

An Architecture for a Federation of Heterogeneous Lexical & Dictionary Databases

Jon Patrick², Jun Zhang²
Xabier Artola-Zubillaga¹

Jon Patrick, *Department of Information Systems, Massey University, Palmerston North, New Zealand.*

AFFILIATION: ² Information Systems Dept. Massey Univ. New Zealand. ¹ Lengoia eta Sistema Informatikoak, Euskal Herriko Unibertsitatea (UPV/EHU), Basque Country.

E-MAIL: ² J.D.Patrick@massey.ac.nz
¹ jiparzux@si.ehu.es

FAX NUMBER: +64 6 350 5611

PHONE NUMBER: +64 6 350 5552

1. Introduction

Dictionaries in electronic form are rapidly emerging, although at the moment they more closely mimic their paper predecessor than exploit the added advantages available through computer devices. In particular, extensive cross-referencing and the rapid access of material relevant to a user's needs can now be delivered at a reasonable cost. However, as user familiarity and competence grows, so does the demand for more elaborate and sophisticated access. It is now possible to acknowledge the need to support and deliver elaborate information systems that enable a user to access multiple dictionaries of many types and languages in a single software environment, much in the same way as that provided by Mosaic on the World Wide Web. Such an ambitious project needs an upper level design that caters for the wide diversity of needs for potential users across a variety of hardware platforms and software environments with broad geographical dispersion. We define the requirements of this project as having four broad categories which in turn can be broken into smaller tasks, many of which can be tackled independently. These categories are:

(1) the conversion of paper dictionaries to an electronic form in a way that provides for the automatic extraction of their implicit structure and for converting the information into lexical databases (Boguraev & Briscoe eds., 1989),

(2) to automate the production of different database structures as required for different computational requirements, for example, lexical databases vis-a-vis dictionary databases vis-a-vis multi-lingual databases whilst at the same time fulfilling the requirements of category (1) above.

(3) to develop techniques for the generation and integration of conventional dictionaries with multi-media dictionaries and encyclopaedic information.

(4) the provision of inter-connecting semantic links between the lexical units of an unlimited number of electronic dictionaries and similar documents to enable browsing and search trails across the various documents

This paper has the aims of:

(i) defining an overarching architecture to support the requirements of categories 1-4 above,

(ii) to introduce a generalized dictionary query language,

(iii) to introduce a new type of database organization for mono-, cross- and multi-lingual dictionaries, that allows a relaxation of demands for a strict physical organisation of the data so as to facilitate federation,

(iv) to report on the implementation of the new organization in (iii) above for Basque-English and English-Basque cross-lingual dictionaries.

2. Description of some of the problems

Electronic versions of dictionaries for human use are emerging everywhere. The search for new ways of representing and retrieving dictionary knowledge represents a challenge. There is a large distance between an electronic dictionary that is merely a word-processor formatted text file and an intelligent dictionary help system such as that described in (Agirre et al., 1993, 1994), for example, where the user is allowed, among other sophisticated access functions, to retrieve words based on semantic constraints formulated in a specially conceived language.

Moreover, lexical databases (LDB) for natural language processing (NLP) share many of the features that dictionaries have, but they also need many other kinds of information that are necessary when the goal is the automatic treatment of the language.

To gather these two kinds of resources into a single system is desirable:

NLP needs real-size LDB's that have to be built with computer support. Much research has been carried out in the last few years (Boguraev & Briscoe eds., 1989) in the sense of taking advantage of conventional dictionaries in machine-readable form to extract from them information needed for NLP tasks. Hence, computational linguists need on-line access to dictionaries when working on their LDB's.

Lexicographers can take advantage of information coded in LDB's when doing their work: much more formal and systematic ways are used in coding the information for NLP than when com-

piling information for the human reader. In addition, multi-lingual societies require bilingual cross-lingual and multi-lingual dictionaries. The language industry needs multi-lingual tools (online dictionaries, writing aids, machine-assisted translation tools, etc.), that obviously require the use of cross- and multi-lingual databases. We define a cross-lingual dictionary as one which provides a match between words and terms in one language with those in another language, whereas a multi-lingual dictionary has a head word in one language and the definendum in one or more other languages.

