Towards the Creation of New Galician Language Resources: From a Printed Dictionary to the Galician WordNet

Fernando Mag·n MuÒoz
"RamÛn PiÒeiro" Center for Research in Humanities
15896 Santiago de Compostela, Galicia, Spain
fmagan@cirp.es

Abstract

The work herein presented concentrates on describing the initial methods for creating and developing new distinct Galician language resources which will be able to work in conjunction with one another. We set out, firstly, to explain the processing of a dictionary by taking an extant printed monolingual Galician dictionary, and converting its data into a new SGML-encoded operational structure while trying, at the same time, to preserve all the original lexical information. Secondly, we introduce the Reference Corpus of the Modern Galician Language (CORGA) and describe the CORGA Encoding Scheme (CORGA-ES). Next, semantic aspect is dealt with within a discussion about the creation of the Galician version of WordNet. Finally, by activating the three previously obtained mechanisms, one inmediate result is the creation of a primitive lexical-semantic database which interacts with the corpus thus improving the database coverage; it will include items and cross-reference data from the semantic network contained within the Galician WordNet. This new Galician language database will become increasingly useful as its design, format, and composition improve. As a consequence, newer resources may be derived from this system through internal processing and interacting cycles among all the previous resources.

1. Introduction

Basic computational Language Resources for the Galician language have been scarce up until now. This scarcity is mainly due to the "special" historical nature of Galician, a milenium-language which did not have an "official" academic grammar until 1982. This situation is changing thanks to the efforts put into promoting and stimulating research into the Galician language, by the Galician universities and research centers. A usefull random sample might be the linguistic research being undertaken in the "Ramón Piñeiro" Center for Research in Humanities (CRPIH).

A language that fails to keep pace with technological innovation if not dead is certanly "unwell". Furthermore, there will never be new-technology applications envolving NLP systems if basic lexical-semantic resources are not created previously. There is therefore a general consensus that Galician should try to catch up with state of the art technological developments (Cole *et al*, 1997) in order to find itself a place among most of the foremost European languages.

This paper is organised as follows. In Section 2, we introduce and develop a method for converting a monolingual Galician dictionary into a Machine-Tractable Dictionary. Section 3 is devoted to presenting a Galician corpus and its encoding scheme. Next, the opportunity for bulding up a Galican version of the WordNet is presented in Section 4. A model for developing a Galician lexical database is outlined in Section 5. The concluding section considers the convenience of using standards in the processes of building language resources.

2. Getting a Machine-Tractable Dictionary

We start utilizing the "reusability" concept (Calzolari, 1991) by taking an extant printed dictionary, the Royal Galician Academy's monolingual Galician dictionary, DRAG (DRAG, 1997), in order to get a Machine-Tractable Dictionary (MTD) (Wilks *et al.*, 1996), in the most automatic way possible. From the lexicographic point of view, DRAG is a traditional dictionary. It contains approximately 25,000 headwords which enclose about 57,500 definitions; from one to 18 definitions per entry, and 2.3 definitions per entry on average. Initially a printed copy of DRAG was scanned and OCR processed to convert it to word-

processor format files preserving the typesetter¥s style and giving, as a result a, conventional Machine-Readable Dictionary (MRD), with entries such as the following[1]:

abater *v.t.* 1. Facer caer, tirar polo chan. *Abateuno dun golpe. O vento abateu o valado.* SIN. ***derrubar.*** 2. Por ext., matar [a alguÈn] con violencia, caendo a vÌctima Û chan. *Abateu o inimigo a balazos.* 3. Baixar [unha cousa] poÒÈndoa en sentido horizontal. *Abate-lo asento do coche.* 4. fig. Facer perder [a alguÈn] o ·nimo ou as forzas. *Todos aqueles fracasos acabaron por abatelo completamente.* SIN. ***aplanar, chafar*** (fam.). CF. ***deprimir, desanimar.*** // *v.p.* 5. Caer polo chan, vir para abaixo. *Co temporal abateuse o teito da casa.* SIN. ***derrubarse.*** 6. fig. Perde-lo ·nimo ou as forzas. *Con tantos desgustos abateuse todo.* CF. ***deprimirse, desanimarse.*** 7. Inclina-lo corpo acerc·ndoo Û chan. *AbatÈronse trala parede para evita-las balas.* CF. ***abaixarse, agacharse, anicarse.***

