M3.1 – Cross-lingual Thesaurus and Controlled Term Translation

Cristina España-Bonet[1], Roland Ramthun[2] and Sophie Henning[1]
[1]Universität des Saarlandes
[2]Leibniz-Zentrum für Psychologische Information und Dokumentation

– v1.2 –
March 2018

Abstract

This document describes the data, resources, methodology and software developed to translate the controlled terms and related text available as metadata in the PubPsych database.

Contents

1 Controlled Terms in PubPsych

The different database segments in PubPsych use controlled terms from different systems (e.g. the Medical Subject Headings or the APA thesaurus) or even no controlled terminology at all. Many databases include indexing terms, meaning these terms are descriptive, but freely assigned and not from a controlled vocabulary.

The relevant fields holding the controlled term and indexing term information are

CT*lan*H: "Controlled term high". These are terms from controlled vocabulary (MeSH, APA/PSYNDEX terms, etc.), not freely assigned terms.

CT*lan*L: "Controlled terms low". As CT*lan*H, but the person who created the record gave these entries a lower importance for describing the content than the ones in CT*lan*H.

IT*lan*H: "Additional descriptor high". As the name says, additional describing terms, which may have been freely chosen by the person who created the record, so they do not need to come from a controlled terminology.

IT*lan*L: "Additional descriptor low". As IT*lan*H, but with lower descriptive relevance.

Table 1 and Table 2 give an overview about the available descriptive term data available for each database segment.

	CTEH/CTEL	CTDH/CTDL	CTFH/CTFL	CTSH/CTSL
MEDLINE	90.4%/96.0%	90.4%/96.0%	90.4%/96.0%	-/-
ERIC	53.2%/98.9%	-/-	-/-	-/-
ISOC	-/-	-/-	-/-	-/-
NARCIS	-/-	-/-	-/-	-/-
NORART	-/94.2%	-/-	-/-	-/-
PASCAL	-/-	-/-	-/-	-/-
PsychData	-/94.3%	-/94.3%	-/-	-/-
PsychOpen	-/-	-/-	-/-	-/-
PSYNDEX	96.0%/93.2%	96.0%/93.2%	-/-	-/-

Table 1: Relative availability of controlled terms for records per database segment

	ITEH/ITEL	ITDH/ITDL	ITFH/ITFL	ITSH/ITSL
MEDLINE	9.9%/8.9%	-/-	-/-	-/-
ERIC	43.6%/20.9%	-/-	-/-	-/-
ISOC	-/-	-/-	-/-	-/95.9%
NARCIS	40.8%/-	-/-	-/-	-/-
NORART	-/34.6%	-/-	-/-	-/-
PASCAL	-/99.4%	-/0%	-/99.4%	-/99.2%
PsychData	-/-	-/-	-/-	-/-
PsychOpen	80.8%/-	-/-	-/-	-/-
PSYNDEX	4.4%/1.3%	4.5%/1.3%	-/-	-/-

Table 2: Relative availability of indexing terms for records per database segment

For a first analysis, we extract some of the controlled terms (CTs) from a frozen Solr instance[1] using the fields `CTDL`, `CTEL`, `CTFL` and `CTSL`. Table 3 shows the statistics per language. We use CT*lan*L for this analysis because it is the field appearing in more records.

Some of the entries have two parts, the descriptor and a class specification in parentheses:

```
Action Potentials
Action Potentials (drug effects)
Action Potentials (genetics)
```

This allows to further split the controlled terms into a descriptor and a specification thereby reducing the number of unique terms to translate as seen in rows *uniq descriptors* and *uniq specifications* of Table 3.

	German	English	French	Spanish
CT*lan*L total	3,659,210	4,639,171	2,371,110	0
CT*lan*L uniq	56,754	60,939	51,759	0
descriptors uniq	23,556	27,734	18,623	0
specifications uniq	393	392	187	0

Table 3: Number of controlled terms per language in the PubPsych Database. See text for the nomenclature. CT*lan*L denotes the name of the language dependent PubPsych fields for CTs, e.g. CTDL for German or CTSL for Spanish

After this preliminary analysis to study the expectable quantity of different terms, we fixed the set of relevant fields to be CT*lan*H, CT*lan*L, IT*lan*H and IT*lan*L. In order to translate these 16 fields (4 fields per language) we create a quadrilingual lexicon as explained in the next section.

2 Quadrilingual Lexicon

The resources described in this section can be found in the project's Seafile in the folder: `CLIR-PubPsych/Code/MT/DBtranslator/models/CT`

2.1 Multilingual MeSH

We extract the largest part of our quadrilingual lexicon from a quadrilingual MeSH version created with MeSHMerger[2] `file MeSH_2017_de+en+fr+es.xml`. The format of the data has been changed to a list format for CT translation. We extract one list per language L1, where for each term (preferred, non-preferred, and permutations) describing a concept in L1 only the preferred term in the other languages L2, L3 and L4 is added as translation. This ensures that any term for any concept in any language is always mapped to the preferred term in the other languages. The identifier of the concept is also added. With this procedure we obtain 175,004 concepts for English, 96,333 for French, 70,694 for German and 66,828 for Spanish. The difference between languages stems from the different number of *synonyms* (permutations and strings) in the MeSH translations.

[1] PubPsych record set as of 4th August 2017 with 1,037,536 entries.
[2] `https://github.com/clubs-project/MeSHMerger`

Example for the English terms for concept `ID:M0000020`. We first show the complete MeSH entry for the concept, and then the four files that are generated were one can see why different languages have different numbers of entries:

```
MeSH_2017_de+en+fr+es.xml:
    <concept id="M0000020">
    <term id="T000045" lang="eng" preferred="true">
    <string>Abomasum</string>
    <permutation>Abomasums</permutation>
    </term>
    <term id="spa0000603" lang="spa" preferred="true">
    <string>Abomaso</string>
    </term>
    <term id="spa0049997" lang="spa" preferred="false">
    <string>Cuajar</string>
    </term>
    <term id="ger0000018" lang="ger" preferred="true">
    <string>Labmagen</string>
    </term>
    <term id="fre0063293" lang="fre" preferred="true">
    <string>Abomasum</string>
    </term>
    <term id="fre0000018" lang="fre" preferred="false">
    <string>Caillette</string>
    </term>
    </concept>

mesh.dekey.txt:
    Labmagen|||en:Abomasum|||es:Abomaso|||fr:Abomasum|||ID:M0000020

mesh.enkey.txt:
    Abomasum|||es:Abomaso|||de:Labmagen|||fr:Abomasum|||ID:M0000020
    Abomasums|||es:Abomaso|||de:Labmagen|||fr:Abomasum|||ID:M0000020

mesh.eskey.txt:
    Abomaso|||en:Abomasum|||de:Labmagen|||fr:Abomasum|||ID:M0000020
    Cuajar|||en:Abomasum|||de:Labmagen|||fr:Abomasum|||ID:M0000020

mesh.frkey.txt:
    Abomasum|||en:Abomasum|||es:Abomaso|||de:Labmagen|||ID:M0000020
    Caillette|||en:Abomasum|||es:Abomaso|||de:Labmagen|||ID:M0000020
```

Notice that within each file/language, the keys are unique but there might be degeneracy when we concatenate the 4 languages into a single file – in this example, *Abomasum* is a key both for English and French.

