Meta
Journal des traducteurs
Translators' Journal

META

The Multilingual Corpus of Survey Questionnaires: A tool for refining survey translation

Diana Zavala-Rojas, Danielly Sorato, Lidun Hareide et Knut Hofland

Volume 67, numéro 1, avril–mai 2022

Pour de nouvelles méthodes en traductologie quantitative
Exploring New Methods in Quantitative Translation Studies

URI : https://id.erudit.org/iderudit/1092191ar
DOI : https://doi.org/10.7202/1092191ar

Aller au sommaire du numéro

Éditeur(s)

Les Presses de l'Université de Montréal

ISSN

0026-0452 (imprimé)
1492-1421 (numérique)

Découvrir la revue

Citer cet article

Zavala-Rojas, D., Sorato, D., Hareide, L. & Hofland, K. (2022). The Multilingual Corpus of Survey Questionnaires: A tool for refining survey translation. *Meta*, *67*(1), 71–93. https://doi.org/10.7202/1092191ar

Résumé de l'article

Cet article décrit la conception et la compilation du *Multilingual Corpus of Survey Questionnaires* (MCSQ), le premier corpus de questionnaires d'enquêtes internationales accessible au public. La version 3.0 (Rosalind Franklin) est compilée à partir des questionnaires de l'Enquête sociale européenne, de l'European Values Study, de l'Enquête sur la santé et le vieillissement et la retraite en Europe, et du WageIndicator Survey dans la langue de départ, anglais (britannique), et leurs traductions en huit langues (catalan, tchèque, français, allemand, norvégien, portugais, espagnol et russe). Les documents du corpus ont été traduits en vue de maximiser la comparabilité des données entre les cultures. Après avoir contextualisé les objectifs et les procédures de traduction d'enquête, cet article présente des exemples de deux types de résultats de traduction problématiques dans les questionnaires d'enquêtes : le premier type concerne le choix de termes idiomatiques ou d'expressions fixes dans la langue de départ. Le deuxième type concerne les cas où la variation sémantique des choix de traduction dépasse la portée autorisée pour maintenir les propriétés psychométriques à travers des langues. Avec ces exemples, nous souhaitons démontrer comment la linguistique de corpus peut être utilisée pour analyser les résultats de traduction passés et pour améliorer la méthodologie de traduction de questionnaire.

The Multilingual Corpus of Survey Questionnaires: a tool for refining survey translation

DIANA ZAVALA-ROJAS
*Universitat Pompeu Fabra, Barcelona, Spain**
diana.zavala@upf.edu

DANIELLY SORATO
*Universitat Pompeu Fabra, Barcelona, Spain***
danielly.sorato@upf.edu

LIDUN HAREIDE
Møreforsking AS/Møre Research AS, Ålesund, Norway
lidun.hareide@moreforsking.no

KNUT HOFLAND
University of Bergen, Norway
knut.hofland@gmail.com

RÉSUMÉ

Cet article décrit la conception et la compilation du *Multilingual Corpus of Survey Questionnaires* (MCSQ), le premier corpus de questionnaires d'enquêtes internationales accessible au public. La version 3.0 (Rosalind Franklin) est compilée à partir des questionnaires de l'Enquête sociale européenne, de l'European Values Study, de l'Enquête sur la santé et le vieillissement et la retraite en Europe, et du WageIndicator Survey dans la langue de départ, anglais (britannique), et leurs traductions en huit langues (catalan, tchèque, français, allemand, norvégien, portugais, espagnol et russe). Les documents du corpus ont été traduits en vue de maximiser la comparabilité des données entre les cultures. Après avoir contextualisé les objectifs et les procédures de traduction d'enquête, cet article présente des exemples de deux types de résultats de traduction problématiques dans les questionnaires d'enquêtes : le premier type concerne le choix de termes idiomatiques ou d'expressions fixes dans la langue de départ. Le deuxième type concerne les cas où la variation sémantique des choix de traduction dépasse la portée autorisée pour maintenir les propriétés psychométriques à travers des langues. Avec ces exemples, nous souhaitons démontrer comment la linguistique de corpus peut être utilisée pour analyser les résultats de traduction passés et pour améliorer la méthodologie de traduction de questionnaire.

ABSTRACT

This article describes the design and compilation of the *Multilingual Corpus of Survey Questionnaires* (MCSQ), the first publicly available corpus of international survey questionnaires. Version 3.0 (Rosalind Franklin) is compiled from questionnaires from the European Social Survey, the European Values Study, the Survey of Health, Ageing and Retirement in Europe, and the Wage Indicator Survey in the (British) English source language and their translations into eight languages (Catalan, Czech, French, German, Norwegian, Portuguese, Spanish and Russian). Documents in the corpus were translated with the objective of maximising data comparability across cultures. After contextualising aims and procedures in survey translation, this article presents examples of two types of problematic translation outcomes in survey questionnaires: The first type relates to the

choice of idiomatic terms or fixed expressions in the source text. The second type relates to cases where the semantic variation of translation choices exceeds the scope allowed to maintain the psychometric properties across languages. With these examples, we aim to demonstrate how corpus linguistics can be used to analyse past translation outcomes and to improve the methodology for translating questionnaires.

RESUMEN

Este artículo describe el diseño y la compilación del *Multilingual Corpus of Survey Questionnaires* (MCSQ), el primer corpus público de cuestionarios de encuestas internacionales. La versión 3.0 (Rosalind Franklin) se compila a partir de cuestionarios de la Encuesta Social Europea, el European Values Study (EVS), la Encuesta de Salud, Envejecimiento y Jubilación en Europa, y el WageIndicator Survey en el idioma de origen (inglés británico) y sus traducciones a ocho idiomas (catalán, checo, francés, alemán, noruego, portugués, español y ruso). Los documentos del corpus se tradujeron con el objetivo de maximizar la comparabilidad de los datos entre culturas. Después de contextualizar los objetivos y procedimientos en la traducción de encuestas, este artículo presenta ejemplos de dos tipos de resultados de traducción problemáticos en cuestionarios de encuestas. El primer tipo se relaciona con la elección de términos idiomáticos o expresiones fijas en el texto original. El segundo tipo se relaciona con los casos en los que la variación semántica de las opciones de traducción excede el alcance permitido para mantener las mismas propiedades psicométricas en todos los idiomas. Con estos ejemplos, nuestro objetivo es demostrar cómo se puede utilizar la lingüística de corpus para analizar los resultados de traducción y mejorar la metodología de traducción de cuestionarios.

MOTS-CLÉS/KEYWORDS/PALABRAS CLAVE

corpus multilingue, traduction de questionnaires, recherche d'enquêtes, TRAPD, linguistique de corpus
multilingual corpus, questionnaire translation, survey research, TRAPD, corpus linguistics
corpus multilingüe, traducción de cuestionarios, investigación por encuestas, TRAPD, lingüística de corpus

1. Introduction

This article describes the design and compilation of the Multilingual Corpus of Survey Questionnaires (MCSQ), in addition to exemplifying its potential applications. It is the very first publicly available multilingual corpus of international survey texts. The MCSQ is an open-access, open-source research resource. Version 3.0 (named Rosalind Franklin)[1] of the corpus is compiled from the British English (source language) versions of the European Social Survey (ESS), the European Values Study (EVS), the Survey of Health, Ageing and Retirement in Europe (SHARE), and the Wage Indicator Survey (WageIndicator) and their translations into eight languages - Catalan, Czech, French, German, Norwegian (Bokmål), Portuguese, Spanish and Russian - and 29 of their language varieties.[2] We included these projects to ensure the corpus had a wide range with respect to the largest survey projects using probabilistic samples in Europe. These projects provide data for secondary research in the social sciences and humanities (SSH), and are widely used in the SSH communities. The current version comprises more than 4 million words and approximately 766,000 sentences. Survey questions, also called *items*, define documents in the MCSQ and constitute its basic unit of analysis. Throughout this article, we will demonstrate that

the MCSQ constitutes an important resource, which has the potential to improve the translation of questionnaires and survey research.

