# Retreading Dictionaries for the 21st Century

**Xavier Gómez Guinovart[1] and Alberto Simões[2]**

1 **Galician Language Technology and Applications (TALG Group)**
**Universidade de Vigo, Galiza, Spain**
xgg@uvigo.es

2 **Centro de Estudos Humanísticos, Universidade do Minho**
**Campus de Gualtar, Braga, Portugal**
ambs@ilch.uminho.pt

## Abstract

Even in the 21st century, paper dictionaries are still compiled and developed using standard word processors. Many publishing companies are, nowadays, working on converting their dictionaries into computer readable documents, so that they can be used to prepare new features, such as making them available online. Luckily, most of these publishers can pay review teams to fix and even enhance these dictionaries. Unfortunately, research institutions cannot hire that amount of workers.

In this article we present the process of retreading a Galician dictionary that was first developed and compiled using Microsoft Word. This dictionary was converted, through automatic rewriting, into a Text Encoding Initiative schema subset. This process will be detailed, and the problems found will be discussed. Given a recent normative that changed the Galician orthography, the dictionary has undergone a semi-automatic modernization process. Finally, two applications for the obtained dictionaries will be shown.

**1998 ACM Subject Classification** I.7.2 Document Preparation

**Keywords and phrases** dictionary, markup language, language processing, lexical information retrieval, Galician language

**Digital Object Identifier** 10.4230/OASIcs.SLATE.2013.115

## 1 Introduction

Until recently lexicographers' typical training would not include information technology, and most of them would use a computer merely as an end-user, using tools such as Microsoft Word and, probably, a little of Microsoft Excel. At the same time, a lot of publishing companies did not have the tools or the mechanisms to maintain and compile a dictionary correctly, and they would propose lexicographers to use text processing tools for that task. These two factors led to the existence of dictionaries that are currently in the press, and that are only available as Microsoft Word, QuarkXPress or even PDF files. Although these formats are suitable to produce a printed document, or even to be made available for download, they are not suited to be processed automatically by computers.

An example of this situation is the "*Diccionario de sinónimos da lingua galega*," a Galician thesaurus published by *Editorial Galaxia* (Vigo) in 1997 [7]. Sixteen years after its publication, the authors wanted to provide a second revised edition of their work, in a format useful both for on-line querying and for Natural Language Processing (NLP) tasks.

Unfortunately, all the authors had was a set of Microsoft Word files that needed extra treatment to be usable by a computer program in NLP tasks, such as the extraction of

2nd Symposium on Languages, Applications and Technologies (SLATE'13).
Editors: José Paulo Leal, Ricardo Rocha, Alberto Simões; pp. 115–126

OpenAccess Series in Informatics
Schloss Dagstuhl – Leibniz-Zentrum für Informatik, Dagstuhl Publishing, Germany

synonyms, antonyms and hypernyms (as presented in [14] or [8]), or to be available as an on-line dictionary.

The solution was to develop a tool to rewrite the Microsoft Word files into a semantic-rich format. Given the existence of specific Extended Markup Language (XML) dialects for dictionaries, we decided to convert the Word files into one of these formats: the Text Encoding Initiative [15] (TEI) definition to encode dictionaries. In fact, given the detailed annotation supported by TEI, we decided to use just a subset[1] of the standard format.

After the automatic conversion of the Microsoft Word files into TEI, the dictionary was revised in a semi-automatic approach, marking the dubious parts of the document, leading the way to the human reviewing process.

Finally, the dictionary obtained is being used for two NLP projects: first, to extract synonyms for a WordNet-like structure for the Galician language (GALNET); and second, to enrich a Portuguese online dictionary (Dicionário-Aberto).

This paper is structured as follows: the next section will discuss the process of rewriting non semantic-structured documents (Microsoft Word files) into a semantic-rich format (TEI). Section 3 explains the spelling normalization of the document, updating it to the latest orthographic normative of the Galician language. Section 4 explains how the dictionary obtained will be used in two NLP projects: enriching a Portuguese on-line dictionary, and extracting concept semantic relations. Finally, we draw some conclusions about this process and the results obtained.

## 2 Word to TEI Conversion

This section describes the process applied to the dictionary Word files, and how they were transformed into well formatted TEI documents:

1. in the first place, Microsoft Word files were converted into very simple text files, with basic XML markup;
2. these text files were then rewritten into well formed XML files, following a TEI subset to encode dictionaries;
3. finally, a set of methods for error detection and correction is developed and applied.

