HAL
open science

Dbnary: Wiktionary as a LMF based Multilingual RDF network

Gilles Sérasset

To cite this version:

Gilles Sérasset. Dbnary: Wiktionary as a LMF based Multilingual RDF network. Language Resources and Evaluation Conference, LREC 2012, Nicoletta Calzolari and Khalid Choukri and Thierry Declerck and Mehmet Uğur Doğan and Bente Maegaard and Joseph Mariani and Jan Odijk and Stelios Piperidis, May 2012, Istanbul, Turkey. hal-00954046

HAL Id: hal-00954046

https://inria.hal.science/hal-00954046v1

Submitted on 25 Mar 2014

HAL is a multi-disciplinary open access archive for the deposit and dissemination of scientific research documents, whether they are published or not. The documents may come from teaching and research institutions in France or abroad, or from public or private research centers.

L'archive ouverte pluridisciplinaire **HAL**, est destinée au dépôt et à la diffusion de documents scientifiques de niveau recherche, publiés ou non, émanant des établissements d'enseignement et de recherche français ou étrangers, des laboratoires publics ou privés.

Dbnary: Wiktionary as a LMF based Multilingual RDF network

Gilles Sérasset

UJF-Grenoble 1, Laboratoire d'Informatique de Grenoble
GETALP Team, 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 article focusses on Wiktionary, the dictionary part of the collaborative resources sponsored by the *Wikimedia* foundation.

In this article we present a word net that has been extracted from French, English and German wiktionaries. We present the structure of this word net and discuss the specific extraction problems induced by this kind of contributive resources and the method used to overcome them.

Then we show how we represent the extracted data as a Lexical Markup Framework (LMF) compatible lexical network represented in Resource Description Framework (RDF) format.

Keywords: Wiktionary, Multilingual Lexical Database, Lexical Networks, LMF, RDF.

1. Introduction

Wiktionary is a huge and free resource available on the web. Its main advantages are the presence of definitions that could help for disambiguation tasks and the large number of translations to many different languages. The drawback of this resource is the fact that the entries are described using a wiki syntax specifying the *form* of the entry rather than its *structure*. Moreover this description is sometime erroneous or heterogeneous.

The goal of the dbnary project is to provide an extraction process that produces a lexical word net as detailed as possible from wiktionary dumps. The extracted data can be used, as is, in another project[1] or the extraction process itself can be integrated into another tool (for example, to have an on-demand extraction using the latest available data) as it is available as part of an open source project [2]. Many efforts have already been attempted to use wiktionary data in NLP applications. Most of them were ad-hoc efforts and some of them provided either an API to, or an XML dump of, the extracted data. But the wiktionary data is an *evolving* resource. It means that the data *along with its encoding* changes while time goes on. Hence, the extraction program has to cope with the evolving usages of the contributors. Moreover, each wiktionary language edition uses its own encoding and usages to represent lexical information. We do believe that we can solve both problems by providing the extracted data *and* the extraction program as an open source system.

In this paper, we will first give a very short and general description of the lexical structure of the main language editions of wiktionary. Then we address the main difficulties we met when extracting data from the different wiktionary dumps. We will then show how the extraction program is organized to allow its maintenance and extension by its users. Finally we will present the structure of the extracted data that is based on Lexical Markup Framework (LMF) standard, and stored as a RDF graph.

[1] Latest extracted data is available at http://kaiko.getalp.org/dbnary/

[2] http://dbnary.forge.imag.fr/

2. Wiktionary and its data

2.1. Overview

Wiktionary[3] is a web based collaborative effort led by the Wikimedia foundation[4] to build a free content dictionary in many languages.

2.2. Macro- and Micro- Structures

(Meyer and Gurevych, 2012) give an extended description of wiktionary. In this section, we will provide an overview of the elements that are pertinent to this study.

Wiktionary organizes its data in a way that may be surprising for a lexicographer. This may be explained by the contributive approach used for building the resources and by the intended user experience. The key concepts used in wiktionary guidelines are also mainly motivated by the technology used to pursue this collaborative effort.

