Building an Electronic Version of the Cuban Basic School Dictionary

Iñaki Alegria, Xabier Arregi, Xabier Artola, Mikel Astiz
Faculty of Computer Science
University of the Basque Country
649 p.k. Donostia
xabier.artola@ehu.es

Leonel Ruiz Miyares
Centro de Lingüística Aplicada
Avda. Raúl Pujols s/n, Parque Zoológico, Vista Alegre
90400, Santiago de Cuba

Abstract

This article describes a work done at the Centre for Applied Linguistics (CLA) of Santiago de Cuba in collaboration with the IXA NLP Research Group of the University of the Basque Country, whose main goals have been (1) the conversion of a monolingual dictionary from a text-processing program format into XML, and (2) the design and development of an electronic dictionary.

The article outlines the process of the conversion of the dictionary from its original format into TEI-conformant XML. This semi-automatic process has been completed with a quality control phase performed by lexicographers who have manually checked and corrected the encoding of every entry in the dictionary.

Moreover, the design of a browser-based electronic dictionary application and its implementation are also explained in the article. Currently, the application has been released and distributed through 200 Cuban schools. An on-line version will be soon published at the CLA website.

1 Introduction

The *Diccionario Básico Escolar* (DBE; Miyares 2003) is a pedagogical dictionary for young students. It is made on basis of lexicology research on a corpus of 700,000 words taken from 7,000 Cuban student compositions, a selection of Cuban popular books and children's literature, as well as relevant Cuban newspapers such as *Granma*, *Juventud Rebelde* and *Trabajadores*. The dictionary contains more than 7,000 entries and 14,000 meanings spread over 1,016 pages in its paper version, and has been awarded with the Laurence Urdang International Award from EURALEX for the support of lexicographical research in 2002.

The electronic DBE project[1] is the result of a collaboration of the Centre for Applied Lin-

[1] The project has been possible thanks to the financial support of the Basque Government (FOCAD project no. PRO-2003K2/0007).

guistics of Santiago de Cuba (CLA; http://www.santiago.cu/hosting/linguistica/) and the IXA NLP Research Group of the Faculty of Computer Science of the University of The Basque Country in Saint-Sebastian (http://ixa.si.ehu.es).

The main goals of the project were (1) to devise and to apply a method to semi-automatically convert a conventional dictionary into an XML-based dictionary database, and (2) to develop an electronic dictionary application for secondary and high school students.

This article outlines the conversion process of the dictionary from its original format into an XML-encoded version, and explains the design and implementation of an electronic dictionary application running on this encoding. After this introduction, the original dictionary and its main features are described in section 2. In section 3, the conversion process carried out to take the dictionary from its original RTF[2] format to TEI-conformant XML is outlined. Next, section 4 is devoted to describe the architecture and GUI of the application, along with some implementation issues. Finally, some conclusions are given and future work depicted in section 5.

2 Structure and features of the dictionary

The DBE is a Spanish dictionary that is intended for use by students in secondary level high schools in the age of 11 to 17. One important feature that distinguishes it from ordinary Spanish dictionaries is that it describes Spanish as it is employed in Cuba; so, besides "common" Spanish entries it also contains many specific Cuban Spanish entries.

The following features can be distinguished in the dictionary entries: headword, typical spelling errors (realized as red letters in the headword), pronunciation (for English words used in Cuban Spanish), part-of-speech, geographic, domain and usage style labels, verbs' inflection model, syllabification of the headword, inflected form(s) of the headword, one or more senses containing definition text, example sentences, often with labels referring to the part-of-speech used in the sentence or other usage notes, and with references to the headword emphasized in bold and underlined meta-linguistic references, synonyms, antonyms, and similar words, usage notes and labels, attached to the definition or to particular examples, and, finally, related subentries, such as locutions and other noun or verb phrases.

The DBE contains 7473 entries including 14013 word senses. Attached to the entries you can find 1380 diminutives, along with plural forms of nouns, feminine and plurals of adjectives, participles of verbs, etc. Regarding lexical relationships, 3601 synonyms, 474 antonyms and 75 similar words can be found in the dictionary, allowing to navigate between related entries. Moreover, 1062 locutions, 651 phraseologisms and 39 sayings are defined and exemplified as subentries of the main entries.