### 2.1 Dealing with Microsoft Word Formats

The dictionary was presented as a set of Microsoft Word files, probably from Office'95, one for each letter. Fortunately, these files had very little formatting other than bold and italic markups, and the revision history. Figure 1 shows an entry as presented by Microsoft Word.

**claro**, **a**, *adx* 1. Iluminado, luminoso. 2. Brillante, limpo, nítido, transparente. 3. Despexado, soleado. 4. Evidente, manifesto, nidio, obvio, patente. 5. Aberto, franco, sincero./ *sm* 6. V **clareiro**. 7. V **clareeira**. ¶

**Figure 1** Entry for the word *claro* as presented by Microsoft Word.

To explore and convert a Microsoft Word file is not easy. There are a number of tools for that task, but many of those just convert the Word file into other formats, such as plain text

[1] Nevertheless, the element names and their nesting rules follow the official format, making it easy for anybody familiar with the TEI format to quickly understand and process our subset.

(losing all markup). Given that we could not avoid performing the conversion, the best tool we could find to convert a Microsoft Word file was Microsoft Word itself.

The recent Microsoft Word versions can save their documents as *Microsoft Word Open XML Document* (known as `docx`). As the name shows, this format is based in XML, and therefore it should be easy to process using standard XML tools.

Although with the `docx` extension, these documents are compressed files, where a set of files (including a group of XML files) are stored in different folders. After expanding one of these documents we could find the main XML file easily, by looking at their file size. It is clearly named `document.xml`, and we went on exploring this file format.

Listing 1 shows the same entry presented above as codified by the `docx` format. There is documentation on this file format, but given the amount of details that can be present in these files, and the simplicity of our documents, we engaged into a quick exploration of the file format. For that purpose, we used a Perl module named `XML::DT` [1], which is able to create a skeleton program to transform specific XML files (that shows the specific entity tags and tag attributes that are really used in the supplied documents).

**Listing 1** Entry for the word *claro* as stored in `docx` format.

```
<w:p w14:paraId="3001E63B" w14:textId="77777777" w:rsidR="006254C9"
w:rsidRDefault="00632CB3"><w:pPr><w:pStyle w:val="SINORMAL"/><w:jc
w:val="both"/><w:rPr><w:b/><w:color w:val="000000"/></w:rPr></w:pPr>
<w:r><w:rPr><w:b/><w:color w:val="000000"/></w:rPr><w:t>claro</w:t>
</w:r><w:r><w:rPr><w:color w:val="000000"/></w:rPr><w:t
xml:space="preserve">, </w:t></w:r><w:r><w:rPr><w:b/><w:color
w:val="000000"/></w:rPr><w:t>a</w:t></w:r><w:r><w:rPr><w:color
w:val="000000"/></w:rPr><w:t xml:space="preserve">, </w:t></w:r><w:r>
<w:rPr><w:i/><w:color w:val="000000"/></w:rPr><w:t
xml:space="preserve">adx </w:t></w:r><w:r><w:rPr><w:color
w:val="000000"/></w:rPr><w:t xml:space="preserve">1. Iluminado,
luminoso. 2. Brillante, limpo, íntido, transparente. 3. Despexado,
soleado. 4. Evidente, manifesto, nidio, obvio, patente. 5. Aberto,
franco, sincero./ </w:t></w:r><w:r><w:rPr><w:i/><w:color
w:val="000000"/></w:rPr><w:t>sm</w:t></w:r><w:r><w:rPr><w:color
w:val="000000"/></w:rPr><w:t xml:space="preserve"> 6. V </w:t></w:r>
<w:r><w:rPr><w:b/><w:color w:val="000000"/></w:rPr><w:t>clareiro</w:t>
</w:r><w:r><w:rPr><w:color w:val="000000"/></w:rPr><w:t
xml:space="preserve">. 7. V </w:t></w:r><w:r><w:rPr><w:b/><w:color
w:val="000000"/></w:rPr><w:t>clareeira</w:t></w:r><w:r><w:rPr><w:color
w:val="000000"/></w:rPr><w:t>.</w:t></w:r></w:p>
```

The next step was to divide the element tags present in these files into three major categories:

- **Elements to ignore:** many elements from the XML documents should be completely ignored, and the content should be discarded. Examples of these kind of elements are the revision markup, which delimits text that was deleted, hidden text, and certain meta-information;
- **Pass-through elements:** a couple of elements delimit strings, which are then annotated with different kind of information, such as whether the string is formatted in any specific way, or whether it is part of a replacement string, etc. For these elements we just return their content (which will then be formatted by the structured elements);
- **Structure elements:** these are the most interesting elements, which mark strings as

bold, italics or paragraphs. Unfortunately, and unlike other markup languages, there are different ways to set strings in bold or italics, and they all needed to be accounted for. This categorization process was completely iterative: looking at the result obtained, comparing it with the original Microsoft Word file, and understanding the real meaning of each element. This result process is a set of paragraphs (annotated with the `entry` tag) and a number of strings annotated in bold or italics, as XML entities (as shown in Listing 2).

**Listing 2** Entry for the word *claro* after the XML processing phase.

```
<entry>
<b>claro</b>, <b>a</b>, <i>adx </i>1. Iluminado, luminoso. 2.
Brillante, limpo, íntido, transparente. 3. Despexado, soleado.
4. Evidente, manifesto, nidio, obvio, patente. 5. Aberto, franco,
sincero./ <i>sm</i> 6. V <b>clareiro</b>. 7. V <b>clareeira</b>.
</entry>
```

In this phase we found tagging failures due to formatting errors in the original text of the dictionary, as human editing was not always consistent. These errors had to be corrected at this stage, as their presence would have complicated the following processing steps. Examples include words that should be completely in bold but had one character that was not formatted accordingly (most at the end of the word, but we found a few cases in the middle of the word as well). Most of these errors were manually corrected in this early stage of the process.

## 2.2 Towards TEI: Enriching a Basic XML Format

The previously mentioned format was rewritten into a TEI subset for dictionaries. We will not discuss this subset here, but a formal definition of the structure (a Document Type Definition (DTD) file) was created, so that one could validate the rewriting process results. This rewriting approach was also implemented in Perl, using a `Text::RewriteRules` module, and applying a similar approach as the one used for Dicionário-Aberto [11]. Nevertheless, this task was harder for the Galician dictionary, as the document structure was scarcer.

The first task was to collect the abbreviations used in the dictionary in order to classify entries regarding their use or according morphological information. These abbreviations are not similar to any other words, and therefore their detection makes it easier to gain some more knowledge on the document structure. Unfortunately, not all abbreviations were correctly listed in the dictionary roll of classification terms. This lead to the manual addition of abbreviations to the list whenever we found a missing term.

Given the amount of rewriting rules, and the fact that most of them can be quite unreadable (regular expressions can be intricate), we will only explain our approach:

1. a first group of substitutions adds specific markup where there is no ambiguity about their annotation (for example, bold words at the beginning of entries that represent entries head words; or specific abbreviations that can be marked right away);
2. after that, by using the added markup, a set of rules tries to find more places where markup can be added with minimal doubt (as bold closing tags, right after the opening of the head word term);
3. this process is repeated, adding more markup, probably in more doubtful places;
4. after validating a number of the resulting TEI documents against the DTD, specific rules were added to treat specific issues and special cases.

A rule of thumb was applied: rewriting rules do not need to be 100% precise. Imagine that a specific rule has 30% of false positives. Nevertheless, the true positive cases can help

other rules to be applied, and later other rules can use the extra markup that was meanwhile added to help fixing the wrong 30% cases.

**Listing 3** Entry for the word *claro* in the defined TEI subset.

```
<entry id="claro">
 <form>
  <orth>claro , a</orth>
 </form>
 <sense>
  <gramGrp>adx</gramGrp>
  <def n="1">Iluminado , luminoso. </def>
  <def n="2">Brillante , limpo , íntido , transparente. </def>
  <def n="3">Despexado , soleado. </def>
  <def n="4">Evidente , manifesto , nidio , obvio , patente. </def>
  <def n="5">Aberto , franco , sincero.</def>
 </sense>
 <sense>
  <gramGrp>sm</gramGrp>
  <def n="6">V <ref target="#clareiro">clareiro</ref>. </def>
  <def n="7">V <ref target="#clareeira">clareeira</ref>. </def>
 </sense>
</entry>
```

In the end, the XML files obtained were compliant with the previously defined DTD. Listing 3 presents a TEI-formatted entry[2].

