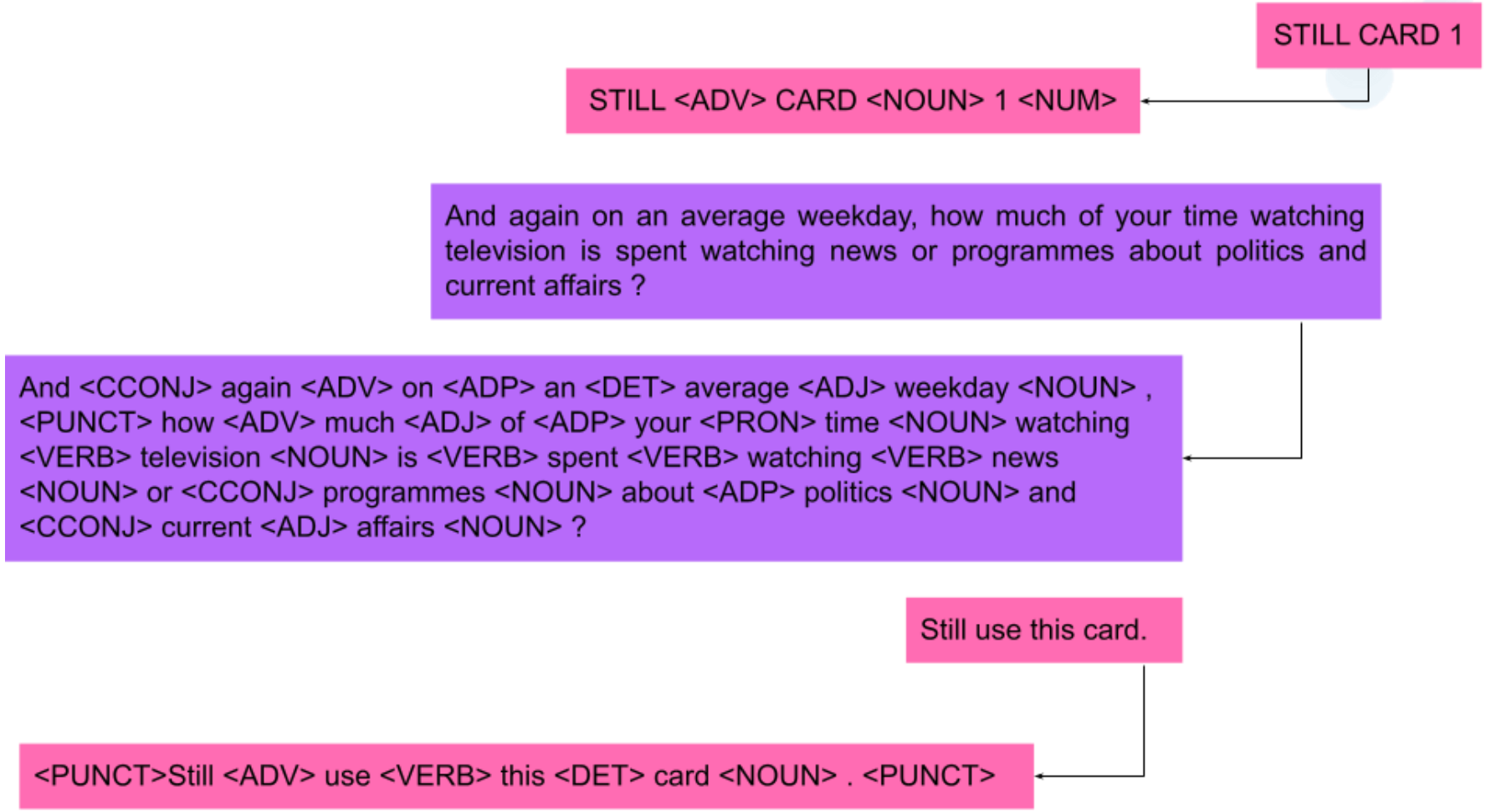
Visualizing the alignment



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



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

What could I do with MCSQ?

- Analyzing translations
- Translation memory
- Data to feed translation engines (machine translation)
- Bilingual dictionaries of survey terms
- Analyzing linguistic patterns of survey items
- (easily) Retrieving past question wordings to use as reference
- (easily) Comparing survey items

