

Dbinary: Extracting Wiktionary Data for 12 Languages with Enhanced Translation Relations

Gilles Sérasset

Andon Tchechmedjiev

UJF-Grenoble 1, Laboratoire d’Informatique de Grenoble
GETALP Team, BP 53, 38051 Grenoble cedex 9, France
`gilles.serasset@imag.fr`

Abstract

After winning the Monnet Challenge in 2012, we continued our efforts in extracting multilingual wiktionary data. This data, made available as Linked Data structured using the LEMON Model, now contains 12 language editions. This short paper presents the current status of the dbnary dataset.

The extracted data is registered at <http://thedatahub.org/dataset/dbnary>.

Keywords: Wiktionary, Multilingual Lexical Resource, Lexical Networks, LEMON, RDF.

1 Introduction

The GETALP (Study group for speech and language translation/processing) team of the LIG (Laboratoire d’Informatique de Grenoble) is in need for multilingual lexical resources that should include language correspondences (translations) and word sense definitions. In this regard, the set data included in the different wiktionary language edition is a precious mine.

Alas, many inconsistencies, errors, difference in usage do exist in the various wiktionary language edition. Hence, we decided to provide an effort to extract precious data from this source and provide it to the community a Linked Data. This dataset won the Monnet Challenge in 2012, when it consisted of 6 language editions. The structure of this dataset, which is intensively based on LEMON is presented in Sérasset [2012]. This short paper purpose is to present the current state our dataset.

2 Extracting Data from Wiktionary

2.1 No Common Approach

Errors and incoherences are inherent to a contributive resource like wiktionary. This has been heavily emphasized in related works (Hellmann et al. [2013], Meyer and Gurevych [2012a]). Still, we succeeded not only in extracting data from 12 different language editions, but we are maintaining these extractor on a regular basis. Indeed, our dataset evolves along with the original wiktionary data. Each time a new wiktionary dump is available (about once every 10/15 days for each language edition), the dbnary dataset is updated. This leads to a different dataset almost every day.

Some language editions (like French and English) have many moderators that do limit the number of incoherence among entries of the same language. Moreover, those languages that

contain the most data, use many *templates* that simplify the extraction process. For instance, the translation section of the French dictionary usually uses a template to identify each individual translation.

This is not true however, with less developed Wiktionary language editions. For instance, in the Finnish edition, some translations are introduced by a template giving the language (e.g. {fr} precedes French translation) and others are introduced by the string "ranska" which is the Finnish translation for "French". In this case the translator needs to know the Finnish translation of all language names to cope with the second case and avoid losing almost half of the available translation data.

Moreover, since 2012, we have added new languages that exhibits a different use of the Wikimedia syntax. For instance, translations in the Russian wiktionary are entirely contained in one unique template, where target languages are a parameter. Moreover, in the Bulgarian Wiktionary, the full lexical entry is contained in one single template where sections are the parameters. In such language editions, templates can not be parsed using regular expressions, as they are inherently recursive (template calls are included in parameter values of other templates). This invalidates our initial approach which was based on regular expressions. In order to cope with these languages, we had to use an advanced parser of the Wikimedia syntax (called Bliki engine¹) to deal with such data.

Our extractors are written in java and are open-source (LGPL licensed, available at <http://dbnary.forge.imag.fr>).

2.2 Tools to Help Maintenance

In this effort, we also had to develop tools to evaluate the extarctor's performance and to maintain it. Our first tool² compares extracted translations with interwiki links. Many of the translations in a wiktionary language edition do point to entries in the Wiktionary edition of the target language. Such inter-wiki links are available through the Wiktionary API. By randomly sampling the extracted data, we are able to compare the extracted data with such links. This gives us an idea of the extractor performance. However, this relies on the availability of inter-wiki links, which is not the case in some language edition.