Large-scale comparative survey projects, such as the ESS, the EVS, the SHARE, and the WageIndicator, provide cross-national and cross-cultural data to the SSH. Empirical social research is often based on data gathered by administering survey questionnaires to representative population samples across countries, using an approach which, in survey terminology, is called *Ask the Same Question* (ASQ). In the ASQ method, the concepts to be measured and the answer options that are used to summarise opinions quantitatively should be kept stable across languages, and cultural adaptation is only implemented to facilitate fluency and the use of locally appropriate terminology. The goal of the translation teams is to make the text functionally equivalent for the purpose of statistical analysis, that is, they should keep the same psychometric properties and capture the same psychological variables (e.g. opinions and attitudes), in relation to the themes in the survey, across linguistic contexts (Harkness, Villar, *et al.* 2010; Mohler and Johnson 2010; Zavala-Rojas, Saris, *et al.* 2018). A rigorous multilingual translation of survey questionnaires has become an important area of methodology for survey design, as evidence suggests that low quality translations hamper data comparability and increase errors in measurement (Davidov and De Beuckelaer 2010; Oberski, Saris, *et al.* 2007).

Linguistic corpora are important tools for both linguistic and sociolinguistic research, both from theoretical and application-oriented perspectives (Izquierdo, Hofland, *et al.* 2008). In addition, meticulously sampled linguistic corpora may also function as archives (Hareide and Hofland 2012). The MCSQ aims to perform both these functions: firstly, to provide a tool for refining the best practices for survey translation procedures in a more scientific way through the application of corpus methodology and, secondly, to preserve source and translated survey texts in a searchable format.

This article is divided into three parts. In the first part, we contextualise survey translation focusing on the Translation, Review, Adjudication, Pretesting, and Documentation (TRAPD) method (Harkness 2003). TRAPD is considered the gold standard for survey translation methodology, and the documents included in the MCSQ were translated using a variant of this method[3]. In the second part, we describe in detail the compilation of the MCSQ. In the third part, we present examples that demonstrate some of the weaknesses of current practices in the use of the TRAPD method. We show how the use of idiomatic or fixed expressions in the source text results in target text output that may create uncertainty in the data collected. We also show how non-consistent use of scales of strength across language varieties may create differences in the measurement objectives within the same language, across languages, and across time. In conclusion, we argue that this multilingual corpus will facilitate a new and more scientific approach to survey translation procedures.

2. Survey translation in context

International survey questionnaires are designed in a source language, which in Europe normally is British English, and are then translated into the target languages as well as language varieties for the participating countries in a given survey project.

Translating survey questionnaires is a challenging task because these questionnaires perform the dual role of being a script for a communicative event and a measurement instrument. The TRAPD method (Harkness 2003) was one of the first attempts to create more scientifically sound survey translation procedures. In such a method, team members provide their specific expertise to arrive at a final translation. At least two translators should produce independent, parallel translations from the source version into the respective target language ('T' in the TRAPD acronym)[4]. In a team meeting, the reviewer assesses and reconciles the translations ('R') and the adjudicator ('A') is responsible for the final decisions on the different translation options. The translated questionnaire is pretested ('P') before being fielded, and the whole process is documented ('D'). Team members combine expertise on survey methodology, linguistics, and knowledge related to the questionnaire topic as well as the culture where it will be administered.

The adoption of the TRAPD method has contributed to solving some seriously problematic practices in survey translation. Firstly, the extended use of back translation as a quality control mechanism (Brislin 1970) was discontinued in favour of incorporating review meetings as a step in the process. Secondly, the lack of emphasis on the translatability of the source text was addressed by incorporating both survey and translation experts into the translation team. Thirdly, the inability to trace back translation decisions was resolved by incorporating a thorough documentation step. Proponents of the TRAPD approach argue that its use results in rich local variations within the written varieties of one language, as well as within groups of related languages (Mohler, Dorer, et al. 2016).

Despite the rigorous methodology Harkness set up, certain weaknesses can be noticed from a translation point of view. Firstly, the implementation of the TRAPD method is human-work intensive, as it requires a multidisciplinary team interacting iteratively to produce a final translation. A second weakness is that each team produces a bilingual translation, thus the translations into the different varieties of a given language (e.g. Swiss-French, Belgian-French) are not necessarily harmonised. This approach often results in deviations between the distinct varieties of one language. These variations do not necessarily reflect linguistic differences between language varieties, but may reflect choices made by each of the translation teams, and therefore may hamper data comparability. Hence, translation options easily multiply in number. Without a corpus or a common repository for accessing previous or finalised versions, the assessment of such translation options is very difficult, thereby hindering replicability. A third weakness is related to the 'Documentation' step. Behr, Dept, *et al.* (2018) defined *input documentation* as texts that are prepared before translation, including source texts and guidelines, and *output documentation* as texts produced during the translation process, such as draft and final translated versions. A completed round of translations in a multilingual survey project would generate an excessive amount of input and output documentation. Although it is possible to analyse translation documentation on a case-by-case basis (see for instance Behr, Dept, *et al.* 2018; Mohler, Hansen, *et al.* 2010; Mohler and Uher 2003), systematic analysis of survey translation documentation is not currently done to a large extent, and it is time consuming. Managing, storing, analysing, and reusing translation documentation in a systematic way is a challenge for teams using the TRAPD method. The MCSQ will primarily constitute a tool for addressing the latter two weaknesses.

Harkness' ambitions went beyond the TRAPD methodology. She viewed the translation process as an integral part of the survey lifecycle, therefore proposing that translation should be better integrated into the survey design:

> [w]henever possible, translation should be integrated into the study design. In practice, however, translation rarely is seen as part of questionnaire design and usually is treated as an addendum. In most instances, translation begins once the source questionnaire has been finalized. (Harkness 2003: 35)

In recent years, some of the survey projects have taken steps in this direction by investing both human and economic resources into translation by adding verification procedures to the TRAPD methodology. Translation has been included in the questionnaire design stage: cf. Dorer (2015); Zavala-Rojas, Saris *et al.* (2018); Fitzgerald and Zavala-Rojas (2020). At the same time, the harmonisation of translation procedures between shared-language countries also takes place in some survey projects, e.g. EVS (EVS 2020). These additional steps have resulted in complex translation procedures, requiring additional steps in the TRAPD methodology before a final translated version is produced. The TRAPD methodology could therefore benefit from some refinement, a greater standardisation, and a more harmonised approach to the translation of multilingual surveys. Harkness (2003) herself envisioned a future in which developments in linguistics and translation studies would contribute to "resolv[ing] and document[ing] inevitable differences across translations," and "refin[ing] available procedures for comparative assessment" (Harkness 2003: 56). We agree that Harkness' ideal to integrate translation into the study design should be embraced and we propose that incorporating corpus methodology into survey translation will serve this purpose.

The launch of the TRAPD method coincides roughly in time with the so-called "empirical turn" in translation studies (Snell-Hornby 2006: 115). Hareide (2019) points out that this shift in translation studies was inspired by the paradigm change in linguistics from prescriptive to descriptive grammar, due to the incorporation of the corpus linguistic method. As the grand old man of English grammar, Leech (1992: 112), so aptly stated: "a significant advantage of the corpus linguistic method is that it allows for the analyst to approach the study of language from the context of the scientific method."

In translation studies, Chesterman pointed out that

> [c]orpus based research into translation universals has been one of the most important methodological advances in translation studies during the past decade or so, in that it has encouraged researchers to adopt scientific methods of hypothesis generation and testing. (Chesterman 2004: 46)

Also, since its inception, corpus-based translation studies (CBTS) has been one of the fastest growing subfields of translation studies (Ji, Hareide, *et al.* 2017: 4).

