Undefined 0 (0) 1
IOS Press

The Apertium Bilingual Dictionaries on the Web of Data

Jorge Gracia [a,*], Marta Villegas [b], Asunción Gómez-Pérez [a], and Núria Bel [b],
[a] *Ontology Engineering Group, Universidad Politécnica de Madrid*
Campus de Montegancedo s/n Boadilla del Monte 28660 Madrid. Spain
E-mail: {jgracia,asun}@fi.upm.es
[b] *Institut Universitari Linguistica Aplicada, Universitat Pompeu Fabra*
Roc Boronat, 138 08018 Barcelona. Spain
E-mail: {marta.villegas,nuria.bel}@upf.edu

Abstract. Bilingual electronic dictionaries contain collections of lexical entries in two languages, with explicitly declared translation relations between such entries. Nevertheless, they are typically developed in isolation, in their own formats and accessible through proprietary APIs. In this paper we propose the use of Semantic Web techniques to make translations available on the Web to be consumed by other semantic enabled resources in a direct manner, based on standard languages and query means. In particular, we describe the conversion of the Apertium family of bilingual dictionaries and lexicons into RDF (Resource Description Framework) and how their data have been made accessible on the Web as linked data. As result, all the converted dictionaries (many of them covering under-resourced languages) are connected among them and can be easily traversed from one to another to obtain, for instance, translations between language pairs not originally connected in any of the original dictionaries.

Keywords: linguistic linked data, multilingualism, Apertium, bilingual dictionaries, lexicons, lemon, translation

1. Introduction

The publication of bilingual and multilingual language resources as linked data (LD) on the Web can largely benefit the creation of the critical mass of cross-lingual connections required by the vision of the multilingual Web of Data [9]. The benefits of sharing linguistic information on the Web of Data have been recently recognized by the language resources community, which has shown increasing interest in publishing their linguistic data and metadata as LD on the Web [2]. As result of interlinking multilingual and open language resources, the Linguistic Linked Open Data (LLOD) cloud is emerging[1], that is, a new linguistic ecosystem based on the LD principles that will allow the open exploitation of such data at global scale.

In this article we will focus on the case of electronic bilingual dictionaries as a particular type of languages resources. Bilingual dictionaries are specialized dictionaries that describes translations of words or phrases from one language to another. They can be unidirectional or bidirectional, allowing translation, in the latter case, to and from both languages. A bilingual dictionary can contain additional information such as part of speech, gender, declination model and other grammatical properties.

Electronic bilingual dictionaries have been typically developed in isolation, in their own formats and accessible through proprietary APIs. We propose the use of Semantic Web techniques to make translations available on the Web to be consumed by other semantic enabled resources in a direct manner, based on standard languages and query means. The result of a principled conversion into RDF is that the LD dictionaries are connected among them [7] and can be easily traversed from one to another to obtain, for instance,

*Corresponding author

[1] An updated picture of the current LLOD cloud can be found at `http://linguistic-lod.org/`

translations between language pairs not originally connected in any of the original dictionaries. Other potential uses of bilingual dictionaries in LD are to enhance LD-based machine translation [10] and crosslingual information access over LD [9].

In particular, we have converted the Apertium family of bilingual dictionaries [4] into RDF, making their data interconnected and accessible on the Web as LD. Thus, the main contributions of this work are:

- We propose a method for converting bilingual dictionaries into RDF and publish them as LD on the Web, that we have particularised in the Apertium case.
- As result, we have contributed to the cloud of LLOD with 22 new linguistic datasets, many of them covering *under-resourced languages* that had little or none presence so far in the Web of Data (e.g., Occitane, Asturian, Aragonese, Esperanto, Basque, etc.).
- We also analyse the resulting Apertium RDF graph and exemplify how to traverse it to obtain direct and indirect translations.
- We have used and evaluated the *one time inverse consultation* algorithm to compute the confidence degree of indirect translations.

The remainder of the paper is organized as follows. In Section 2 we briefly describe the Apertium initiative. Section 3 discusses how lexica and language translations can be represented as LD. Section 4 describes the method we have followed to convert the Apertium dictionaries into RDF and to publish them as LD on the Web. In Section 5 we exemplify how to traverse the Apertium RDF graph to obtain translations, and how to compute a confidence degree for indirect translations. Section 6 analyses related work and, finally, conclusions can be found in Section 7.

2. The source data

Apertium is a free/open-source machine translation platform [4], initially aimed at related-language pairs which currently includes up to 40 language pairs[2]. The system was released under the terms of the GNU General Public License.

The translation engine consists of series of assembled modules which communicate using text streams. One of the modules is the lexical transfer module which reads lexical forms of the source language and delivers the corresponding target language lexical forms. The module uses a bilingual dictionary which contains an equivalent for each source lexical form. Apertium dictionaries are designed so that they can be compiled into letter transducers [6] which are able to process input strings and produce output strings. Accordingly, dictionaries are made of entries consisting of string pairs that correspond to the inputs and outputs of the transducer. The Apertium dictionaries are described in XML. Notice that the Apertium initiative benefits *under-resourced languages* specially, for which it is difficult to apply Statistical Machine Translation techniques due to the lack of large parallel corpora in such languages.