2.2 Multilingual Wikipedia Entries

To increase the amount of psychological term translations, we have extracted multilingual in-domain titles from Wikipedia on psychology and health with the WikiTailor tool[3] [1].

WikiTailor extracts domain articles by exploring the categories graph starting from the category describing the domain (psychology and health in our case) and identifying a subset of related categories and their associated articles[4]. These articles are gathered

[3] `https://github.com/cristinae/WikiTailor`

[4] We use models WT0.5-100 or WT0.5-500 depending on the language. Refer to WikiTailor manual if you want to replicate these models `http://cristinae.github.io/WikiTailor/`

independently for English, German, French and Spanish and, afterwards, the intersection or union of the articles is done. For the intersection, we use the articles that have been identified simultaneously in the four languages. For the union, we expand the set of articles to include all the articles that have been identified as in-domain articles at least in one of the languages with the equivalent article in the other three languages in case it exists. Using the intersection of in-domain articles in the four languages we obtain a high precision/low recall multilingual lexicon with 497 entries. With the union of in-domain articles we gather a low precision/high recall multilingual lexicon with 81,369 entries.

The lexicon contains both single words and phrases related to our domain, but in lots of cases entries correspond to named entities:

En	Es	Fr	De
Perception	Percepción	Perception	Wahrnehmung
Echoic_memory	Memoria_ecoica	Mémoire_auditive	Echoisches_Gedächtnis
Emil_Kraepelin	Emil_Kraepelin	Emil_Kraepelin	Emil_Kraepelin
...	...	...	...

In a similar way, we extract aligned category names from Wikipedia, but this time, we select all of them and not only those related to psychology. 38,038 entries are obtained in this case.

As for the MeSH lexicon, we build 4 files, one per language, with the entries in the four languages aligned. In this case though, there is no associated ID:

```
wp.dekey.txt:
Wahrnehmung|||en:Perception|||es:Percepción|||fr:Perception
Echoisches Gedächtnis|||en:Echoic memory|||es:Memoria ecoica|||fr:Mémoire auditive

wp.enkey.txt:
Perception|||es:Percepción|||de:Wahrnehmung|||fr:Perception
Echoic memory|||es:Memoria ecoica|||de:Echoisches Gedächtnis|||fr:Mémoire auditive

wp.eskey.txt:
Percepción|||en:Perception|||de:Wahrnehmung|||fr:Perception
Memoria ecoica|||en:Echoic memory|||de:Echoisches Gedächtnis|||fr:Mémoire auditive

wp.frkey.txt:
Perception|||en:Perception|||es:Percepción|||de:Wahrnehmung
Mémoire auditive|||en:Echoic memory|||es:Memoria ecoica|||de:Echoisches Gedächtnis
```

2.3 Apertium Dictionaries

Apertium [2] is a free/open-source ruled-based translation engine that uses bilingual dictionaries for lexical transfer. We have used three of their dictionaries[5] (*en-de*, *en-es* and *es-fr*) to extract a quadrilingual dictionary with the overlapping entries. Table 4 shows the number of entries of this multilingual dictionary in comparison with the other sources.

2.4 Post-edited Automatic Translations

Finally, we have selected a set of tokens within our controlled terms not covered by the previous resources and translated them with the automatic translation engine DeepL[6]

[5] http://wiki.apertium.org/wiki/List_of_dictionaries

[6] https://www.deepl.com

	German	English	French	Spanish
MeSH	70,694	175,004	96,333	66,828
WPtitles (health+phsy.)	81,369	81,369	81,369	81,369
WPcategories	38,038	38,038	38,038	38,038
Apertium	7,792	5,935	6,020	5,846
Manual	4,262	4,142	4,047	4,081
Wikidata (Propotype2 only)	5,576,686	5,576,686	5,576,686	5,576,686
Total Propotype1	202,128	304,277	225,607	195,937
Total Propotype2	5,778,814	5,880,963	5,802,293	5,772,623

Table 4: Number of aligned terms per language in our multilingual resources. The row with the total excludes duplicate entries between the sources.

(translator as of 25th January and 1st-2nd February 2018). These ∼4,000 entries have been manually post-edited mainly to improve mistranslations due to ambiguous options, but the post-editor was neither native in the four languages nor in-domain expert. Table 4 shows the exact number of entries depending on the source language in the row ”Manual”. Mismatches in the numbers of one row hint to the availability of synonyms for a language.

2.5 WikiData

For the final version of the lexicon we add multilingual entries from Wikidata[7].

BLABLABLA

2.6 Cleaning and Quad-lexicon Compilation

We have cleaned and compiled two quadrilingual lexicons: one that only consists of entries of the MeSH dictionary and one that consists of the entries of all the other dictionaries. We have separated the sources since we consider the MeSH entries to be of higher quality.

We have applied the following cleaning steps to both dictionaries:

- Lowercase tokens
- Remove diacritics (e.g *rücklauf* → *rucklauf*)
- Replace *ß* with *ss* (that is how Solr deals with this character)
- Delete *[dokumenttyp]* annotation (e.g. *biografie [dokumenttyp]* → *biografie*)
- Remove the whole entry if the source word or a translation is empty
- Remove the whole entry if the source word or a translation is a stopword in any of the languages

Moreover, we eliminated source words that were source words in more than one language, sticking to the following order: English was favoured over German, German over French and French over Spanish. In order to cover different spellings, we have also introduced some duplicates: Source words containing umlauts were duplicated with a version in which the umlaut was replaced by the basic character and an *e* (*rücklauf* was not only changed to *rucklauf*, but also lead to another entry with *ruecklauf* as a source word).

[7] `https://www.wikidata.org`

Similarly, we have added duplicates for words ending with *-ise* respectively *-ize*, *-isation* respectively *-ization* and *-our* and *-or* to account for differences between American and British English. Furthermore, we manually deleted some wrong entries.