When we maintain the extractor, we need to carefully check that the patches we added do not introduce regressions in the extractor. For this, we developped our own **RDFdiff** command line tool which computes the differences between 2 RDF dumps. Such a command is already provided in the JENA toolbox, however, the JENA implementation does not correctly deal with anonymous nodes. Indeed, anonymous nodes are always considered as different by the JENA implementation when the RDF specification states that 2 anonymous nodes that share the same properties should be considered equal. Our version of RDFDiff correctly handles such anonymous node (that are heavily used in the LEMON model). With this implementation, it is now easy to compute the difference between the original extraction and the new one and to decide, based on these differences, if the new version is good enough for production.

From time to time, a Wiktionary language edition drastically changes the way it encodes some data. Actively following the discussions on each Wiktionary edition to anticipate such changes is not an option with so many languages. Hence, with each language extraction update, we compute a set of statistics that gives detailed figures on the size of the data . These statistics are available live on the dbnary web site³. Overall, the most useful statistics are the ones that capture the evolution of the extracted data over time. For instance Figure 1 shows the evolution

¹<https://code.google.com/p/gwtwiki/>

²this heuristic was initially suggested by Sebastian Hellman

³<http://kaiko.getalp.org/about-dbnary>

of the size of the extracted French datasets since its original extraction. This plot allowed us to detect that a major refactoring was happening on the French language edition. This allowed us to patch the extractor for this new organisation right away.

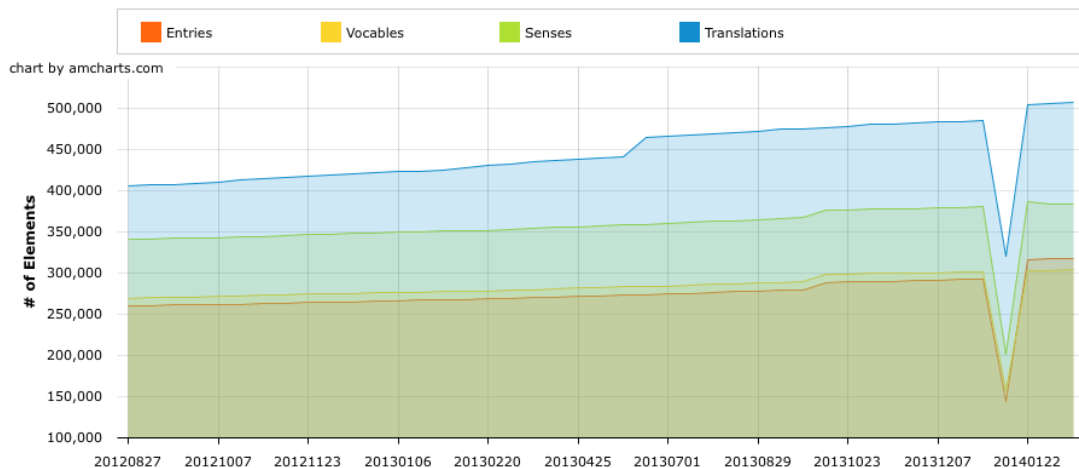


Figure 1: Some of the statistics available about the French extracted data.

3 Extracted Data as a LEMON Lexical Resource

3.1 Extracted Entries

The main goal of our efforts is not to extensively reflect the specific structures and constructs of wiktionary data, but to create a lexical resource that is structured as a set of monolingual dictionaries + bilingual translation information. Such data is already useful for several application, but most importantly it is a sound starting point for a future multilingual lexical database.

Monolingual data is always extracted from its dedicated Wiktionary language edition. For instance, the French lexical data is extracted from the French language edition (the data is available on <http://fr.wiktionary.org>). However, we do not capture as of yet, any of the French data that may be found in other language editions.

We also filtered out some parts of speech in order to produce a result that is closer to existing monolingual dictionaries. For instance, in French, we disregard abstract entries corresponding to prefixes, suffixes or flexions (e.g.: we do not extract data concerning *in-* or *-al* that are prefixes/suffixes and that have a dedicated page in the French language Edition).