In our opinion, the field of survey design and translation would benefit from a shift similar to the empirical turn in translation studies, and the MCSQ is a valuable resource to this end. By functioning both as a repository and a searchable multilingual corpus, the MCSQ allows survey designers and translators to systematically learn from previous successes and errors through the inspection of wording and formulations across languages and language varieties in order to avoid those that have caused problems in previous rounds or similar studies. In addition, the MCSQ

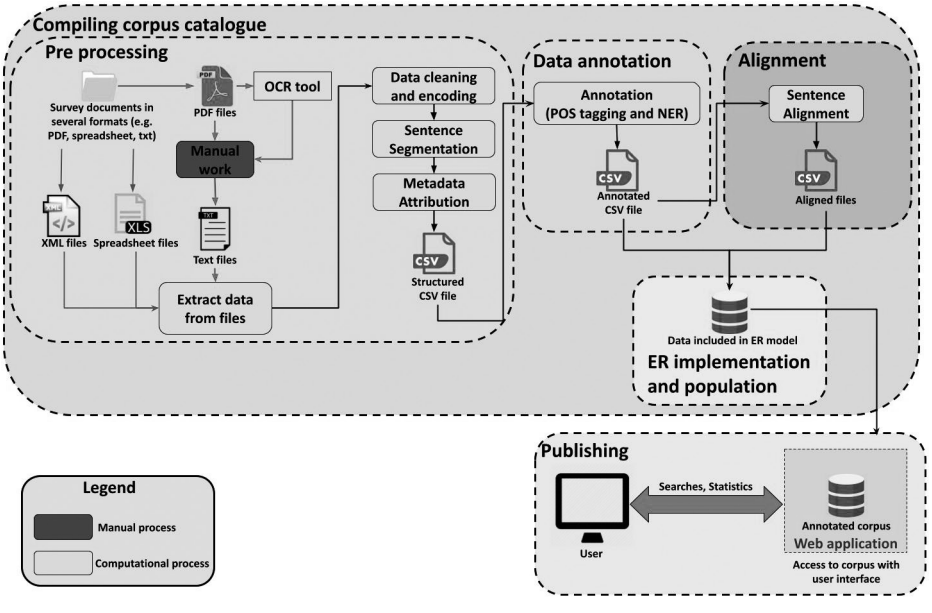
may constitute a resource for a larger degree of standardisation in survey translation across languages, language varieties, and cultures. It is possible to visualise all translation decisions related to a term or an expression in a multilingual setting, allowing for a more collaborative approach to survey translation across different teams. The MCSQ will also be helpful for training translators in the highly technical field of questionnaire translation, by providing examples of phrases, lexicon, response options, etc. In short, the MCSQ is a unique multilingual resource created to provide information, from a translation point of view, on whether all the variants of the survey are indeed asking the same question.

3. Compiling a corpus of survey questionnaires

In order to compile the MCSQ, we developed a framework which follows general best practices for the pre-processing of text data (Sanjurjo González 2017) and the creation of parallel corpora (Doval and Sánchez Nieto 2019, Hareide 2019) and which is suitable for a multilingual corpus with numerous languages and language varieties. As depicted in Figure 1, this framework specifies two main stages: compiling the corpus catalogue and publishing the corpus, each of which involves several steps. Firstly, we conceptualised a database model that adequately represents the domain and yet has a simple structure so as to be efficient (see Section 4.6). Afterwards, the questionnaire's texts were pre-processed and transformed into a comma-separated values (CSV) format. Then, the pre-processed survey items received Part-of-Speech (PoS) Tagging and Named Entity Recognition (NER) annotations (see Section 4.4) and were aligned

Figure 1

Framework for the creation of the MCSQ Source: Own graph, adapted from Sanjurjo González (2017)



Note: processes with a red background were executed using algorithms while processes with a blue background were dependent on manual work.

with respect to the source (see Section 4.5). Afterwards the data was populated in the previously designed Entity-Relationship (ER) model (see Section 4.7), as it was already structured in a convenient format for this purpose. The final stage corresponded to the publishing step (see Section 5). Based on this framework, Figure 1 presents a description of each of the stages and the specific steps needed to create the MCSQ.

4. Compiling the corpus catalogue

4.1. The contents of the corpus

Available questionnaires were retrieved from the websites of the ESS, the EVS, the SHARE and the WageIndicator projects to compile a catalogue of survey questionnaires using the following criteria: study (survey project), edition (called *rounds* on the ESS website and *waves* in other projects), year, country and language. For all the studies listed, a *source questionnaire* version, written in localised British English, exists. All questionnaires are composed of *survey items*. Commonly, a survey item is a *request for an answer* with a set of *answer options*, and may include additional textual elements to guide interviewers and to clarify the information that should be understood and provided by respondents (Saris and Gallhofer 2014). Survey items define documents in the MCSQ. These survey items were divided into sentences which constitute segments in the database.

As of late 2021, the ESS has published nine editions, the EVS has published five, the WIS has published two, and the SHARE has published eight plus a COVID-19 specific questionnaire. The format of the data sources sampled to compile the corpus varies depending on the study and the year of the edition. For ESS round 1 to round 9, questionnaires were retrieved in Portable Document Format (PDF) from the ESS website. For EVS, the source files were obtained either in spreadsheet format (wave 5) or XML format (wave 3 and wave 4). EVS wave 1 and most questionnaires of wave 2 were excluded at this point due to only being available as a scanned images with low quality resolution. They are being retyped to be included in a future version of the database. Likewise, SHARE source files were obtained either in spreadsheet format (COVID-19 questionnaires) or XML format (waves 7 and 8). All Wage Indicator questionnaires were received in spreadsheet format. Appendices 1 and 2 list the questionnaires included in MCSQ version 3 according to study, edition, country, and language, and the number of sentences and the number of tokens in each country-language combination, respectively.

The survey questionnaires included in this corpus are administered as in-person oral interviews. The answers are recorded in a standardised way either on paper or on a Computer Assisted Personal Interview (CAPI) device. A survey questionnaire performs a dual role of being both a guide to a communicative event between two people and, at the same time, an instrument for transforming that communicative event into data. These highly formatted texts are therefore complex, normally featuring scales, boxes that can be easily ticked, columns as well as guidelines for the interviewer. No industry standard exists for the creation of questionnaire documents across survey projects. As such, some files are produced in a word processor, whereas others are created as technical documents for programming the interview on a CAPI-device. The latter contain extensible visible coding and therefore do not exist in printable versions. The interfaces for retrieving and downloading the questionnaires

come in many different formats because the teams for different survey projects have different archiving systems. For some systems, access to the data must be granted, meaning that files cannot be downloaded automatically from their websites. Consequently, gathering the data needed for the compilation of the MCSQ proved time intensive and required extensive manual work.

4.2. Data nomenclature

As the MCSQ data is composed of questionnaires from different survey projects, we had to establish a common nomenclature to specifically identify questionnaire files as well as each of the sentences in the corpus. Such nomenclature also facilitates the process of checking metadata, as it carries information on the survey project (or study; SSS), edition (round or wave; RRR), year (YYYY), language (LLL), and country (CC), with the following sequence: SSS_RRR_YYYY_LLL_CC. Language codes follow the ISO 639-2/B standard (three-digit standard) and country codes follow the ISO 3166 Alpha-2 standard (two digits). To illustrate this, the questionnaire file for ESS round 1, performed in the year 2002, written in the French of France, would be named as indicated in the first example below (*survey*). To refer to a given sentence *i* from that questionnaire (*survey_item_id*), a sequential integer number (*i*) is added to the initial sequence:

- survey = ESS_R01_2002_FRE_FR
- survey_item_id = ESS_R01_2002_FRE_FR_i

4.3. Preprocessing

Pre-processing is one of the crucial tasks in corpus building. It is necessary in order to clean, standardise, and in some cases harmonise data inconsistencies. When mixing data from several sources, such as the case of the MCSQ, special attention is required in this step. In this section, we describe the preprocessing carried out in the MCSQ. Files available in PDF format were converted into plain text format using a combination of both manual work and Optical Character Reader (OCR) tools.[5] However, the corpus data sources contain certain structures that OCR tools are not able to extract correctly, like nested tables. PDFs do not contain the internal indications of structural elements necessary for a computer to correctly interpret such complex structures, which consequently automatically hinders the transformation to plain text format. Due to scenarios like the one mentioned here, a manual conversion of PDF questionnaires to plain text was necessary to ensure that the survey items were correctly structured

After transforming the PDFs into plain text, the text files were converted into CSV format. Questionnaires received in XLS or XML formats are both machine readable and were also converted to CSV format. All texts were normalised (Jurafsky and Martin 2000), which in this context refers to a series of steps to convert texts into a more convenient, standard form. Regardless of how the file formats were converted, all texts went through the following procedure:

a) UTF-8 encoding
b) Removal of unnecessary elements (e.g., trailing spaces, markup tags such as bold and italic, dots sequences)

c) Tokenisation (segmentation) of the words
d) Sentence segmentation
e) Recognition of instructions using regular expressions (*Regex)*
f) Standardised metadata attribution and harmonisation of documents type

Algorithms were designed and implemented for the aforementioned file format conversion, data extraction and preprocessing using the Python programming language.[6] An additional difficulty was that the SHARE questionnaires were designed to be conducted with the aid of an electronic device, therefore they contained *dynamic fields* that were meant to be replaced at the time of the interview based on what the interviewee previously answered (e.g., name of child, year of birth). Dynamic fields were automatically identified and either removed or replaced by fixed values as an additional pre-processing step.