From this initial MRD, the files are translated into HTML format files which make two objectives possible. On the one hand, the separation of presentational information and real data from the whole dictionary; thus, type, style and size font are preserved within the HTML tags. On the other hand, this procedure means that the dictionary's data is converted into ASCII format for easier processing. In the following step Unix tools and the euristic knowledge provided by lexicographers are used to analyse HTML tags and character separators in order to discover the lexical contents. SGML generic identifiers (specific tag names) are introduced by replacing the HTML tag strings which separate individual components including dictionary data. Before assertaining the rules of grammar which facilitate the generation of current file structure, we had to close all open data elements and to introduce an additional markup, and in so doing, leave the structure of the entries well-defined, and as consistent as possible. The next task was to try to find out what Document Type Definition (DTD) could be infered from the texts tagged in this way, i.e., to determine the corresponding grammar. Tests were carried out by using the Fred program, a SGML grammar builder from the Fred Project in OCLC (Fred, 1998). The program had to be run various times and some errors had to be corrected because ambiguous SGML content models[2] still appeared.

Once all structural ambiguity was eliminated, the whole dictionary was SGML-parsed to check its SGML-conformance while at the same time, verifying its consistency, by a SGML parser. The main rules at the top fo the DTD are similar in appearance[3] to the Figure 1.

```
<!ELEMENT DRAG - -    (ENTRY+)                                          >
<!ELEMENT ENTRY            - -    ( HEADWORD, E_GC+, GENERAL? )          >
<!ELEMENT HEADWORD         - -    (#PCDATA)                              >
<!ATTLIST  HEADWORD        HOMOG_NUM  NUMBER  #IMPLIED    >
<!ELEMENT E_GC  - -    ( GRACAT, DEF_STRU+)              >
<!ELEMENT GENERAL          - -    (SIN | ANT | CF | OBS)                 >
<!ELEMENT GRACAT           - -    (#PCDATA)                              >
<!ELEMENT DEF_STRU         - -    ( ( DEF, AFISEM?, DEF_EXAMPLE, CF, OBS, EXAMPLE+ )
                                  | ( DEF, AFISEM?, DEF_EXAMPLE, SIN, ANT?, CF )
                                        | ( DEF, AFISEM?, DEF_EXAMPLE, SIN, REXIST?, CF )
                                              | .....................................................
        <! .... .... .....                                                 >
```

Figure 1. Top elements in the MTD¥s Document Type Definition.

A Machine-Tractable-Dictionary's entry consists of single headword (a simple homograph if there are various homographs), one or more gramatical categories, and a common structure ("general") at the bottom. For each gramatical category, the entry includes one or more definition structures; and each definition structrure includes a unique dictionary definition. The general element encloses information which affects

[1] English meanings for the definitions in the entry example, for the Galician word "abater" could be: 1. to kick down, cut down, demolish; 2. to shoot down, kill somebody; 3. to bring down; 4. to discourage, humble, humiliate; 5. to drop, fall; 6. to be depressed, get discouraged; 7. to prostrate, lay down.

[2] These are "content models for which an element or character string ocurring in the document instance can satisfy more than one *primitive content token* without look-ahead" (Goldfarb, 1994). Such content models can be easily identified if the DTD includes statements of the type of (#PCDATA | ELEMENT_NAME)+ on its *content model declaration*.

[3] The complete DTD for the MTD is not included here, many elements and atributes are not explicity shown.

all dictionary definitions included within the entry. (Figure 2 shows the example entry, previously introduced, into the MTD.).