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

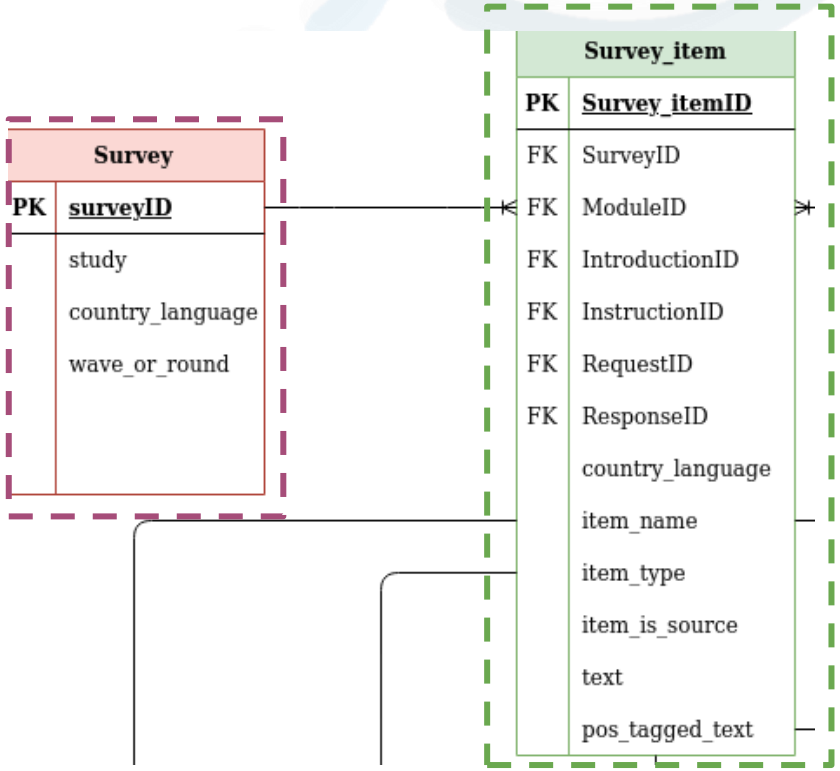
MCSQ design

Think about spreadsheets

- It is a **Entity Relationship** database
  - A representation of data as **tables (entities)** that have attributes (metadata) and **relationships** with other **tables (entities)**

- A **survey** is an **entity** that has a **relationship** with one or more **survey items**

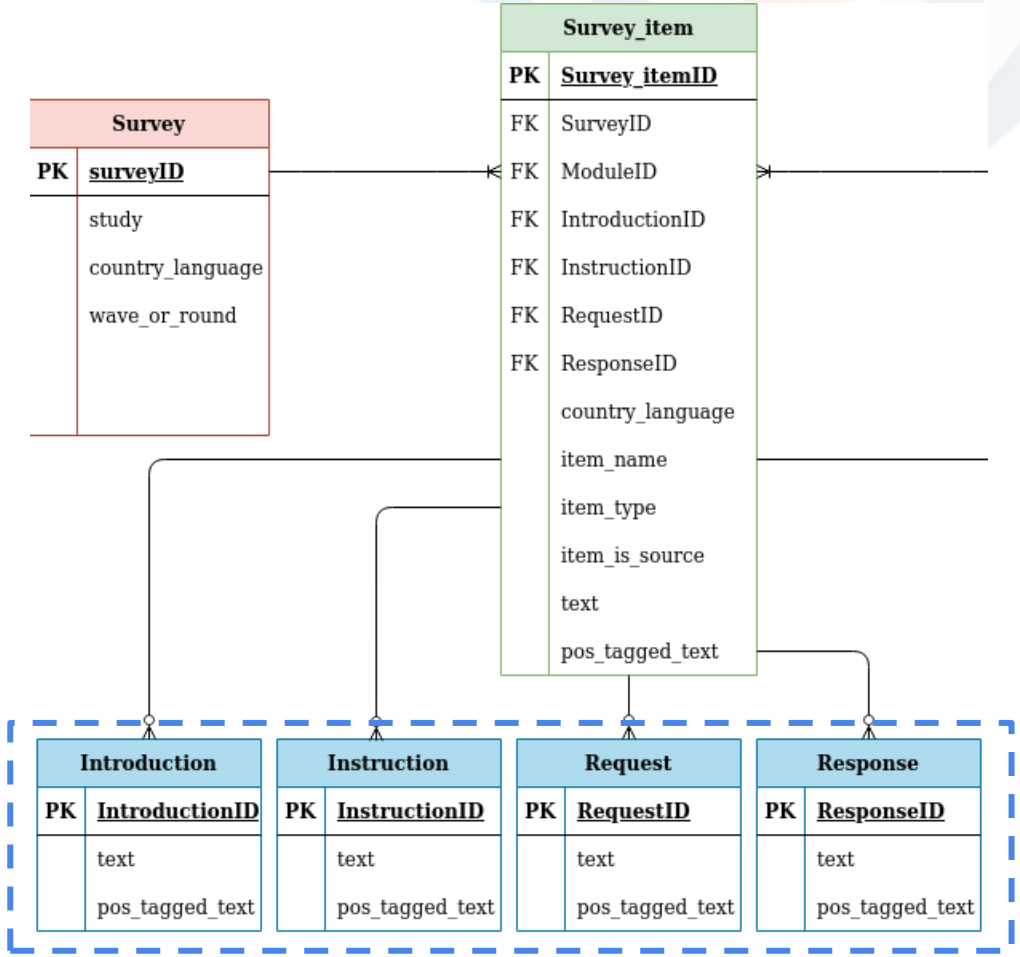
  - A survey is composed by one or more survey items



