The Multilingual Corpus of Survey Questionnaires Query Interface

Danielly Sorato
Universitat Pompeu Fabra
Barcelona, Spain
danielly.sorato@upf.edu

Diana Zavala-Rojas
European Social Survey ERIC
Universitat Pompeu Fabra
Barcelona, Spain
diana.zavala@upf.edu

Abstract

The dawn of the digital age led to increasing demands for digital research resources, which shall be quickly processed and handled by computers. Due to the amount of data created by this digitization process, the design of tools that enable the analysis and management of data and metadata has become a relevant topic. In this context, the Multilingual Corpus of Survey Questionnaires (MCSQ) contributes to the creation and distribution of data for the Social Sciences and Humanities (SSH) following FAIR (Findable, Accessible, Interoperable and Reusable) principles, and provides functionalities for end-users that are not acquainted with programming through an easy-to-use interface. By simply applying the desired filters in the graphic interface, users can build linguistic resources for the survey research and translation areas, such as translation memories, thus facilitating data access and usage.

1 Introduction

In the last decade, research in the Social Sciences and Humanities (SSH) field has consolidated a paradigmatic transformation from theoretically-oriented approaches to data-driven methodologies (Kitchin, 2014; Manovich, 2011). Digitizing SSH data otherwise distributed as non-machine readable formats to convenient electronic ones is a decisive and necessary step in the SSH research life-cycle (Gómez et al., 2016). However, the digitization process of artifacts, documentation, and other research objects produces enormous amounts of data, which must be systematically stored and managed. This presents both a necessity and an opportunity to design and implement tools to manage and use such data in a FAIR (Findable, Accessible, Interoperable, and Reusable) way, allowing not only the preservation of SSH data but facilitating scientifically sound and transparent research processes.

Focusing on strengthening this paradigmatic transformation in the SSH field, the Multilingual Corpus of Survey Questionnaires (MCSQ) provides easy access to textual survey data in a machine-readable format and related services streamlined to the needs of the SSH community. In this context, we present the MCSQ interface: a tool that offers easy-to-use functionalities for end-users unacquainted with programming or data manipulation libraries. Other than common search functionalities offered by off-the-shelf corpus management tools such as IMS Open Corpus Workbench (Evert and Hardie, 2011) and Sketch Engine (Kilgarriff et al., 2014), the MCSQ interface provides features that are convenient for translators and survey practitioners, such as allowing users to build their own translation memories and facilitating the retrieval and comparison of multilingual data.

This paper is organized as follows. In Section 2, we briefly describe the MCSQ, and discuss how it follows the FAIR principles. Subsequently, in Section 3, we describe the interface infrastructure and its functionalities. Finally, in Section 4, we present our conclusions and future work.

2 The Multilingual Corpus of Survey Questionnaires

The Multilingual Corpus of Survey Questionnaires (MCSQ) is the first publicly available corpus of survey questionnaires. The MCSQ is an open-source and open-access resource that was designed and implemented following the FAIR principles. The current version 3 of the corpus (entitled Rosalind Franklin) contains approximately 4 million tokens and includes 306 distinct questionnaires designed in the source (British) English language and their translations into Catalan, Czech, French, German, Norwegian, Portuguese, Spanish, and Russian, adding to 30 country-language combinations (e.g., French from Switzerland). Questionnaires in the MCSQ comprise more than 40 years of

Proceedings of LaTeCH-CLfL 2021, pages 43–48
Punta Cana, Dominican Republic (Online), November 11, 2021.

survey research from large-scale comparative survey projects that provide cross-national and cross-cultural data to the SSH, namely, the European Social Survey (ESS), the European Values Study (EVS), the Survey of Health Ageing and Retirement in Europe (SHARE), and the WageIndicator Survey (WIS).