At this point, the MTD is ready. It may be loaded into a DBMS with a SGML front-end and in order to be able to access, retrieve, and modify data; and further, a new printed dictionary can be managed by introducing new entries, electronically. Furthermore, it constitutes a well-structured language resource which we can be used to build up subsequent resources (which will come back to in Section 5).

```
<!-- ---------------------- Begin (Machine-Tractable Dictionary entry) ---------------------------- -->
<ENTRY>
<HW HOMOG_NUM = "0">abater</HW>
<!-- ---------------------- GRAMATICAL CAT #1 ---------------------------------------------- -->
<E_GC>
<GRACAT>v.t.</GRACAT>
<!-- ------------ Structure #1 ------------- -->
<DEF_STRU>
<DEF NUM="1"><DEF_LIT>Facer caer, tirar polo chan.</DEF_LIT></DEF>
<DEF_EXAMPLE>Abateuno dun golpe. O vento abateu o valado. </DEF_EXAMPLE>
<SIN>derrubar. </SIN>
</DEF_STRU>
<!-- ------------ Structure #2 ------------- -->
<DEF_STRU>
<DEF NUM="2"><AFISEM>Por ext.</AFISEM><DEF_LIT>matar [a algu&eacute;n] con violencia, caendo a
v&iacute;ctima &oacute; chan.</DEF_LIT></DEF>
<DEF_EXAMPLE>Abateu o inimigo a balazos. </DEF_EXAMPLE>
</DEF_STRU>
<!-- ------------ Structure #3 ------------- -->
<DEF_STRU>
<DEF NUM="3"><DEF_LIT>Baixar [unha cousa] po&ntilde;&eacute;ndoa en sentido horizontal.</DEF_LIT></DEF>
<DEF_EXAMPLE>
Abate-lo asento do coche. </DEF_EXAMPLE>
</DEF_STRU>
<!-- ------------ Structure #4 ------------- -->
<DEF_STRU>
<DEF NUM="4"><AFISEM>fig.</AFISEM><DEF_LIT>Facer perder [a algu&eacute;n] o &aacute;nimo ou as
forzas.</DEF_LIT></DEF>
<DEF_EXAMPLE>Todos aqueles fracasos acabaron por abatelo completamente. </DEF_EXAMPLE>
<SIN>aplanar, chafar </SIN>
<REXIST>fam.</REXIST>
<CF>deprimir, desanimar. </CF>
</DEF_STRU>
</E_GC>
<!-- ---------------------- GRAMATICAL CAT #2 ---------------------------------------------- -->
<E_GC TYPE = NEWCAT>
<GRACAT>v.p.</GRACAT>
<!-- ------------ Structure #5 ------------- -->
<DEF_STRU>
<DEF NUM="5"><DEF_LIT>Caer polo chan, vir para abaixo.</DEF_LIT></DEF>
<DEF_EXAMPLE>Co temporal abateuse o teito da casa. </DEF_EXAMPLE>
<SIN>derrubarse. </SIN>
</DEF_STRU>
<!-- ------------ Structure #6 ------------- -->
<DEF_STRU>
<DEF NUM="6"><AFISEM>fig.</AFISEM><DEF_LIT>
Perde-lo &aacute;nimo ou as forzas.</DEF_LIT></DEF>
<DEF_EXAMPLE>Con tantos desgustos abateuse todo. </DEF_EXAMPLE>
<CF>deprimirse, desanimarse. </CF>
</DEF_STRU>
<!-- ------------ Structure #7 ------------- -->
<DEF_STRU>
<DEF NUM="7"><DEF_LIT>Inclina-lo corpo acerc&aacute;ndoo &oacute; chan.</DEF_LIT></DEF>
<DEF_EXAMPLE>Abat&eacute;ronse trala parede para evita-las balas. </DEF_EXAMPLE>
<CF>abaixarse, agacharse, anicarse. </CF>
</DEF_STRU>
 </E-GC>
</ENTRY>
```

Figure 2.Tagged entry into the MTD.

3. Encoding an Unstructured Corpus

Language corpora are essential resources for general linguistic research and crucial in the creation of NLP applications and tools for a limited-diffusion language such as Galician. We present CORGA (Reference Corpus of the Modern Galician Language), as an on-going project that constitutes the basis for the establishment of an Electronic Text Center in the CRPIH. Its ultimate objective is the compilation of an exhaustive collection of modern texts that will be made available for research tasks. Currently, the total number of words for CORGA is about 10 million, and they are contained within 275 electronic documents of different types: journal texts (35), plays (58), narrative (121), essay (41) and oral texts (20). Word counts differ widely depending on the type of text (c.f. Table1).

	Docs	Words	%W
Plays	58	795,902	8.4%
Narrative	121	4137289	43,9%
Essay	41	2,469,200	26,2%
Journal	35	1,818,590	19,3%
Oral Texts	20	204,001	2.2%
Total:	275	9,424,982	100,0%

Table 1. Current word account in CORGA

In the first phase of its development, written texts were introduced by means of scanning and by means of converting composition tapes supplied by the publishing houses; oral texts were converted from audio cassetes by typing their contents. As a second step, texts were regulated; All "incorrect" words were marked in cases of obvious typographical mistakes in the source (<errata>), foreign words (<foreign>) and unexplained errors (<sic>). Next, texts were converted to HTML format to separate the presentational markup from word-processor format, in a similar way as described for the DRAG, and then saving them as files in ASCII format. We conclude this initial phase, by loading all CORGA documents into a database which can operate from a hybrid perspective, because the DBMS used allows the management of mixed content on the same database record: for one, a few relational fields (doc_number, doc_title, doc_type - play, journal, narrative, ...-, author, etc.) are taken in account; for another, all words from the complete document text are included and, therefore, indexed. Thus, CORGA may be accessed, the first results being obtained by carrying out word counts, frequency indexes, word retrieval, word searching in context, etc.. In spite of the extraordinary benefits provided by the accessibility of the early database it was clear that CORGA should be encoded (tagged) in order to take advantage of the explicit and implicit information included therein, and so, an encoding scheme for CORGA was required.