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

MCSQ design

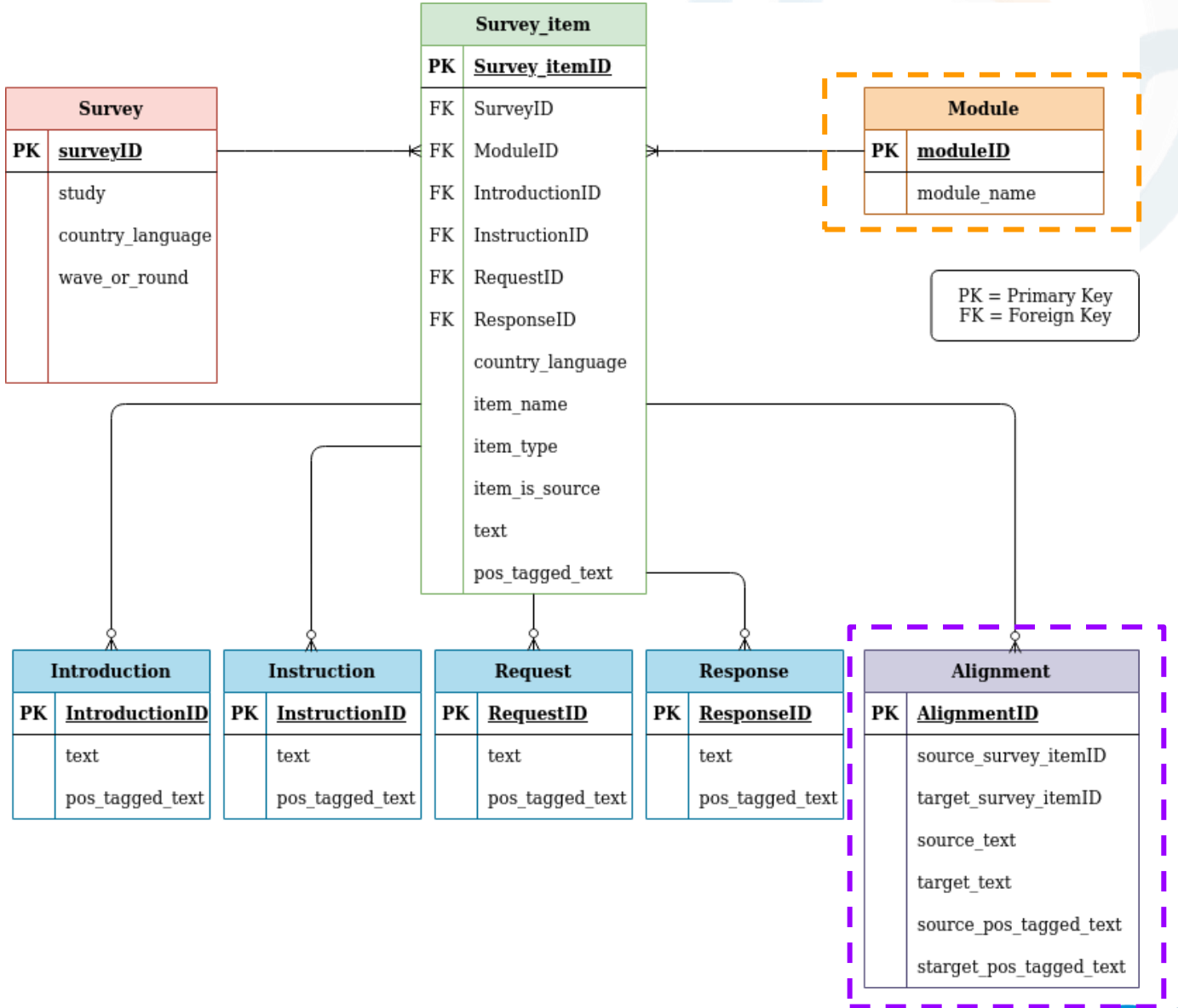
- In its turn, a **survey item** has a **relationship** with one or more **introduction, instruction, request and response entities**
  - a **survey item** can be decomposed into **introduction, instruction, request and response**



