YaMTG: An Open-Source Heavily Multilingual Translation Graph Extracted from Wiktionaries and Parallel Corpora

Valérie Hanoka, Benoît Sagot

YaMTG: An Open-Source Heavily Multilingual Translation Graph Extracted from Wiktionaries and Parallel Corpora

Valérie Hanoka[1,2] **Benoît Sagot**[1]

1. Alpage, INRIA Paris-Rocquencourt & Université Paris 7, bâtiment Olympe de Gouges, 75013 Paris, France
2. Verbatim Analysis, 14 rue Friant, 75014 Paris, France
valerie.hanoka@inria.fr, benoit.sagot@inria.fr

Abstract

This paper describes YaMTG (*Yet another Multilingual Translation Graph*), a new open-source heavily multilingual translation database (over 664 languages represented) built using several sources, namely various wiktionaries and the OPUS parallel corpora (Tiedemann, 2009). We detail the translation extraction process for 21 wiktionary language editions, and provide an evaluation of the translations contained in YaMTG.

Keywords: Translation Graph, Multilingual Lexicon, Wiktionary

1. Introduction

Large-scale general-purpose multilingual translation databases are useful in a wide range of Natural Languages Processing (NLP) tasks. This is especially true for research efforts targeted to under-resourced languages. Translation databases can be used for adapting existing resources in other languages. This has been applied for example for the development of wordnets in languages other than English (see e.g., de Melo and Weikum (2009)).

There is thus a real need in NLP for *open-source* multilingual lexical databases that compile as many translations as can be found on any freely available resource in any language.

In this paper, we introduce YaMTG (*Yet another Multilingual Translation Graph*), a new open-source heavily multilingual translation database (over 664 languages represented). While developing YaMTG, we have put the emphasis on the three following features:

1. Our resource is intended for general purpose;
2. We tried to to find a right balance between 'large-scale' and 'quality', thus leaving a limited amount of noise in the data;
3. No information is inferred from the data in order to avoid noise propagation.
4. The resource is easily accessible, usable and extensible.

The remainder of this paper is organised as follows. We first overview the literature on related work (Section 2). Next, we describe in Section 3 how we extracted automatically a preliminary large-scale set of translation and synonym pairs from 21 wiktionary language editions and from the OPUS parallel corpora (Tiedemann, 2009). We detail the translation extraction process for 21 wiktionary language editions, and provide an evaluation of the extracted translation and synonym pairs. Sections 4 and 5 are then respectively dedicated to the construction of the translation and synonym graph proper and to its filtering, in order to reduce the level of noise in the graph. Finaly, we provide in Section 6 quantitative data about the final (filtered) graph YaMTG, together with an overall evaluation of YaMTG and an assessment of the quality of our filtering step.

2. Related work

Research focused on the creation of heavily multilingual word-based translation databases tends to use collaboratively constructed lexicons such as wiktionaries. By doing so, the community is taking advantage of the 'the wisdom of crowds,' thus overcoming the lack of expert-built lexicons for some languages (Meyer and Gurevych, 2012).

Etzioni et al. (2007) built TransGraph, a multilingual sense-distinguished dictionary extracted from wiktionaries and other dictionaries. It covers 100 languages, with 3 languages having over 100,000 words. This translation graph was later reshaped by Mausam et al. (2009) with the creation of the PanDictionary. It covers over 1000 languages, whose translations are still emanating from wiktionaries and other dictionaries. Yet, the resource doesn't seem to be openly available as we write.

The PanLex project (Baldwin et al., 2010) is a lemmatic translation resource which combines a large number of multilingual, bilingual, and monolingual lexical resources. It includes many inferred translations links, less accurate than directly extracted ones. As we write, it covers 1353 language varieties and 12M expressions extracted from more than 4000 different resources, including wiktionaries in many languages.[1] It is still under development. The data is available on the PanLex website.[2] To our knowledge, it is the most complete heavily multilingual translation database. However, we were not able to find an evaluation of the quality of the data.