The dictionary was originally typed at the CLA and stored in 27 files in RTF, one per letter. The following sample entries might make clearer the usage of the different fields and features:

[2] Rich Text Format.

cerca sf. Valla, tapia o muro generalmente de alambre, estacas o piedras que se pone alrededor de cualquier terreno para resguardarlo o limitarlo. *Pusieron una* ***cerca*** *alrededor del terreno deportivo para evitar que penetren intrusos.*

cer-ca; cercas (pl.); cerquita (dim.)

cerca adv. 1. Próximamente, inmediatamente, a corta distancia. *Mi casa se encuentra cerca de la escuela. Vamos al cine a pie, pues queda* ***cerca****.* Ant. lejos. // loc. adv. **cerca de**. Aproximadamente, más o menos, casi. *Mi abuela está saludable, aunque tiene* ***cerca de*** *noventa años.*

cer-ca; cerquita (dim.)

fábrica sf. 1. Establecimiento o edificio donde se fabrica algo, en el que existen equipos, máquinas, herramientas, etc., necesarios para producir determinado tipo de objetos. *Ibia comenzó a trabajar en una* ***fábrica*** *de zapatos.* Sin. industria, factoría, empresa. 2. fig. Acción de construir o producir algo. *La colmena es la* ***fábrica*** *donde se elabora la miel.*

fá-bri-ca; fábricas (pl.)

3 Conversion process: from RTF to XML

The use of XML to represent the dictionary knowledge in a structured way allowing to explicitly mark-up the different fields in the entries means a radical change with respect to the original RTF version, offering both the lexicographers and the users more and richer possibilities to later on search the information in the dictionary.

The first step, before the conversion, is to define the XML language to encode the dictionary, i.e. the data structure into which the dictionary must be transformed from its original format. We followed the TEI guidelines (2001) for that, adopting a subset of the TEI DTD for dictionaries and adding to it some elements and attributes to deal with the unforeseen features.

The conversion of the dictionary from RTF to XML consisted of three consecutive phases:

- Pre-processing of the original documents in order to get an unambiguous, "normalized" and consistently *edited RTF version* of the dictionary. *Word* macros have been used in this phase.
- Conversion of the edited RTF version into a *preliminary XML version* using a tool named *Ferret* (Patrick *et al.* 2002), whose goal is to semi-automatically learn the structure of entries, based mainly on typographical features, and to encode them into (in our case, preliminary) XML.
- Post-processing of the preliminary XML version in order to correct encoding errors, so getting the *final XML version* of the DBE. XSLT scripting has been used in this phase.

Finally, a human quality control phase has been carried out to guarantee that the final encoding of the dictionary is completely correct. The entries were manually revised at the application prototype, being their encoding manually corrected when necessary and the problems found in its rendering reported. A total of 984 entries (13%) were corrected as a result of this quality control.

4 Electronic DBE: GUI, architecture, and information flow.

The main goals of the electronic version of the DBE have been (1) to provide the students with a useful and modern dictionary tool for the learning and practice of the language, (2) to build first a CD version of the dictionary, and then to make it available on-line in the web, and (3) to make use of XML technology as a basis for the storage and exploitation of the dictionary.

The functionality we wanted for the application included:

- First letter and index-based browsing facilities.
- Normal search or lookup, with closest match help.
- Advanced search: filtering of entries based on selected parts-of-speech.
- Hyperlinking facilities: cross-references between related entries, synonyms, antonyms, etc.
- Orthographic help, based on purposely encoded misspelling feasibility.
- Accessibility of verb paradigms and illustrations directly from the entries.
- Some statistics on the contents of the dictionary and help.

The application has been developed as a web application. The user needs just an ordinary web browser where the GUI of the application is shown. For the CD version, the web server is embedded into the application.

4.1 Graphical User Interface

The graphical user interface (GUI) of the application consists of several HTML pages.

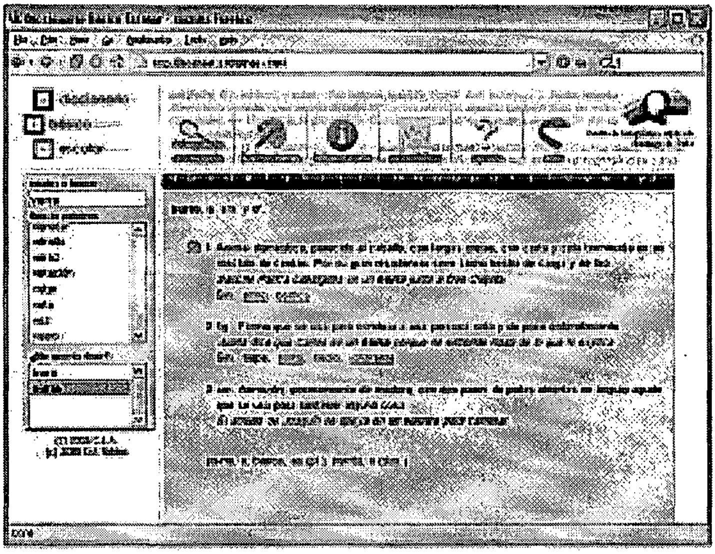


Figure 1. Graphical User Interface of the electronic DBE.

