D3.1 – Cross-lingual Thesaurus and Controlled Term Translation

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

– v0.2.2 –
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

For a first analysis, we extract some of the controlled terms (CTs) in the frozen Solr instance "pubpsych-core"[1] using the fields `CTDL`, `CTEL`, `CTFL` and `CTSL`. Table 1 shows the statistics per language. Notice that there are no `CTSL` (Spanish CTs) but we did not retrieve any result for `CTSH` either, the other field with information related to controlled terms we dealt with these terms at the begining, but they seem to have disappeared now?.

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 1.

	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 1: 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 select the relevant fields to be translated:

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.

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`

[1] PubPsych record set as of 4th August 2017.

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.

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
```

[2] `https://github.com/clubs-project/MeSHMerger`

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 with the WikiTailor tool[3] [1]. Lexicons have been built from aligned articles in the psychology and health domains for English, German, French and Spanish using WikiTailor models WT0.5-100 or WT0.5-500 depending on the language.

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[4] (*en-de*, *en-es* and *es-fr*) to extract a quadrilingual dictionary with the overlapping entries. Table 2 shows the number of entries of this multilingual dictionary in comparison with the other sources.

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

[4] `http://wiki.apertium.org/wiki/List_of_dictionaries`

	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
Total	202,128	304,277	225,607	195,937

Table 2: Number of aligned terms per language in our multilingual resources. The row with the total excludes duplicate entries between the 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[5] (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 2 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.

3 Controlled Term Translation

3.1 Methodoly

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:

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. casing would be better?
   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 in a word-by-word basis.

[5] `https://www.deepl.com`

English

	Source	Parts		Tokens		
		trad (%)	untrad (%)	trad (%)	untrad (%)	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	Parts		Tokens		
		trad (%)	untrad (%)	trad (%)	untrad (%)	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	Parts		Tokens		
		trad (%)	untrad (%)	trad (%)	untrad (%)	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 3: 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.

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 B) 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.

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`). We cannot evaluate the quality of the translation in both cases 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 3 shows the coverage for the CT*lan*L field in the languages of the project. 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 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[6] 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`:

```
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:

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

A Appendix: Basic Rules for Plural Formation

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

English

- ⋆ NOUN-y ⇐ NOUN-ies
- ⋆ NOUN ⇐ NOUN-es
- ⋆ NOUN ⇐ NOUN-s

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

French

- ⋆ NOUN ⇐ NOUN-s

German

- ⋆ NOUN (¨) ⇐ NOUN-er
- ⋆ NOUN ⇐ NOUN-n
- ⋆ NOUN ⇐ NOUN-e
- ⋆ NOUN ⇐ NOUN-s

Spanish

- ⋆ NOUN ⇐ NOUN-es
- ⋆ NOUN ⇐ NOUN-s

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

English

		Parts		Tokens	
	Source	trad (%)	untrad (%)	trad (%)	untrad (%)
M	DFK	198,456 (35.4%)	362,399 (64.6%)	719,892 (81.0%)	169,154 (19.0%)
	PMID	585,430 (74.3%)	202,969 (25.7%)	1.197,228 (99.3%)	7,886 (0.7%)
Q	DFK	348,343 (53.3%)	304,793 (46.7%)	1,032,638 (99.7%)	2,613 (0.3%)
	PMID	590,132 (74.9%)	198,267 (25.1%)	1,203,749 (99.9%)	1,365 (0.1%)

German

		Parts		Tokens	
	Source	trad (%)	untrad (%)	trad (%)	untrad (%)
M	DFK	169,522 (30.0%)	395,236 (70.0%)	483,674 (66.5%)	243,932 (33.5%)
	PMID	587,289 (73.6%)	210,611 (26.4%)	1,070,023 (99.0%)	10,879 (1.0%)
Q	DFK	375,436 (57.1%)	282,199 (42.9%)	846,825 (99.9%)	901 (0.1%)
	PMID	593,543 (74.4%)	204,357 (25.6%)	1080,881 (100.0%)	21 (0.0%)

French

		Parts		Tokens	
	Source	trad (%)	untrad (%)	trad (%)	untrad (%)
M	PMID	545,436 (70.1%)	232,138 (29.9%)	1,327,668 (97.6%)	32,301 (2.4%)
Q	PMID	557,964 (71.8%)	219,610 (28.2%)	1,353,105 (99.5%)	6,864 (0.5%)