## 2.3 Semi-automatic Correction of Conversion Errors

Having files that are compliant with the defined DTD is fine, but does not mean that they are semantically correct. In fact, that is far away from the truth. Taking profit of the existence of a valid DTD, it was straightforward to define a Cascading Style Sheet (CSS) to pretty print the dictionary, making it easier to browse it, and visually detect errors. Figure 2 shows an example of the CSS rendering.

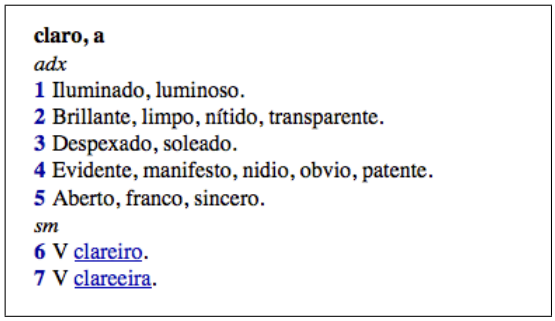


**Figure 2** Entry for the word *claro* as rendered using CSS[3].

[2] In order to enhance readability, we edited spaces and new lines in this snipped.

[3] The image suggests that the CSS rendering includes hyperlinks. Unfortunately that is not possible to obtain using CSS. For this task of visual validation the CSS fakes the link, formatting the text as if it was a valid link.

Errors from conversion are then corrected first by searching a set of error patterns, and secondly by browsing the pretty print of the dictionary. First of all, we identify the most common error patterns by an accurate human review of a number of the TEI files resultant from the automatic conversion, and elaborated a set of regular expressions for the error detection. After that, the patterns are searched in the TEI files, modifying them by applying in each case the necessary corrections in the text or in the tagging. Finally, we browse the files using the CSS rendering, searching and correcting unexpected errors remaining in the dictionary.

To illustrate this point, a common error coming from conversion is the addition of spurious blank spaces into the lemmas, that is, between their letters. The treatment of this error cannot be fully automatized, because the dictionary includes lemmas with genuine blank spaces between words. So the work on error correction after conversion has had to be designed as a regular-expression guided human task.

## 3 Linguistic and Spelling Normalization of Historical Variants

The official Galician orthography was introduced in 1982 and made law in 1983 by the Galician government. In July 2003 the Galician Royal Academy modified the language normative, introducing some important changes in spelling, morphology and lexicon. For the sake of the normalization of the dictionary, written in 1997 according to the normative of 1982, we designed a Perl program (in fact, a large set of regular expressions) to correct the text into the current official Galician normative established in 2003 [10, 5].

This Perl program replaces all "historical variants" of Galician words with their normative equivalent, leaving a mark which points to the type of normalization applied. Each mark implies a different human post-editing. So the "automatic normalization" performed by the Perl program is followed by a detailed post-edition human process.

The different types of automatic normalization and the different post-editing actions performed in each case are as follows:

- `[MX1]` Non-dubious morphological and lexical normalization according to [10]. This category includes more than 50 terms whose endings must be changed, or which should be completely changed, following the new normative: *catalana* > *catalá*, *diferencia* > *diferenza*, *pubertade* > *puberdade*, *anfitriona* > *anfitrioa*, *rector* > *reitor*, *servicio* > *servizo*, *pao* > *pau*, *tribu* > *tribo*, *esto* > *isto*, *nembargantes* > *porén*, *a penas* > *apenas*, *tal vez* > *talvez*, *alomenos* > *polo menos*, etc. Tagging for this type of automatic normalization substitutes the old form of the Galician term by its new normative equivalent, leaving the `[MX1]` tag on its left. Human post-editing process, for this category, is limited to supervising the possible mistakes in automatic normalization due to misspellings or unexpected homographs, and to removing the `[MX1]` tag.
  Example 1 shows this process for a TEI fragment (*sub voce* acaso). Item *a.* shows that TEI fragment in its original state previous to automatic normalization, item *b.* shows the same piece of TEI after the automatic process of normalization, and finally item *c.* shows the remaining fragment after human post-editing. The same convention is used for all examples in what follows.