Metadata was added to the corpus by the attribution of segment level variables (e.g., survey item ID, item name), and the different survey item types found across studies were harmonised. For instance, some of the data sources subdivided requests item types into introduction, request and sometimes even sub requests, whereas other sources did not. As the aim was to create a concise unique model for these sources and minimise manual annotation, we simplified and standardised such labels.

Saris and Gallhofer (2014) decompose survey items in all their possible structural elements, deriving from a model that includes up to eight components: "Introduction," "motivation," "information regarding the content," "information regarding the definition," "instruction for the respondent," "instruction for the interviewer" "request for an answer" and "response scales or options." Although the level of details presented in Saris and Gallhofer (2014) would not be feasible due to the necessity of time-consuming manual annotations in the corpus, we were able to decompose a survey item into *introduction, instruction, request* and *response* segments.

To facilitate the identification of *request*, *introduction* and *response* elements in plain text files, a file specification was developed containing textual tags for such items. These tags are then interpreted during the pre-processing steps. Additionally, a set of language-specific regex based patterns was developed in order to automatically identify instruction segments at the time of pre-processing.

4.4. Data annotation

The MCSQ contains PoS tagging and NER annotations. PoS tagging provides useful information about the sentence structure (syntax), while NER identifies named entities in a text, which are instances of real-world objects such as locations and organisations (e.g., Barcelona is an instance of a location named entity). The applied tagging strategy for both annotations are based on language models learned by neural networks. In the case of PoS tagging annotations, pre-trained models from *Flair*[7] were used for English, Czech, French, German, Norwegian and Spanish. As pre-trained models were not available for the Catalan, Portuguese and Russian languages, custom models were trained to this end, using the (language specific) universal dependencies treebanks and the *Flair* framework. The texts were tagged using the Universal Dependencies tagset[8], which is homogeneous across languages. NER annotations in English, German, French, and Spanish datasets were tagged using pre-trained models from *Flair*, while *Slavic BERT*[9] models were used for Czech and Russian languages.

Finally, Catalan, Norwegian and Portuguese datasets were tagged using pre-trained models from *SpaCy*[10].

4.5. Data alignment

Due to the large amount of data involved and the opportunity of leveraging structural information in the alignment phase, we designed an alignment strategy based on metadata. We developed an algorithm which aligns two given files with respect to their attributes. Firstly, the source and target items are filtered in relation to the *module*, to ensure that only the modules present in both source and target files are considered for alignment. Then, for each module, the question names (*item_name*) are filtered. Again, only survey items present both in the source and the target texts proceed to the alignment step.

Afterwards, the alignment is executed as follows: for all segments attributed to a given *item_name*, aligned candidates are selected for alignment according to their *item_type* metadata. In other words, target response segments are aligned with source response segments, target instruction segments are aligned with source instruction segments and so on. The alignment procedure differs according to the *item_type*.

Aligning *response* segments is the simplest case. Since answer scales in a survey questionnaire obey a sequence in which numerical categories are present, the alignments are defined by checking which numbers associated with a given response category are equal amongst the source and the target response segments. The alignments for country-specific response texts (e.g., questions about political party preference, affiliation to religious denominations, etc.) are excluded by design as they only exist in the target languages and do not have any correspondence to segments in the English source texts.

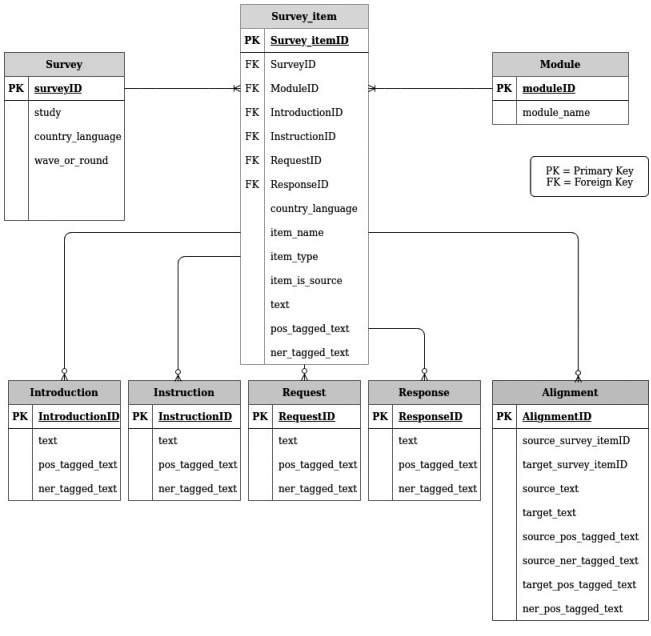
For the remaining segments, the procedure for determining the alignment pairs depends on the number of source and target segments. 0-1 or 1-1 cases are solved trivially due to the filtering of item names. For cases where there are more source segments than target segments (and vice-versa), we calculate the ratio between the candidate source and the target segments. In this first implementation of the algorithm, we define the rule that the ratio closest to 1 is the best alignment pair. The method that defines the best alignment pair is applied recursively until there are no more segments to align in the set that has fewer segments (source or target). Segments that remain unpaired are also included, in the original order in which they appeared. A limitation of this strategy is that errors in automatic sentence segmentation and metadata attribution are further propagated to the alignment phase.

4.6. Entity-Relationship (ER) model

Once the texts have been selected and pre-processed for inclusion in a corpus, a decision must be made regarding how they should be represented in electronic form (Kenny 1998). In order to represent and store the MCSQ data, we designed an *Entity-Relationship* (ER) model, which is a representation of data as interrelated objects of interest (entities). An entity is an abstraction of some aspect of the real world, whereas a relationship between two certain entities specifies how they relate to each other. Designing an ER model to hold the contents of a corpus is a challenging task.

Besides correctness and readability, other factors, such as scalability, performance, and maintainability, also need to be taken into account. There are no specific rules for the design of an ER model, as its conception depends crucially on the specific domain and intended usage. In an ER model, an entity, or *table*, also has *attributes*, also known as *fields*, *metadata* or *paradata*. Metadata describe the characteristics of the data in more detail. Each entity corresponds to a table in the database and each attribute within an entity represents a *column* in such a table. The MCSQ ER model is depicted in Figure 2.

FIGURE 2
MCSQ Entity-Relationship (ER) Model diagram



Eight different tables make up this model, namely *Survey, Module, Survey Item, Introduction, Request, Instruction, Response*, and *Alignment*. One survey consists of several instances of survey items. Therefore, the relationship between *Survey* and *Survey Item* indicates that one *Survey* entity type can have many *Survey Item* entity types associated with it. The tables *Introduction*, *Request*, *Instruction*, and *Response* are elements that may constitute a survey item. The survey item elements have a zero-to-many relationship with survey items because they may not be present, i.e. not all survey items necessarily have all four substructures.

The *Alignment* entity indicates the relationship between text segments (present in the *Survey Item* table) in the English source language and their translations in the target languages. Thus, this table shows which translation segment(s) corresponds to a given segment in the source English language. The segments are aligned at sentence level and the information about any correspondence between source and target sentences can be used, for instance, in a translation memory.

In the context of ER databases, the most common method for specifically identifying the entries of a given table and to create relationships amongst tables is

through the use of Primary Keys (PK) and Foreign Keys (FK). For instance, the attribute *SurveyID* in the table *Survey* (marked with the acronym PK) is responsible for specifically identifying the *Survey* table. The field *SurveyID* in the table *SurveyItem* is an FK, i.e. an PK from another table, in this case, the primary key of *Survey*. The FK acts as a cross-reference between entities, or in other words, it establishes links between the tables. This explanation holds for all fields marked as PK or FK. The model depicted in Figure 2 was developed to represent in a structured manner how a survey questionnaire with its survey items and its elements relate to each other. This design enables the inclusion of new data in the MCSQ, as the database architecture is simple and easy to extend.