Within a questionnaire, a *survey item* (Saris and Gallhofer, 2014) is a unit for data collection. It includes the question, the response options, and other textual elements presented to the participants in the study. The questionnaires in the MCSQ were designed for in-person interviews. Except in the case of the WIS, the translation process was implemented according to the TRAPD (Translation, Review, Adjudication, Pretesting and, Documentation) method (Harkness et al., 2003), a team approach for the translation of survey questionnaires.

In the TRAPD method, two translators ('*T*' in the TRAPD acronym) should produce independent translations into the respective target language. Optionally, they could split a source questionnaire and work independently on one of its parts. In a team meeting with the translators, a reviewer ('*R*') assesses the two translations, and an adjudicator '*A*' takes the final decisions on the different translation options. Teams can opt for one of the translation options or create a new one by combining the strengths of each translation. The translated questionnaire is pretested before fieldwork ('*P*'), and the whole process is documented ('*D*'). Ideally, the team is composed of survey experts and language experts (e.g. professional translators) capable of assessing not only translatability aspects, but also necessary adaptations for the culture where it will be administered and to achieve the measurement objectives.

Questionnaires included in the MCSQ were obtained from the survey projects' archives in distinct formats such as spreadsheets, Extensible Markup Language (XML), and Portable Document Format (PDF) files. The PDF files had to undergo an additional step of conversion to plain texts before going through the preprocessing pipeline. Then, the texts were extracted from the input files and preprocessed, sentence aligned with respect to the English source and annotated with Part-of-speech (POS) and Named Entity Recognition (NER) tags.

During the preprocessing steps, in addition to text cleaning and normalization, we attributed metadata to the text segments to identify and navigate among the items in the MCSQ:

- The study or survey project, among ESS, EVS, SHARE or WIS;
- The round or wave, that is, a sequential number survey projects use to identify the edition of the study;
- The year in which the survey data and questionnaire were published;
- Language and country of the questionnaire. As the survey questionnaires texts are country-localized, this attribute helps to account for the language varieties included in the corpus;
- The name of each survey item, referred to as item name in the corpus;
- The item type, which indicates the role each sentence has in the questionnaire, among introduction, instruction (both for the participants and for the interviewers), request (or question) and response (or answer) (Saris and Gallhofer, 2014)
- The survey item ID, which is a combination of the study (S), round (R), year (Y), language (L), and country (C), with the following digits $SSS\_RRR\_YYYY\_LLL\_CC\_i$[1], where i is a sequential number that uniquely identifies and indicates the text segment order.

To align the survey items, we developed our own sentence alignment algorithm, which leverages metadata available in MCSQ to minimize the search space for candidates of a given alignment. Namely we use the module, item name, and item type metadata to this end. Additionally, we take into account the sentence length and information from domain specific bilingual dictionaries built with *Word2Word*(Choe et al., 2020) as alignment heuristics.

For the POS tagging task, we used *Flair* (Akbik et al., 2019) pre-trained models for the Czech, English, German, French, Norwegian, and Spanish languages, whereas for Catalan, Portuguese and Russian languages we trained our own models, also using the *Flair* framework. The Named Entity Recognition (NER) annotation was executed using pre-trained models from different sources. Namely,

[1] Language codes follow the ISO 639-2/B standard (three-digit standard) and country codes follow the ISO 3166 Alpha-2 standard (two digits).

Flair for English, German, French, and Spanish, *SpaCy*(Honnibal et al., 2020) for Catalan, Norwegian and Portuguese, and *Slavic BERT* (Arkhipov et al., 2019) for Czech and Russian. Due to the domain specificity and nature of the texts, some of the models (e.g., Catalan) performed worse than others, especially in instruction segments.

We developed the MCSQ aiming to be as FAIR (Findable, Accessible, Interoperable, and Reusable) as possible. To be *Findable*, we attributed the rich metadata described above. The code and data used to compile the corpus are publicly available and it has attached a persistent identifier [2]. Furthermore, the MCSQ will be submitted for permanent preservation to a CLARIN ERIC[3] repository in 2021, where the data will receive a persistent identifier.

In order to be *Accessible*, we made the MCSQ data available in open format (CSV with tab separators) and safeguarded accessibly, i.e. through user registration, via the interface hereby described[4] .