Wiktionary is organized as a set of wiktionary *language editions* (one per language) containing a set of *pages* characterized by a *page name*. Each page contains lexical data from different languages. In a wiktionary language edition (say the edition of language l_1), all lexical data (including data from other languages) are described using language l_1. Dictionary articles may be related to other articles in the same language edition (via lexico-semantic or translation links). They may also be related by translation links to articles on another edition. Pages may also be related to pages with the same *page name* in other editions.

Under this organization, each edition is (ultimately) intended to contain all lexical data of all languages described in the edition language.

2.3. Anatomy of a wiktionary page

While the details of the structure of lexical data differs between wiktionary language editions, wiktionary uses a

[3] http://www.wiktionary.org/

[4] http://www.wikimedia.org/

Page	*chat*	
Language*	1 English	*English language comes first in the English wiktionary*
	1.1 Pronunciation[a]	
Etymology*	1.2 Etymology 1	*Etymology is used to distinguish homographs*
Core entry[b]*	1.2.1 Verb	*The core entry contains the definitions of the entry*
Definition*	*To be engaged in informal conversation.*	*list of word senses, described by their definitions, along with examples.*
	...	
LexRels*		*list of lexico-semantic relations*
Translation*	*Danish: snakke, sludre*	*list of translations grouped by word senses when available*
	1.2.2 Noun	*Core entries are distinguished by their part of speech*
	1.3 Etymology 2	
	1.3.1 Noun	
	3 Dutch	*Entries in other languages are structured according to the same principles*
	4 French	
	...	

[a]This particular entry does not follow the rule that states that pronunciation should be given *after* the Etymology.
[b]This term is the one used in English Wiktionary guidelines.

Figure 1: Annotated table of contents of the entry "chat" in the English wiktionary.

common general abstract structure similar to the one found in main paper dictionaries. We illustrate this general structure with the structure of a page on the English server. Figure 1 shows the annotated table of contents of *"chat"* page on the English edition[5].

```
{{also|chăt}}
==English==
{{wikipedia}}

===Pronunciation===
* {{IPA|/tʃæt/}}
* {{audio|en-us-chat.ogg|Audio (US)}}
* {{rhymes|æt}}

===Etymology 1===
Abbreviation of ''[[chatter]]''.
{{rfe|bird and louse also from chatter??}}

====Verb====
{{en-verb|chat|t|ing}}
[[image:Wikimania 2009 - Chatting (3).jpg|thumb|Two people
'''chatting'''. (1) (2)]]

# To be [[engage]]d in informal [[conversation]].
#: ''She '''chatted''' with her friend in the cafe.''
#: ''I like to '''chat''' over a coffee with a friend.''
# To talk more than a few words.
#: ''I met my old friend in the street, so we '''chatted''' for a
while.''
# To exchange [[text]] or [[voice]] [[message]]s in [[real time]]
through a [[computer network]], as if having a [[face-to-face]]
conversation.
#: ''Do you want to '''chat''' online later?''

=====Translations=====
{{trans-top|be engaged in informal conversation}}
* Bulgarian: {{t|bg|бъбря}}
* Catalan: {{t|ca|xerrar}}
* Danish: {{t-|da|snakke}}, {{t|da|sludre}}
* Dutch: {{t+|nl|kletsen}}, {{t+|nl|babbelen}}
* Finnish: {{t+|fi|jutella}}, {{t-|fi|rupatella}},
{{t|fi|turista}}, {{t|fi|horista}}, {{t|fi|lörpötellä}}
* German: {{qualifier|formal}} sich {{t+|de|unterhalten}},
{{qualifier|informal}} {{t+|de|plaudern}},
{{qualifier|dialectal}} {{t-|de|schwatzen}}, {{t-|de|klönen}}
* Greek: {{t|el|ψιλοκουβεντιάζω|tr=psilokouventiázo}},
{{t|el|συνομιλώ|tr=synomiló}}
{{trans-mid}}
* Latin: {{t|la|fabulor}}
* Maltese: {{t|mt|tpaċpiċ}}
* Persian: {{t|fa|گپیدن|tr=gapidan|sc=fa-Arab}}
* Russian: {{t+|ru|разговаривать|tr=razgovárivat'}} {{impf}},
{{t+|ru|болтать|tr=boltát'}} {{impf}}
* Sichuan Yi: {{tø|ii|ꑍꉆ|tr=syp mgep}}
* Spanish: {{t+|es|charlar}}, {{t-|es|platicar}}
* Swedish: {{t|sv|småprata}}, {{t|sv|konversera}},
{{t|sv|tjatta}}, {{t|sv|snacka}}
```