Some researchers have produced multilingual translation data as a side effect of other goals. The success of the Princeton WordNet (Fellbaum, 1998) as a generic language resource spurred the creation of similar resources in other languages. Many researchers in non-English speak-

[1] `http://panlex.org/tech/plrefs.shtml` (consulted in February 2014)

[2] `http://www.panlex.org/`

ing countries launched various projects, sometimes involving several teams and languages, all aimed at creating new wordnets in their respective languages (Hamp and Feldweg, 1997; Farreres et al., 1998; Pianta et al., 2002; Diab, 2004; Sagot and Fišer, 2008). A list of resources which follows the wordnet design can be found on the Global WordNet Association website.[3] Researchers undertook several actions to establish relationship between wordnets in different languages. The EuroWordNet project (Vossen, 1998) devolved the manual creation and linkage of new wordnets in 8 european languages upon several institutions over many years. Atserias et al. (2004) merged lexico-semantic information contained in local wordnets into the EuroWordNet framework. Similar initiatives have been taken for the design and development of multilingual WordNets in others geographic areas, such as BalkaNet (6 languages, (Stamou et al., 2002)) or IndoNet (18 languages, (Bhat et al., 2013)). Their development was manually carried out by experts, event if some tasks were automated. Yet the quality of the lexico-semantic information provided by these resources prevail over the multilingual aspect.

Half-way between large multilingual lexical databases and wordnet-like ontologies, some researchers aimed at unifying multilingual data from various sources into unique knowledge graphs.

De Melo and Weikum (2009) introduced, and further extended (de Melo, 2012), their Universal WordNet, covering concepts in a language-independent fashion. To do so, they used wordnets in many languages, freely available translation dictionaries, wiktionnaries in different language editions, monolingual and multilingual thesauri and ontologies, and parallel corpora (de Melo, 2012, p.35-37). The resulting resource covers over 200 languages, whose translations were extracted from open-source dictionaries, wiktionaries and parallel corpora. The '201012' dump downloaded on February 2014 contains more than 85,000 meanings instantiated by nearly 1.5 million (1,481,412) lexical items. While lexemes in 419 languages are displayed, 46 languages have more than 10,000 instances. This Universal Wordnet is freely available under a Creative Commons license.

Navigli and Ponzetto (2010, 2012) developed BabelNet, a wide-coverage multilingual knowledge resource. They mapped encyclopedic and onto-lexical data (namely Wikipedia and Wordnet). Concepts thus obtained were supplemented with lexicalisations in a large number of languages obtained via a machine translation pipeline. The latest version of BabelNet, namely BabelNet 2.0, contains 44,490,880 lexical items in 50 languages, regrouped under 9,348,287 synsets.

Among the resources mentioned above, the most interesting ones in terms of language coverage (Baldwin et al., 2010; de Melo and Weikum, 2009; Navigli and Ponzetto, 2012) make use of a specific translation graph architecture which links translations to *meaning* nodes rather than translations between them. This data structure tends to amplify noise in the translation data in case of a erroneous link between a term in a language and a meaning node.

[3] `http://globalwordnet.org/?page_id=38`

Compared to these works, YaMTG contains less lexical items (900,000 vs. 1.5M for UWN, 12M for PanLex and 44M for BabelNet 2.0; see Table 1 for details). This is because it was designed using fewer sources, keeping its development cost low. We plan to include more translations to YaMTG in order to increase it's coverage soon. However, only PanLex covers more languages than YaMTG.

Multilingual Resource	# Lexical Entries	# Languages
YaMTG	881,643	664
UWN	1,481,412	419
PanLex	12,000,000	1353
BabelNet 2.0	44,490,880	50

Table 1: Comparison of the basic figures for the main multilingual translation resources and YaMTG

3. Translation Extraction

3.1. Wiktionaries

3.1.1. Extraction

As described by (Meyer and Gurevych, 2012), Wiktionary is a multilingual online dictionary encoding linguistic knowledge in multiple languages. There exists independent wiktionaries for each language. Those are called language editions. In each language edition, a wiki page corresponds to a character string which can represent one or more lexical items. These items can either be in the editions' language, in other languages or both. Language editions dumps are made available by the Wikimedia Foundation at `http://dumps.wikimedia.org/`. We extracted translations from the dumps[4] available in September 2013.

Wiktionaries are organised as a set of formatted pages, whose title may correspond to an *article* describing one or more lexical entries. The page content of the different language editions displays different structures for the linguistic information of the lexical entries. Sérasset (2012) plans to build and provide open-source access to an a language-independent automatic extractor for a comprehensive set of linguistic information encoded in different wiktionary language editions.

It is possible to custom a set of language specific triggers (Figure 1) for the identification and extraction of translations and synonyms from wiktionary pages.