The *Interoperability* aspect of the MCSQ is a work in progress. Currently, our metadata is a simplified subset of the DDI codebook[5], adapted for CSV files instead of XML ones, and including linguistic metadata such as part-of-speech tags. We use the Universal POS tags[6] for part-of-speech metadata tags. Furthermore, the data model that describes and structures the MCSQ metadata will be submitted to FAIRsharing[7].

To ensure *re-usability* aspects, we are currently acquiring a license for the MCSQ that will allow users to remix, adapt, and build upon the work done in MCSQ. Moreover, we upload publicly available documentation and materials about the corpus with persistent identifiers[8][9][10] in the Zenodo repository[11].

Making data FAIR is a process. Although we consider that findability, accessibility, and reusability facets are well underway, interoperability is yet a work in progress and needs enhancement, such as producing mappings of the MCSQ metadata to the DDI standard.

3 The interface of the Multilingual Corpus of Survey Questionnaires

Aiming at combining searches in the MCSQ contents and functionalities that facilitate the usage of the MCSQ data for end-users, we developed the MCSQ interface[12]. To have complete control over the technology stack and freedom to include innovative functionalities, instead of using off-the-shelf corpus management tools, we developed our infrastructure from scratch. Users must register to access all functionalities in the interface[13]. The corpus and all its metadata are available for visualization and download through the interface.

3.1 Infrastructure

The MCSQ is hosted in a virtual machine provided by Universitat Pompeu Fabra, which runs a Debian Linux Operating System. The data is stored in an Entity-Relationship (ER) database, implemented in PostgreSQL[14]. The user interface of the MCSQ is a Flask application[15], that runs on top of the ER database. We use the SQLalchemy library (Bayer, 2012) to facilitate the manipulation of data and SQL objects in a high-level programming language. The interface is currently in the Alpha stage since new functionalities are still being added as we receive users' feedback. Nonetheless, it is fully functional.

3.2 Functionalities

In order to define which functionalities should be implemented, we consulted with corpus linguists, survey practitioners, translators with experience in questionnaire translation, and computational linguists. Our objective was to develop functionalities that would allow for data usage in real research contexts, such as questionnaire design, multilingual resources for domain-specific machine translation, translation verification, among others. The functionalities of the MCSQ interface can be divided into three main blocks, which are presented in the following subsections.

[2] doi.org/10.5281/zenodo.5153300

[3] https://www.clarin.eu/

[4] Permanent preservation in a CLARIN repository aims at improving accessibility as well

[5] https://ddialliance.org/Specification/DDI-Codebook/

[6] https://universaldependencies.org/u/pos/

[7] https://fairsharing.org

[8] doi.org/10.5281/zenodo.4785196

[9] doi.org/10.5281/zenodo.4555099

[10] doi.org/10.5281/zenodo.4696181

[11] https://zenodo.org/

[12] http://easy.mcsq.upf.edu

[13] The MCSQ interface is European Union GDPR compliant (General Data Protection Regulation), registration requires an email used exclusively for managing access to the MCSQ. Users about the purpose of the processing and the data controller. They can request to delete their account at any given time.

[14] https://www.postgresql.org

[15] https://flask.palletsprojects.com

3.2.1 Linguistic

The linguistic functionalities in the MCSQ interface are those commonly found in other corpus query engines. This block of functionalities consists of (i) word searches; (ii) retrieving and comparing word collocations; (iii) word frequencies and; (iv) POS-tag sequence search. *Word searches* can be conducted both in aligned and non-aligned data, comprising partial[16], single and multiple word search. By applying the metadata filters, one can search for words or word sequences restricting results by item type (e.g., introductions, response options), language variation, year, study, etc.