Figure 2: Excerpt of the entry "chat", as available in the English language edition, in mediawiki syntax.

2.4. Internal representation of a wiktionary page

All collaborative efforts led by the Wikimedia foundation involves the use of wikis to collect semi-structured collaborative data. The software used for these wikis is MediaWiki (WikiMedia, 2012). Under MediaWiki, each page is defined using a specific syntax that describes its *formatting*. This formatting language may be extended by defining *templates* that will be expanded during page rendering. Most wiktionary language editions use specific templates to help users format their contributed lexical data in a coherent way. Figure 2 shows a sample of the internal representation of the entry "chat", in the English language edition, where one can note some templates (surrounded by double curly braces) and links (surrounded by double square brackets). Figure 3 gives an excerpt of the entry "chat" in the French language edition.

```
{{voir|Chat}}
== {{=fr=}} ==

{{-étym-}}
: ''([[#Nom commun 1|Nom 1]])'' Du bas
{{étyl|la|fr|mot=cattus|sens=chat ([[domestique]])}}. Le
classique ''[[felis]]'' désignait le [[w:Chat sauvage
d'Europe|chat sauvage d'Europe]] (''Felis silvestris
silvestris'').
: ''([[#Nom commun 2|Nom 2]])'' Emprunté à
l'{{étyl|en|fr|mot=chat|sens=[[discussion]], [[bavardage]]}}.

{{-nom-|num=1|fr}}
{{fr-rég|ʃa}}
[[Fichier:Kamee01.jpg|thumb|Un '''chat''' domestique attentif.
(1,2)]]
'''chat''' {{pron|ʃa|fr}} {{m|équiv=chatte}}
# {{animaux|fr}} [[mammifère|Mammifère]] [[carnivore]] [[félin]]
de taille moyenne, au [[museau]] court et arrondi, domestiqué,
apprivoisé ou encore à l'état sauvage.
#* ''Entre autres manies, mon oncle avait celle de tuer tous les
'''chats''' qu'il rencontrait. Il faisait, à ces pauvres bêtes,
une chasse impitoyable, une guerre acharnée de trappeur.''
{{source|{{w|Octave Mirbeau}}, ''Mon oncle'',}}
#* ''Soudain, d'un seul élan, ''cela'' se précipita sur lui, avec
un miaulement plaintif et la queue droite. C'était un jeune
'''chat''', menu et décharné, qui frottait sa tête contre les
jambes de Bert, en ronronnant.'' {{source|[[w:H.G. Wells|H.G.
Wells]], ''{{w|La Guerre dans les Airs}}'', 1908 - Traduit par
Henry-D. Davray & B. Kozakiewicz, page 335, Mercure de France,
1921}}
# Individu [[mâle]] de cet animal.
#* ''Le '''chat''' ronronne sur mes genoux.''
#* ''Le '''chat''' en regardant passer la '''chatte''' du voisin
se sent tout chose.''
# {{par ext}} [[félin|Félin]].
#* '''Chat-tigre''' Espèce de félin sauvage de l'Amérique
méridionale.''
```

Figure 3: Excerpt of the entry "chat", as available in the French language edition, in mediawiki syntax.