Therefore, the problem posed in this paper is related mainly to lexical and dictionary knowledge representation and retrieval issues, for either NLP applications or human users, and takes into account the need to deal with a great diversity of electronic lexical sources. The solution proposed is intended to gather, in a single integrated computer architecture, all the aspects mentioned above as well as some initial work on practical solutions.

3. Variety and complexity of source materials

Currently, a diversity of lexical resources can be found in electronic form, including Roget's Thesaurus and different dictionaries (monolingual and/or bilingual) of different languages and of different types (explanatory dictionaries with definitions, synonym dictionaries, thesauri, etc.). Moreover, lexical databases developed mainly for NLP, which contain information needed for the automatic treatment of different languages, are also available.

Obviously, this variety of sources contains a great diversity and complexity in formats in which this information is actually stored from plain texts, marked up or not, to conventional database management systems or even sophisticated knowledge-based systems. This high diversity of formats or representations provides for very differentiated levels of retrieval functionality. Furthermore, the potential for candidate dictionaries to be in geographically disparate locations must be considered. The need to provide for reusability of these sources is evident.

Therefore, the requirements of a system that would serve as a basis for all these sources include a flexible architecture which provides a Generalized Lexical and Dictionary Description Language (GLDDL) along with a powerful query language for retrieval.

4. Architecture of the system

The architecture proposed in this paper aims to provide a solution to the problem of gathering such a diversity of lexical and dictionary sources into a well-integrated system. Its most important

characteristics are the integration of all the different sources or tools in a single federation of databases (the term database is here employed in its widest sense), and the conception of a generalized lexical and dictionary description language that would provide a platform for the exchange of information between the databases and the users. Two aspects are discussed here: the federation itself, and the query language.

A Federation of Heterogeneous Lexical & Dictionary Databases (FHLDDDB). It is not conceivable to compel the providers of the different lexical and dictionary sources to convert their information bases to a single standard, and it is actually unnecessary and indeed unworkable. Our proposal aims to accept any source in any format, and integrate them in a so-called "federation of heterogeneous lexical and dictionary databases". We distinguish here the term Lexical Database (LDB), that stands for those databases built as support for NLP tasks and, so, source of a variety of computational lexica, from the term Dictionary Database (DDB), that stands for the computer implementations of dictionaries for human use, be they explanatory dictionaries, synonym dictionaries, thesauri, bilingual dictionaries, etc. The structure of each database in the federation has to be described by means of the generalized lexical and dictionary description language mentioned above.

Universal Query Language: to design a suitable query language that, based on the common lexical description language, will allow the end user, either human or program, to communicate with the federated system.

5. A Generalized Lexical & Dictionary Description Language (GLDDL)

A GLDDL is a common description language for lexical and dictionary knowledge that, placed on top of the FHLDDDB, will facilitate the exchange of information between the physical information stores and the end user.

To that end, some pilot studies on real dictionaries and lexical databases, such as the huge one built in Japan (EDR, 93), must be considered. Surveys of lexicographers could also be considered. Moreover, the standards being proposed for the representation of lexical and dictionary knowledge in different projects recently finished or still ongoing must be analysed, e.g. as *Acquilex I* (Common Lexical Entry) and *II*, *Genelex*, *Comlex*, *Multilex*, *Eagles*, *Cambridge Language Survey*, etc. (Copestake, 1992; Sperberg-McQueen & Burnard eds., 1994).

The TEI Guidelines have been drawn up through a lengthy study of these projects and arrived at an integration of most of the aforementioned studies. Although the TEI guidelines were drawn up for essentially encoding paper dictionaries, we propo-

se that the TEI standard for Feature Structures encoding be the first version of the GLDDL for terms and their definitions. We expect that in time and with more experience this version will be modified.

5.1. A Representation Formalism for a GLDDL

A GLDDL must be implemented with a general representation formalism. Feature Structures (FS) as definable in an Object-Oriented DBMS are general purpose data structures that have been used in many systems to encode linguistic information. There exists a well-developed theoretical framework for them, and their applicability to encode the information found in dictionaries, or in lexical databases for NLP, is quite natural (Ide et al., 93).

We present here a different proposal to feature structures and other approaches. We start from the point that a TEI-like standard supplies a comprehensive set of terminology and operationalisable definitions and criteria for us to describe entries in a lexical or dictionary database. One of the limitations in describing a document as complex and diverse as a dictionary is the rich variety of organisational structures used by lexicographers from entry to entry. However, such diversity almost always has its own idiosyncratic organization that enables a set of parse rules to be defined for the structures of the greater part of the dictionary.

