Semantic Web 0 (0) 1
IOS Press

Dbnary: Wiktionary as a Lemon Based RDF Multilingual Lexical Resource

Gilles Sérasset *
GETALP-LIG, UJF-Grenoble 1
BP 53, 38051 Grenoble cedex 9, France
gilles.serasset@imag.fr

Abstract. Contributive resources, such as wikipedia, have proved to be valuable in Natural Language Processing or Multilingual Information Retrieval applications. This work focusses on Wiktionary, the dictionary part of the collaborative resources sponsored by the *Wikimedia* foundation. In this article, we present our effort to extract Multilingual Lexical Data from wiktionary data and to provide it to the community as a Multilingual Lexical Linked Open Data (MLLOD). This lexical resource is structured using the LEMON Model.
This data, called *dbnary*, is registered at http://thedatahub.org/dataset/dbnary.

Keywords: Wiktionary, Multilingual, Lexical Resource, LEMON, Multilingual Linguistic LOD

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 data included in the different wiktionary[1] language edition is a precious mine.

Alas, many inconsistencies, errors, difference in usage do exist in the different wiktionary language edition. Hence, we decided to provide an effort to extract precious data from this source and provide it to the community as Linked Data. After a first version that used an RDF version of the LMF model [3,2] and described in [6], we decided to adapt our extractors to LEMON model [5]. This linked dataset has won the "Monnet-Challenge" in 2012.

*E-mail: Gilles.Serasset@imag.fr.
[1] http://www.wiktionary.org

2. Extracting data from wiktionary

2.1. Motivation

Errors and incoherence are inherent to a contributive resource like wiktionary. Some language editions (like French and English) have many moderators that do limit the number of incoherence among entries of the same language. Moreover, such languages, which 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 anymore 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 translations) and others are introduced by the string "ranska" which is the Finnish translation for "French". In this case the extractor 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.

Many such inconsistencies and some errors in the data renders the development of an extractor quite te-

dious. as many people in NLP is trying to use this data for different applications, we decided that we would extract as much data as we can from wiktionary and provide it to the community while ensuring interoperability with other lexical data.

The dbnary extractor is written in java and is open-source (LGPL licensed, available at `http://dbnary.forge.imag.fr`). Anyone way contribute to this extraction effort by taking contact with the author.

2.2. Scope of the extracted data

The main goal of our efforts is not to extensively reflect 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 it is merely a starting point for a future multilingual lexical database.

The monolingual data is always extracted from its own wiktionary language edition. For instance, the French lexical data is extracted from French language edition[2]. Hence, we completely disregard the French data that may be found in other language editions.

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

Our work did focus only on the lexical data. Hence, we do not provide any reference to any ontology.

2.3. Availability of the extracted data

Dbnary data is provided using Creative Commons Attribution-ShareAlike 3.0 license (). It may be downloaded from the dbnary website[3] as a set of turtle files (one per language).

As the wiktionary language editions constantly evolve with entries modifications and additions, the dbnary also evolves. Each time the wikimedia foundation provides a new dump[4] of a wiktionary language edition, dbnary data is extracted with the new dump and made available online. At the time of writing, the dumps are updated about once every ten days for each language.

Dbnary data is also available as Linguistic Linked Open Data (LLOD). Hence, all dbnary URIs are dereferencable and a SPARQL endpoint is available at `http://kaiko.getalp.org/sparql`. However, as the dbnary data changes almost everyday, the data that is available this way is not necessarily up to date.

3. Extracted Data as a LEMON Lexical Resource

3.1. Using LEMON for legacy lexical data

LEMON model itself is not sufficient to represent lexical data that is currently available in classical monolingual and bilingual dictionaries. For instance, LEMON does not contain anything to represent translations between languages as it assumes that such a translation will be handled by the ontology description. Moreover, LEMON assumes that all data is well-formed and fully specified. As an example, synonymy relation is a property linking a LexicalSense to another LexicalSense. While this is correct to assume as a *principle*, this does not account for the huge amount of legacy data that is available in dictionaries and lexical databases.