The Wikimedia foundation provides a monthly dump[6] of all pages of a wiktionary in this formatting language.
The dumps are extracted and encoded as UTF-16 text file. The extraction process goes through all pages in the dump file (using the provided xml based structure). Then each page is parsed using a finite state automaton implemented

[5]This example has been extracted in february 2012. It is likely that the entry changed since then.

[6]http://dumps.wikimedia.org/

in java. The general abstract structure of entries is taken into account by a common abstract class while language specific details are refined through a language specific implementation class.

3. Specific Problems and Extraction Process

3.1. Related work and motivation of our approach

Many projects addressed wiktionary data extraction. For instance, (Sajous et al., 2010) uses *WiktionaryX*, an XML version of a 2010 wiktionary dump for French and English, available at (Sajous, 2010). (Zesch et al., 2008) provide a free to use, but closed-source, java library to programmatically access the data of the English and German wiktionaries. Other projects did use wiktionary based data in NLP applications without providing details on the way this data were extracted.

As stated in (Sajous et al., 2010), "*When merging information extracted from several languages, the homogenisation of the data structure often leads to the choice of the poorest one, resulting in a loss of information.*". In this work we did not try to provide a uniform entry representation for all languages but rather used a simple lexical network model to represent as much data as we can extract correctly from the wiktionary dumps. We also chose to ignore some of the structure to ease the extraction process.

Lexical Entry The wiktionary unit of information is a *page*. While classical resources often create different lexical entries for homonyms, we chose to keep the wiktionary approach. Hence, each lexical entry in the extracted data corresponds to a unique page in wiktionary.

Homonymy and Polysemy If two words are homonyms, they will be described in the same page. In wiktionary the homonyms will be distinguished by different *etymologies*. It is quite difficult to coherently extract homonyms (and gather word senses under the correct etymology) as entries in different language editions are very incoherent on this aspect. For instance, in the French edition, there should be only 1 etymology section (each etymology being numbered), and other part of speech sections will make reference to the corresponding etymology. On the other hand, in the English edition, several etymology sections will be used, each one preceding the part of speech (and word senses) it covers. Hence, we chose to ignore etymology and gather all word senses in a flat list in the lexical entry.

Lexico Semantic Relations and translation links For the very same reasons, it is most of the time not possible to reliably attach a lexicon semantic relation to its correct word sense. For instance, in the French entry "chat", "matou" is a synonym of the word sense defined by "chat mâle" while "minet" is a synonym of the general word sense defined by "chat domestique". The same goes for translation links. We chose to attach such relations to the lexical entry rather than to its word senses, as the current attempts led to too many errors in the extracted data. However, whenever possible, we kept in the extracted structure the different *hints*[7] that are given in wiktionary. With such an approach we may be able to later re-attach the correct translations to the correct word senses by processing the extracted data, while a subset of entries may be processed with an ad-hoc extractor tailored to extract a gold standard for this task.

All the above mentioned project do stress that the data is sometimes erroneous and most of the time heterogeneous. Among errors and incoherences one may find:

Unconventional encoding of structuring elements For instance, in French, the main language section titles are encoded using a set of templates, named using the ISO 639-1 2 letters language codes (ISO639-1, 2002). Here, `== {{=fr=}} ==` encodes the section heading "Français". Some French contributors did not use this templates but used `== Français ==` which leads to the very same rendering.

Multiple templates may encode the same information For instance, translation equivalents are gathered in boxes which are titled using a summary of a preceding definition. Such boxes represent a word sense for which the translations are valid. In the French language edition, such boxes may be created either with the `{{boîte début|...}}` template or with the `{{(|...}}` template. As both templates are quite common, the extraction process must recognize both of them.

Syntactically incorrect elements Some entries do contain templates that are syntactically ill-formed (e.g. a template is opened with curly braces and closed by square brackets).

Order of the sections Even when the templates are correctly used, the order of the different sections does not necessarily follows the recommendations available in the documentation. For instance, in the English wiktionary, contributors are asked to put pronunciation section after the etymology. In the entry "chat" that we gave as an example in Figure 2, this order is inverted.