The *word collocation* functions allow searching for bigram or trigram collocations of a word, ranked by raw frequency. It is possible to compare the collocations of two words, being that the words can be selected from any sub-portion of the corpus. The interface allows users to retrieve and compare up to 30 collocations. The *word frequencies* can be computed for single or multiple words. Word frequency results can be filtered by item type, language variation, year, and study. Finally, *POS-tag sequence search* is available as well, since the corpus is POS annotated, which can be useful for analyzing syntactic patterns.

The application encapsulates all queries to the database, hiding them completely from the users. Hence, users build their queries by simply selecting the desired filters in the graphic interface, which facilitates the access and use of the data. An example of the aforementioned filters is shown in the search for word usage in the sentence aligned data, in Figure 1. This functionality allows users to search for source andor target words in the aligned sentences of the MCSQ to see how the source segments were translated to the target languages, filter results by metadata, see POS-tagging and NER annotations, and optionally, download the results.

Most of the results, except for the the ones that are exclusively downloadable, are presented as depicted in Figure 2. In addition to the text and POS tag annotation, which is shown when the display annotation option is selected, the survey item IDs of the segments are also exhibited. The IDs simplify the identification of the texts by providing immediate reference to the study, edition and year of publication of every segment. In the case of functionalities that calculate statistics or information about the texts, such as frequencies and collocations, the results are shown in a table similar to the one depicted in Figure 3. The aforementioned functionalities allow researchers to analyze linguistic aspects of the questionnaires, for instance, to investigate linguistic patterns, or assess translation equivalence.

3.2.2 Data resources

The *data resource* functionalities allow for the selection of any given subset of the corpus data to visualize and/or download fully customized datasets, which is useful for expanding the analysis of the data, or use the texts as input to design new questionnaires. Mainly, data resources are downloaded as CSV files with tab separators, with the exception of downloadable files to build a translation memory. We specifically chose the CSV format because it is easy to use across different domains of research, other file formats such as XML require knowledge about parsing to be used. In the case of customized translation memories, a TMX file is built in the back-end and outputted for the user. Translators can customize their translation memories and upload them into a CAT tool. It is possible to create subsets of the data to build a translation memory by filtering the corpus according to (i) the language; (ii) the language variation (language/country); (iii) the study and; (iv) the year.

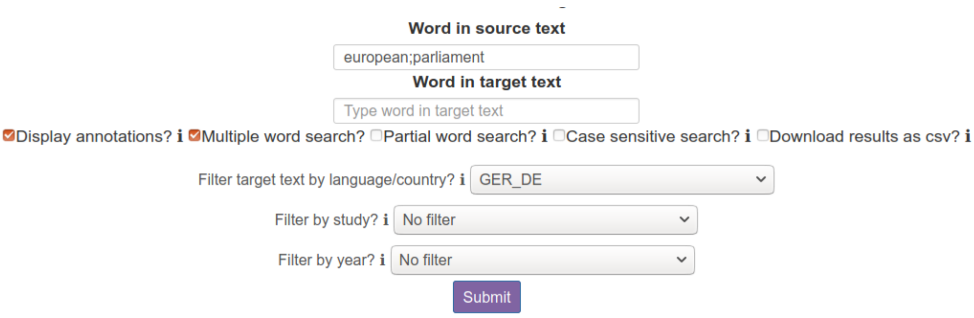
3.2.3 Comparison

This feature allows comparing questionnaires in multiple languages, without filters, by item type, or by word. Features to compare texts are useful for survey practitioners and questionnaire designers who can examine entire questionnaires in multiple languages, without having to search for and compare various files separately.

4 Conclusion

The MCSQ is a valuable contribution to open-source, open-access, and high-quality SSH data that follows FAIR principles, whilst its interface provides useful functionalities for researchers across different disciplines, such as linguists, survey practitioners, computational linguists, and translators. Aside from traditional corpus search functionalities, we developed features to support end-users that are not acquainted with programming, thus facilitating the accessibility of the data and related resources to all audiences. Through this work, we hope to stimulate the creation of FAIR

[16] A partial word search for the word “run” would retrieve “runs”, “running”, etc, for instance