These triggers are instantiated by regular expressions. For instance, English edition triggers are:

- Translation_ON: `{{trans-top.*`
- Translation_OFF: `{{trans-bottom}}`
- Synonyms_ON: `====Synonyms====`
- Synonyms_OFF: `{{checksyns}}`
- All_swiches_OFF: `^\s*$`
- Ignore: `{{trans-[^tb].+}}`

[4] They can be downloaded directly at `http://dumps.wikimedia.org/*L*wiktionary/latest/*L*wiktionary-latest-pages-articles.xml.bz2`, replacing the '*L*' part in the URL by the ISO-639-1 language code of the desired language edition.

Language Specific Triggers

TRANSLATIONS_**ON**	TRANSLATIONS_**OFF**
SYNONYMS_**ON**	SYNONYMS_**OFF**
IGNORE	ALL_SWITCHES_**OFF**
SOURCE_LANGUAGE	IMPLICIT_TRANSLATION

Swiches

TRANSLATIONS
SYNONYMS

Side Effects for triggers

IGNORE: - *do nothing, go to next line.*

TRANSLATIONS_ON: - *print* "#TRANS#"
- *switch on* TRANSLATION

TRANSLATIONS_OFF: - *switch off* TRANSLATION

SYNONYMS_ON: - *print* "#SYNO#"
- *switch on* SYNONYMS

SYNONYMS_OFF: - *switch off* SYNONYMS

IMPLICIT_TRANSLATION: - *print* "#IT#" + *current line*

NEW_SOURCE_LANGUAGE: - *print* "#LANG#" + *current line*

Side effects for switches

If switches are ON : - *print current line*

Figure 1: Language independent set of *triggers*, *switches* and *side effects* for translation and synonymy extraction from wiktionaries.

- Source_Language: `^[^>]*>?==([^=]+)==$`
- Implicit_translations: `^#\s*\[\[.+\]\].?$`

For the Hindi edition, we instantiated the triggers as follows:

- Translation_ON: `^\({{-trans-}}\)|\( *==* * अन्य भाषाओं में *==*\)|\( *==* * अनुवाद *==*\)`
- `Translation_OFF: ^\( *{{-.*\)|\( *==\)|\(^\[\[\)`
- `Synonyms_ON: ^ *==* * समानार्थी! *==*`
- `Synonyms_OFF: ^\[\[.*`
- `All_swiches_OFF: ^=`
- `Ignore:`
- `Source_Language: space="preserve">{?{?-?(.[^-]*)-?}?}?`
- `Implicit_translations: ^ *#\s*\(([^)]*)\)\s*\[\[.*`

Using this customized triggers for every article page in a wiktionary edition enable the extraction of raw translation data. Figure 2 shows an example of raw data extracted from the Hindi edition with the set of Hindi triggers presented above.

Noisy elements (e.g., punctuation characters) are then removed. Languages names (e.g., हिन्दी[5]) and language codes (e.g., 'en' for English, 'gu' for Gujarati) are converted in ISO 639-2 (alpha-3) language codes.[6] The information is put in a normalised format (Table 2), keeping track of the origin of the translation at the end of each line (e.g., `#WY_hi` stands for Hindi wiktionary; see Table 2).

[5]हिन्दी is the Hindi term for 'Hindi.'

[6]List available at `http://www.loc.gov/standards/iso639-2/php/code_list.php`

```
#ITEM#  आसमान  #LANG# == हिन्दी ==
#IT# # [[ अम्बर ]]
#IT# # [[ आकाश ]]
#IT# # [[ गगन ]]
#IT# # [[ नभ ]]
#TRANS#
* {{en}} : [[sky]]  [[:en:sky]]
* {{gu}} : [[ આકાશ ]] [[:gu:  આકાશ ]]
* {{bn}} : [[ আকাশ ]] [[:bn:  আকাশ ]]
* {{es}} : [[firmamento]] [[:es:firmamento]]
* {{fr}} : [[ciel]] [[:fr:ciel]]
* {{fa}} : [[  آسمان ]] [[:fa:  آسمان ]]
```

Figure 2: Fragment of the raw extraction of translation for the Hindi wiktionary.

Source language		**Target language**		**Wiktionary**
Lang	Term	Lang	Term	**edition** (lang.)
hin	आसमान	hin	अम्बर	hi
hin	आसमान	hin	आकाश	hi
hin	आसमान	hin	गगन	hi
hin	आसमान	hin	नभ	hi
hin	आसमान	eng	sky	hi
hin	आसमान	guj	આકાશ	hi
hin	आसमान	ben	আকাশ	hi
hin	आसमान	spa	firmamento	hi
hin	आसमान	fra	ciel	hi
hin	आसमान	fas	آسمان	hi

Table 2: Fragment of the normalized translation extracted from the Hindi wiktionary.