During the METANET4U Project[3], a good number of lexicons were converted into Lexical Markup Framework [5] (LMF) in an effort to upgrade existing resources to agreed standards and guidelines. Many Apertium lexicons were included in that process[4].

Thus, following the LMF model, for each bilingual Apertium lexicon a new `LexicalResource` was created. Each `LexicalResource` contains two `Lexicons` (for source and target languages) and a set of `SenseAxis` elements that are used to link senses in different languages. Only open categories were considered (nouns -including proper nouns-, verbs, adjectives and adverbs). The corresponding IDs were generated by concatenating the word form, the part of speech (PoS) tag and the language tag. The following XML code exemplifies the LMF representation of a single translation ("bench"@en → "banco"@es):

```
<LexicalResource>
<Lexicon>
  <feat att="language" val="en"/>
  <LexicalEntry id="bench-n-en">
    <feat att="partOfSpeech" val="n"/>
          <Lemma>
            <feat att="writtenForm" val="bench"/>
          </Lemma>
    <Sense id="bench_banco-n-1"/>
  </LexicalEntry>
</Lexicon>

<Lexicon>
  <feat att="language" val="es"/>
  <LexicalEntry id="banco-n-es">
          <feat att="partOfSpeech" val="n"/>
          <Lemma>
            <feat att="writtenForm" val="banco"/>
          </Lemma>
```

[2] http://wiki.apertium.org/wiki/Main_Page

[3] http://www.meta-net.eu/projects/METANET4U/

[4] A complete list of available lexicons can be found at http://lod.iula.upf.edu/types/Lexica/by/standards

```
        <Sense id="banco_bench-n-r"/>
  </LexicalEntry>
</Lexicon>
<SenseAxis id="bench_banco-n-banco_bench-n"
     senses="bench_banco-n-l␣banco_bench-n-r"/>
</LexicalResource>
```

As it is shown in the above example, every resulting `LexicalEntry` includes the lemma, the part of speech and a `Sense` element. `Sense` elements are needed as a place holder to encode translation equivalents in the `SenseAxis` elements. Each `LexicalEntry` has as many senses as target equivalences. `Sense` elements only have an ID which is formed by concatenating the source form, the target form, the PoS tag and the 'l' or 'r' tags (which in the original dictionaries indicate "left" and "right" respectively for the source and target languages). Finally, the corresponding `SenseAxis` element is generated. Here the `senses` attribute collects the related senses.

3. The representation model

For representing lexical content in RDF, we adopt the LExicon Model for ONtologies (*lemon*) [12] as basis, which is a de facto standard for representing ontology lexica. Such a model is meant for creating lexica and machine readable dictionaries in multiple natural languages as LD, usually for describing (or accompanying) an ontology. The model allows to keep separate linguistic descriptions from the ontological model they accompany. Linguistic annotations (data categories or linguistic descriptors, e.g., to denote gender, number, part of speech, etc.) are not captured in the model, but have to be specified for each lexicon by dereferencing their URIs as defined in some external catalog of data categories. In particular we use LexInfo[5] to that end, which is an ontology of types, values and properties partially derived from ISOcat[6]. The core of the *lemon* model consists of the following elements[7]:

`LexicalEntry`. An entry in the lexicon (word, multiword expression or even affix) that is assumed to represent a single lexical unit with common properties (e.g., PoS) across all its forms and meanings.

`LexicalForm`. A form represents a particular version of a lexical entry, for example a plural or some other inflected form.

`LexicalSense`. The sense refers to the usage of a lexical entry with a specific meaning and can also be considered as a reification of the relation between a lexical entry and the ontological entity that characterizes its meaning in a certain context.

`Reference`. The reference is an entity in the ontology that defines the formal semantics associated to the lexical entry.

The *lemon* model did not consider the representation of explicit *translations* initially. To that end, an extension of *lemon* was proposed for representing translations on the Web of Data: the *lemon translation module* [8]. The translation module consists essentially of two OWL classes: `Translation` and `TranslationSet`. The latter is a set of translations sharing some common properties such as provenance, language pair, etc. `Translation` is a reification of the relation between two *lemon* lexical senses that point to two lexical entries in two different languages, and has the following OWL properties:

`translationSource` and `translationTarget`. They point to the *lemon* lexical senses that act as source and target of the translation, respectively.

`translationConfidence`, to assign a confidence value to the translation pair.

`context`, which is an unrestricted property intended to express/determine, if needed, the specific application context in which a pair of lexical senses are translation of each other.

`translationCategory`, that points to an external registry of translation categories or types[8].