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

Complete ER diagram

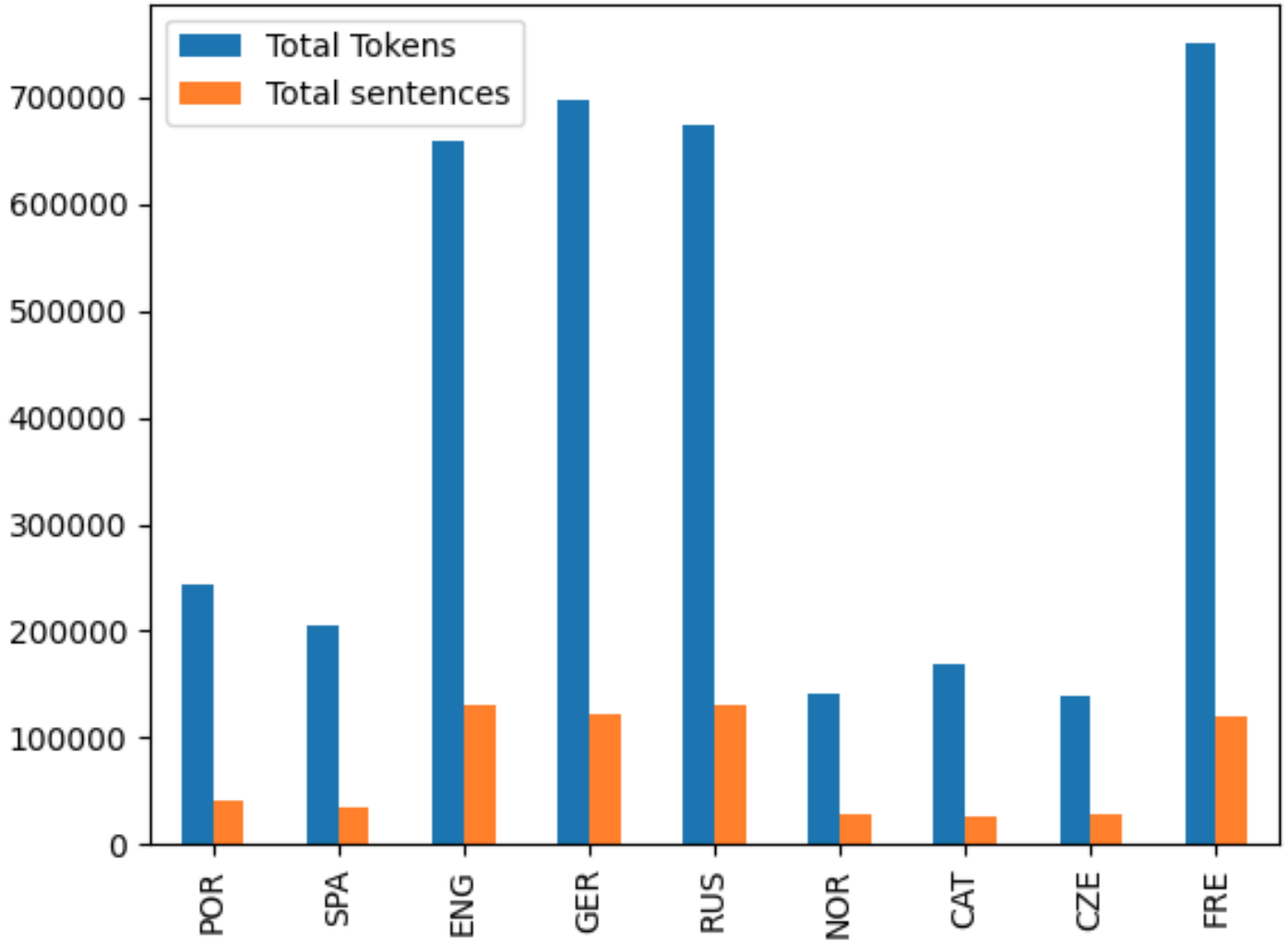
- Additional tables to store module and alignment information