3.1.2. Evaluation of the extraction process

Table 3 presents the proportion of wikipedia entries from which translations or synonyms have been extracted for 21 languages edition, together with the number of raw (unfiltered) translation pairs extracted.

The biggest wiktionary editions (English and French) have an extraction for roughly a quarter of their non-empty entries (respectively 18.4% and 17.8%). The best extraction rates (90.1% and 80.9%) have been achieved for the Portugese and Romanian editions. Greek and Polish are the two edition having worst extraction rates (11.7% and 11.4%).

For Greek, that poor score seem to be mainly due to implicit translations which are not extracted from articles in another language. More precisely, we retain only one trigger for implicit translations (`^#\s*\[\[.+\]\].?$`). Yet many implicit translations are marked by the pattern `^* \[\[`. We chose not to identify this trigger because, as it is also used to mark related terms (*συγγενικά*). Therefore, it would bring much noise in the data.

We evaluated the wiktionary extraction process for the 21 languages edition we processed by randomly selecting 3150 translation pairs (150 extracted from each language edition) and manually checking their adequacy against the corresponding wiktionary.

Table 4 presents the result of this evaluation with an error analysis. Accuracy scores range from 78% (Spanish) to 97.3% (Dutch). We identified 4 major error classes:

1. translations whose source or target language hasn't been successfully identified;
2. definitions mistaken for translations;

Wikt. ed. (lang)	#entries	#non-empty entries	#entries we extracted transls. or syns. from	%
BUL	821873	85096	14812	17.4
CES	53448	53442	16549	31
DAN	16708	14974	10672	71.3
DEU	341474	341006	60951	17.9
ELL	428746	421762	49335	11.7
ENG	3609456	3599203	661166	18.4
FRE	2399766	2393544	426952	17.8
HIN	24266	23901	4853	20.3
HUN	205646	203401	154893	76.2
ITA	119223	119126	48180	40.4
JAP	108800	105942	57635	54.4
NLD	373532	373531	79433	21.3
POL	393627	391626	44743	11.4
POR	188905	188728	170053	90.1
RON	115753	111305	89999	80.9
RUS	594425	488631	146296	29.9
SLK	4518	4517	1906	42.2
SPA	482886	482690	63807	13.2
SWE	368885	367964	62075	16.9
VIE	218032	216313	117085	54.1
TUR	307225	305437	106535	34.9

Table 3: Proportion of wiktionary entries having at least one raw translation and/or synonym pair extracted by our system.

3. translation containing residual 'noise' (mainly non alphanumeric characters);

4. miscellaneous errors.

Most instances of class 1 to 3 errors will be discarded by the filtering step described in Section 5. Type 4 ('miscellaneous') errors are the most detrimental, as most of them will remain in the data.

Table 4 indicates a mean score of 87.6% for the unfiltered extraction. However, this score is biased because smallest languages edition are overrepresented. Weighting the scores by the size of the extraction (Table 3) leads to a more reliable score of 89.4%.

3.2. Opus

OPUS (Tiedemann, 2009) is a collection of translated texts from the web aimed at providing the community with a publicly available parallel corpus in a wide variety of languages. We used the word alignments issued on October 9, 2013 to add new translations links in the graph, retaining only alignments with a frequency at least 10. With this threshold, we retained 576131 terms (or lexical items) for 3598227 translations in 31 languages from OPUS translations. As OPUS translations are prone to alignment errors, their quality may not be as good as desired. Yet they contains a substantial amount of inflected forms which are not necessarily present in other sources. We rely on the filtering steps (Section 5) to avoid retaining the worst translations.

4. Building the Translation Graph

Previous steps resulted in more than 18M translations or synonyms in 4,324 languages.[7]
Note that while wiktionary translations are mainly lemmas, OPUS translations consists in a lot of inflected terms' translations. We kept them in the graph, but it is easy to get rid of them as we kept translations origins in the data. Discarding translations originating from OPUS only thus removes inflected terms. All these translations contains a lot of noise such as errors identified in Section 3.1.2 plus alignment flaws arising in OPUS data.

5. Filtering

Unfiltered translations gathered in previous steps need to be filtered in order to decrease the noise. We removed many translations matching one of the following criteria:

- They are 'too long' (usually Section's 3.1.2 class 3 errors, i.e., definitions mistaken for translations) or 'too short'[8] (usually stopwords);
- They contain punctuations and/or numbers;
- Their charset does not correspond to their declared language;
- They appear in less than X other translations;
- They appear in more than Y other translations (when too many translations are available for the same word, it turns out it is very often the result of an error).