After applying this procedure to each other dictionary from the sections 2.2, 2.3 and 2.4, we merged them into one dictionary, while, in the case of duplicates, following this priority setting: The lexicon built with WikiTailor was favoured over the dictionary made of Wikipedia category name alignments, that one over the post-edited automatic translations and those over the Apertium dictionary.

3 Controlled Term Translation

3.1 Methodology

We use the resources described in the previous section to translate the controlled terms appearing in the articles of the PubPsych database (Section 1). Notice that the most accurate translation would be achieved with the multilingual MeSH alone. The other three resources add noise to the translations but significantly increase the coverage of the engine.

We follow the strategy below and sketched in Figure 2:

1. A CT is splitted into the descriptor and the class specification (Section 1). Both parts are subsequently cleaned and translated independently. Ex: *Action Potentials (genetics)* ⇒ *Action Potentials, genetics*
2. Part Translation
   2.1. All possible capitalisations of the part (*Action Potentials*, *action potentials*, *Action potentials*) are looked up in the corresponding quadrilingual lexicon and, in case the entry exists, the translations into the other three languages are obtained.
   Ex: *Action Potentials*|||*es:Potenciales de Acción*|||*de:Aktionspotentiale*|||*fr:Potentiels d'action*
   2.2. The original capitalisation is restored.
3. Token Translation. If a part is not found in the dictionary, it is translated on a token-by-token basis.
   3.1. The part is split into tokens and words are translated independently.
   3.2. All possible capitalisations of the token are looked up in the corresponding quadrilingual lexicon and, in case the entry exists, the translations into the other three languages are obtained.
   3.3. If the entry is not available, some basic rules regarding the formation of plural nouns (see Appendix A) are applied to obtain a singular form for the entry. In case the entry exists, the translations into the other three languages for the singular form are obtained and used to translate it.
   3.4. If the entry is not available, we copy the source token as translation for the three other languages.
   3.5. The original capitalisation is restored.
4. Tokens and parts are joined with the appropriate punctuation to build the final translation of the original CT.

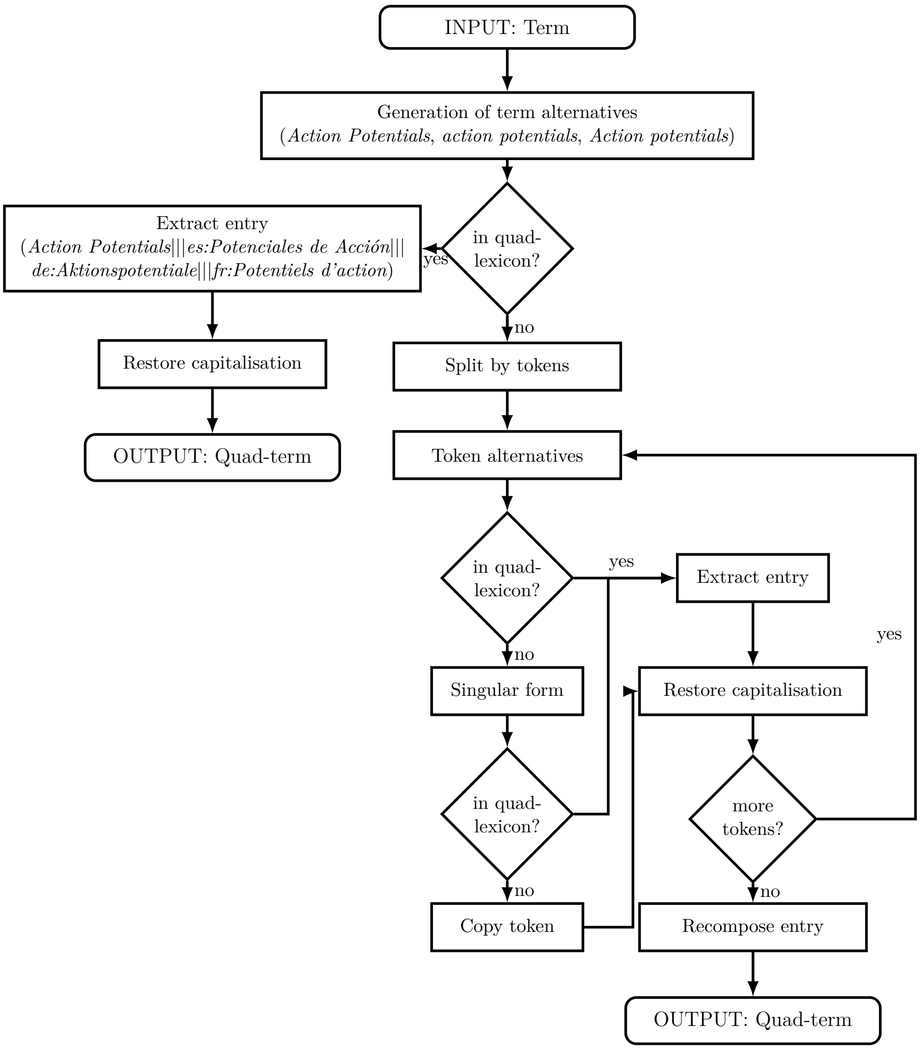

Figure 1: Flux diagram for the controlled term translation. If a complete term cannot be matched, a token by token translation is applied.

We apply the previous methodology to translate the CTs using two different lexicons: the multilingual MeSH (named `MeSH` or `M` in tables), and the union of the MeSH, Wikipedia, Apertium and manual multilingual lexicons (`QuadLex` or `Q`). For the final system (*S2*) we also add the 5 million Wikidata entries. We cannot evaluate the quality of the translation because we do not have a subset of multilingual controlled terms other than MeSH itself, so we quantify the effect of our resources by the number of entries they are able to translate. Table 5 shows the coverage for the CT*lan*L field in the languages of the project. In Table 6, we show the improvements achieved by $S2$ with the new processing (Section 2.6) and with the new resource (Section 2.5). The MeSH thesaurus alone covers between 25%-87% of the all the controlled terms depending on the database and language. These numbers improve as we consider the coverage we can obtain by using also word-level mappings. The word-level mappings together with the usage of the extended lexicon allows us to reach almost a 100% in all the cases. Results for CT*lan*H, IT*lan*H and IT*lan*L are shown in Appendix B.