Additionally other broadly used vocabularies such as Dublin Core[9] can be used to attach information about provenance, authoring, versioning, and licensing. Finally, the Data Catalogue Vocabulary[10] (DCAT) can be used to represent other metadata information associated to the publication of the RDF dataset.

In Figure 1 we illustrate a translation that is represented using *lemon* and the translation module. In short, `lemon:LexicalEntry` and their associated properties are used to account for the lexical informa-

[5] `http://www.lexinfo.net/ontology/2.0/lexinfo`
[6] `http://www.isocat.org/`
[7] See figure and a short description at `http://lemon-model.net/learn/5mins.php`
[8] As for instance the ones contained at `http://purl.org/net/translation-categories`
[9] `http://purl.org/dc/elements/1.1/`
[10] `http://www.w3.org/TR/vocab-dcat/`

tion, while the tr:Translation class puts them in connection through lemon:LexicalSense. Other options would be possible, of course, such as connecting the lexical entries directly without defining "intermediate" senses. Nevertheless, we understand that *translations occur between specific meanings of the words* and the class lemon:LexicalSense allows us to represent this fact explicitly.

The models presented in this section (*lemon* and its translation module) have been the basis for the new *lemon-ontolex* model[11] and its *vartrans* module. These have been discussed and defined by the W3C Ontology Lexica (Ontolex) community group[12] and are near to be released[13]. All the *lemon* ingredients used in Apertium RDF have a direct mapping into the new model, so the conversion into *lemon-ontolex*, which we leave as future work, should be straightforward.

4. RDF Generation Methodology

Some guidelines have been proposed to produce and publish high quality multilingual LD on the Web [17]. As a further step, the W3C Best Practices for Multilingual Linked Open Data (BPMLOD) community group[14] has recently published specific guidelines for generating and publishing certain types of language resources as LD (e.g., bilingual dictionaries, WordNets, terminologies in TBX, etc). The methodology we used for the Apertium RDF case, which we describe in this section, has served as basis for the development of the *guidelines for bilingual dictionaries*[15] that were discussed in the BPMLOD group.

The conversion into RDF of the Apertium dictionaries started with the *analysis of the data* and *selection of relevant vocabularies*, already discussed in Sections 2 and 3 respectively. These were followed by the steps described in the remainder if this section: modelling, URIs design, generation, linking, and publication.

4.1. Modelling

Every Apertium bilingual dictionary, which came originally in a single LMF file, was converted into three different objects in RDF, namely: source lexicon, target lexicon, and translation set. This division fits naturally in the *lemon* translation module scheme. As result, two independent monolingual lexicons in RDF are created, along with a set of translations that connects them. The publication of a number of bilingual dictionaries that follow the same scheme leads to the creation of a pool of online monolingual lexicons that grows with time, all of them connected within the same global RDF graph by sets of translations.

4.2. URIs design

Among the different patterns and recommendations for defining URIs we follow the one proposed by the ISA Action[16] for European governmental data [1]. In short, the pattern is as follows: http://{domain}/{type}/{concept}/{reference}, where {type} should be one of a small number of possible values that declare the type of resource that is being identified. Typical examples include: 'id' or 'item' for real world objects; 'doc' for documents that describe those objects; 'def' for concepts; 'set' for datasets; or a string specific to the context, such as 'authority' or 'dcterms'. For example, in Apertium RDF, the English-Spanish translation set is named as: http://linguistic.linkeddata.es/id/apertium/tranSetEN-ES

In order to construct the URIs of lexical entries, senses, and other lexical elements, we have preserved the identifiers of the original LMF data whenever possible, propagating them into the RDF representation. Some minor changes have been introduced, though. For instance, in the original LMF data the identifier of the lexical entries ended with the particle "-l" or "-r" depending on their role as "source" or "target" in the translation (see Section 2). In our case, directionality is not preserved at the level of lexicon but in the Translation class, so these particles were removed from the identifier. In addition, some other suffixes were added for readability: "-form" for lexical forms, "-sense" for lexical senses, and "-trans" for translation.

4.3. Generation

This activity deals with the transformation into RDF of the selected data sources using the chosen representation scheme and modelling patterns. There are a

[11] https://www.w3.org/community/ontolex/wiki/Final_Model_Specification

[12] https://www.w3.org/community/ontolex/

[13] At the time of writing this, eptember 2015

[14] http://www.w3.org/community/bpmlod/

[15] http://www.w3.org/2015/09/bpmlod-reports/bilingual-dictionaries/

[16] http://ec.europa.eu/isa/actions/01-trusted-information-exchange/1-1action_en.htm