The first three filtering criteria are designed to avoid errors from class 1 to 3 as identified in Section 3.1.2 The two last criteria were rather designed to get rid of types 4 errors (miscellaneous errors) and OPUS misalignment. We empirically chose $X = 3$. Concerning Y, we assumed that it may be somehow related to the number of languages present in the graph. After applying the first filtering criteria, the graph displays terms (lexical items) in 3,739 languages. For those languages, only 468 appears 100 times or more. We assumed that a term should reasonably not have more than an average of 2 translations for all of these 468 languages. In other words, we decided that a term should not be involved in more than $Y = 936$ translations.

6. Results and evaluation

6.1. Graph properties

The filtered graph retains nearly a tenth of the terms and 15% of the languages present in the unfiltered graph. Yet, 32.3% of the translation/synonym links, or edges, are retained (cf. Table 6). 51.2% of these remaining translations are extracted solely from OPUS, 42,1% from wiktionaries only, whereas 3.2% were extracted from both OPUS and at least one of the 21 wiktionaries we processed.

[7] As a standard of comparison, the *Ethnologue* association states living languages number more than 7,105 (`http://www.ethnologue.com/`).

[8] This filter does only concern less than 3 letters words written in Cyrillic, Greek and Latin script.

	Before Filtering						After Filtering						
Lang.	**Erroneous**					**Good**	**Erroneous**					**Good**	**Kept**
edition	Misc.	lang.	def.	noise	raw	%	Misc.	lang.	def.	noise	raw	%	%
bg	8	1	1	4	136	90.7	3	0	1	0	93	95.9	64.7
cs	4	0	19	5	122	81.3	0	0	2	1	16	84.2	12.7
da	2	0	3	2	143	95.3	1	0	0	0	65	98.5	44
de	4	1	0	2	143	95.3	1	0	0	0	46	97.9	31.3
el	4	0	5	2	139	92.7	1	0	2	1	65	94.2	46
en	1	5	0	4	140	93.3	1	1	0	0	111	97.4	76
es	5	0	15	13	117	78	0	0	0	1	46	97.9	31.3
fr	4	0	8	0	138	92	2	0	3	0	95	95	66.7
hi	9	3	3	4	131	87.3	1	0	0	0	42	97.7	28.7
hu	3	9	1	1	136	90.7	1	2	0	0	56	91.8	40.7
it	0	4	14	2	130	86.7	0	1	0	0	58	96.7	40
ja	2	10	26	15	97	64.7	1	3	8	0	38	71.7	35.3
nl	0	0	0	4	146	97.3	0	0	0	0	68	100	45.3
pl	3	3	0	2	142	94.7	0	0	0	0	68	100	45.3
pt	5	3	18	4	120	80	2	0	0	0	89	97.8	60.7
ro	0	0	13	3	134	89.3	0	0	0	0	116	100	77.3
ru	10	1	4	2	133	88.7	1	0	0	0	95	99	64
sk	5	7	12	0	126	84	2	0	0	0	39	95.1	27.3
sv	4	4	9	0	133	88.7	2	3	0	0	69	89.6	51.3
tr	2	11	4	0	133	88.7	0	1	1	0	83	96.5	57.3
vi	9	0	20	0	121	80.7	0	0	0	0	31	100	20.7
Total	84	62	175	69	2760	—	19	11	17	3	1389	—	—
%	2.7	1.9	5.6	2.2	87.6	—	1.3	0.8	1.2	0.2	96.5	—	—

Table 4: Evaluation figures for the unfiltered extraction of translations/synonymy from 21 wiktionary editions (sample size for each language: 150), and for the filtered graph.

	Wiktionaries	**OPUS**	**Union**
Terms	5,246,987	487,251	7,903,646
Transl./Syn. links	14,449,087	3,631,229	18,080,316
Languages	4,324	32	4,324

Table 5: Number of unfiltered translations and synonyms links in wiktionaries and OPUS and the union of both

	Before filtering	**After Filtering (YaMTG)**
Terms	7,903,646	881,643
Transl./Syn.	18,080,316	5,842,279
Languages	4,324	664

Table 6: Number of terms, translations and synonyms before and after applying filters.