Even if at this point we cannot evaluate the quality of the translations, note that copying the source word into the output does not necessarily correspond to a wrong translation because in most cases the unknown words are named entities. Equivalently, using

English

	Source	Full descriptors, classes trad (%)	Full descriptors, classes untrad (%)	Tokens trad (%)	Tokens untrad (%)	Tokens uniq
MeSH	ACCNO	344,453 (30.4%)	787,342 (69.6%)	1,325,648 (70.9%)	545,113 (29.1%)	1834
	DFK	544,275 (33.3%)	1,092,037 (66.7%)	2,043,618 (77.2%)	603,889 (22.8%)	2051
	NORART	5,630 (24.6%)	17,223 (75.4%)	34,048 (86.9%)	5,128 (13.1%)	86
	PDID	197 (43.9%)	252 (56.1%)	623 (80.5%)	151 (19.5%)	87
	PMID	2,987,945 (86.9%)	448,879 (13.1%)	5,007,120 (97.1%)	151,482 (2.9%)	242
QuadLex	ACCNO	586,440 (51.8%)	545,355 (48.2%)	1,861,278 (99.5%)	9,483 (0.5%)	33
	DFK	900,396 (55.0%)	735,916 (45.0%)	2,640,589 (99.7%)	6,918 (0.3%)	57
	NORART	5,779 (25.3%)	17,074 (74.7%)	38,941 (99.4%)	235 (0.6%)	9
	PDID	287 (63.9%)	162 (36.1%)	771 (99.6%)	3 (0.4%)	1
	PMID	3,094,379 (90.0%)	342,445 (10.0%)	5,155,774 (99.9%)	2,828 (0.1%)	30

German

	Source	Full descriptors, classes trad (%)	Full descriptors, classes untrad (%)	Tokens trad (%)	Tokens untrad (%)	Tokens uniq
MeSH	DFK	480,050 (29.1%)	1,172,023 (70.9%)	1,328,236 (61.7%)	823,705 (38.3%)	3528
	PDID	182 (38.0%)	297 (62.0%)	425 (64.4%)	235 (35.6%)	132
	PMID	2,915,784 (84.5%)	535,085 (15.5%)	4,321,857 (94.7%)	240,222 (5.3%)	160
QuadLex	DFK	1,002,373 (60.7%)	649,700 (39.3%)	2,150,866 (100.0%)	1,075 (0.0%)	30
	PDID	319 (66.6%)	160 (33.4%)	660 (100.0%)	0 (0.0%)	0
	PMID	3,067,454 (88.9%)	383,415 (11.1%)	4,561,948 (100.0%)	131 (0.0%)	13

French

	Source	Full descriptors, classes trad (%)	Full descriptors, classes untrad (%)	Tokens trad (%)	Tokens untrad (%)	Tokens uniq
M	PMID	2,520,288 (75.3%)	824,711 (24.7%)	5,508,721 (92.9%)	419,105 (7.1%)	961
Q	PMID	2,648,537 (79.2%)	696,462 (20.8%)	5,737,329 (96.8%)	190,497 (3.2%)	334

Table 5: Number of CT*lan*L translated with the multilingual MeSH and the full QuadLexicon for English, German and French. There are no entries for Spanish. A CT term is splitted into two parts (the descriptor and the class specification), and in case of no-matching it is further splitted into tokens.

the quadrilingual lexicon to translate an entry does not assure a correct translation because, besides of the existing noise, the concatenation of word translations does not need to correspond to the term translation. However, we followed the proposed approach to maximize retrieval quality and not translation quality.

3.2 Software

A python script takes care of the CT translation. It can be found in the `DBtranslator` package[8] together with all the software developed to translate the different components of the PubPsych database. The complete translation pipeline going from downloading the field data for all the documents in the database, to translate them and uploading the translations is run by `tradCTs.sh`:

[8]https://github.com/clubs-project/DBtranslator

English

Source	Full descriptors, classes		Tokens	
	trad (%)	untrad (%)	trad (%)	untrad (%)
S1 ACCNO	586,440 (51.8%)	545,355 (48.2%)	1,861,278 (99.5%)	9,483 (0.5%)
S2 ACCNO	598647 (52.9%)	533148 (47.1%)	1869206 (99.9%)	1555 (0.1%)
S1 DFK	900,396 (55.0%)	735,916 (45.0%)	2,640,589 (99.7%)	6,918 (0.3%)
S2 DFK	919275 (56.2%)	717177 (43.8%)	2646966 (100.0%)	224 (0.0%)
S1 NORART	5,779 (25.3%)	17,074 (74.7%)	38,941 (99.4%)	235 (0.6%)
S2 NORART	7,263 (31.8%)	15,590 (68.2%)	39,168 (100.0%)	8 (0.0%)
S1 PDID	287 (63.9%)	162 (36.1%)	771 (99.6%)	3 (0.4%)
S2 PDID	287 (63.9%)	162 (36.1%)	774 (100.0%)	0 (0%)
S1 PMID	3,094,379 (90.0%)	342,445 (10.0%)	5,155,774 (99.9%)	2,828 (0.1%)
S2 PMID	3,103,478 (90.2%)	337,578 (9.8%)	5,158,593 (100.0%)	37 (0.0%)

German

Source	Full descriptors, classes		Tokens	
	trad (%)	untrad (%)	trad (%)	untrad (%)
S1 DFK	1,002,373 (60.7%)	649,700 (39.3%)	2,150,866 (100.0%)	1,075 (0.0%)
S2 DFK	1,010,788 (61.2%)	641,286 (38.8%)	2,150,863 (99.9%)	1,078 (0.1%)
S1 PDID	319 (66.6%)	160 (33.4%)	660 (100.0%)	0 (0.0%)
S2 PDID	318 (66.4%)	161 (33.6%)	658 (99.7%)	2 (0.3%)
S1 PMID	3,067,454 (88.9%)	383,415 (11.1%)	4,561,948 (100.0%)	131 (0.0%)
S2 PMID	3,054,860 (88.5%)	397,321 (11.5%)	4,556,463 (99.8%)	7352 (0.2%)

French

Source	Full descriptors, classes		Tokens	
	trad (%)	untrad (%)	trad (%)	untrad (%)
S1 PMID	2,648,537 (79.2%)	696,462 (20.8%)	5,737,329 (96.8%)	190,497 (3.2%)
S2 PMID	2,982,535 (87.1%)	442,789 (12.9%)	5,670,349 (98.6%)	77,790 (1.4%)

Table 6: Number of CT*lan*L translated with the multilingual MeSH and the full QuadLexicon of System 2 for English, German and French. There are no entries for Spanish.

```
user@machine:~/home/DBtranslator/scripts/$  bash tradCTs.sh -h
tradCTs.sh -f CTH|CTL|ITH|ITL [-h]

where:
-h  show this help text
-f  field to translate [CTH|CTL|ITH|ITL]

Example:
bash tradCTs.sh -f CTH
```

If you want to consider a new field, add it to `preproField4trad.py`. If you want to use the script only on a subset of the database, please, modify the Solr query accordingly in the same file.