In order to cope with this legacy data, we extended the LEMON model by adding new classes and properties. However, when a piece of data is representable as a LEMON entity, then it is done so. Moreover, when possible, we did use the ISOcat registry [7] to identify standard elements in the lexical data.

3.2. Dbnary extension to LEMON

The LEMON model has been extended to cope with legacy lexical data. Added classes and properties are:

Vocable: Several lexical entries may be contained in a single wiktionary page. And most lexical relations are simply targeted to wiktionary pages. Hence, we added the dbnary:Vocable class to represent a wiktionary page. This class is a subclass of lemon:LexicalEntry. Instances of this class are related to their lexical entries through the dbnary:refersTo property.

LexicalEntity: Lexical relations should usually link two *Lexical Senses*. However, most relations found in legacy lexical data is underspecified.

[2] We use the term "French language edition" to refer to the data available on `http://fr.wiktionary.org`

[3] `http://kaiko.getalp.org/about-dbnary`

[4] dumps are available at `http://dumps.wikimedia.org/`.

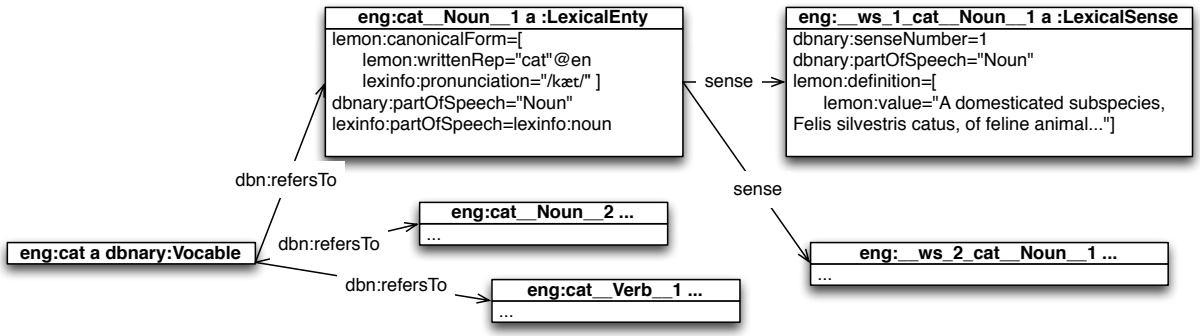

Fig. 1. An extract of the dbnary entry "cat" in English, showing the respective roles of Vocable, LexicalEntry and LexicalSense in the dbnary dataset.

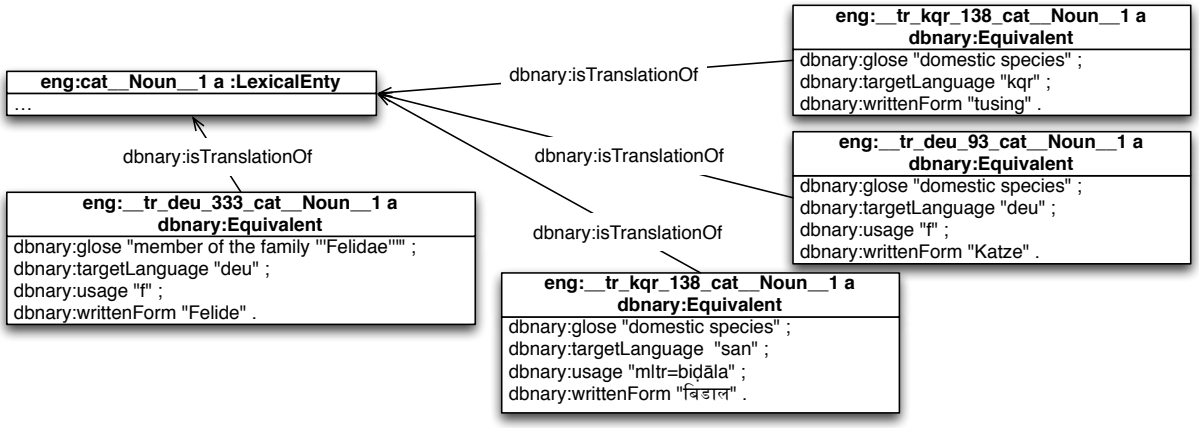

Fig. 2. A subset of the translations related to the lexical entry eng:cat_Noun_1.