Table 4: 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	Parts		Tokens	
		trad (%)	untrad (%)	trad (%)	untrad (%)
M	DFK	1,226 (9.0%)	12,400 (91.0%)	14,535 (73.2%)	5,325 (26.8%)
	NBN	42,314 (20.8%)	161,394 (79.2%)	199,607 (61.6%)	124,180 (38.4%)
	PMID	38,238 (35.6%)	69,069 (64.4%)	121,994 (79.6%)	31,287 (20.4%)
	POID	2 (3.6%)	53 (96.4%)	60 (59.4%)	41 (40.6%)
Q	DFK	2,647 (19.4%)	10,979 (80.6%)	18,359 (92.4%)	1,501 (7.6%)
	NBN	74,130 (36.4%)	129,578 (63.6%)	259,977 (80.3%)	63,810 (19.7%)
	PMID	43,165 (40.2%)	64,142 (59.8%)	141,317 (92.2%)	11,964 (7.8%)
	POID	7 (12.7%)	48 (87.3%)	80 (79.2%)	21 (20.8%)

German

	Source	Parts		Tokens	
		trad (%)	untrad (%)	trad (%)	untrad (%)
M	DFK	909 (6.7%)	12,645 (93.3%)	9,139 (58.9%)	6,386 (41.1%)
Q	DFK	2,575 (19.0%)	10,979 (81.0%)	12,236 (78.8%)	3,289 (21.2%)

Table 5: 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.

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

		Parts		Tokens	
	Source	trad (%)	untrad (%)	trad (%)	untrad (%)
M	DFK	427 (10.3%)	3733 (89.7%)	4181 (71.5%)	1664 (28.5%)
	INIST	530892 (35.0%)	986834 (65.0%)	1554603 (67.7%)	741218 (32.3%)
	NORART	3414 (21.3%)	12647 (78.7%)	17008 (70.3%)	7199 (29.7%)
	PMID	34598 (24.1%)	108745 (75.9%)	163090 (68.6%)	74490 (31.4%)
Q	DFK	870 (20.9%)	3290 (79.1%)	5439 (93.1%)	406 (6.9%)
	INIST	936934 (55.1%)	764453 (44.9%)	2381113 (92.5%)	191716 (7.5%)
	NORART	6198 (38.6%)	9863 (61.4%)	22422 (92.6%)	1785 (7.4%)
	PMID	56195 (39.2%)	87148 (60.8%)	214683 (90.4%)	22897 (9.6%)

German

		Parts		Tokens	
	Source	trad (%)	untrad (%)	trad (%)	untrad (%)
M	DFK	192 (4.7%)	3,851 (95.3%)	2,569 (56.0%)	2,015 (44.0%)
	INIST	0 (0%)	0 (0%)	0 (0%)	0 (0%)
Q	DFK	698 (17.3%)	3,345 (82.7%)	3,512 (76.6%)	1,072 (23.4%)
	INIST	0 (0%)	6 (100.0%)	5 (62.5%)	3 (37.5%)

French

		Parts		Tokens	
	Source	trad (%)	untrad (%)	trad (%)	untrad (%)
M	INIST	583,688 (40.9%)	844,195 (59.1%)	1,569,921 (71.5%)	626,879 (28.5%)
Q	INIST	891,522 (55.6%)	710,891 (44.4%)	2,237,974 (90.8%)	225,493 (9.2%)

Spanish

		Parts		Tokens	
	Source	trad (%)	untrad (%)	trad (%)	untrad (%)
M	INIST	442,354 (32.7%)	910,874 (67.3%)	1,366,275 (67.2%)	667,708 (32.8%)
	ISOC	73,362 (27.4%)	194,717 (72.6%)	294,892 (70.5%)	123,488 (29.5%)
Q	INIST	736,027 (48.4%)	783,455 (51.6%)	2,082,259 (91.2%)	201,189 (8.8%)
	ISOC	131,846 (49.2%)	136,233 (50.8%)	388,298 (92.8%)	30,082 (7.2%)

Table 6: Number of IT*lan*L translated with the multilingual MeSH (M) and the full QuadLexicon (Q) for English, German, French and Spanish.