The up-to-date instructions for installing and using the software can be found in the git repository:

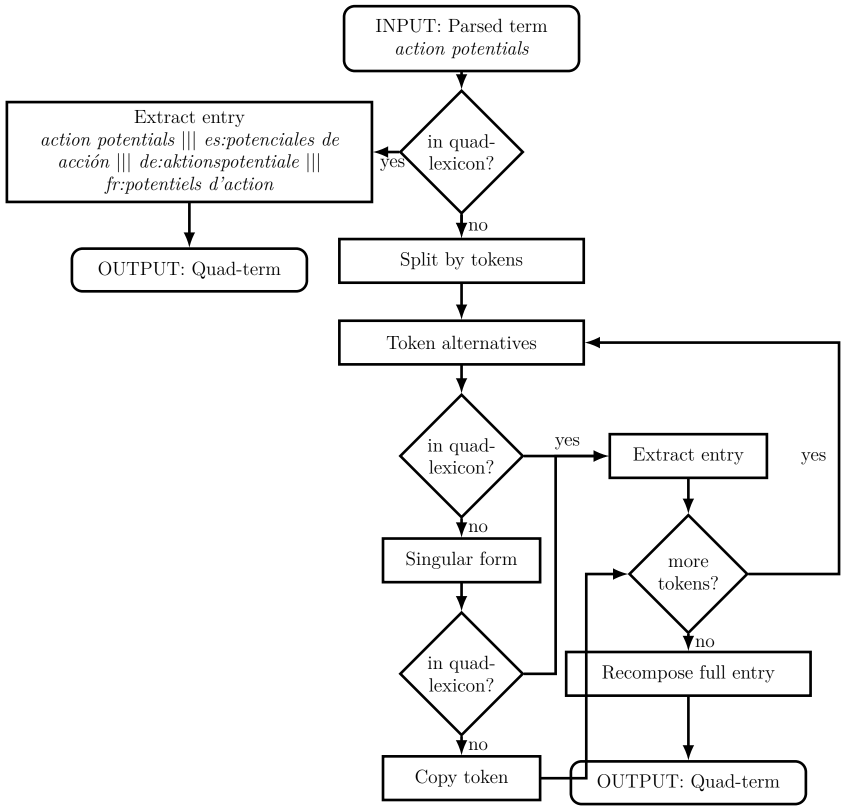

Figure 2: Flowchart for query term translation. If a complete term cannot be matched, a token by token translation is applied.

`https://github.com/clubs-project/DBtranslator`

4 Query Translation

4.1 Methodology

4.2 Off-line Software

4.3 Online Integration into PubPsych

The approach was implemented with respect to Solr 6.6.5. We added four classes to the existing PubPsych backend:

- **QueryFieldRewriter**: This is the main class of the translation module where the actual translation takes place. In a *QueryFieldRewriter* object, the two dictionaries (high-quality and low-quality) and a mapping of field names to language-specific field names are stored. We map field names to language-specific field names, because if a language-specific version of a field exists, we can directly query the correct field with the translation. Since the database schema does not follow a strict pattern in the mapping of these language-specific field names, we need to store the mapping.
- **QueryNode**: The purpose of the *QueryNode* class is to provide an interface to manipulate the queries more easily than with Solr's *Query* class. The main advantage is that one only has to deal with one class for all query objects (and not with the many subclasses of *Query* that all provide different methods/fields). This way, one does not have to build new *Query* objects for each change, but change the respective

QueryNode object instead and then map it back to a *Query* object once all changes due to translation have been performed.

- **Translation**: The *Translation* class is used to store information about the translation of a specific field and string. More precisely, we store the original field name and a mapping of language-specific field names to translation strings in the respective target language. For some fields (e.g. *text*), there are no language-specific field names. In this case, we have to use dummy field names that do not exist in the database schema to have different keys for the different languages. We store in an additional map whether a field name actually exists in the schema in order to search only existing fields in the final query.

- **PreTranslationInfo**: This classed is used to store information on the query before the actual translation happens. We concatenate the strings of all subqueries that fulfill certain conditions and store to which subqueries the concatenation belongs.

For more details on fields and methods of the classes, see `https://github.com/clubs-project/documentation/blob/master/sofware/onlineQueryTranslation.html`.

4.4 Evaluation of the Queries

We gathered approximately 100 real-life queries in each of the languages of the PubPsych database (English, German, French and Spanish) and another 100 real-life queries whose language could not be classified. These queries are the same as in `https://link.springer.com/chapter/10.1007/978-3-030-14401-2_4`.

The experiments were run on a machine with 96 cores at 2.4 gigahertz and 1 terabyte memory, but processes were not parallelized and the Java VM was only given 10 gigabyte memory each time a new Solr instance was initialized (which had to be done for every collection of 100 queries to reset the statistics, see below). We had seven different settings. For each setting, the JAR files had to built anew, and we let the software translate all 500 queries in each setting. Running a script that performed all these actions took a bit more than half an hour.

In each of the seven settings, we used a different combination of dictionaries. If not mentioned otherwise, the dictionaries were used in their "non-diff" version, which means that entries that are the same in all languages were kept (e.g. "Ich bin ein Berliner" is translated in all languages as "Ich bin ein Berliner" since it is the title of a Wikipedia article), whereas they were deleted in the "diff" version. In addition to the MeSh dictionary and the quad-lexicon described in sections 2.2, 2.3 and 2.4, we built another dictionary by extracting entities in the Wikidata dump that exist simultaneously in the four languages (see section 2.5).