4.7. Implementation and population

The MCSQ ER model was implemented in PostgreSQL.[11] The design of the questionnaires makes use of text segments that are repeated across languages, rounds or even inside the same questionnaire. For instance, the instruction *Use the same card to answer* is a segment commonly used in the ESS to prompt using a show card that was presented for a previous question. Aiming to avoid excessive repetition, unique segments were identified throughout the questionnaires and included only once in the tables *Introduction*, *Instruction*, *Request*, and *Response*. Repeated segments in the *Survey Item* table are linked back to the IDs of the unique segments in the aforementioned tables. To populate the database, scripts were developed to insert the preprocessed, annotated, and sentence aligned questionnaires into the ER model.

5. Publishing the corpus

The MCSQ database is stored in a virtual machine provided by Universitat Pompeu Fabra (UPF) in Barcelona.[12] For easy access and search of the data, a public domain has been made available: <https://www.upf.edu/web/mcsq/>. From this domain, users can access an interface to query the database, download data, and build customised translation memories. The corpus is also permanently stored in the Common Language Resources and Technology Infrastructure Norway (Clarino) Bergen repository https://repo.clarino.uib.no/xmlui/handle/11509/142. The MCSQ interface encapsulates SQL queries, allowing the users to perform their searches by simply applying filters or typing the words they are interested in without having to worry about SQL syntax.[13]

6. Facilitating analysis of past translations with the MCSQ

We have argued that the MCSQ was designed to facilitate the adoption of corpus methodology in the various stages of survey design and translation (Section 2). At the design stage of a questionnaire, before entering into the TRAPD process, previous translation decisions can be retrieved from the corpus, analysed and incorporated into the survey design. We argue that the corpus will prove beneficial in every stage of the TRAPD method or other committee approaches to questionnaire translation. In the translation step, the corpus will constitute a reference for translators, allowing for the analysis of past successes and errors in order to refine the survey text. At the review meeting and during adjudication, the team can use corpus methodology,

including statistical analysis of usage of linguistic terms, to decide what the best translation option is. Word frequency, collocational patterns and localisation information can be gained from queries in the corpus to inform the decisions of the adjudication team. Since teams across survey projects work on different time schedules, information from localised surveys from other teams sharing the same language can be retrieved from the corpus and made available to the adjudication teams for intra-language harmonisation.

In this section, we present a few examples from the MCSQ to demonstrate how it can be used to analyse past translation outcomes. We do this by extracting information on commonly used terms in the response scales of survey questions and exemplifying two types of problematic translations: The first type relates to cases where the choice of terms in the source document has resulted in poor translations, specifically in terms of idioms and fixed expressions. The second type relates to cases where the semantic variation of the translation choices exceeds the scope allowed to maintain the psychometric properties across languages, namely, in the intensity attached to verbal labels of response scales, also called *qualifiers*.

6.1. Idiomaticity in the source language text

Since translatability assessment is often not integrated into the study design, certain structures in the source texts may create problems in the translations (Harkness 2003: 35). As the source texts are localised for a specific country, thus allowing them to be administered to this specific population, the use of culture specific terms, idioms and fixed expressions in the source text may lead to poor translations. Traditionally, idioms or fixed expressions have been defined as "frozen patterns of language which allow little or no variation in form, and in the case of idioms, often carry meanings which cannot be deduced from their individual components" (Baker 1992: 63).[14] In this definition, idioms are characterised by semantic non-compositionality because they must be dealt with as a unit, since their intended meaning represents somewhat more than the sum of meanings of their components (Taylor 2002: 549). However, cognitive grammar presents an alternative approach to idiomaticity, pointing out that every language user also has an enormous repertoire of 'fixed expressions' or formulaic language stored in their memory (Jackendoff 1997: 155-156) and these expressions are not necessarily characterised by idiosyncrasies (Taylor 2002: 541-542). For many expressions, their idiomaticity resides in their formal properties, such as collocational requirements that are not fully predictable from general principles, and therefore must be learned, for instance expressions such as *by and large* and *for better or for worse* (Taylor 2002: 543). Other expressions are formulaic or conventionalised, catch phrases or clichés that often allow for some variation in form, such as *add oil/fuel to the flames* and *hit the hay/sack*. An alternative definition of idioms is therefore:

> "multi-word expressions that speakers have learned as conventionalized associations of a phonological form with a semantic representation, irrespective of the 'regularity' of such expressions." […] A person's knowledge of a language consists, precisely, in knowledge of idioms, that is, conventionalized form-meaning relations, at varying levels of generality. (Taylor 2002: 541)

By the definitions described above, the source text expression, *a great deal of time* in Example 1, extracted from the MCSQ, is clearly idiomatic. The texts are extracted from the MCSQ using the interface, and the unique identification of the sentence is shown in brackets. The example demonstrates how the intensity of the qualifiers in the questionnaire answers are affected by the use of idiomatic or fixed expressions in the source language. The translations are presented with their literal back-translations (glosses) for examples in German.

1) A great deal of time.
(ESS_R01_2002_ENG_SOURCE_2273)
(ESS_R04_2008_ENG_SOURCE_2272 for b, d, e and f)

a) Vraiment beaucoup de temps.
(ESS_R01_2002_FRE_FR_2083)

b) Enormément de temps.
(ESS_R04_2008_FRE_FR_2557)

c) Enormément de temps.
(ESS_R01_2002_FRE_CH_2572)

d) Une grande partie du temps.
((ESS_R04_2008_FRE_CH_2705)

e) Enormément de temps.
(ESS_R04_2008_FRE_BE_2705)

f) Sehr viel Zeit. [Very much time.]
(ESS_R01_2002_GER_DE_2571)

g) Enorm viel Zeit. [Enormously much time.]
(ESS_R01_2002_GER_CH_2591)

h) Sehr viel Zeit. [Very much time.]
(ESS_R04_2008_GER_CH_2779)

An analysis of the instances of *a great deal of time* in the English source language (63 in total) reveals that the expression has been translated into expressions that correspond to *very much time*, *a large part of the time, enormous amounts of time*, et cetera. The different choices of wording are not problematic in the TRAPD method; however, what is problematic in this example is the difference in the intensity of the qualifiers. If the same intensity is not maintained across languages and language varieties, the data collected may not be comparable. Here we see that not only does the intensity of the translated expression vary across languages and language varieties, it also varies across rounds of the survey. In round 1 in 2002, the version using French from France consistently used vraiment beaucoup de temps, whereas Swiss French used the clearly more intense expression enormément de temps. In round 4 in 2008, however, French from France (FR) increased the intensity to enormément de temps whereas the Swiss French (CH) team decreased the intensity to une grande partie du temps. In German from Germany (DE), the expression **sehr viel Zeit** [very much time / a lot of time] is consistently used across the two rounds, whereas the Swiss German (CH) team decreased the intensity from **enorm viel Zeit** [enormously much time] in round 1 to **sehr viel Zeit** [very much time / a lot of time] in round 4, in line with the reduction in intensity in the Swiss French version. If the use of an idiomatic or fixed expression causes problems in one language, problems tend to

replicate in others. This concept is acknowledged by Behr, Dept, *et al.* (2018: 349): "Truly problematic items are likely to be problematic in many countries." This study reveals that the expression proved problematic in French, German, Russian and Portuguese, where the intensity of the expression was not kept stable across rounds and varieties.

6.2. Challenges of not using scales of strength consistently or mistranslation of intensity of scales

Problems of intensity do not only arise in the translation of idioms. Survey questions are commonly made up of at least a *request for an answer* and an *answer scale.* The answer scale given to the respondent allows the concept being asked in the *request for an answer* to be quantified. Respondents are given a set of possible answers, ordered on a scale of strength, that are assumed to be consistent across languages and language varieties. When the scales are not consistent across languages, the answers that respondents give may be dependent on group membership (in this case, language-country questionnaire version), thereby causing measurement errors in the survey data. As a consequence, indicators may present differences because of the characteristics of its design, and not because there are truly differences in respondents' opinions, as we will see in Examples 2-5 and 9-12. In Examples 2-5, one can observe that the English terms *extremely, definitely, completely*, represent the maximum of some attribute that is being measured, e.g. *extremely satisfied*. Saris and Gallhofer (2014) suggested that using this type of adverb, one representing the extreme point of response options, provides a fixed reference point in the mind of respondents, thus facilitating an understanding of response scales and improving the quality of the answers. In example 2, the text extracted from the MCSQ is presented in brackets along with the unique identification codes for each sentence.