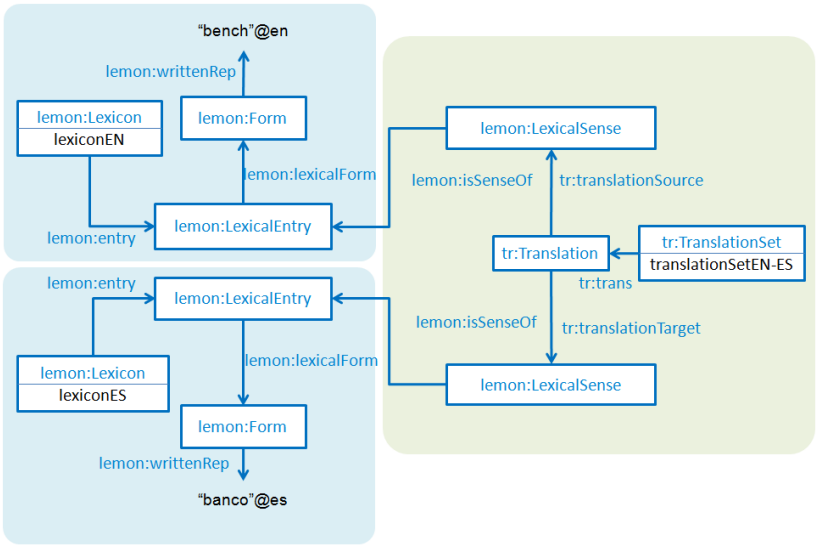

Fig. 1. Representation of the translation of "bench" from English to Spanish

number of tools that can be used to assist the developer in this task, depending on the format of the data source. In our case, Open Refine[17] was used for defining the transformations from XML into RDF.

As result of the transformation, three RDF files were generated, one per component (lexicons and translation set). The code in Figure 2 contains the RDF of the single translation illustrated in Figure 1. The first two parts of the code come from the EN and ES lexicons and the last one from the EN-ES translation set.

All our scripts for the Apertium RDF generation from LMF have been recorded, stored, and made available online to enable their later analysis and reuse[18].

4.4. Linking

The Apertium RDF data have been linked to two external datasets: LexInfo and BabelNet[19]. In particular, 690,650 links have been established to LexInfo and 277,089 to BabelNet. In the first case, LexInfo has been used as an external catalog to provide definitions to the PoS of the Apertium lexical entries. As for BabelNet, links were established between the Apertium lexical senses and the BabelSynsets. In that way the meaning underlying every possible lexical sense is better defined and can be enriched with additional context such as glosses or images coming from BabelNet[20].

4.5. Publication

Once the RDF data was generated, they were loaded in a Virtuoso[21] triple store, where they are accessible through a single SPARQL endpoint[22]. Pubby[23] was used to develop a LD interface. In that way, all the data from the Apertium bilingual dictionaries were made accessible as LD on the Web in a unified graph with lexical entries, senses, translations, etc. as nodes. All the nodes were identified with dereferenceable URIs.

Regarding the publication of the metadata, we considered that DCAT suffices for the purposes of describing the elements generated in the RDF conversion of bilingual dictionaries. Furthermore, some data management platforms such as Datahub use DCAT in a preferred way for representing metadata. The RDF version of the Apertium dictionaries was published in Datahub[24]. The Datahub platform created a metadata file based on DCAT for every Apertium RDF dataset. We extended such metadata file with some additional

[17] http://openrefine.org/
[18] http://dx.doi.org/10.6084/m9.figshare.1342816
[19] http://babelnet.org/

[20] Due to space limitations, the Apertium-BabelNet linking strategy will be analysed in a different document.
[21] http://virtuoso.openlinksw.com/
[22] http://linguistic.linkeddata.es/sparql/
[23] http://wifo5-03.informatik.uni-mannheim.de/pubby/
[24] A list of the Apertium RDF dictionaries available in Datahub can be found at http://datahub.io/dataset?q=apertium+rdf&organization=oeg-upm

```
@prefix lemon: <http://www.lemon-model.net/lemon#>
@prefix tr: <http://purl.org/net/translation#>
@prefix lexinfo: <http://www.lexinfo.net/ontology/2.0/lexinfo#>
@prefix dc: <http://purl.org/dc/elements/1.1/>
@prefix apertium: <http://linguistic.linkeddata.es/id/apertium/>

apertium:lexiconEN a lemon:Lexicon ;
        dc:source <http://hdl.handle.net/10230/17110> ;
        lemon:entry apertium:lexiconEN/bench-n-en .
apertium:lexiconEN/bench-n-en a lemon:LexicalEntry ;
        lemon:lexicalForm apertium:lexiconEN/bench-n-en-form ;
        lexinfo:partOfSpeech lexinfo:noun .
apertium:lexiconEN/bench-n-en-form a lemon:Form ;
        lemon:writtenRep "bench"@en .

apertium:lexiconES a lemon:Lexicon ;
        dc:source <http://hdl.handle.net/10230/17110> ;
        lemon:entry apertium:lexiconES/banco-n-es .
apertium:lexiconES/banco-n-es a lemon:LexicalEntry ;
        lemon:lexicalForm apertium:lexiconES/banco-n-es-form ;
        lexinfo:partOfSpeech lexinfo:noun .
apertium:lexiconES/banco-n-es-form a lemon:Form ;
        lemon:writtenRep "banco"@es .

apertium:tranSetEN-ES a tr:TranslationSet ;
        dc:source <http://hdl.handle.net/10230/17110> ;
        tr:trans apertium:tranSetEN-ES/bench_banco-n-en-sense-banco_bench-n-es-sense-trans .
apertium:tranSetEN-ES/bench_banco-n-en-sense a lemon:LexicalSense ;
        lemon:isSenseOf apertium:lexiconEN/bench-n-en .
apertium:tranSetEN-ES/banco_bench-n-es-sense a lemon:LexicalSense ;
        lemon:isSenseOf apertium:lexiconES/banco-n-es .
apertium:tranSetEN-ES/bench_banco-n-en-sense-banco_bench-n-es-sense-trans a tr:Translation ;
        tr:translationSource apertium:tranSetEN-ES/bench_banco-n-en-sense ;
        tr:translationTarget apertium:tranSetEN-ES/banco_bench-n-es-sense .
```