CORGA-Encoding Scheme (CORGA-ES) aims at identifying the encoding structure for the potential Galician corpora from the point of view of their representation (structural and linguistic markup) and from the computational point of view (paying attention to consistency, recuparability, and above all, reusability) bearing in mind the specific peculiarities of the Galician language. Design criteria for the CORGA-ES follow the recommendations expounded in the Encoding Scheme of the Corpus Encoding Standard (CES) (CES, 1996) -a jointly MULTEXT, EAGLES, and CNRS/Vassar initiative-. Like CES, CORGA-ES draws a distinction between two encoding levels:

1. Structural Markup level, which corresponds to the minimal level of encoding. It provides features which make the CORGA a "reusable" corpus.

2. Linguistic Annotation, which includes the linguistic information which is generated from some kind of linguistic analysis processing (taggers, lemmatizers, lexical analyzers, etc.) and added to the structural markup.

Linguistic annotation and structural markup are maintained in different SGML documents and generated from two distinct SGML DTDs, corgaDoc.dtd and corgaAna.dtd, respectively. The relationship between the corresponding elements from both encoding levels are marked by special SGML link structures.

3.1 Structural Markup

Within the structural markup, at the very top of the hierarchy, a CORGA corpus (corgaCor) consists of a CORGA header (corgaCab) which is followed by a series of CORGA documents (corgaDoc). Each corgaDoc has its own header which uses the same structure as the corpus header. All documents have this header followed by a "body" of text (corpo).

Paragraph-structure is the reference element in CORGA-ES, so we can establish two structural encoding levels: the encoding at paragraph level and the encoding within the paragraph. Since the text will often contain elements which represent structures which are higher or on the same level as that of paragraphs, the divison element will be used to introduce parts, chapters, sections, poems, etc. The divisions thus appear wherever they are needed (also nested).

After a division, just two types of elements can ocurr: elements on the same level as the paragraph (paragraph, spoken paragraph[4], poem, note, etc.) and new division elements.

Conceptually, the same elements which appear in a phrase can appear in a paragraph. This is why this level of structural encoding is also called phrase-level. Paragraphs may contain the following element categories: elements of linguistic interest, highlighted elements, elements for identifying orthographical sentences (s-units), pointers or reference elements, gaps and potential ambiguities; and elements identifying speech transcription phenomena (sic, foreign, and errata elements would also be included within a paragraph).

Tags of linguistic interest serve to pre-identify specific elements (dates, numbers, measures, proper nouns, time, ...), thus these are used to help software tools to process the corpus materials (lemmatizers, sintactic taggers, etc).

Spoken paragraphs include, at structural level, pause, vocalization, event, and sincronymy tags.

3.2 Linguistic Annotation

The strategy that CORGA-ES adopts means that the linguistic annotation is not added to the structural information in the same corpus document. Both information types are maintained separatelly in two different SGML documents. Documents which are linguistically annotated point to the corresponding structurally marked documents by including SGML pointers from their annotations.

A CORGA linguistically annotated document (corgaAna) is generated from the fixed specifications within the corgaAna.dtd. Since the marking up of the linguistic annotation of CORGA is performed by a linguistic tagger[5] (Vilares *et al.*, 1996), CORGA-ES is, actually, specifying a recommended output from this tool.

The top of a linguistically annotated CORGA document consists of a CORGA header (corgaCab) and a list of chunks (chunkList). Each chunk includes an analysis paragraph (a series of orthographical sentences). A sentence element includes token elements, but it may include non-token elements (information elements) which consist of non-analysed text.

A token consists of: the orthographical form of a token (ortho), a disambiguation element (disamb) which includes one or more disambiguated CORGA tags, elements of mophosyntactic information (lex) with one or more morphosyntatic information groups (one group for each possible token analysis). A *lex* element includes *lemma* (a base form, or lemma), morphosyntatic description (msd) (following the EAGLES recommendations (EAGLES, 1996)) and, CORGA tag (ctag) (a corpus tag with the morphosyntatic data that it corresponds to).