Given that the scope and main focus of our work is the production of lexical data, we do not provide any reference or alignment to any ontology (toward top-level concepts for example).

3.2 LEMON and non-LEMON modelled Extracted Data

All of the extracted data could not be structured using solely the LEMON model. For instance, LEMON does not contain any mechanisms that allow to represent translations between languages, as the underlying assumption is that such translation will be handled by the ontology

description. Moreover, LEMON further assumes that all data is well-formed and fully specified. As an example, the synonymy relation is a property linking a *Lexical Sense* to another *Lexical Sense*. While this is a correct assumption in *principle*, it does not account for the huge amount of legacy data that is available in dictionaries and lexical databases and that isn't disambiguated.

In order to cope with such legacy data, we introduced several classes and properties that are not LEMON entities. However, we make sure that whenever a piece of data can be represented as a LEMON entity, it is indeed represented as such. Most of these points have already been covered in Sérasset [2012].

3.3 Links to other datasets

The dbnary dataset makes use of other datasets. Firstly, while all extracted lexical entries are associated with a language-specific part of speech that is given by its original Wiktionary language edition, we also add, when available a `lemon:partOfSpeech` relation to a standard value defined in the *lexinfo* ontology⁴ (Buitelaar et al. [2009]). Secondly, while the LEMON model uses a string value to represent languages, we additionally use the property `dcterms:lang` to point to a language entity defined in the *lexvo* ontology (de Melo and Weikum [2008]).

3.4 Disambiguation of translation sources

Many of the translations present in Wiktionary are associated with a hint used by human users to identify the sense of the source of the translation. Depending on the language, this hint may take the form of a sense number (e.g. in German and Turkish), of a textual gloss (e.g. English) or of both a sense number and a textual gloss (e.g. French, Finnish).

By using an adaptation of various textual and semantic similarity techniques based on partial or fuzzy gloss overlaps, we were able to disambiguate the translation relations. We obtained F-measures of the order of 80% (on par with similar work on English only, such as Meyer and Gurevych [2012b]), across the three languages where we could generate a gold standard (French, Portuguese, Finnish). We have shown that most of the disambiguation errors are due to inconsistencies in Wiktionary itself that cannot be detected at the generation of DBnary (shifted sense numbers, inconsistent glosses, etc.).

The relations between translations and lexical senses has also been made part of this dataset.

3.5 Size of the involved data

At the time of the writing, the statistic about extracted data from the most up-to-date dump files are the following:

4 Conclusion and Perspectives

The present article exhibits some preliminary results on what is essentially an open source tool to extract a LEMON based lexical network from various Wiktionary language editions. Such a work is interesting for many users that will be able to use the extracted data in their own NLP systems. Moreover, as the extracted resource uses the Resource Description Framework (RDF) standard and the LEMON model, the extracted data is also directly usable for researchers in the field of the Semantic Web, where it could be used to ease the construction of ontology alignment systems when terms in different languages are used to describe the ontologies of a domain.

⁴<http://www.lexinfo.net/ontology/2.0/lexinfo>

Language	Entries	Vocables	Senses	Translations	Total
Bulgarian	18,831	27,071	18,798	13,888	78,588
English	553,499	528,341	447,073	1,332,332	2,861,245
Finnish	50,813	50,488	59,612	122,724	283,637
French	318,101	304,465	383,242	507,359	1,513,167
German	211,564	282,902	102,468	390,938	987,872
Italian	34,363	101,525	45,022	62,305	243,215
Japanese	25,492	25,637	29,679	87,906	168,714
Modern Greek (1453-)	246,211	241,845	137,072	57,615	682,743
Portuguese	45,788	45,968	81,807	267,801	441,364
Russian	130,879	143,653	116,925	365,389	756,846
Spanish	58,679	65,854	85,852	114,951	325,336
Turkish	64,899	69,383	91,418	66,928	292,628
Total	1,759,119	1,887,132	1,598,968	3,390,136	8,635,355

Table 1: Number of lexical elements in the graphs.