Example 2

Texts extracted from the MCSQ using the interface, in brackets identification key in the MCSQ

2) Definitely.
(ESS_R05_2010_ENG_SOURCE_1959)

a) Tout à fait.
(ESS_R05_2010_FRE_BE_1599)

1) Completely democratic.
(ESS_R06_2012_ENG_SOURCE_190)

a) Tout à fait (démocratique).
(ESS_R06_2012_FRE_CH_219)

2) Very.
(EVS_R04_2008_ENG_GB_878)

a) Tout à fait.
(EVS_R04_2008_FRE_LU_671)

3) (Agree) strongly.
(EVS_R03_1999_ENG_GB_74)

a) Tout à fait (d'accord).
(EVS_R03_1999_FRE_LU_55)

In the French language versions of the corpus, extreme qualifiers have been translated as tout à fait. However, the English adverb *strongly*, which is not a fixed reference point given that different respondents can assign different values to define what *strong* represents, is also translated as tout à fait. Here the intensity of the qualifiers is changed, creating a problem in the scales of strength across languages. A similar case can be observed in Example 4, where the British source language expressions *extremely* and *fully* have been extracted from the MCSQ along with their translations. As one may note, these expressions have been translated as très, tout à fait and extrêmement respectively. The first adverb clearly varies in intensity compared to the latter two.

Source language extreme qualifiers

4) Extremely (good).

(ESS_R08_2016_ENG_SOURCE_327 for a)
(ESS_R08_2016_ENG_SOURCE_1021 for b)

a) Très (bon).

(ESS_R08_2016_FRE_CH_176)

b) Extrêmement (bon).

(ESS_R08_2016_FRE_CH_718)

5) Extremely (satisfied).

(ESS_R02_2004_ENG_SOURCE_299)

a) Tout à fait (satisfait).

(ESS_R02_2004_FRE_FR_317)

6) (Trust them) completely.

(EVS_R03_1999_ENG_GB_2240)

a) Confiance complète.

(EVS_R03_1999_FRE_LU_1774)

In Examples 7-10, similar problems can be observed in German, where the extreme qualifiers *extremely* or *completely* are translated as **Äußerst** [utmost], **extrem** [extreme], and **voll und ganz** [fully]. To complicate matters further, however, *very* is also translated into both **sehr** [highly] and **voll und ganz** [fully].

Source language extreme qualifiers

7) Extremely (good).

(ESS_R06_2012_ENG_SOURCE_231 for a)
(ESS_R07_2014_ENG_SOURCE_725 for b)

a) Äußerst (gut) [utmost (good)].

(ESS_R06_2012_GER_DE_251)

b) Extrem (gut) [Extremely (good)].

(ESS_R07_2014_GER_CH_964)

8) Extremely (easy).

(ESS_R06_2012_ENG_SOURCE_774)

a) Voll und ganz [fully].

(ESS_R06_2012_GER_CH_1008)

9) Completely

(ESS_R07_2014_ENG_SOURCE_74)

a) Voll und ganz [fully].

(ESS_R07_2014_GER_CH_75)

10) Very (true)

(ESS_R02_2004_ENG_SOURCE_2349)

a) (Trifft) voll und ganz (zu) [(meets) fully].

(ESS_R02_2004_GER_AT_2267)

Lastly, Example 11 presents a statement from the MCSQ and translations of the same sentence into three different French language varieties, from Belgium (BE), France (FR) and Switzerland (CH). The examples are presented with their unique identification code.

Example 5

Texts extracted from the MCSQ using the interface, in brackets identification key in the database

11) Most people can be trusted.

(ESS_R06_2012_ENG_SOURCE_31)

a) La plupart des personnes sont dignes de confiance.

(ESS_R06_2012_FRE_BE_30)

b) On peut faire confiance aux gens.

(ESS_R06_2012_FRE_FR_28)

c) On peut faire confiance à la plupart des personnes.

(ESS_R06_2012_FRE_CH_30)

We observe in the example that while the sentence is rendered with equivalent psychometric properties in Belgium French and Swiss French, the version for France omits the qualifier *most*, leaving the sentence as *people can be trusted*. We question if the two statements *most people can be trusted* and *people can be trusted,* provide the same concept, as the semantic content is clearly changed.

In these examples we have seen that where the freedom of choice inherent in the TRAPD method allows for localisation, it may also cause inconsistent translation of qualifiers that is not necessarily justified by cultural differences across countries, or differences in local language varieties. A more stringent use of qualifiers will reduce measurement errors related to translation choices. A more standardised approach to the translation of the verbal labels of the answer scales with regard to intensity will enhance the statistical comparability of the data across rounds, countries and languages in the survey, and for this the MCSQ will constitute a great tool.

7. Conclusions

This article presents the Multilingual Corpus of Survey Questionnaires (MCSQ), a multilingual corpus of survey texts originally written in British English and translated into eight languages and 29 language varieties. Aside from the WageIndicator questionnaires, texts in this corpus were translated using the TRAPD method, currently considered the gold standard of survey translation. In the examples in Section 6, we have demonstrated that the TRAPD could benefit from some refinement. Just

as the introduction of corpus linguistics caused the scientific turn in translation studies and various subfields of linguistics, we propose it is now time for a refinement of the field of survey translation.

Survey researchers assume that by implementing the ASQ method, the translated questions will be functionally equivalent for statistical analysis, that is, the data will be statistically comparable across linguistic groups and across time (Harkness, Villar, *et al.* 2010; Mohler and Johnson 2010, Zavala-Rojas, Saris, *et al.* 2018). Nonetheless, from our examples we conclude that to fulfil the goal of statistical comparability of the ASQ, a larger degree of standardisation in survey translation across languages, language varieties, cultures and time is needed. In addition, we have shown that greater attention to the translatability of the source language text is paramount to avoid problematic structures such as idioms or fixed expressions, which are open to interpretation.

The MCSQ functions both as a repository of previous rounds of surveys and a tool for systematic analysis of previous errors and successes. Before the compilation of the MCSQ, no method for tracing translation decisions systematically in multilingual surveys had been in place. The corpus allows for the retrieval and preservation of source and translated questionnaires, provides textual data for survey translation activities and research, and facilitates the visualisation and statistical analysis of previous translation decisions across languages. It is also possible to assess in a comparative perspective how a term or a collection of terms have been translated across different languages and in different contexts, and analyse retrospectively whether this decision was appropriate to communicate the intended source text message. Source-language terms that have proven problematic may be avoided in new rounds, and consequently the MCSQ also allows for the integration of translation analysis into the design of the source questionnaire, as suggested by Harkness (2003). In addition, the MCSQ provides valuable training material for the highly-specialised field of survey design and translation. By constituting a tool for the improvement of best practices, both during the design and translation phases of survey questionnaires, the MCSQ allows for a more scientifically refined TRAPD methodology in a way that Harkness, the creator of the TRAPD, had envisioned for the future. Furthermore, the MCSQ provides valuable corpus resources on the highly specialised domain of surveys for minority languages such as Catalan, as well as for the 29 language varieties represented, thus constituting an important resource for cross-linguistic comparisons of specialised use of language. The MCSQ is representative of the largest European survey projects using probabilistic samples that provide data for secondary research in the SSH, thus being highly representative of the domain of survey texts. In line with the focus on open-source, open-access principles, this corpus is openly accessible in CSV format. Furthermore, the interface implemented for users to interact with the MCSQ allows for the creation of translation memories compatible with CAT tools.

Compiling a corpus is a complex interdisciplinary activity. The creation of this corpus required the collaboration of survey experts, statisticians, computational linguists, corpus linguists and translation scholars, as well as a combination of intensive manual and computational tasks. The MCSQ was designed using an Entity-Relationship model as it aims to represent in a structured manner how a survey questionnaire, its survey items and its linguistic elements relate to each other. This design enables the inclusion of new data, as the database architecture is simple and easy to extend, should the funding become available.

To sum up: the MCSQ constitutes a powerful instrument for the further development of best practices both for the design of the source questionnaire and for the improvement of questionnaire translation methodologies. It contributes to the consolidation and the improvement of the translation procedures in multilingual survey projects by providing an open, searchable, aligned and annotated corpus of such questionnaires. Overall the MCSQ facilities a more scientific approach to survey translation and research.