3.2. Organization of the extractor

As stated above, the errors and incoherence that are inherent to this contributive resource make things rather complicated for building of a generic extractor. Moreover, we want to use the many wiktionary language editions as a whole, interoperable, lexical network. Hence we need a tool that will be easy to adapt to a new language edition.

We also want to keep the evolving nature of wiktionary, so that the available data will stay as synchronized as possible with the evolution of the resource. For this, we need to adapt the extractor to the evolving usages of the contributors that add new information but also change the templates themselves. This means that the extractor should be easy to

[7] e.g. many translations are grouped under an annotation that is usually a summary of a previous definition.

change. This aspect is crucial in the context of a multilingual extractor.
For all these reasons, we focus on building an extractor that is:

- open-source: so that it is very easy to anyone to adapt it to his own needs,
- based on LGPL license: so that we encourage users to provide their own tunings and heuristics to the main code base,
- efficient: so that one may be able to do the extraction process on the fly, either from a dump, or directly from the online data,
- simple: we do not require any additional software installation as we do not use any database and we provide a simple build method based on `maven`[8] that takes care of library dependencies.
- with several tools: along with the extractor, we provide several development tools (e.g. a "grep in wiktionary" tool where you can find all wiktionary entries containing a certain pattern, or a "get the raw entry" of a page in the dump which may come handy when your dump file is around 4 Gb).

The extractor itself is a java program containing 2 kind of classes:

- `WiktionaryExtractor` which parse the wiktionary entry and
- `WiktionaryDataHandler` which store the extracted data

The Wiktionary extractor is an abstract class that handles general WikiMedia syntax (links, templates, etc.) and language independent processing. Language specific classes inherit from it and define the different *patterns* that are used to structure the lexical entries in the language edition. This way the addition of a new language mainly consists in identifying regular expressions that match the different elements structuring the entry (section headers) and the different elements containing data to be extracted (translation templates, definition patterns, ...).
The *WiktionaryExtractor* class and its children classes parse the entry and trigger methods of the *WiktionaryDataHandler*. It is the responsibility of the data handler to structure and store the extracted data. This way, it is possible to adapt the extraction process to a new extracted data organization.
In this extraction process some heuristics are used to capture heterogeneous or erroneous data. But, as the wiktionary evolves (along with its conventions) and as the extraction program is adapted with new heuristics, one has to ensure that the extraction does not regress. For this, we use the Mulling tool provided by (Archer, 2010) to compute the differences between extracted graphs. Such difference may be quickly evaluated and the extraction heuristics may be adopted or rejected accordingly.