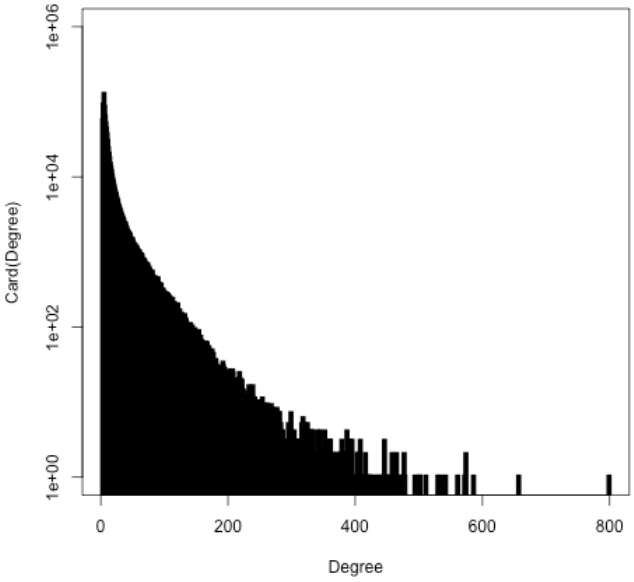

Figure 3: Degree distribution in the filtered graph YaMTG.

Figure 3 displays the degree distribution in the graph. This distribution resembles that of a small-world graph, a type of graph often encountered when modeling real-world data. Small-world graphs have tightly interconnected node clusters, and a shortest mean path length that is similar to that of a random graph that has the same number of nodes and the same number of edges. We checked the 'small-world-ness' of our translation graph against the definition of that notion as presented by Humphries and Gurney (2008).[9] It turns out that our translation graph indeed qualifies as a small-world graph.

Figure 4 illustrates the distribution of translation or synonymy pairs w.r.t. the languages at both ends of the pair.

[9]Humphries and Gurney (2008) define a graph G as being small-world by constrast with an Erdős-Rényi random graph G_{rand} that has the same number of nodes and edges than G in the following way. Let us call L_G (resp. $L_{G_{rand}}$) the average shortest path length in G (resp. G_{rand}). Moreover, let n_G^{Δ} the number of triangles in G and $n^{(}2)_G$ the number of paths of length 2 in G. We then define the transitivity-based clustering coefficient in G as $C_G^{\Delta} = 3n_G^{\Delta}/n^{(}2)_G$. The same definition holds for G_{rand}. Then, G

Source	total	Erroneous				Good	
		Misc.	lang.	def.	noise		
OPUS	99	10	0	0	0	89	90%
Wiktionaries	85	2	4	0	2	77	91%[10]
Both	16	0	0	0	0	16	100%
Total	200	12	4	0	2	182	91%

Table 7: Evaluation figures for the final (filtered) translation/synonymy graph, i.e., YaMGT based on a randomly selected sample of 200 edges.

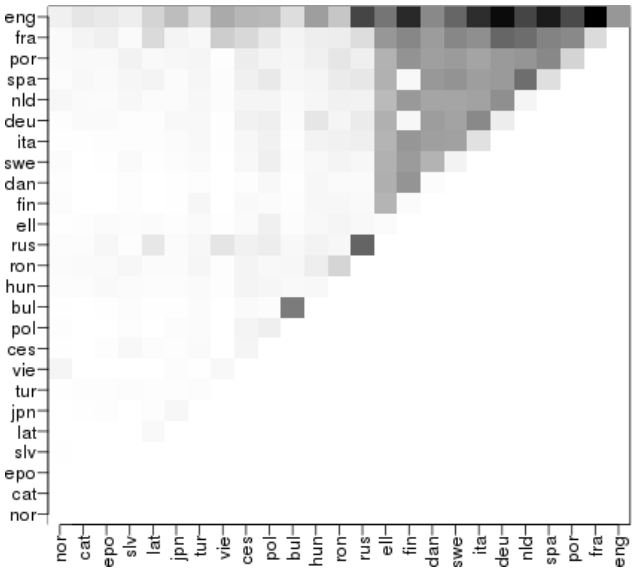

Figure 4: Language matrix for the 25 most frequent languages in the graph (darker means more translation or synonymy pairs).

We evaluated the quality of the final graph by looking over 1,639 edges selected as follows:

- 1,439 edges which were evaluated in Section 3.1.2 and still present in the final graph. This will allow us to evaluate the reliability of the filtering described in Section 5;
- 200 more edges randomly chosen in the translation graph, in order to assess its overall accuracy in a non biaised way.

These edges cover 3,332 terms in 72 languages.