1. High-quality dictionary: MeSh, low-quality dictionary: the concatenation of the quad-lexicon from sections 2.2, 2.3 and 2.4 and the Wikidata dictionary

2. High-quality dictionary: none, low-quality dictionary: Wikidata dictionary

3. High-quality dictionary: none, low-quality dictionary: Wikidata dictionary ("diff" version)

4. High-quality dictionary: none, low-quality dictionary: quad-lexicon from sections 2.2, 2.3 and 2.4

5. High-quality dictionary: none, low-quality dictionary: quad-lexicon from sections 2.2, 2.3 and 2.4 ("diff" version)

	muw	mum	muq	buw	bum	buq	cw	cm	cq	suw	sum	suq
1e	102	1	16	161	6	24	34	0	0	2	0	0
1d	49	1	10	106	0	21	72	0	16	8	0	0
1s	103	5	17	183	2	19	53	0	1	22	0	0
1f	85	8	13	185	2	14	65	0	2	18	0	0
1n	13	0	8	119	4	58	31	0	17	0	0	0
2e	0	0	0	231	6	54	67	0	2	7	0	0
2d	0	0	0	116	0	24	112	0	29	5	0	0
2s	0	0	0	258	4	52	87	0	2	45	0	3
2f	0	0	0	247	7	40	94	0	4	22	0	1
2n	0	0	0	128	4	68	34	0	19	0	0	0
3e	0	0	0	178	5	33	120	0	8	11	0	0
3d	0	0	0	92	0	22	134	0	34	9	0	0
3s	0	0	0	184	4	27	159	0	5	37	0	3
3f	0	0	0	173	6	20	169	0	10	21	0	1
3n	0	0	0	52	2	13	112	0	58	1	0	0
4e	0	0	0	230	2	52	77	0	1	6	0	0
4d	0	0	0	117	0	32	110	0	29	10	0	1
4s	0	0	0	205	2	32	144	0	3	24	0	2
4f	0	0	0	193	4	29	156	0	4	20	0	0
4n	0	0	0	23	0	5	147	0	84	2	0	0
5e	0	0	0	230	2	52	77	0	1	6	0	0
5d	0	0	0	117	0	32	110	0	29	10	0	1
5s	0	0	0	205	2	32	144	0	3	24	0	2
5f	0	0	0	193	4	29	156	0	4	20	0	0
5n	0	0	0	23	0	5	147	0	84	2	0	0
6e	106	1	16	0	0	0	201	0	27	0	0	0
6d	52	1	10	0	0	0	171	0	55	3	0	0
6s	106	5	17	0	0	0	233	0	25	7	0	0
6f	89	8	14	0	0	0	247	0	28	3	0	1
6n	14	0	8	0	0	0	154	0	87	1	0	0
7e	105	1	16	135	2	12	65	0	1	4	0	0
7d	50	1	10	80	0	16	95	0	19	9	0	0
7s	103	5	17	108	2	15	126	0	1	17	0	1
7f	86	8	13	110	1	7	140	0	4	10	0	0
7n	13	0	8	37	2	8	116	0	62	3	0	2

Table 7: The number refers to the setting (see beginning of section 4.4), the letter indicates the set of queries that has been translated (*e*: English, *d*: German, *s*: Spanish, *f*: French and *n*: language could not be classified).

6. High-quality dictionary: MeSh, low-quality dictionary: none
7. High-quality dictionary: MeSh, low-quality dictionary: quad-lexicon from sections 2.2, 2.3 and 2.4

We did not only evaluate the actual translations of the queries, but also acquired some statistics for each combination of query collection and setting (see Table 7). The numbers collected are the following:

- **MeSh usage word level** (muw): Number of words that are translated with the MeSh dictionary (inside *translate*). This number is incremented when a string that either is the result of a concatenation of subqueries or a multi-token phrase is split into tokens (because it cannot be translated as a whole) and a token or a possible singular form of it can be translated using the MeSh dictionary. Whenever the MeSh dictionary is not used in a setting, this number is 0.

- **MeSh usage multi-token level** (mum): Number of whole multi-token strings that are translated with the MeSh dictionary. The *multi-token strings* include concatenations of several subqueries (according to the criteria mentioned in the detailed implementation explanation) and multi-word phrases which were explicitly marked as a phrase by the user. If a concatenation cannot be translated as a whole (meaning a single lookup of the whole string in the dictionary), but all its tokens can be translated using only the MeSh dictionary, this number **is not** incremented. If a phrase cannot be translated as whole, but all its tokens can be translated using only the MeSh dictionary, this number **is** incremented. The reasoning behind this difference in counting is that if a user explicitly marks several tokens as a phrase, it is likely that these tokens actually form a phrase, whereas the simple automatic concatenation of all tokens in a query does not take into account any semantic information.

  Whenever the MeSh dictionary is not used in a setting, this number is 0.

- **MeSh usage query level** (muq): Number of queries that are entirely (including all subqueries) translated using only the MeSh dictionary. If a token can not be translated, but a possible singular form of it and this singular form is found in MeSh, then the whole query still counts as "translated only with MeSh" (assuming that everything else or a singular form of each token is found in MeSh). Whenever the MeSh dictionary is not used in a setting, this number is 0.

- **Backoff usage word level** (buw): Number of words that are translated with the low-quality dictionary. The conditions are the same as for MeSh usage word level. The only difference is that the MeSh dictionary is always preferred over the low-quality dictionary: if something can be translated using MeSh, it is not looked up at all in the low-quality dictionary. This means that if MeSh is used, this number is only a lower bound for the number of words translatable with the low-quality dictionary. Whenever no low-quality dictionary is used in a setting, this number is 0.

- **Backoff usage multi-token level** (bum): Number of whole multi-token strings that are translated with the low-quality dictionary. The conditions are the same as for MeSh usage multi-token level. The only difference is that the MeSh dictionary is always preferred over the low-quality dictionary: if something can be translated using MeSh, it is not looked up at all in the low-quality dictionary. This means that if MeSh is used, this number is only a lower bound for the number of multi-token strings translatable with the low-quality dictionary. Whenever no low-quality dictionary is used in a setting, this number is 0.

- **Backoff usage query level** (buq): Number of queries that are entirely (including all subqueries) translated only using the low-quality dictionary. This implies that none of the strings or tokens could be found in the MeSh dictionary, since we always try that one first. The conditions are the same as for MeSh usage query level. The only difference is that the MeSh dictionary is always preferred over the low-quality dictionary: if something can be translated using MeSh, it is not looked up at all in the low-quality dictionary. This means that if MeSh is used, this number is only a

	BLEU	BLEU-1	BLEU-2	BLEU-3	BLEU-4
1	**16.92**	**53.88**	**30.24**	**15.49**	9.16
2	6.99	40.86	23.56	11.03	3.96
3	5.83	39.4	18.27	6.4	2.49
4	13.62	50.49	27.57	13.16	6.74
5	13.62	50.49	27.57	13.16	6.74
6	12.83	41.99	24.39	13.32	**10.04**
7	16.63	53.24	29.69	14.95	9.19

Table 8: BLEU scores of the different settings (see beginning of section 4.4) computed on translations and gold standard containing stopwords.