Some relations link a *Lexical Sense* to a *Vocable* or to a *Lexical Entry*. Others even link two *Lexical Entries*. In order to cope in a standard way with such underspecified relations, we introduced the dbnary:LexicalEntity class which is the union of *Lexical Entry* and *Lexical Sense*.

Nyms: Most wiktionary language edition do provide "nym" relations (mainly synonym, antonym, hypernym, hyponym, meronym and holonym), which are almost always underspecified. Hence, DBnary introduces 6 new "nym" properties (in dbnary name space). These relations domains and ranges are *Lexical Entities*.

Translations: As there is no way to represent bilingual translation relation in LEMON, we introduced the dbnary:Equivalent class that collects translation information contained in wiktionary. This class admits several properties:

- dbnary:isTranslationOf relates the equivalent to its source *lexical Entity*.
- dbnary:targetLanguage is a data property whose type is a string containing the target language code as defined in ISO639-3 standard.
- dbnary:writtenForm gives the written form of the translation in the target language. Here, we decided not to relate to a vocable as some translations are not to be defined as lexical entries in the target language.
- dbnary:glose is a string property that contains any available information used to dentate the lexical sense of the source of the equivalent.
- dbnary:usage is a string property that contains any available information concerning this equivalent object. It usually gives additional information on the target entry.

3.3. Structure of the extracted lexical data

Figure 1 illustrates the main elements characterizing a lexicon entry in the dbnary data. Each wiktionary page is represented by a dbnary:Vocable element that refers to its corresponding lemon:LexicalEntry. Each lexical entry corresponds to one lemma, one etymology and one part of speech. Each lexical entry is related to its lemon:LexicalSense by the lemon:sense property. A lexical sense corresponds to one definition.

Each lexical entry is related to its canonical form and eventually to alternate forms (that are represented using lemon:LexicalForms). The part of speech is available through dbnary:partOfSpeech property that gives the part of speech as defined by the wiktionary language edition, and through the isocat:partOfSpeech property that points to a standard ISOcat part of speech value.

Figure 2 illustrates the *dbnary* extension to LEMON that is used to represent the many translations that are available in wiktionary. Each translation is represented as a dbnary:Equivalent instance that is associated to a *lexical entry* by the isTranslationOf property. The translation is given as a string through the writtenForm property, as some translations may not necessarily correspond to vocables in the target language (e.g. explanatory translations). The target language of the translation is given using ISO 639-3 3 letter language code [4]. When available, the glose property value gives an indication concerning the lexical sense that is translated. In the current dataset, the translation is linked to an usually ambiguous lexical entry and the *glose* is kept for further attachment to the correct word sense.

Wiktionary contains many lexical relations (like synonymy, antonymy, etc.) that are represented using the above mentioned "nym" properties which relate a *lexical entity* with another *lexical entity*. In the current dataset, most of the property subjects are *lexical entries*. However, in case of monosemous lexical entries, the synonymy relation is attached to the unique lemon:LexicalSense. Figure 3 illustrates how dbnary encodes "nym" relations with examples of the dbnary:synonym property. In the upper part of the figure is an example of a dbnary:synonym property that is related to the ambiguous lexical entry eng:cat__Noun__1. The lower part shows that, in the monosemous lexical entry eng:voyager__Noun__1, the dbnary:synonym property is related to the unique lexical sense of the entry. The object of such properties are always dbnary:Vocable.