6.2. Overall evaluation

We estimated the overall accuracy of the links in the filtered graph based on a manual evaluation of the 200 randomly selected pairs. The overall accuracy result is 91% (see Table 7). Amongst the 18 incorrect links in this random evaluation dataset (9%), 4 are incorrect because one of the terms have been assigned a wrong language, 2 still contain noise and 13 are miscellaneous errors (mostly alignment errors from OPUS). Note that all wikipedia definitions mistaken for translations were successfully removed by the filter.

6.3. Filter's accuracy

Amongst the 3,150 translation links evaluated in Section 3.1.2, 1,439 (45.3%) were still present in the final graph. Table 4, restricted to links extracted from wiktionaries, compares the figures before and after filtering. It displays an average accuracy score of 95.8% for the filtered extraction. Weighting the scores by the proportion of translations/synonyms from the different language editions present in the final graph (Table 8) gives us a final accuracy score of 97.2% for translations extracted from wiktionaries. On average, the filter managed to improve accuracy scores by nearly 8 points the (raw and weighted).

7. Conclusions

YaMTG is an Open-Source Heavily Multilingual Translation Graph, freely available online.[11] Its aim is mainly to provide a ready-to-use translation graph. Further versions including other translation sources are under development and will be released soon. The version of YaMTG presented in this paper contains 5.8 million translation or synonymy pairs (edges in the graph) covering 664 languages. Those translations were extracted from 21 wiktionary language editions and from the OPUS Parallel Corpora, and then filtered using generic and graph-based heuristics. We evaluated the quality of the initial set of translations extracted from our various source, the precision of our filtering step and the accuracy of its output, namely YaMTG. We estimated overall quality of the translation/synonymy links within YaMTG as approximatively 91%. If one only considers translation/synonymy links extracted from wiktionaries, this figure reach as much as 97.2%.
In the future, these scores could be improved with a two-fold strategy:

- Adding new translations to the unfiltered graph. The filtering step removes lexical items which are not present in at least X other translations. Thus, satisfying translations which tend to occur less in translation databases are more likely to be removed by the filter. More pairs in the unfiltered graph should lead to less overfiltering.

is a small-world graph whenever $L_G \geq L_{G_{rand}}$ and $C^{\Delta}_{G} \gg C^{\Delta}_{G_{rand}}$.

[10] This figure is lower than one would expect given the results displayed in Figure 4. This is an artifact related to the low number of pairs randomly selected for evaluation, resulting in a even lower number of errors among pairs extracted from wiktionaries only (8 errors). As we shall see below, the accuracy of translation pair extracted from wiktionaries is closer to 97%.

[11] `http://alpage.inria.fr/~hanoka/yamgt.html`

Source type	Language edition	#Links
Wiktionary	Slovak	826
	Hindi	4,012
	Danish	18,281
	Italian	38,031
	Japanese	38,332
	Swedish	42,842
	Czech	61,798
	Bulgarian	68,809
	Turkish	72,037
	Greek	80,328
	Vietnamese	97,622
	Polish	137,058
	German	139,136
	Romanian	147,717
	Spanish	151,981
	Hungarian	152,564
	Dutch	160,144
	Portuguese	166,669
	Russian	309,209
	French	362,404
	English	866,895
	Total	3,116,695
OPUS		3,178,247

(a)

#Sources	#Edges
1	5,449,753
2	333,179
3	58,192
4	1,142
5	13

(b)

Table 8: Number of translations and synonyms by number of distinct sources it was extracted from (subtable a). Note that a same translation link (or edge) may have been extracted from more than one sources (e.g., wiktionary language editions) and can be therefore counted more than once (see subtable b).

- Similarly, playing with the thresholds X and Y (see section 5) may improve the quality of the graph. But increasing X requires to have a sufficient amount of translations.

8. References

Atserias, J., Villarejo, L., Rigau, G., Agirre, E., Carroll, J., Magnini, B., and Vossen, P. (2004). The meaning multilingual central repository. In *In Proceedings of the Second International WordNet Conference*, pages 80--210.

Baldwin, Timothy, Pool, Jonathan, and Colowick, Susan M. (2010). Panlex and lextract: Translating all words of all languages of the world. In *COLING 2010, 23rd International Conference on Computational Linguistics, Demonstrations Volume, 23-27 August 2010, Beijing, China*, pages 37--40.