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

MCSQ in numbers: sentences and tokens

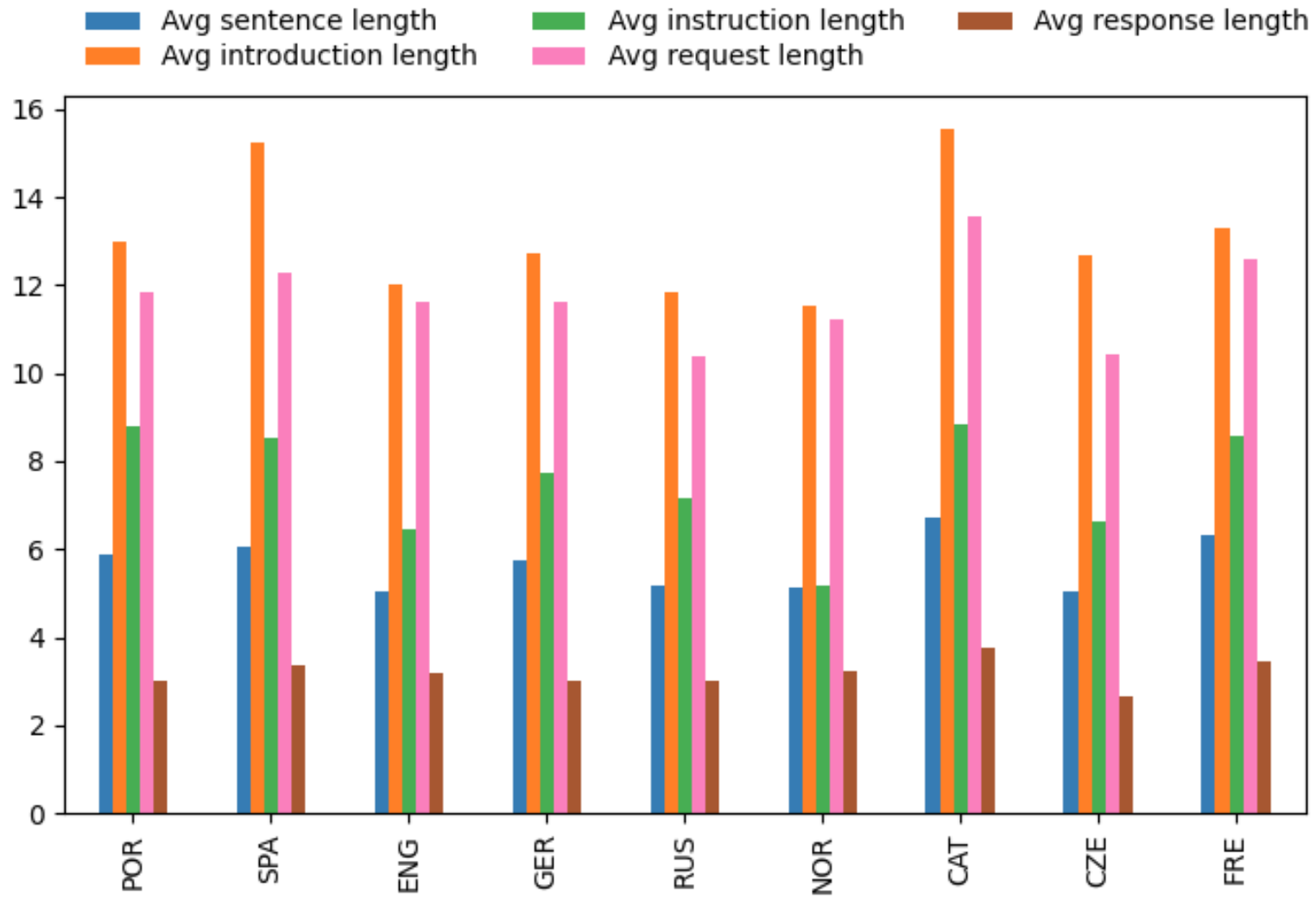
Think about words



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: average text segment length



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

How to cite and relevant links

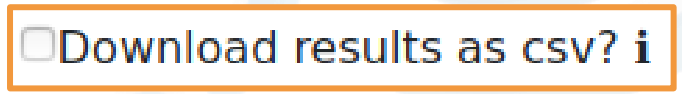
- Official website https://www.upf.edu/web/mcsq/
- Open source
  - Github repository containing developed code https://github.com/dsorato/MCSQ_compiling
  - Technical documentation in Read the Docs https://mcsq-compiling.readthedocs.io/en/latest/
- META paper

Zavala-Rojas, D., Sorato, D., Hareide, L., & Hofland, K. (forthcoming 2021). [MCSQ] Multilingual Corpus of Survey Questionnaires. Meta: Journal Des Traducteurs. @article{Zavala-Rojas,author = {Zavala-Rojas, Diana and Sorato, Danielly and Hareide, Lidun and Hofland, Knut},journal = {Meta: Journal des traducteurs},title = {{[MCSQ] Multilingual Corpus of Survey Questionnaires}}}

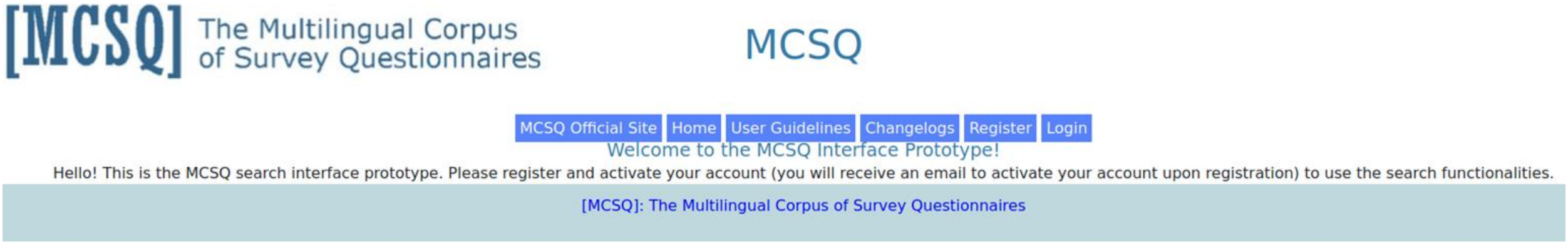
This project is funded from the EU Horizon 2020 Research and Innovation Programme (2014-2020) under Grant Agreement No. 823782

Accessing the data

- Preferably using the search interface (prototype stage) in http://easy.mcsq.upf.edu/
  - Free registration
  - Register and activate account to use functionalities



- Through email contact danielly.sorato@upf.edu
- Futurely in CLARIN repository



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

Interface main functionalities

- Word frequencies
  - to design language experiments, carry out research on lexical semantics, psycholinguistics, etc

- Collocations
  - provide information on word meaning and usage, following the idea that "you can know a word by the company it keeps".