[8] `http://maven.apache.org/`

4. Extracted Data

4.1. Macro- and Micro- Structures

Each language edition is extracted as a lexical network. This network is represented using the W3C standard *Resource Description Framework* (RDF), as described in (Klyne and Carroll, 2004). This structure is stored using the *Turtle* textual syntax which is compact and easy to read (Beckett et al., 2011), hence easing the debugging process. The English language network describes the English lexical entries (giving their part of speech, definitions, lexical relations and translations) while the French one describes the French lexical entries. Each lexical network is stored in a single Turtle file.
The nodes and relations in this lexical network are typed using the classes defined in the Lexical Markup Framework (ISO/TC37/SC4) specification (LMF, 2008; Francopoulo et al., 2006). LMF defines a set of classes using UML notation. These classes have been converted in RDF concepts under a specific name space (e.g. LMF *Lexical Entry* class is described by the RDF `http://www.lexicalmarkupframework.org/lmf/r14\#LexicalEntry` resource).
Unlike the previous effort to create an RDF version of the LMF standard, we did not reified the relation between classes, but rather used simple RDF statements (properties) to avoid cluttering the extracted resource.
Every lexical network node is identified by an IRI, an internationalized URI that allows the use of non ASCII *letters*, (Duerst and Suignard, 2005). For instance, the English lexical entry "chat" is identified by `http://getalp.org/dbnary/eng#chat` while the French lexical entry "chat" is identified by `http://getalp.org/dbnary/fra#chat`.
The extracted network contains nodes from the LMF core package (*Lexical Entry*, *Sense*, *Definition*), the LMF morphology package (*Lemma*) the LMF Machine Readable Dictionary extension (*Equivalent*). The standard is not strictly applied here, as one node named *LexicalEntryRelation* has been introduced (an equivalent of the LMF Sense Relation class, but relating lexical entries rather than senses). Moreover, the *Equivalent* relation relates a lexical entry rather than a sense as stated in the LMF standard.
A lexical entry node has a single `partOfSpeech` property. It may have several values, as the English entry "chat" which may be a verb and a noun. All sense nodes are related to their lexical entry by an `isPartOf` property. They also have a `partOfSpeech` property, which should have only one value. A definition is related to its corresponding sense by the `isPartOf` property. Lemmas are related to the lexical entry by an `isPartOf` property. A lemma is always created with the wiktionary page name as its `writtenForm` property. Additional lemmas are created when we detect alternative spellings in the wiktionary data.
Equivalence nodes are related to their corresponding lexical entry by an `isPartOf` property. They have a mandatory `language` property which value is the normalized ISO 639-3 3 letter language code (ISO639-3, 2007) of the translation. They also have a mandatory `writtenForm`

property whose value is the written form of the translation. They may also have a `glose` property that contains the hint given in wiktionary to identify the sense for which the translation is valid. Finally, they may contain a `usage` property that contains some elements that are associated to individual translations. Its value depends on the language edition usages and the language of the translation. For instance, the French language edition sometimes give an indication on the usage or level of language. It also consistently gives the transliteration of all russian equivalents in Roman writing system.
Lexical entry relation nodes represent several types of relations:

- `ant` relation which relates lemmas to their antonyms,
- `holo` relation which relates lemmas to their holonyms,
- `hyper` relation which relates lemmas to their hypernyms,
- `hypo` relation which relates lemmas to their hyponyms,
- `mero` relation which relates lemmas to their meronyms,
- `syn` relation which relates lemmas to their synonyms,
- `qsyn` relation which relates lemmas to their quasi-synonyms (this one is only used in the French language edition),

Using this lexical network structure, we ignore many of the information available in wiktionaries, either because we do not want to use it in later processing or because it involves far more heuristics during data extraction.

4.2. Example of an extracted lexical entry

Figure 4 gives an excerpt of the lexical network for the English entry “chat” in Turtle format. Figure 5 gives an UML like overview of the same excerpt.