(1) *s. v.* acaso

a. <def n="2">Quizabes, quizais, se cadra, seica, tal vez.</def>

b. <def n="2">Quizabes, quizais, se cadra, seica, [MX1]talvez.</def>

c. <def n="2">Quizabes, quizais, se cadra, seica, talvez.</def>

- `[MX2]` Morphological and lexical normalization with exceptions according to [10]. This category includes only three terms which must be changed on most occasions, but not always: *estudio* > *estudo* (unless with the meaning of room), *vocal* > *vogal* (unless relating to the voice), and *flota* > *frota* (only as noun and not as verb). The procedure used for the tagging of this category is the same as that used in the previous type: the process of automatic normalization replaces the old form of the Galician term with its new normative equivalent, leaving a `[MX2]` tag on its left. Human post-editing now must confirm or reverse the automatic substitution in each case, and remove the tag.

  (2) *s. v.* armada
  a. <def n="1">Escuadra, flota. </def>
  b. <def n="1">Escuadra, [MX2]frota. </def>
  c. <def n="1">Escuadra, frota. </def>

- `[MX3]` Morphological normalization stated in [10] which cannot be accomplished in an automatic way because it requires the gender identification of the term. This category includes two terms: *triple* > *triplo* (or *tripla*), and *cuádruple* > *cuádruplo* (or *cuádrupla*). Again, the procedure used for the tagging of this category is the same as that used in the previous types, using the masculine gender for the substitution and leaving a `[MX3]` tag on its left. Human post-editing now must confirm or replace the term with the feminine form, and remove the `[MX3]` tag.

  (3) *s. v.* trino
  a. <def n="1">Ternario, triple. </def>
  b. <def n="1">Ternario, [MX3]triplo. </def>
  c. <def n="1">Ternario, triplo. </def>

- `[*CC]/[*CT]` Compulsory spelling reduction (*cc* > *c*, *ct* > *t*) of consonant groups before *i/u* vowels, taking into account the list of exceptions stated in [10]. The tagging of this type of automatic normalization replaces the consonant group with the Galician term by its new reduced normative spelling, leaving the `[*CC]` or the `[*CT]` tag on its left in each case. Human post-editing in this category is limited to supervising the possible unexpected mistakes in automatic normalization.

  (4) *s. v.* abducción
  a. <orth>abducción</orth>
  b. <orth>abdu[*CC]ción</orth>
  c. <orth>abdución</orth>

  (5) *s. v.* aboiar
  a. <def n="1">Flotar, fluctuar. </def>
  b. <def n="1">Flotar, flu[*CT]tuar. </def>
  c. <def n="1">Flotar, flutuar. </def>

- `[INI]` Removing of question and exclamation initial marks according to [10]. This kind of normalization only takes place in the examples present in the entries of the dictionary. Human post-editing in this category is limited to supervising the possible unexpected mistakes.

(6) *s. v.* aviado

a. <def n="2"><lbl>fig</lbl> Amolado, apañado, fastidiado <quote>(¡estouche aviada con esta febre!)</quote>

b. <def n="2"><lbl>fig</lbl> Amolado, apañado, fastidiado <quote>([INI]estouche aviada con esta febre!)</quote>

c. <def n="2"><lbl>fig</lbl> Amolado, apañado, fastidiado <quote>(estouche aviada con esta febre!)</quote>

- `[LX1=*mistake]` Non-dubious spelling, morphological and lexical normalization stated in [5]. This category includes more than 700 terms (mostly, but not only, castilianisms and anglicisms) which are marked with an asterisk in the listing of [5] and which are changed following its normative suggestion: *almíbar* > *caldo de azucre*, *altavoz* > *altofalante*, *armonía* > *harmonía*, *avantaxe* > *vantaxe*, *basoira* > *vasoira*, *coruxa* > *curuxa*, *fumo* > *fume*, *obscuro* > *escuro*, *reptil* > *réptil*, *pranchar* > *pasar o ferro*, *prohome* > *home de prol*, *playback* > *son pregravado*, *antidóping* > *antidopaxe*, *feed-back* > *retroacción*, etc. The tagging of this type of automatic normalization replaces the wrong form of the Galician term with its normative equivalent, leaving the `[LX1]` tag on its left along with the changed term. Human post-editing in this category is focused on supervising the possible mistakes in automatic normalization due to misspellings or unexpected homographs, confirming or reversing the automatic substitution in each case.