In Figure 1 a screenshot of the main window of the GUI is provided. As can be seen, the GUI contains the following elements: a menu bar that gives access to the different functionalities of the application (statistics, illustrations, help, advanced search, etc.), an alphabet bar for browsing the dictionary based on the first letter, an input textbox or search frame, a headword selection listbox, a suggestion listbox, and a main area for displaying the entries. In some entries, standard-look hyperlinks are displayed allowing the user to navigate from there to other related entries, such as synonyms, antonyms, and so on. Moreover, in advanced search mode a special checkbox bar allows the user to select the parts-of-speech wanted to be used as a filter when filling up the headword selection box.

The interface developed meets both the wishes of the members of the CLA and the requirements of the application, especially regarding the target user group analysis.

4.2 Client-server architecture and information flow.

The original dictionary was separated into 27 files corresponding to the 27 letters of the Spanish alphabet. After the conversion into XML the dictionary still consists of 27 separate documents located on the server, which constitute the basis of the application. These documents are encrypted in the CD version. Moreover, indexes (one per letter, containing the headword and the different parts-of-speech it belongs to in any of its senses), 80 verb paradigms, and 685 illustrations can be found as well on the server side of the application.

As mentioned above, the electronic DBE has been developed as a web application, where the server serves XML content (entries, indexes and verb paradigms) in response to requests made by the user through the client GUI. XSLT stylesheets are used on the server to search the information requested and fetch it from the dictionary. Moreover, they are also used, in combination with CSS stylesheets, to convert the XML data into HTML before sending them to the browser.

One important remark regarding the underlying infrastructure must be made here. On the one hand, when searching or browsing the dictionary, both in normal and advanced search modes, just the index documents are downloaded in a first moment, in order to populate the headword selection listbox. Filtered index documents are served when in advanced search mode, or when proposing orthographical corrections to the user (in this case to fill up the suggestion listbox). The biggest index file (corresponding to the *c* letter) is less than 100 kilobytes in size, so it is not very costly, in terms of time and space, to completely download one of these each time it is required (with the proviso that, if it is already in the browser's cache, there is no need to download it again). On the other hand, dictionary information is served on a per-entry basis, meaning that, when a particular entry is requested, just the XML element encoding it is fetched from the corresponding document.

Index documents are fetched from the server, filtered if necessary, and rendered at the browser as HTML option elements in the headword selection listbox's select. In the case of entries and verb paradigms, the server must extract the information (the entry or superentry, or the verb paradigm) from the corresponding document.

On the client side, the application reacts to the user's actions and to the events occurring when using the dictionary, submitting the requests to the server and displaying the information received from it. This is implemented using JavaScript.

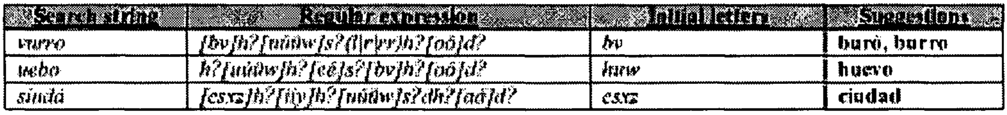
4.2.1 Correction of typical spelling errors

If no exact match is found in the index downloaded (once the user has ended typing), a regular expression representing the possible correct spelling intended by the user is built. This task is based on orthographic criteria (no typing errors are corrected in this version of the dictionary) issued from previous research carried out at the CLA in the area of patterns and likelihood of spelling errors (Miyares 1996).

Two phases can be distinguished here. The first one, the construction of the regular expression, is performed on the client. The regular expression is then submitted to the server along with the possible starting letters of the correctly spelled candidates and the list of POS selected (in the case of advanced search). The function invoked on the server examines the appropriate index documents, filtered according to the list of POS provided, matching the entry headwords against the regular expression. The server processes the index documents, and builds a list of candidates that are then proposed to the user as spelling suggestions.