4.3. Size of the involved data

At the time of writing, we extracted data from the most up to date dump files of English (4.1 Gb), French (3.1 Gb) and German (631 Mb) wiktionaries. The full extraction of the English wiktionary takes around 4 min. on a 2.67 GHz Intel Xeon processor with enough memory to avoid swapping ($\sim$ 800 Mb) as the lexical network is stored in memory during extraction.
Table 1 shows the size of the resulting networks.
As can be seen in table 1 the number of relation is quite surprising in the German wiktionary. At the time of writing we do not have explanations on this figure and we still have to figure out if these relations are errors in the extraction process or problems in the wiktionary data itself. Errors are quite likely as the German wiktionary makes extensive use of nested macros which are difficult to correctly parse with our current automata based architecture.
Table 2 gives more details on the translation equivalents that have been extracted from the 3 wiktionary language editions. It lists the number of translation to the 17 largest language editions of wiktionaries, as found in the English, French and German language editions.

```
@prefix lmf:     <http://www.lexicalmarkupframework.org/lmf/r14#> .
@prefix dbnary:  <http://getalp.org/dbnary/eng#> .

dbnary:chat
      a       lmf:LexicalEntry ;
      dbnary:partOfSpeech "Verb" , "Noun" .

dbnary:__lem_chat
      a       lmf:Lemma ;
      dbnary:writtenForm "chat" ;
      lmf:isPartOf dbnary:chat .

dbnary:__ws_1_chat
      a       lmf:Sense ;
      dbnary:partOfSpeech "Verb"^^<http://www.w3.org/2001/XMLSchema
      dbnary:senseNumber "1"^^<http://www.w3.org/2001/XMLSchema#int
      lmf:isPartOf dbnary:chat .

dbnary:__def_1_chat
      a       lmf:Definition ;
      dbnary:text "To be engaged in informal conversation." ;
      lmf:isPartOf dbnary:__ws_1_chat .

dbnary:__tr_rus_11_chat
      a       lmf:Equivalent ;
      dbnary:glose "be engaged in informal conversation" ;
      dbnary:language "rus" ;
      dbnary:usage "tr=boltát'" ;
      dbnary:writtenForm "болтать" ;
      lmf:isPartOf dbnary:chat .

dbnary:__tr_fra_67_chat
      a       lmf:Equivalent ;
      dbnary:language "fra" ;
      dbnary:writtenForm "discussion" ;
      lmf:isPartOf dbnary:chat .

dbnary:__ws_2_chat
      a       lmf:Sense ;
      dbnary:partOfSpeech "Verb"^^<http://www.w3.org/2001/XMLSchema
      dbnary:senseNumber "2"^^<http://www.w3.org/2001/XMLSchema#int
      lmf:isPartOf dbnary:chat .

dbnary:__def_2_chat
      a       lmf:Definition ;
      dbnary:text "To talk more than a few words." ;
      lmf:isPartOf dbnary:__ws_2_chat .
```

Figure 4: Excerpt of the extracted network for the English entry “chat”, in turtle syntax.

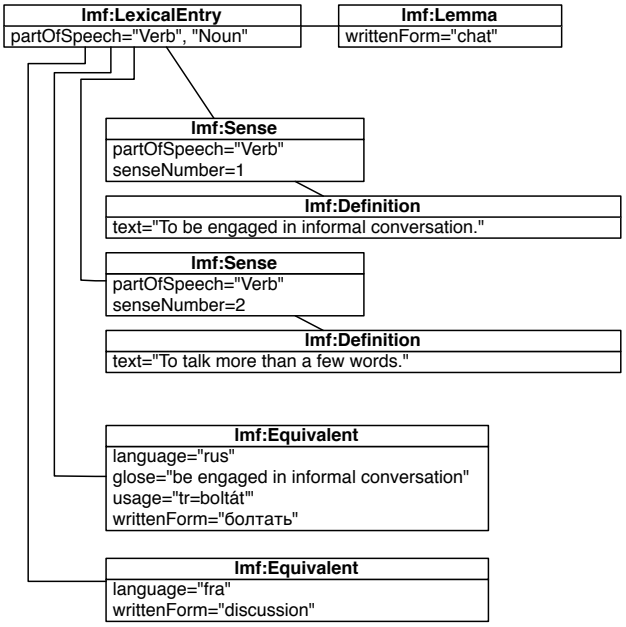

Figure 5: Excerpt of the extracted network for the English entry “chat”, as an UML like network.

5. Conclusion

The current paper shows preliminary results on an open source tool to extract a LMF 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 Lexical Markup Framework (LMF) structure, the extracted data is also directly usable

from/to	deu	ell	eng	fin	fra	hun	ido	ita	lit	nor	pol	por
deu	5	2362	68306	4429	57401	6874	340	33146	1663	4017	14918	8344
eng	32084	10718	1	48577	33291	13225	1496	27160	1899	7837	12156	15652
fra	31590	6465	73169	6934	25[a]	4926	11548	15310	1611	4129	7126	16735

from/to	rus	tam	tur	vie	zho	others	Total
deu	19533	216	4192	426	7926	183589	234098
eng	22314	165	6525	2250	51	261803	235401
fra	6783	565	3667	1333	4030	199281	195946

[a]These are errors in the wiktionary data, where, for instance, the entry "crocodile nain" contains translations to Lingala that are tagged as French translations.