---

[4] Elements including text writen as spoken, and written to be spoken (such as in the transcription of audio tapes or theatre plays) are considered spoken paragraphs; they have the same structural value as normal paragraphs.

[5] A linguistic-tagging tool to CORGA corpus is being developed in collaboration, between the Computer Science Department of the University of A Coruña and the Linguistics Department of the CRPIH.

3.2 Effective Marking Up

Effective structural marking up of CORGA is carried out according to the CORGA-ES specifications by automatic and semi-automatic means. Before performing the structural markup processing, CORGA docs (in raw text format, including the tags for "incorrect" words) are converted to HTML format to separate the presentational markup while leaving the files in ASCII format, as in the DRAG dictionary processing.

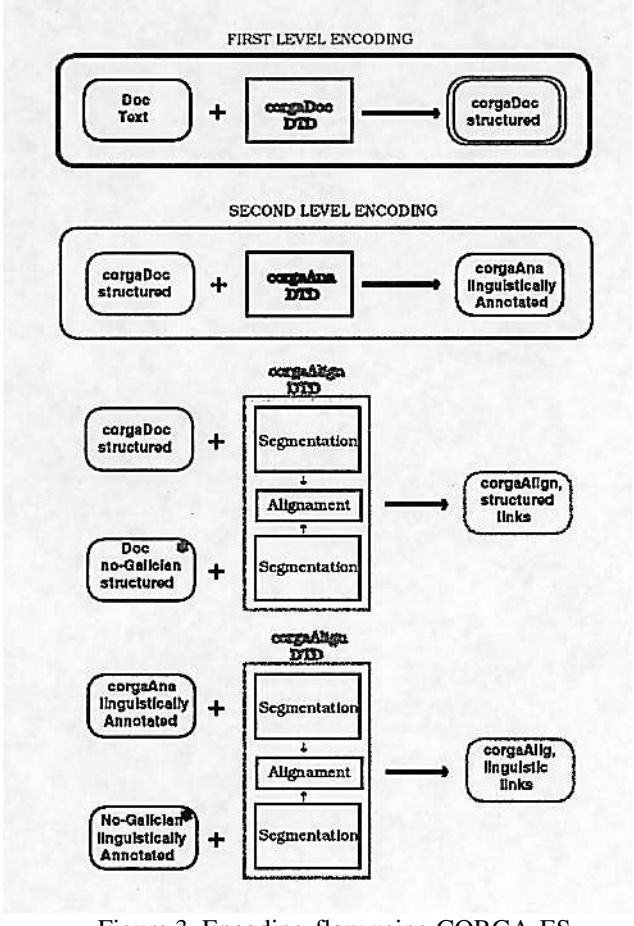

Figure 3. Encoding-flow using CORGA-ES.

Prior to the initiation of a semiautomatic tagging process, we should try to markup the maximum number of elements through automatic processing as possible. In addition to paragraphs, sentences, -and the highligted tags and sic, errata, and foreign elements, tagged in advance- further candidates for "murk-up" should be considered. Furthermore, methods for proper noun identification (McDonald; Mani and MacMillan; Paik *et al.*; 1996) and tagging will be set up because of its relevance to the analysis procedures.

With CORGA documents already structurally marked, the corpus may be considered as being ready to be processed by a linguistic tagger; the output of this kind of processing constitutes CORGA linguistically annotated. The rest of the applications, such as parallel corpora, must be derived from the structurally marked version of CORGA (See Figure 3).

4. The Galician WordNet

WordNet (Miller *et al.*, 1993) is a semantic database for the English language which may be viewed as a linguistic rather than a conceptual ontology. Wordnet database stores word meanings with semantic relations among them such as hyponymy, meronymy, causal relations, etc.. WordNet is organised around the concept of "synset", i.e., a structure comprising one or more word senses that are considered to be identical in meaning, and a "gloss" defining that meaning. WordNet contains nouns, verbs, adjectives and adverbs. It has been conceived as a computational resource, so removing some dicitionary drawbacks, such as the circularity of the definitions and the ambiguity of sense references.

EuroWordNet (EWN) is an EC project that aims at building a multilingual database consisting of wordnets from several European languages (Dutch, British English, Italian and Spanish[6]) via interlingual indexes.

The evident problem with this type of database is that they are language specific. The language must be integrated into EWN by creating the corresponding wordnet for the specific language. We intend to include the Galician language into the multilingual EWN, and we thus need to build the Galician WordNet (GalWordNet).