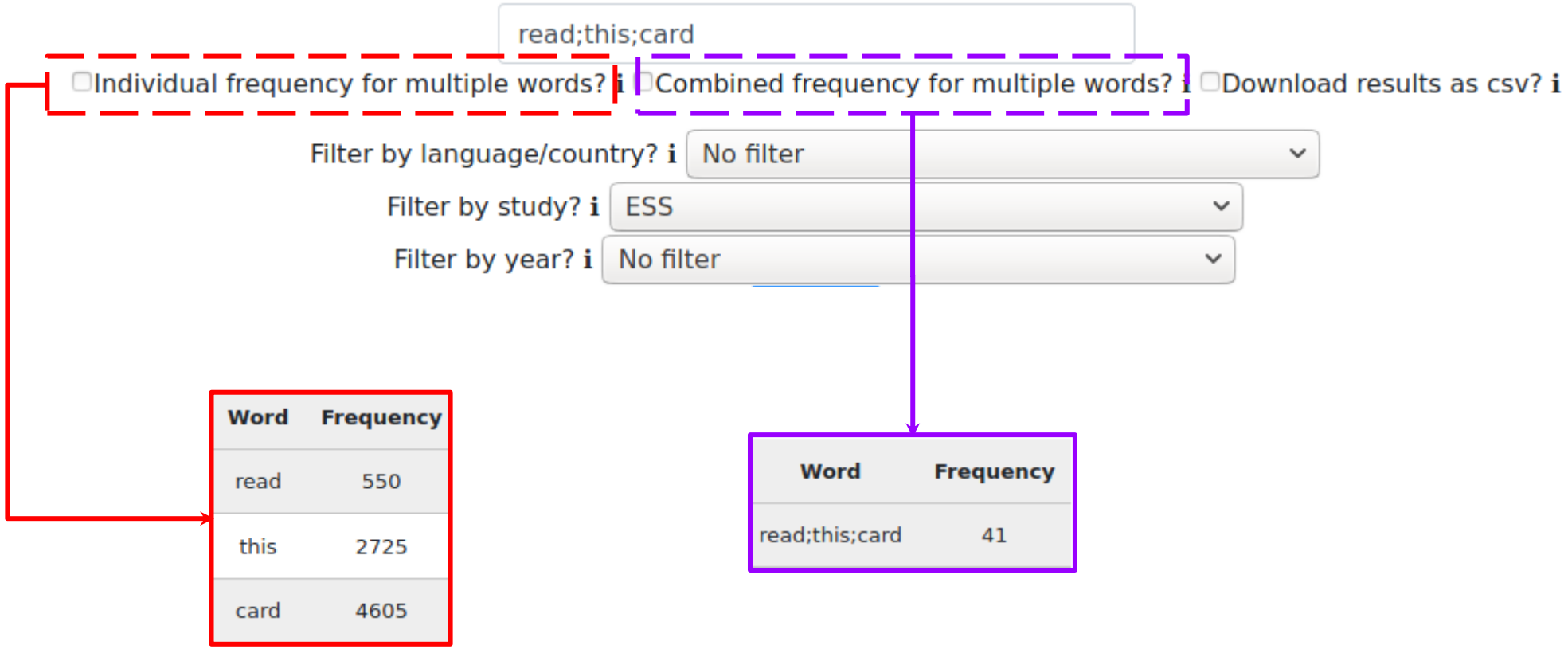
- Word searches
  - To get alignment information or to filter texts

- Compare survey items (up to 8 language varieties, same study/year)
  - To easily retrieve/compare survey items
  - By item type, word occurrence and whole questionnaire
  - **Not aligned!! If you want to see alignments use the alignment table**

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

Word frequency

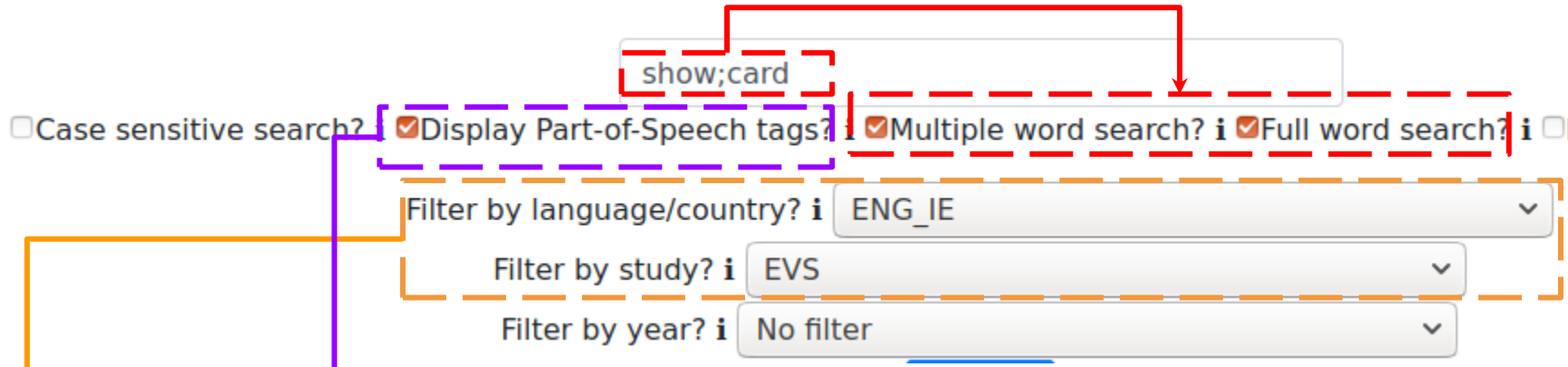
- Compute the frequency of the words ‘read’, ‘this’ and ‘card’ on ESS questionnaires



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

Word search example

- Locating all instruction segments in English from Ireland EVS questionnaires were the words “show card” appear