4. Size of the involved data

All sizes indicated in this sections reflect the state of the dbnary data at the time of writing (December 2012). Table 1 give the number of resources available in dbnary.

	Entries	Vocables	Senses	Equivalents
eng	502493	481311	402815	1021430
fra	264803	274854	347076	419168
deu	106337	171517	90474	446563
fin	42813	43158	51297	114279
por	42042	43028	76124	197931
ita	24473	30568	34133	58383

Table 1

Number of resources by type and language, sorted by number of lexical entries.

As the extraction is performed each time a wiktionary dump is available, this numbers are constantly evolving, as the wiktionary data is evolving and as the extractor itself may be improved.

Table 2 gives an overview of the number of lexico-semantic relations available in each language edition.

	syn	ant	hyper	hypo	mero	holo
eng	30273	6621	830	961	103	0
fra	29986	6471	7356	3428	900	1797
deu	25889	13756	27771	9011	0	0
ita	6430	2027	0	0	0	0
por	3493	571	6	2	0	0
fin	1665	0	0	3	0	0

Table 2

Number of lexicon-semantic relations, sorted by number of synonymy relations.

Table 3 shows the number of translation equivalents available in each language edition, for the major target languages[5]. It also gives the total number of translations and the number of different target language with translations. Not surprisingly, English language edition shows the most translation equivalent to more than 1000 different languages. Surprisingly, the French language edition shows a relatively small number of tar-

[5]The target languages are the 21 having the largest page count in their wiktionaries, i.e. (by increasing order of the ISO 639-3 language codes): German (deu), Greek (ell), English (eng), Finnish (fin), French (fra), Hungarian (hun), Ido (ido), Kannada (kan), Korean (kor), Kurdish (kur), Lithuanian (lit), Malagasy (mlg), Dutch (nld), Polish (pol), Portuguese (por), Russian (rus), Swedish (swe), Tamil (tam), Turkish (tur), Vietnamese (vie), Chinese (zho)

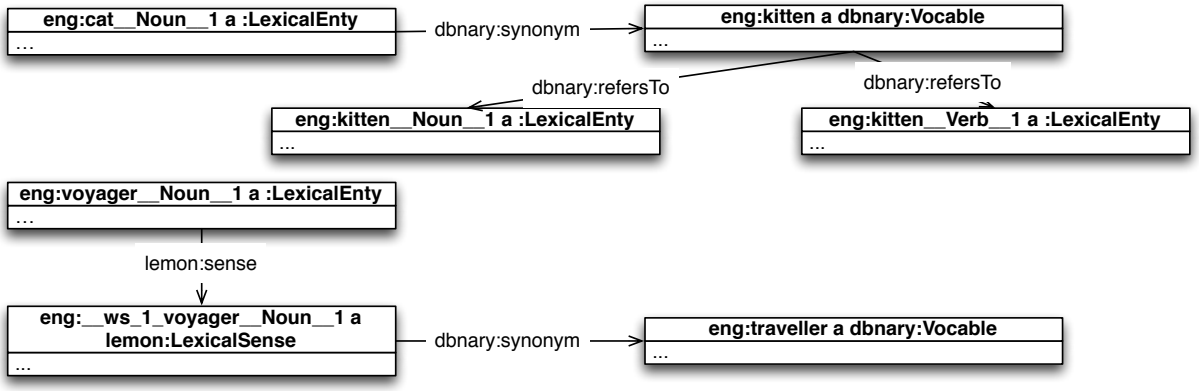

Fig. 3. An example of a synonymy relation for the lexical entry eng:cat_Noun_1.