Fig. 2. Example of the RDF (in turtle syntax) generated for the translation represented in Figure 1.

missing information such as provenance, license, and related resources[25]. The extended metadata was published as part of the Apertium RDF Datahub entries.

Finally, in order to improve the visibility and human access to the Apertium RDF dictionaries, a web portal was developed with pointers to the published individual dictionaries and with some query facilities[26].

In summary, we have transformed all the Apertium dictionaries with available versions in LMF, resulting in a total of 22 Apertium RDF bilingual dictionaries. In Table 1 we show the list of datasets along with their size in terms of number of triples and number of translations. As result of the generation and publication process, the 22 Apertium RDF dictionaries were included in the LLOD cloud[27]. The datasets are available under a GNU general public license.

[25] As it is described at `http://tinyurl.com/py6ro9l`

[26] `http://linguistic.linkeddata.es/apertium/`

[27] A picture of the LLOD cloud as of May 2015 can be found at `http://linguistic-lod.org/llod-cloud-may2015`. The Apertium datasets appear on the left hand side and their links to Lexinfo and BabelNet are also pictured.

5. Traversing the Apertium RDF graph

As result of the generation of the Apertium dictionaries as LD, a large unified graph of linked lexical entries, senses and translations was created and made accessible on the Web. The URIs of all these elements can be seen as the nodes of such a network. Every URI is dereferenceable, meaning that when it is accessed a response is obtained and their attributes and links to other elements get in RDF. In this section we explore, by means of examples, how to get direct and indirect translations from the graph. We also describe a method to calculate the confidence degree of the indirect ones.

5.1. Exploring the graph

As result of publishing the bilingual dictionaries in a unified graph by following consistent naming rules, a *multilingual dictionary* has automatically emerged for the Apertium data. Now, querying for translations from one language to one or many languages can be made in an straightforward manner in SPARQL and through a

Dataset	Num. triples	Num. translations
CA-IT	180,851	7,869
EN-CA	759,601	33,029
EN-ES	576,316	25,830
EN-GL	425,117	20,034
EO-CA	426,301	19,964
EO-EN	617,772	31,474
EO-ES	380,198	17,212
EO-FR	726,281	35,791
ES-AN	71,997	3,110
ES-AST	825,540	36,096
ES-CA	730,501	31,291
ES-GL	206,284	8,985
ES-PT	279,245	12,054
ES-RO	400,366	17,318
EU-ES	262,336	11,838
EU-EN	265,466	13,089
FR-CA	152,002	6,550
FR-ES	495,614	21,475
OC-CA	346,346	15,983
OC-ES	317,162	14,561
PT-CA	163,149	7,111
PT-GL	234,065	10,144

Table 1

Apertium RDF datasets with their size in number of triples and translations. The language codes in the table follow the ISO-639 standard.

single access point[28]. For instance, a SPARQL query[29] can be build to retrieve the direct translations of the English term "bank", along with their part of speech, into Spanish. The result of the query is shown in Table 2.

translated_written_rep	POS
"banco"@es	lexinfo:noun
"orilla"@es	lexinfo:noun
"ribera"@es	lexinfo:noun
"agolpar"@es	lexinfo:verb
"amontonar"@es	lexinfo:verb
"apelotonar"@es	lexinfo:verb
"hacinar"@es	lexinfo:verb

Table 2

Direct translations of "bank" from English to Spanish along with their part of speech (POS).

[28] http://linguistic.linkeddata.es/apertium/sparql-editor/

[29] We omit the query here, for the sake of space, but it can be found at http://files.figshare.com/2201195/ApertiumRDF_ExampleQuery4.txt

In addition to obtain explicitly declared translations (as in the above query), it is possible to infer indirect translations by traversing the graph through pivot lexical entries. For instance, a direct translation cannot be obtained (with a query similar to the previous one) between "bank" and the term in Portuguese, because there exists no English-Portuguese (or vice versa) Apertium bilingual dictionary yet. However, the graph can be traversed to reach indirect translations from English to Portuguese through an intermediate language (e.g., Spanish). This is illustrated in Figure 3, which shows an oversimplified fragment of the graph that results from publishing both the EN-ES and ES-PT dictionaries as LD.