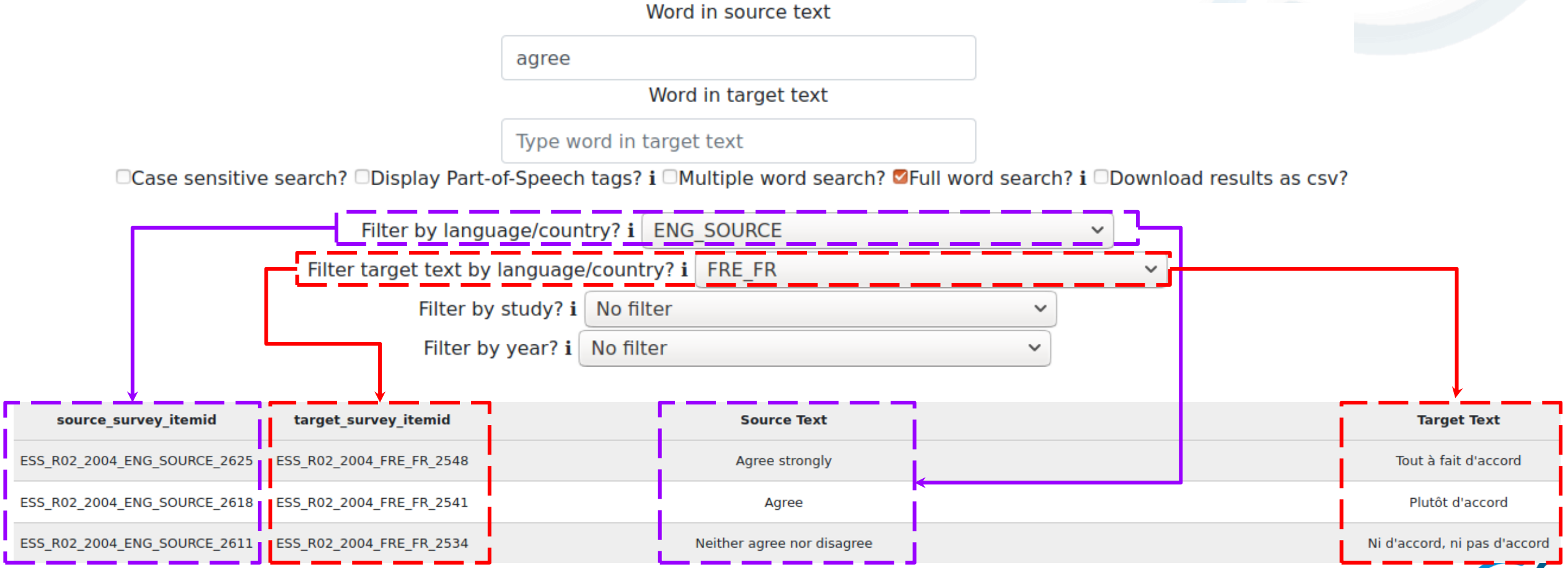


survey_itemid	Text	POS Tagged Text	item_name
EVS_R03_1999_ENG_IE_92	Show card 2	Show <NOUN> card <NOUN> 2 <NUM>	Q5a
EVS_R03_1999_ENG_IE_143	Show card 2	Show <NOUN> card <NOUN> 2 <NUM>	Q5b
EVS_R03_1999_ENG_IE_229	Show card 3	Show <NOUN> card <NOUN> 3 <NUM>	Q7
EVS_R03_1999_ENG_IE_352	Show card 4	Show <NOUN> card <NOUN> 4 <NUM>	Q11
EVS_R03_1999_ENG_IE_365	Show card 4	Show <NOUN> card <NOUN> 4 <NUM>	Q12
EVS_R03_1999_ENG_IE_376	Show card 5	Show <NOUN> card <NOUN> 5 <NUM>	Q13

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

Alignment search example

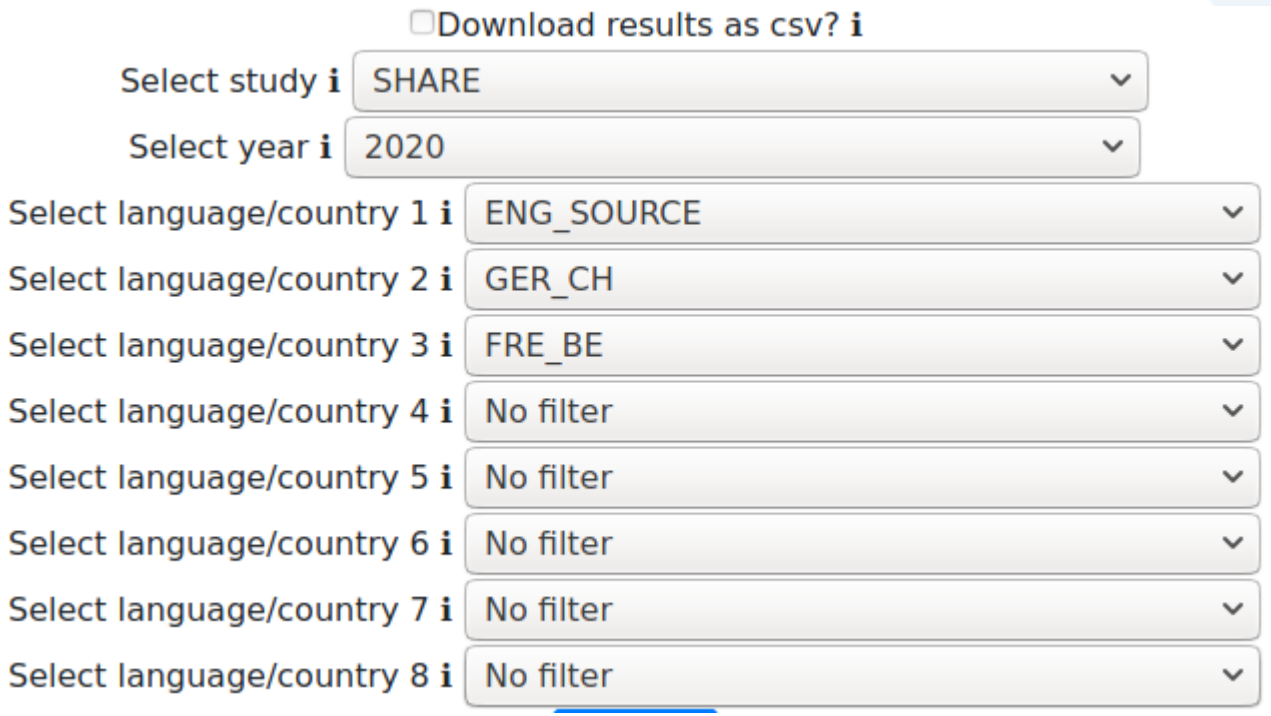
- Searching how the word ‘agree’ was translated to French (from France) questionnaires across all survey projects