Bhat, Brijesh, Poddar, Lahari, and Bhattacharyya, Pushpak. (2013). IndoNet: A Multilingual Lexical Knowledge Network for Indian Languages. In *in Proceedings of the Annual Meeting of the Association for Computational Linguistics (ACL)*, Sofia, Bulgaria.

de Melo, Gerard and Weikum, Gerhard. (2009). Towards a universal wordnet by learning from combined evidence. In *Proceedings of the 18th ACM conference on Information and knowledge management*, CIKM '09. ACM.

de Melo, Gerard. (2012). *Graph-based Methods for Large-Scale Multilingual Knowledge Integration*. universaar - Saarland University Press, Saarbrücken, Germany.

Diab, Mona T. (2004). The feasibility of bootstrapping an arabic wordnet leveraging parallel corpora and and an english wordnet. In *Proceedings of the Arabic Language Technologies and Resources*, Cairo.

Etzioni, Oren, Reiter, Kobi, Soderland, Stephen, and Sammer, Marcus. (2007). Lexical translation with application to image search on the web.

Farreres, Xavier, Rigau, German, and Rodríguez, Horacio. (1998). Using wordnet for building wordnets. *Computing Research Repository*, cmp-lg/980.

Fellbaum, Christiane, editor. (1998). *WordNet: An Electronic Lexical Database*. MIT Press, Cambridge, Massachusetts.

Hamp, Birgit and Feldweg, Helmut. (1997). Germanet - a lexical-semantic net for german. In *In Proceedings of ACL workshop Automatic Information Extraction and Building of Lexical Semantic Resources for NLP Applications*, pages 9--15.

Humphries, Mark D and Gurney, Kevin. (2008). Network `small-world-ness': a quantitative method for determining canonical network equivalence. *PLoS One*, 3(4):e0002051.

Mausam, Soderland, Stephen, Etzioni, Oren, Weld, Daniel S., Skinner, Michael, and Bilmes, Jeff. (2009). Compiling a massive, multilingual dictionary via probabilistic inference. In *Proceedings of the 47th Annual Meeting of the ACL and the 4th International Conference on NLP of the AFNLP*, ACL '09, pages 262--270, Suntec, Singapore.

Meyer, Christian M. and Gurevych, Iryna. (2012). Wiktionary: a new rival for expert-built lexicons? exploring the possibilities of collaborative lexicography. In Granger, Sylviane and Paquot, Magali, editors, *Electronic Lexicography*, chapter 13, pages 259--291. Oxford: Oxford University Press, November.

Navigli, Roberto and Ponzetto, Simone Paolo. (2010). BabelNet: Building a Very Large Multilingual Semantic Network. In *Proceedings of the 48th Annual Meeting of the Association for Computational Linguistics*, pages 216--225, Uppsala, Sweden.

Navigli, Roberto and Ponzetto, Simone Paolo. (2012). BabelNet: The automatic construction, evaluation and application of a wide-coverage multilingual semantic network. *Artificial Intelligence*, 193:217--250.

Pianta, Emanuele, Bentivogli, Luisa, and Girardi, Christian. (2002). Multiwordnet: developing an aligned multilingual database. In *Proceedings of the First International Conference on Global WordNet*.

Sagot, Benoît and Fišer, Darja. (2008). Building a free

french wordnet from multilingual resources. In *Ontolex 2008*, Marrakech, Morocco.

Sérasset, Gilles. (2012). Dbnary: Wiktionary as a lmf based multilingual rdf network. In Chair), Nicoletta Calzolari (Conference, Choukri, Khalid, Declerck, Thierry, Doğan, Mehmet Uğur, Maegaard, Bente, Mariani, Joseph, Odijk, Jan, and Piperidis, Stelios, editors, *Proceedings of the Eight International Conference on Language Resources and Evaluation (LREC'12)*, Istanbul, Turkey, may. European Language Resources Association (ELRA).

Stamou, Sofia, Oflazer, Kemal, Pala, Karel, Christoudoulakis, Dimitris, Cristea, Dan, Tufis, Dan, Koeva, Svetla, Totkov, George, Dutoit, Dominique, and Grigoriadou, Maria. (2002). Balkanet a multilingual semantic network for the balkan languages. *Proceedings of the International Wordnet Conference, Mysore, India*, pages 21--25.

Tiedemann, Jörg. (2009). News from OPUS - A Collection of Multilingual Parallel Corpora with Tools and Interfaces. In Nicolov, N., Bontcheva, K., Angelova, G., and Mitkov, R., editors, *Recent Advances in Natural Language Processing (vol V)*, pages 237--248. John Benjamins, Amsterdam/Philadelphia.

Vossen, Piek, editor. (1998). *EuroWordNet: a multilingual database with lexical semantic networks*. Kluwer Academic Publishers, Norwell, MA, USA.