Language	<i>syn</i>	<i>qsyn</i>	<i>ant</i>	<i>hyper</i>	<i>hypo</i>	<i>mero</i>	<i>holo</i>	Total
Bulgarian	17632	0	34	0	0	0	0	17666
English	31762	0	6980	1252	1212	112	0	41318
Finnish	2478	0	0	0	0	0	0	2478
French	31655	2133	6879	9402	3739	970	1898	56676
German	29288	0	15079	33251	10413	0	0	88031
Italian	9662	0	3425	0	0	0	0	13087
Japanese	3828	0	1578	9	14	0	0	5429
Greek	4990	0	1428	0	0	0	0	6418
Portuguese	3350	0	556	6	4	0	0	3916
Russian	24941	0	9888	22832	5140	0	0	62801
Spanish	15087	0	1525	741	560	0	0	17913
Turkish	3260	0	220	483	164	0	0	4127
Total	177933	2133	47592	67976	21246	1082	1898	319860

Table 2: Number of lexico-semantic relations in the graphs.

Our next objectives are to better generalize the specific processes of the current extractors, so that it will be easier to create extractors for other languages in a more systematic manner. We are currently working on the Russian language edition and we welcome all initiatives aiming at the addition of new language to this open-source tool.

5 Acknowledgements

The work presented in this paper was conducted in the Videosense project, funded by the French National Research Agency (ANR) under its CONTINT 2009 programme (grant ANR-09-CORD-026).

References

Paul Buitelaar, Philipp Cimiano, Peter Haase, and Michael Sintek. Towards linguistically grounded ontologies. In Lora Aroyo, Paolo Traverso, Fabio Ciravegna, Philipp Cimiano, Tom Heath, Eero Hyvönen, Riichiro Mizoguchi, Eyal Oren, Marta Sabou, and Elena Simperl, edi-

- tors, *The Semantic Web: Research and Applications*, volume 5554 of *Lecture Notes in Computer Science*, pages 111–125. Springer Berlin Heidelberg, 2009. ISBN 978-3-642-02120-6. doi: 10.1007/978-3-642-02121-3_12. URL http://dx.doi.org/10.1007/978-3-642-02121-3_12.
- Gerard de Melo and Gerhard Weikum. Language as a Foundation of the {Semantic Web}. In Christian Bizer and Anupam Joshi, editors, *Proceedings of the Poster and Demonstration Session at the 7th International Semantic Web Conference (ISWC)*, volume 401 of *CEUR WS*, Karlsruhe, Germany, 2008. CEUR.
- Sebastian Hellmann, Jonas Brekle, and Sören Auer. Leveraging the Crowdsourcing of Lexical Resources for Bootstrapping a Linguistic Data Cloud. *Semantic Technology*, pages 191—206, 2013. URL <http://svn.aksw.org/papers/2012/JIST\Wiktionary/public.pdf>.
- Christian M. Meyer and Iryna Gurevych. Wiktionary: a new rival for expert-built lexicons? exploring the possibilities of collaborative lexicography. In Sylviane Granger and Magali Paquot, editors, *Electronic Lexicography*, page (to appear). Oxford: Oxford University Press, 2012a. (pre-publication draft at the date of LREC).
- Christian M Meyer and Iryna Gurevych. To Exhibit is not to Loiter: A Multilingual, Sense-Disambiguated Wiktionary for Measuring Verb Similarity. In *Proceedings of COLING 2012*, pages 1763–1780, Mumbai, India, 2012b. The COLING 2012 Organizing Committee. URL <http://www.aclweb.org/anthology/C12-1108>.
- Gilles Sérasset. Dbnary: Wiktionary as a Lemon-Based Multilingual Lexical Resource in RDF. *Semantic Web Journal - Special issue on Multilingual Linked Open Data*, 2012. URL <http://www.semantic-web-journal.net/content/dbnary-wiktionary-lemon-based-rdf-multilingual-lexical-resource>.