Word in source text

european;parliament

Word in target text

Type word in target text

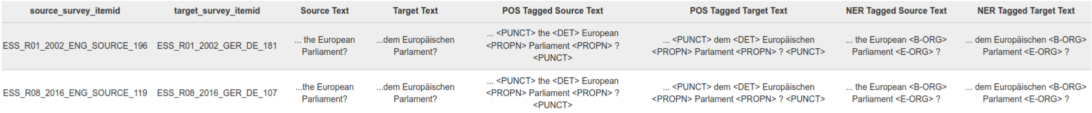
☑Display annotations? i ☑Multiple word search? □Partial word search? i □Case sensitive search? i □Download results as csv? i

Filter target text by language/country? i GER_DE

Filter by study? i No filter

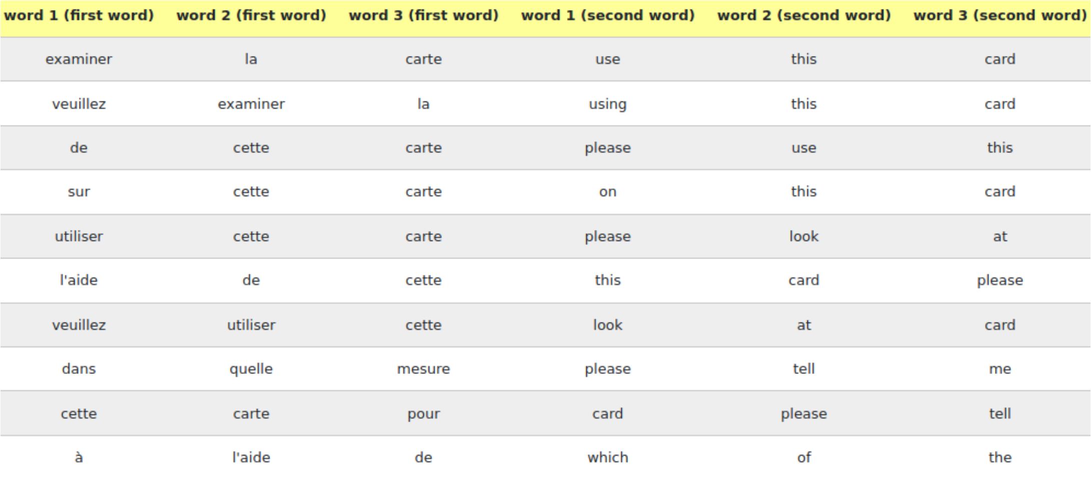
Filter by year? i No filter

Figure 1: Searching for translations of the words "European Parliament" in German from Germany MCSQ alignments.



source_survey_itemid	target_survey_itemid	Source Text	Target Text	POS Tagged Source Text	POS Tagged Target Text	NER Tagged Source Text	NER Tagged Target Text
ESS_R01_2002_ENG_SOURCE_196	ESS_R01_2002_GER_DE_181	... the European Parliament?	...dem Europäischen Parlament?	... <PUNCT> the <DET> European <PROPN> Parliament <PROPN> ? <PUNCT>	... <PUNCT> dem <DET> Europäischen <PROPN> Parlament <PROPN> ? <PUNCT>	... the European <B-ORG> Parliament <E-ORG> ?	... dem Europäischen <B-ORG> Parlament <E-ORG> ?
ESS_R08_2016_ENG_SOURCE_119	ESS_R08_2016_GER_DE_107	...the European Parliament?	...dem Europäischen Parlament?	... <PUNCT> the <DET> European <PROPN> Parliament <PROPN> ? <PUNCT>	... <PUNCT> dem <DET> Europäischen <PROPN> Parlament <PROPN> ? <PUNCT>	... the European <B-ORG> Parliament <E-ORG> ?	... dem Europäischen <B-ORG> Parlament <E-ORG> ?