EWN suggests that EuroWordnet databases are built, in as much as possible, from available existing resources. This fact limits the approach which the development of the Galician wordnet must follow. The most general method for building a wordnet involves working with two different parts of the vocabulary: the core vocabulary (first phase) and the extended vocabulary (next phases).

Core vocabulary consists of the so-called Base Concepts (BC) those which contain the highest number of conceptual relationships, and ocuppy the most important positions in the wordnets. In order to determine the Galician Base Concepts we use a common set of BCs of the most advanced wordnets. The original list of BCs include about 1,000 English synsets (nouns and verbs) selected from Miller's WordNet (version 1.5). Galician is assigned Galician synsets for the EWN BCs through a process of linkage (by hand) whilst in so doing a harmony with the rest of the EWN groups is maintained. Each base concept record is formed by an offset_synset, an English_words_synset, a Galician_words_synset; as in the example that follows (gloss is omitted):

01558020-v | come_down precipitate fall | abater derrubar

The extended vocabulary is more specific and depends on the core vocabulary. The availability of existing resources plays a determining role in its construction. Most EWN groups develop the extended vocabulary (automatically) using the *expand model* which is put into practice by selecting WordNet1.5 synsets; using bilingual dictionaries, translating them into equivalent synsets in the specific language; and verifying the EWN relations (adapting them, and using monolingual dictionaries to check the relations imposed on the specific language synset, if neccessary).

As there are no bilingual English-Galician dictionaries the expand model cannot be used for the construction of the extended Galician vocabulary. At present, we are discussing whether we ought to follow the so-called *merge model*, which is based on selecting the vocabulary from a local resource (a monolingual dictionary, for example), creating the synsets from this source, developing the internal relations independently, and then, generating the equivalence relations for WordNet1.5; or instead whether we should develop a hybrid model by using a list of English-Galician words, the monolingual DRAG, and the existing Galician synonym dictionaries, along with other sources, in order to develop the extended vocabulary.

5. A primitive Lexical-Semantic Database

MRDs contain much of the data which is needed to create computational lexicons. It is commonly accepted that a lexicon would have to include information about morphology, either in a form which enables the generation of all potential word-forms, or as a list of word-forms, or as a combination of the two (Grishman and Calzolari, 1998). With this criteria in mind, it is particularly difficult to find purpose-free (neutral) lexicons because most of them were application-specific; this trend continued until the standartization initiatives appeared (EAGLES, 1993).

We aim to construct, firstly, a Galician lexical database, and by a lexical database we mean a reference system that accumulates information on the lexical items of a language. In this way, Machine-Tractable Dictionaries can also be regarded as primitive lexical databases (Gómez-Hidalgo and Buenagana, 1997). Initially, therefore the MTD obtained from the DRAG, should be purpose-free; it will not be a lexicon (because, initially, it will not cover morphological developments) but it will be a primitive lexical database (pLDB) with the following features:

[6] The database is being extended with German, French, Swedish, Estonian, and Czech. At the same time, some minoritary languages as Catalan and Basque Language are envolved in the creation of their respective WordNets.

1. Each MTD entry is divided into different database records. Each definition constitues an independent record, and the primary key for each record consists of three fields: word (from the corresponding headword), grammatical category, and definition.
2. All database records preserve, through diverse SGML tags and attributes, their in-origin data from the MTD (and, consequently, from the MRD).
3. All the explicit dictionary information can be searched and accesed by querying the database.
4. The format used in the representation of data will contribute to effective language processing.

This above state for the pLDB is reached by processing a set of PERL programs. Some applications would be able to exploit this pLDB as a language resource. Example (in Figure 4) should make the structure of one of these records clearer.

```
<DB1_RECORD NUM = "862">
<W>abater</W><HOMOG_NUM>0</HOMOG_NUM><MTD_ENTRY>1</MTD_ENTRY>
<GRACAT>v.t.</GRACAT><DEF_LIT>Facer caer, tirar polo chan.</DEF_LIT>
<DEF_EXAMPLE  NUM = "1">Abateuno dun golpe. O vento abateu o valado. </DEF_EXAMPLE>
<SIN>derrubar</SIN>
</DB1_RECORD>
.....
<DB1_RECORD NUM = "865">
<W>abater</W><HOMOG_NUM>0</HOMOG_NUM><MTD_ENTRY>4</MTD_ENTRY>
<GRACAT>v.t.</GRACAT><DEF_LIT>Facer perder [a algu&eacute;n] o &aacute;nimo ou as
forzas.</DEF_LIT>
<DEF_EXAMPLE NUM = "1">Todos aqueles fracasos acabaron por abatelo completamente.
</DEF_EXAMPLE>
<SIN>aplanar, chafar </SIN>
<REXIST>fam.</REXIST>
<CF>deprimir, desanimar</CF>
</DB1_RECORD>

<DB1_RECORD NUM = "866">
<W>abater</W><HOMOG_NUM>0</HOMOG_NUM><MTD_ENTRY>5</MTD_ENTRY>
<GRACAT>v.p.</GRACAT><DEF_LIT>Caer polo chan, vir para abaixo.</DEF_LIT>
<DEF_EXAMPLE NUM = "1">Co temporal abateuse o teito da casa. </DEF_EXAMPLE>
<SIN>derrubarse</SIN>
</DB1_RECORD>
```