ACKNOWLEDGEMENTS

The MCSQ has been developed in the SSHOC, "Social Sciences and Humanities Open Cloud," project. It has received funding from the European Union's Horizon 2020 project call H2020-INFRAEOSC-04-2018, grant agreement #823782. Author Lidun Hareide was supported by the research institution Møreforsking, Norway. We would like to thank Gregorio Giacomin, Olga Kushch, Patricia Melo Nogueira, Elsa Peris Monsonís, Maria Rubio Juan, Sophia Bortsov and Julia Furtado de Barros for their valuable contribution in supporting the manual conversion of PDF questionnaires into raw texts for MCSQ. We would also like to thank Prof. Willem Saris for his comments on an early draft of this article.

NOTES

* European Social Survey ERIC, Research and Expertise Centre for Survey Methodology.
** Research and Expertise Centre for Survey Methodology.

1. Version 1 (Ada Lovelace) was a prototype version of the MCSQ dated June, 2020, which only included ESS and EVS questionnaires. Version Rosalind Franklin was released in August, 2021.
2. The languages and language varieties are: Catalan, Czech, French (localized language varieties for France, Switzerland, Belgium and Luxembourg), German (localised for Austrian, German, Switzerland and Luxembourg), Norwegian (Bokmål), Portuguese (localised for Portugal and Luxembourg), Spanish (localised for Spain) and Russian (localised for Azerbaijan, Belarus, Estonia, Georgia, Israel, Latvia, Lithuania, Moldavia, Russia and Ukraine).
3. Except for the Wage Indicator survey project, which implements another committee approach to questionnaire translation.
4. Optionally, the questionnaire is split and each translator works on one of the parts.
5. OCR tools transform PDFs and images to plain text.
6. All aforementioned preprocessing steps were performed algorithmically with Python and auxiliary NLP libraries, such as the Natural Language Toolkit (NLTK). Python Software Foundation (October 2020): Python. *Python.org*. Consulted on 6 April 2022, <https://www.python.org/>. NLTK Project (October 2020): Documentation. Natural Language Toolkit. *Nltk.org*. Consulted on 6 April 2022, <https://www.nltk.org/>. Python scripts and other code used for developing the MCSQ can be accessed at the following repository. Sorato, Danielly (2014-2020): Multilingual Corpus of Survey Questionnaires (MCSQ) Compiling. *Github.com*. Consulted on 6 April 2022, <https://github.com/dsorato/MCSQ_compiling>.
7. FlairNLP (2022): A very simple framework for state-of-the-art NLP. *Github.com*. Consulted on 6 April 2022, <https://github.com/flairNLP/flair>.
8. Universal Dependencies contributors (2014-2021): Universal POS tags. *Universaldependencies.org*. Consulted on 6 April 2022, <https://universaldependencies.org/u/pos/>.
9. Slavic-BERT-NER (2022): Shared BERT model for 4 languages of Bulgarian, Czech, Polish and Russian. Slavic NER. *Github.com*. Consulted on 6 April 2022, <https://github.com/deepmipt/Slavic-BERT-NER>.
10. Explosion (2016-2022): Industrial-Strength Natural Language Processing. *Spacy.io*. Consulted on 6 April 2022, <https://spacy.io/>.
11. With the aid of SQLAlchemy. See references hereinafter. The PostgreSQL Global Development Group (October 2020): PostgreSQL: The World's Most Advanced Open Source Relational Database. *Postgresql.org*. Consulted on 6 April 2022, <https://www.postgresql.org/>. Bayer,

Michael (October 2020): The Python SQL Toolkit and Object Relational Mapper. *Sqlalchemy.org*. Consulted on 6 April 2022, <https://www.sqlalchemy.org/>.

12. Which runs with a Debian Operating System Linux distribution.
13. Universidad Pompeu Fabra (n.d.): Welcome to the MCSQ Interface! *The Multilingual Corpus of Survey Questionnaires*. Consulted on 6 April 2022, <http://easy.mcsq.upf.edu/>.
14. Katz and Postal (1964); Quirk, Greenbaum, *et al*. (1985: 1162); and Cruse (1986: 37) have similar definitions.
15. Language abbreviations in this table follow the ISO 639 2/B international standard, this is three digits to abbreviate a language.
16. Country abbreviations in this table follow the ISO 3166-1/Alpha 2 international standard, this is two digits to abbreviate a country.

REFERENCES

Baker, Mona (1992): *In Other Words. A Coursebook on Translation*. New York: Routledge.

Behr, Dorothée, Dept, Steve, and Krajčeva, Elica (2018): Documenting the Survey Translation and Monitoring Process. *In*: Timothy P. Johnson, Beth-Ellen Pennell, Ineke A. L. Stoop and Brita Dorer, eds. *Advances in Comparative Survey Methods: Multinational, Multiregional, and Multicultural Contexts (3MC)*. Hoboken, NJ: John Wiley and Sons, 341-355.

Brislin, Richard W. (1970): Back-Translation for Cross-Cultural Research. *Journal of Cross-Cultural Psychology*. 1(3):185-216. DOI: <https://doi.org/10.1177/135910457000100301>.

Chesterman, Andrew (2004): Beyond the particular. *In*: Anna Mauranen and Pekka Kujamäki, eds. *Translation Universals: Do They Exist?* Amsterdam/Philadelphia: John Benjamins, 33-49.

Cruse, David Alan (1986): *Lexical Semantics*. Cambridge: Cambridge University Press.

Davidov, Eldad, and De Beuckelaer, Alain (2010): How harmful are survey translations? A test with Schwartz's human values instrument. *International Journal of Public Opinion Research*. 22(4):485-510. DOI: <https://doi.org/10.1093/ijpor/edq030>.

Dorer, Brita (2015): Carrying out 'advance translations' to detect comprehensibility problems in a source questionnaire of a cross-national survey. *In*: Karin Maksymski, Silke Gutermuth and Silvia Hansen-Schirra, eds. *Translation and Comprehensibility*. Berlin: Frank and Timme, 77-112.

Doval, Irene, and Sánchez Nieto, María Teresa (2019): Parallel corpora in focus: An account of current achievements and challenges. *In*: Irene Doval and María Teresa Sánchez Nieto, eds. *Parallel Corpora for Contrastive and Translation Studies. New Resources and Applications*. Amsterdam/Philadelphia: John Benjamins, 1-15.

Fitzgerald, Rory and Zavala-Rojas, Diana (2020): A model for cross-national questionnaire design and pretesting. *In*: Paul C. Beatty, Debbie Collins, Lyn Kaye, José Luis Padilla, Gordon B. Willis, and Amanda Wilmot, eds. *Advances in Questionnaire Design, Development, Evaluation and Testing*. Hoboken, NJ: John Wiley and Sons, 493-520.

Sanjurjo González, Hugo (2017): *Creación de un Framework para el tratamiento de corpus lingüísticos*. Doctoral dissertation, unpublished. León: Universidad de León.

Harkness, Janet A. (2003): Questionnaire translation. *In*: Janet A. Harkness, Fons J. R. van de Vijver and Peter Ph. Mohler, eds. *Cross-cultural Survey Methods*. Hoboken, NJ: John Wiley and Sons, 35-56.

Harkness, Janet A., Villar, Ana and Edwards, Brad (2010): Translation, adaptation, and design. *In*: Janet A. Harkness, Michael Braun, Brad Edwards, Timothy P. Johnson, Lars Lyberg, Peter Ph. Mohler, Beth-Ellen Pennell and Tom W. Smith, eds. *Survey Methods in Multinational, Multiregional, and Multicultural Contexts*. Hoboken, NJ: John Wiley and Sons, 115-140.

Hareide, Lidun (2019): Comparable parallel corpora: A critical review of current practices in corpus-based translation studies. *In*: Irene Doval and María Teresa Sánchez Nieto, eds. *Parallel Corpora for Contrastive and Translation Studies. New Resources and Applications*. Amsterdam/Philadelphia: John Benjamins, 19-38.

Hareide, Lidun and Hofland, Knut (2012): Compiling a Norwegian-Spanish parallel corpus: methods and challenges. *In*: Michael Oakes and Meng Ji, eds. *Quantitative Methods in Corpus Based Translation Studies: A Practical Guide to Descriptive Translation Research*. Amsterdam: John Benjamins, 75-114.

Izquierdo, Marlén, Hofland, Knut and Reigem, Øystein (2008): The ACTRES parallel corpus: an English-Spanish translation corpus. *Corpora*. 3(1):31-41. DOI: <https://doi.org/10.3366/E1749503208000051>.