The construction of the regular expression is based on a set of "clusters" of typical spelling errors. We call cluster in this context to a set of letters and/or pairs of letters that can be misused at a particular position of the word. These clusters can be used to replace all the letters of the search string that are part of one of them by the other letters in the same cluster, so creating all possible combinations of the search string. The clusters that have been implemented in the current version include the following cases, among others: *b/v* confusion, improper use or omission of *h* before or between vowels, misuse or omission of accent on vowels, incorrect substitution of sibilant consonants (*s*, *c*, *z*, *x*), confusion of *y* and *ll*, misuse of *n* instead of *m* before *p* and *b*, etc. So, when building the regular expressions, every letter or pair of letters in the search string that is member of one of the clusters is replaced by an expression representing the other elements in the same cluster. For example, every occurrence of *b* or *v* in the search string will be replaced in the regular expression by *[bv]*, meaning "*b* or *v* can be matched at this position of the word".

In Figure 2, you can see the regular expressions constructed after applying these operations to the search strings *vurro*, *uevo*, and *siudá*, along with the initial letters indicating, in each case, the index files that the server must check to build the list of candidates, and the candidates finally suggested:



Search string	Regular expression	Initial letters	Suggestions
vurro	*[bv]h?[uúüw]s?(l|r|rr)h?[oó]d?*	*bv*	**buró, burro**
uebo	*h?[uúüw]h?[eé]s?[bv]h?[oó]d?*	*huw*	**huevo**
siudá	*[csxz]h?[iíy]h?[uúüw]s?dh?[aá]d?*	*csxz*	**ciudad**

Figure 2. Building regular expressions to correct spelling.

As can be seen, only one regular expression is needed for one search string, and the matching operation of this regular expression against all the headwords in the possible index files is very efficient. Indeed, much more efficient and "intelligent" than applying typical spelling correction algorithms, such as the replacement of letters, the transposition of a pair of contiguous letters, and so on.

We think that this approach suits better the didactic character of the electronic DBE, as regards its aim as a tool to help in the learning of the language. Other features already mentioned, such as the use of red letters in the headwords to indicate the likelihood of orthographical errors, were also conceived, already for the printed version, with the same idea in mind.

5 Conclusions and future work

In this project, a methodology for the semi-automatic conversion of a dictionary from RTF format into XML has been devised and conducted. As a result, the DBE is now a real lexicographical database where the information is represented and structured in a suitable way, encoded accordingly to a well-recognized standard such as the TEI guidelines.

Moreover, an electronic dictionary application has been designed and developed. A CD-ROM version has been distributed at the Cuban schools and is already being used, whereas the on-line version will be available pretty soon. We would like to emphasize here two aspects: on the one hand, the architecture of the application that has been developed as a web application, even for the CD version; on the other, its student-orientation, which follows the research conducted at the CLA over the years on orthography and other language learning aspects (Ruiz & Miyares 1999).

Another point to highlight is the collaboration between the two partners, on the one hand the CLA, involved in applied linguistics for more than 30 years, and the IXA NLP Research Group.

Short-term future work includes the development and setting up of the on-line version of the dictionary at the CLA website. At a longer term, the production of a second version of the electronic DBE is envisaged. This version should include, among other features, full lemmatization of definitions and example texts, in such a way that possibly dynamic hyperlinks over every single word occurring in these texts would be made feasible. Finally, a dictionary editing environment (Alegria *et al.* 2006) has been designed and an operative prototype has already been implemented.

References

A. Dictionaries

Miyares Bermúdez, E. (dir.) (2003), *Diccionario Básico Escolar.* Centro de Lingüística Aplicada, Santiago de Cuba (Cuba).

B. Other Literature

Alegria, I., Arregi, X., Artola, X., Astiz, M., Ruiz Miyares, L. (2006), 'A Dictionary Content Management System'. *Proceedings of EURALEX 2006. Turin* (Italy).

Carroll (eds.), *Traitement automatique des langues. Les dictionnaires électroniques*, vol. 44-2. ATALA (Association pour le Traitement Automatique des Langues), Paris (France), pp. 107-124.

Miyares Bermúdez, E. (dir.) (1996), *Léxico Activo Funcional del Escolar Cubano*, Centro de Lingüística Aplicada, Santiago de Cuba (Cuba). (unpublished.)

Patrick, J., Palko, D., Munro, R., Zappagina, M. (2002), 'User driven example-based training for creating lexical knowledgebases'. *Proceedings of the 2002 Australasian Natural Language Processing Workshop.* Canberra (Australia).

Ruiz Hernández, J. V., Miyares Bermúdez, E. (1999), *Vacuna Ortográfica VAL-CUBA. Metodología* *para prevenir y erradicar las faltas de ortografía (Nivel Primario).* Editorial Academia, La Habana (Cuba), tercera edici6n.

Text Encoding Initiative Consortium (2001), *The XML version of the guidelines for Electronic Text Encoding and Interchange* (http://www.tei-c.org/).