Figure 4. Primitive Lexical Database records

Because information in computacional language resources derived from MRDs, is static, we now intend, to make our pLDB more dynamic. In order to achieve this, the database must be completed by taking data from distinct resources and applications. CORGA corpus stores "live" Galician language data. First, we can make the lexical comparison of contents between both resources, the pLDB and CORGA. Since their correspondence is asymetric, many words belonging to the corpus will not be recorded in the database. Second, we could add the word frequency in CORGA to those word ocurrences existing in the database.

The pLDB will play a singular role in the GalWordNet construction (in the absence of another bilingual resource). Therefore synony data included in the database will have to be cosidered. In a manual approach, the WordNet1.5 synsets would be linked to the Galician wordnet by including the corresponding *offset_synset* for each pLDB record (Figure 5). At this point, the pLDB stops being "primitive" becomig instead, a new resource which can be called, simply, Lexical Database (LDB).

```
<DB2_RECORD NUM = "862">
<W>abater</W><HOMOG_NUM>0</HOMOG_NUM><MTD_ENTRY>1</MTD_ENTRY>
<GRACAT>v.t.</GRACAT><DEF_LIT>Facer caer, tirar polo chan.</DEF_LIT>
<DEF_EXAMPLE  NUM = "1">Abateuno dun golpe. O vento abateu o valado. </DEF_EXAMPLE>
<SIN>derrubar</SIN>
<EWN_SYNSET TYPE = V>01558020</EWN_SYNSET>
</DB2_RECORD>
```

Figure 5. A LDB record includes data from the Miller's WordNet.

A second kind of improvement comes through the internal processing of the three original resources (CORGA corpus, the GalWordNet, and the own LDB). LDB can provide grammatical information but, also, semantic information from synonymy data, definition declaration -relations of the type "is_a" and "part_of"-, examples, etc.. Much of the data included in the database, mainly the definitions, can be used for Word Sense Disambiguation (WSD) (Rigau *et al.*, 1997) and will aid the semantic disambiguation of CORGA. Software tools such as lemmatizers and form generators could provide new database fields of morphological information such as inflectional behaviour, derivation, internal composition, etc. Semantic data extraction techniques, used on the definitions in order to obtain the genus term (a word more general than the headword from each definition field) constitute an important step in the WSD procedure. In achieving this, our LDB really becomes a Lexical-Semantic Database.

New LDB records could be added by extracting collocations, idioms and phrases from the corpus. This would involve extending the database model with wider record-structures. Domain-specific knowledge, present in the corpus, would be extracted and, then, added into domain-specific database records.

Validation techniques (Underwood and Navarreta, 1997) must always be taken in consideration in order to verify the correctness of methods and results.

6. Conclusion

In this paper, we have presented an approach for building up various resources for the Galician language. The line which the paper tried to follow is that of integrating lexical and semantic data in a lexical database by using an MRD, a corpus, and WordNet. A method for obtainig an MTD via the extension of an MRD has been carried out through the application of euristic knowledge and simple programs, with a minimum of manual intervention. We introduced the Reference Corpus for the Modern Galician Language (CORGA) and its encoding-scheme. We then took an overview of the initial drawbacks in the construction of the GalWordNet. The potential benefits of a "neutral" Galician lexical-semantic database were put forward. Two channels of procedure are thus: first, the effective bulding up of the GalWordNet which will be carried out as a three year project; second, the internal processing of each language resource developed with the aim of improving and enriching the database. Although these proceedures will undoubtedly take up a considerable amount of time, it is believed that this is the direction to follow not least because of the extensive utilization of standards (which are taken into account at all times).