A relatively simple SPARQL query[30] can be constructed to get the indirect translations from *"bank"* into Portuguese using Spanish as pivot language. Table 3 shows the result.

pivot_translation_written_rep	indirect_translation_written_rep
"banco"@es	"banco"@pt
"orilla"@es	"orla"@pt

Table 3

Indirect translations of "bank" into Portuguese along with the pivot Spanish translations.

5.2. One Time Inverse Consultation

When using a third pivot language to construct a bilingual dictionary, it is necessary to discriminate inappropriate equivalences between words caused by ambiguities in the pivot language, as Figure 4 illustrates. In fact, when using EN as intermediate language between ES and CA, some wrong translations can be inferred, as for instance "banco"@es $\rightarrow$ "riba"@ca, in addition to the correct ones.

A method to identify such incorrect translations was proposed by Tanaka and Umemura [15] when constructing bilingual dictionaries intermediated by a third language. The method, called *one time inverse consultation* (OTIC), was adapted by Lim et al. [11] in the creation of multilingual lexicons from bilingual lists of words. Given the two languages (source and target) to be connected and a third intermediate language (pivot), given a lexical entry s in the source language, the OTIC method works as follows:

[30] http://files.figshare.com/2133242/ApertiumRDF_ExampleQuery2.txt

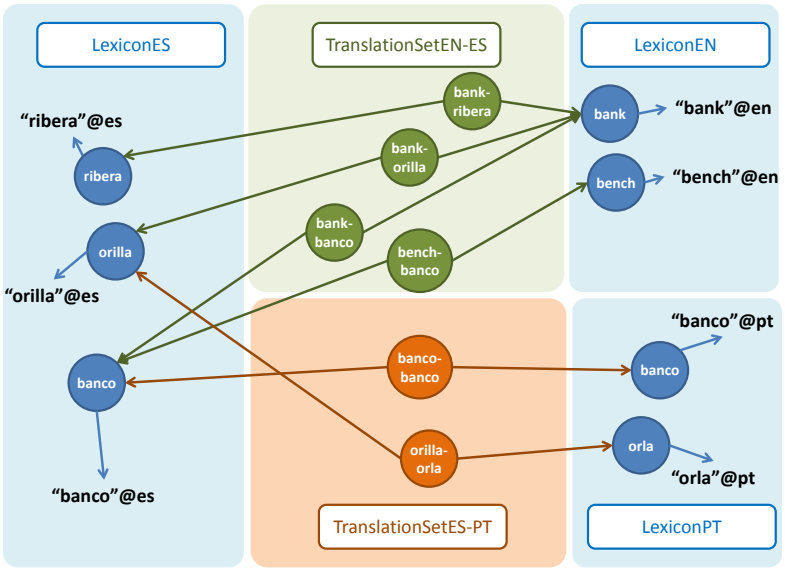

Fig. 3. Simplified representation of the path between some lexical entries in EN and PT (disconnected in the original Apertium dictionaries).

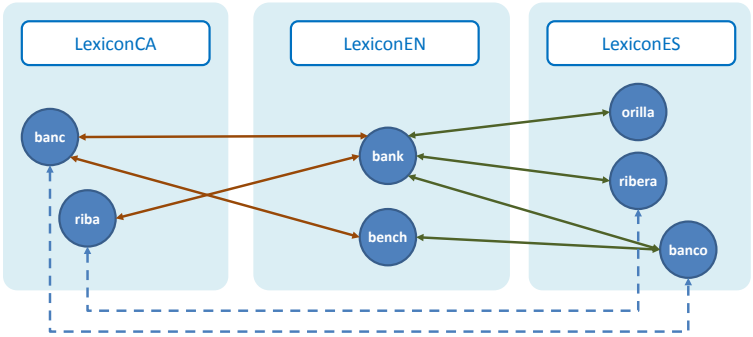

Fig. 4. Example of translation candidates between ES and CA (EN as pivot language) in Apertium RDF. We do not represent here the translation sets for simplicity. The dashed connectors show the direct ES-CA translations.

1. Obtains P_s, which is the set of all direct translations of s in the pivot language
2. For every $p \in P_s$, looks up its translations in the target language (T_p)
3. For every $t \in T_p$,
   (a) looks up its set of translations in the pivot language (P_t)
   (b) calculates the score for t. This is computed with Equation 1 that measures how many translations in P_t match those in P_s. The more matches between P_s and P_t the higher the score of t, which measures how good it is as a candidate translation of s.

$$score(t) = 2 * \frac{P_s \cap P_t}{|P_s| + |P_t|} \quad (1)$$

Following the example in Figure 4, let us suppose that we want to obtain translations for "banco" from Spanish into Catalan using English as pivot language. The application of Equation 1 results in the following scores for the candidate translations:

score("banc"@ca) = 1.0
score("riba"@ca) = 0.5

The correct translation ("banco"@es → "banc"@ca) is higher ranked in this example, while the wrong one could be filtered out with a threshold higher than 0.5.