	BLEU	BLEU-1	BLEU-2	BLEU-3	BLEU-4
1	**18.38**	**59.27**	**36.18**	**18.56**	**8.28**
2	7.7	43.77	24.94	12.06	4.02
3	4.7	42.47	19.63	7.02	1.56
4	12.13	54.47	30.81	14.78	5.1
5	12.13	54.47	30.81	14.78	5.1
6	6.22	42.4	22.07	9.26	4.16
7	17.76	57.94	34.63	16.86	7.96

Table 9: BLEU scores of the different settings (see beginning of section 4.4) computed on translations and gold standard without stopwords.

lower bound for the number of queries translatable with the low-quality dictionary. Whenever no low-quality dictionary is used in a setting, this number is 0.

- **Copies at word level** (cw): Number of words that cannot be translated and thus are copied.
- **Copies at multi-token level** (cm): Number of whole multi-token strings where nothing can be translated. Thus, the respective string is copied.
- **Copies at query level** (cq): Number of queries where nothing can be translated. Thus, the entire query is copied.
- **Singular usage word level** (suw): Number of words that cannot be translated in their original form, but a possible singular form can be translated.
- **Singular usage multi-token level** (sum): Number of phrases where at least one token cannot be translated, but a possible singular form of it. Note that this excludes concatenations and that the counting is different to other numbers on the multi-token level (not all tokens of the phrase have to be translated using singular forms).
- **Singular usage query level** (suq): Number of queries that are entirely translated only using possible singular forms. This also means that no token is copied at all.

The quality of the automatic translations was measured using BLEU scores. The scores were calculated with a Moses script which can be found at `https://github.com/moses-smt/mosesdecoder/blob/master/scripts/generic/multi-bleu.perl`. As Table 8 and Table 9 show, the best translations could be obtained with MeSh and the concatenation of the quad-lexicon from sections 2.2, 2.3 and 2.4 and the Wikidata dictionary, although the low-quality dictionaries introduce some noise for 4-grams. Therefore, we used this dictionary in the subsequent experiments.

	BLEU	BLEU-1	BLEU-2	BLEU-3	BLEU-4
Google with stopwords	14.14	53.18	28.56	16.09	9.35
DeepL with stopwords	14.31	51.59	28.45	**16.21**	**11.95**
Our best system with stopwords	**16.92**	**53.88**	**30.24**	15.49	9.16
Google without stopwords	11.2	54.72	26.08	12.58	7.32
DeepL without stopwords	12.06	52.88	25.93	14.79	**10.15**
Our best system without stopwords	**18.38**	**59.27**	**36.18**	**18.56**	8.28

Table 10: Comparison of BLEU scores obtained with Google, DeepL and our best system

	BLEU	BLEU-1	BLEU-2	BLEU-3	BLEU-4
With stopwords	8.57	40.8	11.97	5.57	2.6
Without stopwords	10.95	45.33	14.53	6.5	3.4

Table 11: BLEU scores of our best system on the XLIFF queries

In order to assess how good our system performs compared with other systems, we also translated the same 500 queries with Google and DeepL.[9] Table 10 displays that our best system performs better on these queries with respect to overall BLEU score, BLEU-1 and BLEU-2. DeepL performs better regarding trigrams (with stopwords) and 4-grams (both cases). It is not surprising that our system generally obtains better scores when the evaluation is performed after the removal of stopwords and that Google and DeepL in this case perform worse than in the other, because our simple dictionary-based approach cannot handle stopwords, whereas the much more sophisticated algorithms of Google and DeepL can.

Another test set we evaluated consists of 261 English queries (XLIFF queries). We let our best system translate them into German, French and Spanish and compared the automatic translations to the translations of two human translators for German and Spanish and one human translator for French. The results can be seen in Table 11.

5 Conclusions

Usage of indexing terms, either controlled or uncontrolled, differs vastly between different database segments contained in PubPsych. By just using the simplest mapping approach of the quadrilingual MeSH thesaurus, we were able to map between 30%–75% of all controlled terms and 9%–40% of free indexing terms between the four languages. A more refined mapping approach and an out-of-domain extension of the quadrilingual lexicon resulted in increased mapping of 67%–100% (controlled terms) and 62%–94% (free terms) respectively. We did not check actual translation quality, but just coverage. The worst mapping performance in all scenarios was exhibited with the PSYNDEX database segment, which uses controlled terminology from the thesaurus of the American Psychological Association. We did not include data from that thesaurus into this evaluation, because it is not available in all four languages.

A Basic Rules for Plural Formation

In order to obtain the a possible singular form of unseen tokens we apply the following basic rules:

[9] The translations were obtained on 06.06.2019. In the meantime, Google and DeepL might have improved their algorithms.

English

- $\star$ NOUN-y $\Leftarrow$ NOUN-ies
- $\star$ NOUN $\Leftarrow$ NOUN-es
- $\star$ NOUN $\Leftarrow$ NOUN-s

French

- $\star$ NOUN $\Leftarrow$ NOUN-s

German

- $\star$ NOUN ($\ddot{-}$) $\Leftarrow$ NOUN-er
- $\star$ NOUN $\Leftarrow$ NOUN-n
- $\star$ NOUN $\Leftarrow$ NOUN-e
- $\star$ NOUN $\Leftarrow$ NOUN-s

Spanish

- $\star$ NOUN $\Leftarrow$ NOUN-es
- $\star$ NOUN $\Leftarrow$ NOUN-s

B Translation Coverage for CT*lan*H, IT*lan*H and IT*lan*L

References

[1] Alberto Barrón-Cedeño, Cristina España-Bonet, Josu Boldoba, and Lluís Màrquez. A Factory of Comparable Corpora from Wikipedia. In *Proceedings of the 8th Workshop on Building and Using Comparable Corpora (BUCC)*, pages 3–13, July 2015.

[2] Mikel L. Forcada, Mireia Ginestí-Rosell, Jacob Nordfalk, Jim O'Regan, Sergio Ortiz-Rojas, Juan Antonio Pérez-Ortiz, Felipe Sánchez-Martínez, Gema Ramírez-Sánchez, and Francis M. Tyers. Apertium: A free/open-source platform for rule-based machine translation. *Machine Translation*, 25(2):127–144, June 2011.