(7) *s. v.* adianto

a. <def n="3">Avantaxe, mellora, progreso. </def>

b. <def n="3">[LX1=*Avantaxe]vantaxe, mellora, progreso. </def>

c. <def n="3">Vantaxe, mellora, progreso. </def>

- `[LX2=*mistake]` Spelling, morphological and lexical substitutions based on the asterisked terms in the listing of [5], but which have two or more normative solutions (not always clearly synonyms) in this normative reference work. This category is formed by 41 terms, and the choice for the Perl program was the solution more frequent in its usage or more general in its meaning: *bucear* > *mergullarse*, *carcoma* > *caruncho*, *inasequible* > *inaccesible*, *rincón* > *recuncho*, etc. The tagging of this type of automatic normalization replaces the wrong form of the Galician term with the selected normative equivalent, leaving the `[LX2]` tag on its left along with the term changed. As in the previous type, human post-editing in this category is focused on supervising the possible mistakes in automatic normalization due to misspellings or unexpected homographs, confirming, modifying or reversing the automatic substitution.

(8) *s. v.* abstruso, a

a. <def n="1">Escuro, inasequible, recóndito. </def>

b. <def n="1">Escuro, [LX2=*inasequible]inaccesible, recóndito. </def>

c. <def n="1">Escuro, inaccesible, recóndito. </def>

- **[LX3=correction?]** Polysemic words which need correction according to [5] but only in some of their meanings. The wrong meaning is unusual so most times it should not be corrected. For this reason, the Perl program does not perform the substitution, but only marks the term suggesting the possible corrected form and its intended meaning: *bolo* > *birlo* (xogo), *cru* > *crup* (doenza), *racha* > *refacho* (de vento), *tanto* > *punto* (no xogo), *demais* > *de máis* (de sobra), *berrón* > *verrón* (porco semental), *solar* > *soar* (terreo),

*vencello* > *birrio* (ave), etc. In general, human post-editing in this category is limited to removing the tag. In some cases, it will be necessary to adopt the corrected form, or even providing a better one.

(9) *s. v.* abstruso, a

a. <def n="5"><lbl>fig</lbl> Cáustico, corrosivo, cru, esgueiro, ferinte, mordaz, ofensivo, punxente, punzante, sedizo.

b. <def n="5"><lbl>fig</lbl> Cáustico, corrosivo, [LXE3=crup (doenza)?]cru, esgueiro, ferinte, mordaz, ofensivo, punxente, punzante, sedizo.

c. <def n="5"><lbl>fig</lbl> Cáustico, corrosivo, cru, esgueiro, ferinte, mordaz, ofensivo, punxente, punzante, sedizo.

## 4 Applications: Galnet and Dicionário Aberto

The process described was performed not just to create a new and valuable resource (a Galician dictionary in a semantic-rich format) but also with some applications of this dictionary in mind.

We have two case studies where we want to take advantage of the Galician dictionary: extracting synonyms for Galnet project, and adding a new language to Dicionário-Aberto. This section will describe how we intend to perform these two ideas.

### 4.1 Extracting Synonyms for Galnet

The aim of the Galnet project [2] is building a WordNet for Galician aligned with the ILI (the inter-lingual index) generated from the English WordNet 3.0. WordNet [6] is a lexical knowledge base structured as a semantic network. In this lexical-semantic network, each node is a concept, and the edges which connect them are the semantic relations (hyponymy, meronymy, etc.) that are established between the concepts. Each concept in the network is represented by the group of synonymic lemmas that can express this concept. In terms of WordNet, each group of synonyms is a *synset*, and each synonym part of this group is a variant (or a lexical variation of the same concept).

Today, WordNet is, probably, the most important computational resource with lexical-semantic information, especially in the field of natural language processing (NLP), where it is used in tasks of automatic semantic disambiguation, information retrieval, automatic text classification and automatic summarization, among others.

Most of the versions of WordNet in languages other than English follow the design model of EuroWordNet [16], where *synsets* that are part of the WordNet for one of the languages, are linked to the *synsets* from other languages, through an ILI that is unique to each concept and which is mainly based on the *synsets* of the English WordNet. Therefore, the set of WordNet lexicons in different languages allows the connection between the *synsets* of any two languages via the ILI, thus constituting a very useful feature in applications of linguistic technologies which deal with multilingual processing, such as automatic translation or cross-language information retrieval.