Both direct and indirect translations can be obtained not only by means of SPARQL queries but through the Apertium RDF user interface[31], which computes also the OTIC values associated to the indirect translations.

In order to confirm whether the application of the OTIC method is beneficial to filter out wrong indirect translations, we set up the following validating exercise[32]. First, we chose a set of three Apertium lexicons which are directly connected among them in the Apertium RDF graph but also can be indirectly connected through the third language in the set: {EN, CA, ES}. The idea is to use the direct translations as golden standard (they have been made by human experts) and compare them with the translations that can be inferred indirectly. We computed the OTIC value for such indirect translations, and filter out those with a value under a certain *threshold*. *Precision* and *recall* were computed for each language pair to measure the quality of the translations[33]. Notice that recall entirely depends on the degree of coverage of the original Apertium dictionaries, which is determined by construction, and cannot be improved by the OTIC method in any way.

A table summarizing the experimental results can be found online[34]. The results show that translations between lexical entries in the studied languages can be obtained indirectly with a good precision. For $threshold = 1$ (i.e. only selecting translations with maximum score), the obtained precision ranges from 62% to 83%. These values are in the same range of the results obtained by Lim et al. [11], although a direct comparison is difficult due to the different experimental setup. The effect of applying the OTIC-based filtering process makes precision increase in up to a 10% while it has an effect on recall, which decreases in a 3-5% range.

6. Related Work

There have been other remarkable efforts to convert and expose multilingual linguistic data as LD on the Web. For instance DBnary [14] extracts multilingual lexical data from Wiktionary data and provide it to the community as linked open data. It covers 21 languages currently and uses *lemon* to represent lexical information. The translation relation however has been defined in their own domain. BabelNet [13] is a wide-coverage multilingual encyclopedic dictionary and ontology that was converted and published as LD [3] also by using *lemon* as core representation mechanism. The result is an interlinked multilingual lexical database on the Web, suitable to be used for enriching existing datasets with linguistic information, or to support the process of mapping datasets across languages. We consider that DBnary, BabelNet, and other similar multilingual LD resources, are complementary approaches to the Apertium RDF set of bilingual dictionaries. For instance, Apertium RDF and BabelNet mutually benefit from the links established between them: terms in Apertium RDF can be expanded with definitions and factual knowledge from BabelNet, while an amount of additional translations could be added into BabelNet, specially for some minority languages.

In this paper we have also introduced a way to infer new translations between initially disconnected languages by traversing the Apertium RDF graph. Notice that the original Apertium framework includes the apertium-dixtools[35] to execute different process on a dictionary file or on several dictionary files. These include the 'crossdics' tool that are used to cross two language pairs a/b and b/c to generate a new language pair for a/c as defined in [16]. By default, the 'crossdics' tool uses a simple cross model (based on transitive rule) defining very simple rules for crossing two sets of dictionaries. In our case, the ability to extend this dictionary crossing technique to the whole set of available dictionaries implies a substantial improvement.

There have been previous efforts in combining existent bilingual dictionaries to create new bilingual [15] or multilingual [11] ones. However, differently to these approaches, Apertium RDF has been developed by applying Semantic Web techniques, and its lexical information is available on the Web to be consumed by humans or by other semantic enabled resources in a direct manner, based on standard languages and query means. Further, the Apertium RDF is now part of the much larger LLOD cloud, thus enabling easier combination with data from other LD sources.

[31] http://lider2.dia.fi.upm.es:8080/lld-search/

[32] All the experimental data are available online at http://dx.doi.org/10.6084/m9.figshare.1344821

[33] Notice that we adopt an "information retrieval" focus in the experiment, thus just trying to check whether the expected translations are obtained or not and measuring this effect in terms of precision/recall, rather than using other machine translation specific metrics.

[34] http://figshare.com/download/file/2201205/1

[35] http://wiki.apertium.org/wiki/Apertium-dixtools

7. Conclusions

In this paper we have described the transformation of 22 Apertium bilingual dictionaries into RDF and their publication as LD on the Web. The proposed methodology is general enough to be applied to other bilingual dictionaries. We have also described and evaluated the *one time inverse consultation algorithm* to compute the confidence degree of indirect translations in the Apertium RDF graph. In our view, the publication of Apertium RDF contributes to the critical mass of cross-lingual connections required by the multilingual Web of Data to be truly useful and operational.