Figure 2: Sample of results for the search in Figure 1. Response segments in the source language appear next to their German alignments. It is possible to examine the study, round, year, and language variation of the segments by examining the *source_survey_itemid* and *target_survey_itemid* columns.



word 1 (first word)	word 2 (first word)	word 3 (first word)	word 1 (second word)	word 2 (second word)	word 3 (second word)
examiner	la	carte	use	this	card
veuillez	examiner	la	using	this	card
de	cette	carte	please	use	this
sur	cette	carte	on	this	card
utiliser	cette	carte	please	look	at
l'aide	de	cette	this	card	please
veuillez	utiliser	cette	look	at	card
dans	quelle	mesure	please	tell	me
cette	carte	pour	card	please	tell
à	l'aide	de	which	of	the

Figure 3: Comparing trigram collocations for the words "card" in the English source and "carte" in French from France sub portions of the MCSQ.

and user-friendly infrastructures for secondary research.

As future work, we aim to explore the following three aspects: i) the possibility of allowing users to add new data to the corpus through the interface; ii) functionalities for the construction of domain-specific bilingual dictionaries and, iii) the comparison of similar questions across rounds and studies. Additionally, we aim at producing hands-on tutorials with a comprehensive overview of the interface functionalities and guidance on how to use them.

Acknowledgements

The MCSQ is part of the SSHOC project, an EU Horizon 2020 Research and Innovation Programme (2014-2020) under Grant Agreement No. 823782. It was developed as part of the European Social Survey ERIC work in the SSHOC, and it had the support of the European Values Survey team.

References

Alan Akbik, Tanja Bergmann, Duncan Blythe, Kashif Rasul, Stefan Schweter, and Roland Vollgraf. 2019. Flair: An easy-to-use framework for state-of-the-art nlp. In *NAACL 2019, 2019 Annual Conference of the North American Chapter of the Association for Computational Linguistics (Demonstrations)*, pages 54–59.

Mikhail Arkhipov, Maria Trofimova, Yuri Kuratov, and Alexey Sorokin. 2019. Tuning multilingual transformers for language-specific named entity recognition. In *Proceedings of the 7th Workshop on Balto-Slavic Natural Language Processing*, pages 89–93, Florence, Italy. Association for Computational Linguistics.

Michael Bayer. 2012. Sqlalchemy. In Amy Brown and Greg Wilson, editors, *The Architecture of Open Source Applications Volume II: Structure, Scale, and a Few More Fearless Hacks*. aosabook.org.

Yo Joong Choe, Kyubyong Park, and Dongwoo Kim. 2020. word2word: A collection of bilingual lexicons for 3,564 language pairs. In *Proceedings of the 12th International Conference on Language Resources and Evaluation (LREC 2020)*.

Stefan Evert and Andrew Hardie. 2011. Twenty-first century corpus workbench: Updating a query architecture for the new millennium.

Nancy-Diana Gómez, Eva Méndez, and Tony Hernández-Pérez. 2016. Social sciences and humanities research data and metadata: A perspective from thematic data repositories. *El profesional de la información*, 25(4):545–555.

Janet A Harkness, Fons JR van de Vijver, Peter Ph Mohler, and John Wiley. 2003. *Cross-cultural survey methods*, volume 325. Wiley-Interscience Hoboken, NJ.

Matthew Honnibal, Ines Montani, Sofie Van Landeghem, and Adriane Boyd. 2020. spaCy: Industrial-strength Natural Language Processing in Python.

Adam Kilgarriff, Vít Baisa, Jan Bušta, Miloš Jakubíček, Vojtěch Kovář, Jan Michelfeit, Pavel Rychlý, and Vít Suchomel. 2014. The sketch engine: ten years on. *Lexicography*, pages 7–36.

Rob Kitchin. 2014. Big data, new epistemologies and paradigm shifts. *Big data & society*, 1(1):2053951714528481.

Lev Manovich. 2011. Trending: The promises and the challenges of big social data. *Debates in the digital humanities*, 2(1):460–475.

Willem E Saris and Irmtraud N Gallhofer. 2014. *Design, evaluation, and analysis of questionnaires for survey research*. John Wiley & Sons.