References

Artale, A., B. Magnini and C. Strapparava (1997). Lexical Discrimination with the Italian Version of WordNet. In Proceedings of the Workshop on Automatic Information Extration and Building of Lexical Semantic Resources. 35th Annual Meeting of the Association for Computational Linguistics (ACL'97), pp. 32-38. Madrid.

Atserias, I., S. Climent, X. Farreres, G. Rigau, H. RodrÌguez (1997). Combining Multiple Methods for the Automatic Construction of Multilingual WordNets. In Proceedings of Recent Advances in Natural Language Processing (RANLP'97).

Calzolari (1991). Lexical Databases and Textual Corpora: Perspectives of Integration for a Lexical Knowledge Base. In Lexical Acquisition, Uri Zernic (Ed.). Hillsdale, NJ: Lawrence Erlbaum Publishers (pp.195).

CES (1996). Corpus Encoding Standard, Document CES 1, Version 1.4, October, 1996. Nancy Ide (Coord.). In http://www.cs.vassar.edu/CES/CES1.html#Contents, Febraury, 1998.

Cole, R., J. Mariani, H. Uszkoreit, G Battista and A. Zampolli (Eds.) (1997). Survey of the State of the Art in Human Language Technology. Studies in Natural Language Processing. Cambridge: Cambridge University Press

DRAG (1997). Diccionario da Real Academia Galega. Real Academia Galega, EdiciÛns Xerais de Galicia, & Editorial Galaxia (Eds.). Vigo: Xerais-Galaxia.

EAGLES (1993). Lexicon Architecture. Expert Advisory Group on Language Engineering Standards. EAGLES Document EAG-CLWG-LEXARCH/B.

EAGLES (1996). Synopsis and Comparision of Morphosintactic Phenomena Encoded in Lexicons and Corpora. A Common Proposal and Applications to European Languages. Expert Advisory Group on Language Engineering Standards. EAGLES Document EAG-CLWG-MORPHSYN/R.

Fred (1998). Fred: The SGML Grammar Builder Project. http://www.oclc.org/fred/docs/about-fred.html; Febraury, 1998.

GÛmez-Hidalgo, J.M. and M. de Buenagana (1997). Integrating a Lexical Database and a Training Collection for Text Categorization. In Proceedings of the Workshop on Automatic Information Extration and Building of Lexical Semantic Resources. 35th Annual Meeting of the Association for Computational Linguistics (ACL'97), pp. 39. Madrid.

Grishman, R. and N. Calzolari (1997). Lexicons. In Survey of the State of the Art in Human Language Technology. Cole, R., J. Mariani, H. Uszkoreit, G Battista and A. Zampolli (Eds.). Cambridge: Cambridge University Press.

McDonald, D.D. (1996). Internal and External Evidence in the Identification and Semantic Categorization of Proper Nouns. In Corpus Proccesing for Lexical Acquisition, B. Boguraev and J. Pustejovsky (Eds.). Cambidge (USA), London (UK): The MIT Press.

Miller, G.A., R. Bechwith, C. Fellbaum, D. Gross, K. Miller, and R. Tengi (1993). Five Papers on WordNet. Tech. Rep. CSL Report 43, Cognitive Science Laboratoy, Princeton University.

Paik, W., E.D. Liddy, E. Yu, M. McKenna (1996). Categorizing and Standardizing Propers Nouns for Efficient Information Retrieval. In Corpus Proccesing for Lexical Acquisition, B. Boguraev and J. Pustejovsky (Eds.). Cambridge (USA), London (UK): The MIT Press.

Rigau, G., J. Atserias and E. Aguirre (1997). Combining Unsupervised Lexical Knowledge Methods for Word Sense Disambiguation. In Proceeding of the 35th Annual Meeting of the Association for Computacional Linguistics, (ACL'97). Madrid.

Underwood, N. and C. Navarreta (1997). A Draft Manual for the Validation of Lexica (Final Report). Center for Sprogteknologi, Denmark.

Vilares, M., J. Graña, A. Pan (1996). Building Friendly Architectures for Tagging. In Proceedings of the Sociedad Española de Procesamiento de Lenguaje Natural (SEPLN'96). pp. 127-132, Seville.

Vossen, P. (1997). Introduction to EuroWordNet, Paper 1, Version 1 (Draft) University of Amsterdam.

Wilks, Y.A., B.M. Slator and L.M. Guthrie (1996). Electronic Words: Dictionaries, Computers, and Meanings. ACL-MIT Press Series in Natural Language Proccesing. Cambridge (USA), London (UK): The MIT Press.