Table 2: Number of translation equivalent (for the considered languages) in the 17 largest wiktionary editions (sorted by alphabetical order on the 3 letters language code).

Nodes in graphs	English	French	German
entries	414929	260467	155258
lemmas	402442	246168	90207
definitions[ab]	354359	330681	80934
relations	79487	106151	215085
equivalents	497204	395227	417687
Total	2102780	1669375	1040105
Relations in graphs			
syn	65103	55434	76606
qsyn[c]	-	2666	-
ant	9964	8760	34691
holo	0	5415	0
mero	224	4996	0
hyper	1047	11272	49051
hypo	3144	17601	54733

[a]The current English extraction program does not yet correctly recognize inflected forms. Hence, many lexical entries represent word forms and many of them are not related to a definition.

[b]There is exactly one definition node per sense node. Hence, sense node are not shown here, but they are counted in the total number of nodes.

[c]This relation is only available in French language edition. Other language editions do not distinguish between synonyms and quasi synonyms.

Table 1: Size of the extracted lexical networks.

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.

As the lexical network is formatted in RDF format, it is immediately usable by many existing tools (Ontology builders, Sparql query engines, reasoners...).

Our final objective is to create a tool that will be to wiktionary what dbpedia (Auer et al., 2007) is to wikipedia.

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

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).

7. References

Vincent Archer. 2010. MuLLinG: MultiLevel Linguistic Graphs for Knowledge Extraction. In *Proceedings of TextGraphs-5 - ACL-2010 Workshop on Graph-based Methods for Natural Language*, pages 69–73. Association for Computational Linguistics.

Sören Auer, Christian Bizer, Georgi Kobilarov, Jens Lehmann, and Zachary Ives. 2007. Dbpedia: A nucleus for a web of open data. In *In 6th Int'l Semantic Web Conference, Busan, Korea*, pages 11–15. Springer.

David Beckett, Tim Berners-Lee, and Eric Prud'hommeaux. 2011. Turtle - terse RDF triple language. W3c working draft 09 august 2011, W3C, August 2011.

M. Duerst and M. Suignard. 2005. Internationalized resource identifiers (iris). Rfc 3987, IETF.

Gil Francopoulo, Monte George, Nicoletta Calzolari, Monica Monachini, Nuria Bel, Mandy Pet, and Claudia Soria. 2006. Lexical Markup Framework (LMF). In *International Conference on Language Resources and Evaluation - LREC 2006*, Gênes/Italie. elra. LIRICS.

ISO639-1. 2002. Codes for the representation of names of languages — part 1: Alpha-2 code. ISO 639-1:2002.

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

Graham Klyne and Jeremy J Carroll. 2004. Resource Description Framework (RDF): Concepts and Abstract Syntax. *Structure*, 10(February):1–20.

LMF. 2008. Language resource management – lexical markup framework. ISO 24613:2008.

Christian M. Meyer and Iryna Gurevych. 2012. 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. (pre-publication draft at the date of LREC).

Franck Sajous, Emmanuel Navarro, Bruno Gaume, Laurent Prévot, and Yannick Chudy. 2010. Semi-automatic Endogenous Enrichment of Collaboratively Constructed Lexical Resources: Piggybacking onto Wiktionary. In Hrafn Loftsson, Eiríkur Rögnvaldsson, and Sigrún Helgadóttir, editors, *Advances in Natural Language Processing*, volume 6233 of *Lecture Notes in Computer Science*, pages 332–344. Springer Berlin / Heidelberg.

Franck Sajous. 2010. WiktionaryX: XML version of the free collaborative dictionary. `http://redac.univ-tlse2.fr/lexiques/wiktionaryx_en.html`.

WikiMedia. 2012. Mediawiki. `http://www.mediawiki.org/`.

Torsten Zesch, Christof Müller, and Iryna Gurevych. 2008. Extracting Lexical Semantic Knowledge from Wikipedia and Wiktionary. In *Proceedings of the Conference on Language Resources and Evaluation (LREC)*.