Due to the novelty of the Apertium RDF datasets, third party reuse of the generated LD is still limited. However, the BabelNet team is currently exploring the potential improvement of their translations in cases in which BabelNet does not provide translations that however can be found in Apertium, which is particularly interesting for certain minority languages. This will lead to establishing links back from BabelNet into Apertium. Further, the Apertium original data is already extensively used by a broad community of deveopers and users. For instance the different versions of the Apertium code and data have reached 125,922 downloads in the last ten years[36], which illustrates the interest and impact of the Apertium project and its data. In fact, one of the objectives of Apertium is to favour the interchange and reuse of existing linguistic data. The Apertium RDF graph described here constitutes an important step towards that direction.

As future work, we plan to enrich the Apertium RDF graph with new datasets as soon as new LMF versions of the existent (and future) Apertium dictionaries will appear. Also, an in depth analysis of the coverage and quality of the translations among all the possible language pairs in the graph will deserve a separate study.

Acknowledgments. This work is supported by the FP7 European project LIDER (610782), by the Spanish Ministry of Economy and Competitiveness (project TIN2013-46238-C4-2-R), and by IULA-UPF-CC-CLARIN.

References

[1] P. Archer, S. Goedertier, and N. Loutas. Study on persistent URIs. Technical report, Dec. 2012.

[2] C. Chiarcos, S. Nordhoff, and S. Hellmann, editors. *Linked Data in Linguistics - Representing and Connecting Language Data and Language Metadata.* Springer, 2012.

[3] M. Ehrmann, F. Cecconi, D. Vannella, J. P. McCrae, P. Cimiano, and R. Navigli. Representing multilingual data as linked data: the case of BabelNet 2.0. In *Proceedings of the Ninth International Conference on Language Resources and Evaluation (LREC'14)*, Reykjavik, Iceland, May 2014. ELRA.

[4] M. Forcada, M. Ginestí-Rosell, J. Nordfalk, J. O'Regan, S. Ortiz-Rojas, J. Pérez-Ortiz, F. Sánchez-Martínez, G. Ramírez-Sánchez, and F. Tyers. Apertium: a free/open-source platform for rule-based machine translation. *Machine Translation*, 25(2):127–144, 2011.

[5] G. Francopoulo, N. Bel, M. George, N. Calzolari, M. Monachini, M. Pet, and C. Soria. Lexical markup framework (LMF) for NLP multilingual resources. In *Proc. of the Workshop on Multilingual Language Resources and Interoperability*, pages 1–8, Sydney, Australia, July 2006. ACL.

[6] A. Garrido-Alenda, M. L. Forcada, and R. C. Carrasco. Incremental construction and maintenance of morphological analysers based on augmented letter transducers. *Proceedings of TMI*, 2002:53–62, 2002.

[7] J. Gracia. Multilingual dictionaries and the web of data. *Kernerman Dictionaries News*, (23):1–4, June 2015.

[8] J. Gracia, E. Montiel-Ponsoda, D. Vila-Suero, and G. Aguado-de Cea. Enabling language resources to expose translations as linked data on the web. In *Proc. of 9th Language Resources and Evaluation Conference (LREC'14), Reykjavik (Iceland)*, pages 409–413. ELRA, May 2014.

[9] J. Gracia, E. M. Ponsoda, P. Cimiano, A. Gómez-Pérez, P. Buitelaar, and J. McCrae. Challenges for the multilingual web of data. *Journal of Web Semantics*, 11:63–71, Mar. 2012.

[10] D. Lewis. Interoperability challenges for linguistic linked data. In *Proc. of Open Data on the Web ODW'13*, Apr. 2013.

[11] L. T. Lim, B. Ranaivo-Malançon, and E. K. Tang. Low cost construction of a multilingual lexicon from bilingual lists. *Polibits*, 43:45–51, 2011.

[12] J. McCrae, G. A. 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*, 46(4):701–719, 2012.

[13] R. Navigli and S. P. Ponzetto. BabelNet: The automatic construction, evaluation and application of a wide-coverage multilingual semantic network. *Artificial Intelligence*, 193:217–250, 2012.

[14] G. Sérasset. DBnary: Wiktionary as a Lemon-Based Multilingual Lexical Resource in RDF. *Semantic Web Journal*, 6(4), 2015.

[15] K. Tanaka and K. Umemura. Construction of a bilingual dictionary intermediated by a third language. In *COLING*, pages 297–303, 1994.

[16] A. Toral, M. Ginestí-Rosell, and F. Tyers. An italian to catalan rbmt system reusing data from existing language pairs. In *Proc. of the Second International Workshop on Free/Open-Source Rule-Based Machine Translation, Barcelona (Spain)*, 2011.

[17] D. Vila-Suero, A. Gómez-Pérez, E. Montiel-Ponsoda, J. Gracia, and G. Aguado-de Cea. Publishing linked data: the multilingual dimension. In P. Cimiano and P. Buitelaar, editors, *Towards the Multilingual Semantic Web*, pages 101–118. Springer Verlag, 2014.

[36] Counting from 2005-10-01 to 2015-10-01.