Source/Target	deu	ell	eng	fin	fra	hun	ido	kan	kor	kur	lit	mlg	nld
deu	0	2596	76248	4695	63909	7066	345	78	1361	1262	1761	60	11055
eng	56880	22047	1	70803	53702	21367	2454	629	13294	1639	5027	211	38066
fin	7626	892	27827	0	6560	2164	905	9	242	159	597	17	2011
fra	32897	6799	71560	7360	7	5096	11442	144	3966	1772	1826	90	29454
ita	4600	529	16614	964	4403	831	175	19	284	112	331	8	2016
por	6186	2492	9795	4046	7635	3192	1831	161	2026	841	2264	293	4687

Source/Target	pol	por	rus	swe	tam	tur	vie	zho	others	Total	*target languages*
deu	15551	8677	15451	46765	222	4726	436	5528	178771	446563	*358*
eng	21989	28049	66023	30198	1329	17271	7589	2	562860	1021430	*1081*
fin	2447	2177	7945	9029	27	1590	94	248	41713	114279	*304*
fra	7507	17206	7483	12152	639	3926	1602	4483	191757	419168	*179*
ita	908	1254	973	908	73	531	198	583	22069	58383	*319*
por	4346	3	3867	4394	540	2845	1000	2933	132554	197931	*678*

Table 3

Number of translation equivalents in each language edition, detailed by target language, for the 21 biggest wiktionary languages, sorted by alphabetic order on ISO language code. The number of different target language is also given.

get languages, while the number of equivalents stays relatively high.

5. Conclusion and Perspectives

The current paper presents the dbnary dataset that is a LEMON based lexical network from different wiktionary language editions. Such a work is interesting for many users that will be able to use the extracted data in their own NLP system. 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 on the Semantic Web, where it could be used to ease the ontology alignment systems when terms in different languages are used to describe ontologies of a domain.

This resources describes a significative number of entries for at least English and French languages that makes it comparable to Wordnet [1]. Moreover, it contains many translation equivalents with certain language pairs that makes it also comparable to many existing multilingual Wordnets

Our next objectives are to better generalize the treatments of the current extractors, so that it will be easier to create and maintain extractors for other languages. We are currently working on the Russian and we welcome all initiative aiming at the addition of new language to this open-source tool.

We will also enhance the *dbnary* data by providing more lexico-semantic relations and translations aligned on Lexical Senses.

6. 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

[1] C. Fellbaum. *WordNet: An Electronic Lexical Database*. MIT Press., Cambridge, MA, 1998.

[2] G. Francopoulo, N. Bel, M. George, N. Calzolari, M. Monachini, M. Pet, and C. Soria. Lexical markup framework (LMF) for NLP multilingual resources. In *Proceedings of the Workshop on Multilingual Language Resources and Interoperability*, pages 1–8, Sydney, Australia, July 2006. Association for Computational Linguistics. URL http://www.aclweb.org/anthology/W/W06/W06-1001.

[3] G. Francopoulo, M. George, N. Calzolari, M. Monachini, N. Bel, M. Pet, and C. Soria. Lexical Markup Framework (LMF). In *International Conference on Language Resources and Evaluation - LREC 2006*, Gênes/Italie, 2006. elra. URL http://hal.inria.fr/inria-00121468/en/. LIRICS.

[4] ISO639-3. Codes for the representation of names of languages — part 3: Alpha-3 code for comprehensive coverage of languages. ISO 639-3:2007, 2007.

[5] J. McCrae, G. Aguado-de Cea, P. Buitelaar, P. Cimiano, T. Declerck, A. Gómez-Pérez, J. Gracia, L. Hollink, E. Montiel-Ponsoda, D. Spohr, and T. Wunner. Interchanging lexical resources on the semantic web. *Language Resources and Evaluation*, May 2012. ISSN 1574-020X. . URL http://dx.doi.org/10.1007/s10579-012-9182-3.

[6] G. Sérasset. Dbnary: Wiktionary as a LMF based Multilingual RDF network. In *Language Resources and Evaluation Conference, LREC 2012*, Istanbul, 2012. URL http://www.lrec-conf.org/proceedings/lrec2012/pdf/387_Paper.pdf.

[7] M. Windhouwer and S. Wright. Linking to linguistic data categories in ISOcat. *Linked Data in Linguistics*, pages 99–107, 2012.