English

	Source	Full descriptors, classes trad (%)	Full descriptors, classes untrad (%)	Tokens trad (%)	Tokens untrad (%)
M	ACCNO	104,815 (30.2%)	241,866 (69.8%)	426,013 (76.1%)	133,654 (23.9%)
	DFK	462,302 (35.5%)	838,814 (64.5%)	1,670,304 (81.0%)	391,538 (19.0%)
	PMID	699,993 (74.3%)	242,666 (25.7%)	1,423,324 (99.4%)	8,840 (0.6%)
Q	ACCNO	162,867 (47.0%)	183,814 (53.0%)	557,624 (99.6%)	2,043 (0.4%)
	DFK	694,203 (53.4%)	606,913 (46.6%)	2,056,459 (99.7%)	5,383 (0.3%)
	PMID	705,345 (74.8%)	237,314 (25.2%)	1,430,707 (99.9%)	1,457 (0.1%)

German

	Source	Full descriptors, classes trad (%)	Full descriptors, classes untrad (%)	Tokens trad (%)	Tokens untrad (%)
M	DFK	393,698 (30.0%)	916,459 (70.0%)	1,123,046 (66.5%)	565,367 (33.5%)
	PMID	701,120 (73.5%)	252,731 (26.5%)	1,273,245 (99.0%)	12,373 (1.0%)
Q	DFK	747,809 (57.1%)	562,348 (42.9%)	1,686,612 (99.9%)	1,801 (0.1%)
	PMID	708,180 (74.2%)	245,671 (25.8%)	1,285,596 (100.0%)	22 (0.0%)

French

	Source	Full descriptors, classes trad (%)	Full descriptors, classes untrad (%)	Tokens trad (%)	Tokens untrad (%)
M	PMID	649,818 (70.0%)	278,327 (30.0%)	1,572,917 (97.6%)	38,389 (2.4%)
Q	PMID	665,293 (71.7%)	262,852 (28.3%)	1,603,726 (99.5%)	7,580 (0.5%)

Table 12: Number of CT*lan*H translated with the multilingual MeSH (M) and the full QuadLexicon (Q) for English, German and French. There are no entries for Spanish.

English

	Source	Full descriptors, classes		Tokens	
		trad (%)	untrad (%)	trad (%)	untrad (%)
MeSH	ACCNO	34,613 (26.0%)	98,309 (74.0%)	176,057 (80.4%)	42,829 (19.6%)
	DFK	3,017 (9.1%)	30,288 (90.9%)	35,389 (73.1%)	13,041 (26.9%)
	NBN	42,314 (20.8%)	161,394 (79.2%)	199,607 (61.6%)	124,180 (38.4%)
	PMID	39,300 (35.4%)	71,565 (64.6%)	126,201 (79.4%)	32,820 (20.6%)
	POID	818 (14.3%)	4,906 (85.7%)	5,687 (61.5%)	3,553 (38.5%)
QuadLex	ACCNO	44,560 (33.5%)	88,362 (66.5%)	204,644 (93.5%)	14,242 (6.5%)
	DFK	6,628 (19.9%)	26,677 (80.1%)	44,678 (92.3%)	3,752 (7.7%)
	NBN	74,130 (36.4%)	129,578 (63.6%)	259,977 (80.3%)	63,810 (19.7%)
	PMID	44,686 (40.3%)	66,179 (59.7%)	146,631 (92.2%)	12,390 (7.8%)
	POID	1,878 (32.8%)	3,846 (67.2%)	8,117 (87.8%)	1,123 (12.2%)

German

	Source	Full descriptors, classes		Tokens	
		trad (%)	untrad (%)	trad (%)	untrad (%)
M	DFK	2,246 (6.8%)	30,890 (93.2%)	22,254 (58.7%)	15,640 (41.3%)
Q	DFK	6,415 (19.4%)	26,721 (80.6%)	29,861 (78.8%)	8,033 (21.2%)

Table 13: Number of IT*lan*H translated with the multilingual MeSH (M) and the full QuadLexicon (Q) for English and German. There are no entries for Spanish or French.

English

	Source	Full descriptors, classes trad (%)	Full descriptors, classes untrad (%)	Tokens trad (%)	Tokens untrad (%)
MeSH	ACCNO	0 (0%)	62,937 (100.0%)	69,785 (64.5%)	38,327 (35.5%)
	DFK	1,035 (10.0%)	9,305 (90.0%)	10,316 (71.2%)	4,171 (28.8%)
	INIST	1,084,844 (34.7%)	2,042,320 (65.3%)	3,208,156 (67.6%)	1,537,860 (32.4%)
	NORART	3,740 (21.4%)	13,719 (78.6%)	18,554 (70.4%)	7,802 (29.6%)
	PMID	36,784 (24.2%)	115,237 (75.8%)	173,377 (68.8%)	78,807 (31.2%)
QuadLex	ACCNO	16,682 (26.5%)	46,255 (73.5%)	108,112 (100.0%)	0 (0%)
	DFK	2,152 (20.8%)	8,188 (79.2%)	13,419 (92.6%)	1,068 (7.4%)
	INIST	1,716,205 (54.9%)	1,410,959 (45.1%)	4,389,674 (92.5%)	356,342 (7.5%)
	NORART	6,734 (38.6%)	10,725 (61.4%)	24,382 (92.5%)	1,974 (7.5%)
	PMID	59,711 (39.3%)	92,310 (60.7%)	228,096 (90.4%)	24,088 (9.6%)

German

	Source	Full descriptors, classes trad (%)	Full descriptors, classes untrad (%)	Tokens trad (%)	Tokens untrad (%)
M	DFK	450 (4.5%)	9,626 (95.5%)	6,272 (55.3%)	5,075 (44.7%)
	INIST	1 (6.2%)	15 (93.8%)	7 (35.0%)	13 (65.0%)
Q	DFK	1,723 (17.1%)	8,353 (82.9%)	8,614 (75.9%)	2,733 (24.1%)
	INIST	2 (12.5%)	14 (87.5%)	10 (50.0%)	10 (50.0%)

French

	Source	Full descriptors, classes trad (%)	Full descriptors, classes untrad (%)	Tokens trad (%)	Tokens untrad (%)
M	INIST	1,186,988 (40.9%)	1,713,817 (59.1%)	3,214,926 (71.7%)	1,268,994 (28.3%)
Q	INIST	1,618,921 (55.8%)	1,281,884 (44.2%)	4,077,461 (90.9%)	406,459 (9.1%)

Spanish

	Source	Full descriptors, classes trad (%)	Full descriptors, classes untrad (%)	Tokens trad (%)	Tokens untrad (%)
M	INIST	896,891 (32.6%)	1,851,974 (67.4%)	2,784,473 (67.3%)	1,355,955 (32.7%)
	ISOC	105,405 (28.0%)	271,601 (72.0%)	413,212 (70.4%)	174,065 (29.6%)
Q	INIST	1,331,980 (48.5%)	1,416,885 (51.5%)	3,779,837 (91.3%)	360,591 (8.7%)
	ISOC	187,681 (49.8%)	189,325 (50.2%)	544,690 (92.7%)	42,587 (7.3%)

Table 14: Number of IT*lan*L translated with the multilingual MeSH (M) and the full QuadLexicon (Q) for English, German, French and Spanish.