Jackendoff, Ray (1997): *The Architecture of the Language Faculty*. Cambridge, MA: The MIT Press.

Ji, Meng, Hareide, Lindun, Li, Defeng and Oakes, Michael (2016). *Corpus Methodologies Explained: An Empirical Approach to Translation Studies*. London/New York: Routledge.

Jurafsky, Dan and Martin, James H. (2000): *Speech and Language Processing: An Introduction to Natural Language Processing, Computational Linguistics, and Speech Recognition*. Hoboken, NJ: Prentice-Hall.

Katz, Jerrold J. and Postal, Paul M. (1964): An Integrated Theory of Linguistic Descriptions. *Journal of Linguistics*. 2(1):119-126.

Kenny, Dorothy (1998): Corpora in translation studies. *In*: Mona Baker, ed. *Routledge Encyclopedia of Translation Studies*. London/New York: Routledge, 50-53.

Leech, Geoffrey Neil (1992): 100 million words of English: The British National Corpus (BNC). *Second Language Research*. 28:1-13.

Mohler, Peter Ph., Dorer, Brita, De Jong, Julie and Hu, Mengyao (2016): Adaptation. *In*: Survey Research Center, ed. *Guidelines for Best Practice in Cross-Cultural Surveys*. Michigan: Survey Research Center, Institute for Social Research, University of Michigan, 378-391. Consulted on 6 April 2022, <https://ccsg.isr.umich.edu/wp-content/uploads/2019/06/CCSG_Full_Guidelines_2016_Version.pdf>.

Mohler, Peter Ph., Hansen, Sue Ellen, Pennell, Beth-Ellen, Thomas, Wendy L., Wackerow, Joachim, and Hubbard, Frost (2010): A Survey Process Quality Perspective on Documentation. *In*: Janet A. Harkness, Michael Braun, Brad Edwards, Timothy P. Johnson, Lars Lyberg, Peter Ph. Mohler, Beth-Ellen Pennell and Tom W. Smith, eds. *Survey Methods in Multinational, Multiregional, and Multicultural Contexts* Hoboken, NJ: John Wiley and Sons, 299-314.

Mohler, Peter Ph. and Johnson, Timothy P. (2010): Equivalence, Comparability, and Methodological Progress. *In*: Janet A. Harkness, Michael Braun, Brad Edwards, Timothy P. Johnson, Lars Lyberg, Peter Ph. Mohler, Beth-Ellen Pennell and Tom W. Smith, eds. *Survey Methods in Multinational, Multiregional, and Multicultural Contexts* Hoboken, NJ: John Wiley and Sons, 17-29.

Mohler, Peter Ph., and Uher, Rolf (2003): Documenting comparative surveys for secondary analysis. *In*: Janet A. Harkness, Fons J. R. van de Vijver and Peter Ph. Mohler, eds. *Cross-cultural Survey Methods*. Hoboken, NJ: John Wiley and Sons, 311-328.

Oberski, Daniel, Saris, Willem E. and Hagenaars, Jacques (2007): Why are there differences in measurement quality across countries? *In*: Geert Loosveldt, Marc Swyngedouw and Bart Cambré, eds. *Measuring Meaningful Data in Social Research*. Leuven/Voorburg: Acco, 1-17.

Quirk, Randolph, Greenbaum, Sidney, Leech, Geoffrey and Svartvik, Jan (1985): *A Comprehensive Grammar of the English Language*. New York: Longman.

Saris, Willem. E., and Gallhofer, Irmtraud N. (2014): *Design, Evaluation, and Analysis of Questionnaires for Survey Research*. Hoboken, NJ: John Wiley and Sons.

Snell-Hornby, Mary (2006): *The Turns of Translation Studies: New Paradigms or Shifting Viewpoints?* Amsterdam/Philadelphia: John Benjamins.

Taylor, John R. (2002): *Cognitive Grammar*. Oxford: Oxford University Press.

Zavala-Rojas, Diana, Saris, Willem E. and Gallhofer, Irmtraud N. (2018): Preventing differences in translated survey items using the survey quality predictor. *In*: Timothy P. Johnson, Beth-Ellen Pennell, Ineke A. L. Stoop and Brita Dorer, eds. *Advances in Comparative Survey Methods: Multinational, Multiregional, and Multicultural Contexts (3MC)*. Hoboken, NJ: John Wiley and Sons, 357-384.

APPENDICES

Appendix 1: Contents of the MCSQ by Study, Edition, Language[15] and Country[16]

	ESS									EVS				SHARE			WIS	
Language and country	**1**	**2**	**3**	**4**	**5**	**6**	**7**	**8**	**9**	**2**	**3**	**4**	**5**	**7**	**8**	**COVID-19**	**Salary Survey**	**COVID-19**
CAT_ES	X	X	X	X	X	X	X	X	X					X	X			
CZE_CZ	X	X		X	X	X	X	X	X		X	X	X	X	X	X	X	X
ENG_GB	X	X	X	X	X	X	X	X	X	X	X	X	X				X	X
ENG_IE	X	X	X	X	X	X	X	X	X	X	X	X	X					
ENG_MT												X	X	X	X	X		
ENG_ SOURCE	X	X	X	X	X	X	X	X	X		X	X		X	X	X	X	X
ENG_LU		X																
FRE_BE	X	X	X	X	X	X	X	X	X		X	X	X	X	X	X		
FRE_CH	X	X	X	X	X	X	X	X	X			X	X	X	X	X		
FRE_FR	X	X	X	X	X	X	X	X	X	X	X	X	X	X	X	X	X	X
FRE_LU	X	X									X	X	X	X	X	X		
GER_AT	X	X	X	X	X		X	X	X		X	X	X	X	X	X		
GER_CH	X	X	X	X	X	X	X	X	X			X	X	X	X	X		
GER_DE	X	X	X	X	X	X	X	X	X	X	X	X	X	X	X	X	X	X
GER_LU		X										X	X	X	X	X		
NOR_NO	X	X	X	X	X	X	X	X	X				X				X	X
POR_PT	X	X	X	X	X	X	X	X	X	X	X	X	X	X	X	X	X	X
POR_LU											X	X	X	X	X			
RUS_AZ												X	X					
RUS_BY											X	X	X					
RUS_EE		X	X	X	X	X		X	X		X	X	X	X	X	X		
RUS_IL	X			X	X	X		X						X	X	X		
RUS_LT				X	X			X	X				X					
RUS_LV			X	X					X		X	X	X	X	X	X		
RUS_RU			X	X	X	X		X			X	X	X				X	X
RUS_UA		X	X	X	X	X					X	X	X					
SPA_ES	X	X	X	X	X	X	X	X	X	X	X	X	X	X	X	X	X	X

Number of sentences, tokens and unique tokens per language and country combination contained in the MCSQ database. Punctuation characters were removed from the token count.

Appendix 2: Number of sentences, tokens and unique tokens per language/country in MCSQ

Language Variety	Number of Sentences	Number of Tokens*	Number of Unique Tokens
CAT_ES	30,724	201,702	6511
CZE_CZ	36,845	179,720	10,082
ENG_GB	37,440	174,671	4254
ENG_IE	33,767	155,213	3936
ENG_LU	5880	25,171	2178
ENG_MT	10,805	77,457	3452
ENG_NIR	4796	18,531	1618
ENG_SOURCE	59,279	299,375	11,454
FRE_BE	35,971	218,624	7169
FRE_CH	37,554	230,469	6847
FRE_FR	41,911	244,887	7822
FRE_LU	16,880	114,359	5915
GER_AT	37,508	203,930	7544
GER_CH	38,516	214,375	7104
GER_DE	44,450	243,533	8366
GER_LU	12,836	85,637	5598
NOR_NO	31,844	156,755	4662
POR_LU	9963	62,191	3933
POR_PT	41,270	229,159	7218
RUS_AZ	4699	20,824	2569
RUS_BY	6094	22,572	2476
RUS_EE	33,398	176,623	10,031
RUS_GE	45,58	19,212	2562
RUS_IL	21,226	126,774	8316
RUS_LT	16,117	80,872	5075
RUS_LV	20,361	113,867	8542
RUS_MD	3428	14,777	2058
RUS_RU	22,974	110,759	6367
RUS_UA	21,330	99,080	5929
SPA_ES	43,573	252,583	7807
Total	**765,997**	**4,173,702**	**177,395**

*Excluding punctuation characters.