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

Compare survey items example

- Comparing the SHARE COVID questionnaires in English, German (Switzerland) and French (Belgium)



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

Compare survey items example

- Comparing the SHARE COVID questionnaires in English, German (Switzerland) and French (Belgium)

survey_itemid	Text	item_name	item_type	survey_itemid	Text	item_name	item_type	survey_itemid	Text
SHA_COVID_2020_ENG_SOURCE_0	Some time ago, we sent you an invitation letter, which also included a data protection statement.	CAA001_	REQUEST	SHA_COVID_2020_GER_CH_0	Vor einiger Zeit haben wir Ihnen einen Einladungsbrief für diese Befragung geschickt.	CAA001_	REQUEST	SHA_COVID_2020_FRE_BE_0	Nous vous avons envoyé il y a quelque temps une lettre d'information sur SHARE qui incluait une déclaration sur la protection de la vie privée.
SHA_COVID_2020_ENG_SOURCE_1	Have you received the statement?	CAA001_	REQUEST	SHA_COVID_2020_GER_CH_1	Dort dabei ist auch eine Erklärung zum Datenschutz gewesen.	CAA001_	REQUEST	SHA_COVID_2020_FRE_BE_1	Avez-vous bien reçu cette déclaration?
SHA_COVID_2020_ENG_SOURCE_2	Yes	CAA001_	RESPONSE	SHA_COVID_2020_GER_CH_2	Haben Sie diese Erklärung erhalten?	CAA001_	REQUEST	SHA_COVID_2020_FRE_BE_2	Oui
SHA_COVID_2020_ENG_SOURCE_3	No	CAA001_	RESPONSE	SHA_COVID_2020_GER_CH_3	Ja	CAA001_	RESPONSE	SHA_COVID_2020_FRE_BE_3	Non
NaN	NaN	NaN	NaN	SHA_COVID_2020_GER_CH_4	Nein	CAA001_	RESPONSE	NaN	NaN
SHA_COVID_2020_ENG_SOURCE_4	In this case, I will then summarise the most important points of the statement for you.	CAA002_	REQUEST	SHA_COVID_2020_GER_CH_5	In diesem Fall werde ich die wichtigsten Punkte der Erklärung für Sie zusammenfassen.	CAA002_	REQUEST	SHA_COVID_2020_FRE_BE_4	Dans ce cas, je vais vous en résumer les points les plus importants.

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

Limitations

- Alignments are not manually checked
- Routing instructions and interviewer notes excluded by design
- Interface is in prototype stage
- No new corpus data after last iteration (not a MCSQ exclusive limitation)
  - By design could grow indefinitely, but depends on funds

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

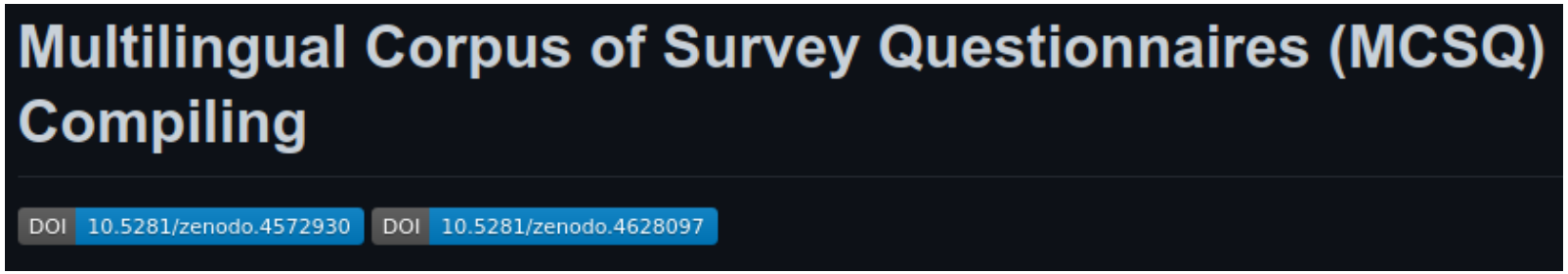
Next steps

- Adding more data
- New annotation (Named Entity Recognition)
- New interface functionalities
- Permanent archiving in CLARIN repository

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

Conclusion

- MCSQ is a multilingual corpus (9 languages) of survey questionnaires
- Survey items stored as structured data with valuable metadata and annotations
- It is open-source and open access (from scratch)
- Follows FAIR (Findable Accessible Interoperable Reproducible) principles



- News about the corpus are posted in the official webpage: 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

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