Galnet is distributed under a Creative Commons license[4] (CC BY 3.0) as part of the Multilingual Central Repository, currently available in version 3 (MCR 3.0). The MCR 3.0 integrates the WordNet for English, Spanish, Catalan, Basque and Galician in the framework

[4] Details on the Creative Commons licenses can be found at `http://creativecommons.org/`.

of EuroWordNet. The ILI index allows the connection between words which are equivalent in different languages. The current version of ILI corresponds to the English WordNet 3.0 developed at Princeton University. The MCR also integrates the WordNet Domains, new versions of the Base Concepts and the Top Ontology, and the AdimenSUMO ontology. Thus, the MCR is a multilingual semantic resource of broad range suitable for use in language processing tasks that require large amounts of multilingual knowledge [4].

In its current state, Galnet reaches a lexical coverage of about one-fifth of the English WordNet, as shown in detail in Table 1.

**Table 1** Galnet current state.

| | WN30 | | Galnet | |
|---|---|---|---|---|
| | Vars | Syns | Vars | Syns |
| N | 117798 | 82115 | 18949 | 14285 |
| V | 11529 | 13767 | 1416 | 612 |
| Adj | 21479 | 18156 | 6773 | 4415 |
| Adv | 4481 | 3621 | 0 | 0 |
| TOTAL | 155287 | 117659 | 27138 | 19312 |

The goal of the Galnet project is to reach a lexical coverage similar to the English WordNet, in order to facilitate language technologies for Galician. One of the methodologies used to extend that coverage is lexical information acquisition from human-oriented electronic dictionaries and thesaurus. In fact we have yet applied that methodology in a previous phase of the project, using the WN-Toolkit [9] to expand Galnet from two existing bilingual English-Galician resources: Wikipedia and the English-Galician CLUVI Dictionary [3].

The automatic extraction techniques applied to these two lexical resources had two distinct objectives: on one hand, expand Galnet with proper names spelled in the same way in English and Galician from the material provided by Wikipedia; on the other hand, expand Galnet with the Galician variants included in Wikipedia and in the CLUVI Dictionary as translations of English words included in the *synsets* of WordNet and not coded yet in Galnet. As for the *Diccionario de sinónimos da lingua galega*, we aim to expand Galnet with the Galician variants included as synonyms in entries whose lemma or companion synonyms are part of Galnet.

## 4.2 Adding Galician Definitions to Dicionário Aberto

Dicionário-Aberto[5] is a website that allows the user to query a continuously updated version of a Portuguese dictionary from 1913. The details on its construction and the main goals for this project can be read in [14, 13, 12]. In order to improve its funcionality, Dicionário-Aberto has been learning new features in the last few years.

Given the proximity between the Galician language and the Portuguese language, and the fact that researchers from these languages often look into the other language resources to check for definitions, terminological solutions, idiomatic expressions or word origins, the Dicionário-Aberto team is interested in using the Galician dictionary presented in this article to enrich the user experience:

- for each dictionary entry, show a list of possible Galician translations,
- for each translation, present the Galician dictionary entry.

[5] Available at `http://dicionario-aberto.net`.

Although the second goal is easy to achieve, the first is quite difficult, namely because there are no freely available good translation dictionaries between Portuguese and Galician, and the ones available cover only a small percentage of the domain of the dictionary (see Simões and Guinovart, this proceedings, for some work on this subject).

Taking advantage of the strong connection available between these two languages, the Galician dictionary entries will also be subject to relation extraction (synonyms, hypernyms, hyponyms, etc) that can be used to enrich the base ontology of Dicionário Aberto, and enhance the user experience when searching or browsing the dictionary.

## 5 Final Remarks

At the beginning of the 21st century we have a lot of useful information that cannot be used because it is encoded in paper or, more recently, in non semantic-rich formats. The definition of procedures to retread these documents into structured documents that can be easily processed by computer programs is relevant. In this paper we present the work done with a Galician dictionary stored in a Microsoft Word file, and how it was processed and converted into a structured format (Text-Encoding Initiative schema subset).

The process used for the dictionary conversion is not universal, and cannot be applied blindly to any dictionary in Microsoft Word format. Nevertheless, the approach is universal, and can be easily adapted for other dictionaries, just by adjusting the rewriting rules for the specific formatting details of the original files.

At the end of the conversion process we obtained a dictionary with more than 24571 entries (41923 meanings or groups of synonyms), written in modern Galician ortography, and annotated using a semantic-rich format, that can be easily explored for different tasks. At the moment the dictionary is not available for complete download, but its content will soon be available for querying using a simple web interface, as an independent resource, or complementing the Dicionário-Aberto dictionary of the Portuguese language.

**Acknowledgments** This work was partially supported by Grant TIN2012-38584-C06-04, supported by the Ministry of Economy and Competitiveness of the Spanish Government on "*Adquisición de escenarios de conocimiento a través de la lectura de textos: Desarrollo y aplicación de recursos para el procesamiento lingüístico del gallego (SKATeR-UVIGO)*"; and by the Xunta de Galicia through the "*Rede de Lexicografía (Relex)*" (Grant CN 2012/290) and the "*Rede de Tecnoloxías e análise dos datos lingüísticos*" (Grant CN 2012/179).

## References

1 José João Almeida and José Carlos Ramalho. XML::DT a Perl down-translation module. In *XML-Europe'99, Granada - Espanha*, May 1999.

2 Xavier Gómez Guinovart, Xosé María Gómez Clemente, Andrea González Pereira, and Verónica Taboada Lorenzo. Galnet: WordNet 3.0 do galego. *Linguamática*, 3(1):61–67, 2011.

3 Xavier Gómez Guinovart, Alberto Álvarez Lugrís, and Eva Díaz Rodríguez. *Dicionario moderno inglés-galego*. 2.0 Editora, Ames, 2012.

4 Aitor González, Egoitz Laparra, and German Rigau. Multilingual central repository version 3.0: upgrading a very large lexical knowledge base. In *6th Global WordNetConference*, Matsue, Japan, 2012.

**5** Manuel González González and Antón Santamarina Fernández. *Vocabulario Ortográfico da Lingua Galega (VOLGa)*. Real Academia Galega/Instituto da Lingua Galega, A Coruña/Santiago, 2004.

**6** George A. Miller, Richard Beckwith, Christiane Fellbaum, Derek Gross, and Katherine Miller. Wordnet: An on-line lexical database. *International Journal of Lexicography*, 3:235–244, 1990.

**7** Camiño Noia Campos, Xosé María Gómez Clemente, and Pedro Benavente Jareño. *Diccionario de sinónimos da lingua galega*. Galaxia, Vigo, 1997.

**8** Hugo Gonçalo Oliveira and Paulo Gomes. Onto.PT: automatic construction of a lexical ontology for Portuguese. In *5th European Starting AI Researcher Symposium (STAIRS 2010)*, August 2010.

**9** Antoni Oliver González. WN-Toolkit: un toolkit per a la creació de wordnets a partir de diccionaris bilingües. *Linguamática*, 4(2):93–101, 2012.

**10** Real Academia Galega. *Normas ortográficas e morfolóxicas do idioma galego*. Editorial Galaxia, Vigo, 2004.

**11** Alberto Simões and José João Almeida. Processing XML: a rewriting system approach. In Alberto Simões, Daniela da Cruz, and José Carlos Ramalho, editors, *XATA 2010 — 8ª Conferência Nacional em XML, Aplicações e Tecnologias Aplicadas*, pages 27–38, Vila do Conde, May 2010.

**12** Alberto Simões, José João Almeida, and Rita Farinha. Processing and extracting data from Dicionário Aberto. In Nicoletta Calzolari et al., editor, *Seventh International Conference on Language Resources and Evaluation (LREC2010)*, pages 2600–2605, Valletta, Malta, May 2010. European Language Resources Association (ELRA).

**13** Alberto Simões and Rita Farinha. Dicionário Aberto: Um novo recurso para PLN. *ViceVersa*, 16:159–171, December 2011.

**14** Alberto Simões, Álvaro Iriarte Sanromán, and José João Almeida. Dicionário-aberto – a source of resources for the portuguese language processing. *Computational Processing of the Portuguese Language, Lecture Notes for Artificial Intelligence*, 7243:121–127, April 2012.

**15** TEI Consortium, editor. *TEI P5: Guidelines for Electronic Text Encoding and Interchange*, chapter 9. Dictionaries. TEI Consortium. `http://www.tei-c.org/release/doc/tei-p5-doc/en/html/DI.html` (January 2012), Version 2.0.1 edition, December, 22nd 2011.

**16** Piek Vossen. Wordnet, eurowordnet and global wordnet. *Revue française de linguistique appliquée*, 7:27—-38, 1990.