We advocate that the dictionary entries' structures should not be dismantled just to pack them into a formal database schema whether it be relational or feature-structured as each have their own limitations. Rather we propose to keep the original entries intact as a single one field string in a database record just as they appear in a paper dictionary. However, in our system we parse the entries into their component structures and store in a separate field in the database record the parse node number (equivalent to its feature name) for each structure and pointers into the record where each feature commences. This strategy requires a parser to be written for every dictionary and the parse structure states to be identified in terms of the LDDL. The resulting database organization we call a Parse State Structure (PSS).

This strategy has a number of benefits over previous solutions to lexical and dictionary database designs.

- (i) The data structure is indefinitely recursive and hence more flexible,
- (ii) It is not possible to waste space due to empty schema fields in the database records.
- (iii) The particular schema for a database does not need to be learnt for either human or machine usage,
- (iv) The database is directly portable between

DBMSs without restructuring although indexing mechanisms will need to be recompiled,

(v) A TEI marked up version of the database can be automatically generated at any time as it is a structural subset of the parse tree and parse states map directly to the TEI nomenclature in the LDDL.

5.2. Generalized Lexical & Dictionary Query Language (GLDQL)

When a query from the user is presented to the system, it must be addressed to one or several units in the FHLDDDB. These queries, expressed in, say, SQL or Object-SQL for units resident in relational or object-oriented databases, or in specific query languages designed for the specific lexical stores, must then be broadcast to the retrieval modules of each unit. Once the answers are obtained in a variety of forms, they must be translated and possibly unified, into the GLDDL, in order to provide a unified answer. A translation process must take place when translating an answer from the system into the answer ultimately presented to the user. The PSS offers a number of advantages for querying a single database or a federation of databases. The basic form of a Generalized Lexical & Dictionary Query Language (GLDQL), which mimics SQL is:

```
SELECT <select_list>
FROM <dictionary_list>
WHERE <predicate>
```

where *select_list* is the list of structural elements (features, TEI attributes) required to be extracted from each entry; *dictionary_list* is a list of dictionaries to be searched; and, *predicate* is the set of restriction conditions as formed by SQL-like logical expressions that is applied to any features in an entry.

If such a query is broadcast to a federation of databases it would have to be mapped to the DBMS structure and data definition terminology for each database that had to respond to the request. A mapping scheme has been defined to operate with the SQL of the Oracle DBMS.

6. Work to date

We have stored the entries for each of a Basque-English (Aulestia, 1990) and an English-Basque (Aulestia & White, 1990) dictionary in their own Oracle Databases. Implementing the PSS inside a standard DBMS – like Oracle gives advantages such as, all querying functionalities provided by Oracle can be used to provide the intermediate results, for example, wild card matching in conditions; database security is automatically handled and all user interface software available with the DBMS is readily accessible.

Included in Table 1 are some statistics for the two

dictionaries that we have processed. The source files were created from an OCR process from the paper dictionaries and therefore contained errors that have not yet been corrected. The parsing hit rate should therefore increase significantly when the erroneous records are corrected and resubmitted to the parser. The figures show that a source file of 4 Kbytes and 43,545 entries can be loaded in 60 minutes into an Oracle DBMS running on a Sparc Server 1000. Although this is a long period for a small file it represents only an initial overhead for creating the database and a large variety of indices on many attributes. This also explains the 10-fold increase in the size of the Oracle database compared to the source file. However, once the database is created the efficiency of the data representation is demonstrated in that the whole dictionary can be written to a TEI conformant file in 4 minutes. We also have found that response to queries over a large number of fields or wild card patterns in the definendum field is very nearly immediate. Individual entries can be requested in a Marked-up TEI conformant format. To test our claims of transportability a sample component of the databases has been provided to another site and readily installed without problem.

	Aulestia Basque-English	Aulestia & White English-Basque
Source File Size (ASCII text)	3,966 Kb	1,666 Kb
Parsable Data File Size	3,353 Kb	1,473 Kb
Parsable Entry Number	43,545	22,663
Parsing Hit Ratio	90.5%	92.7%
Space Used in Database	44,520 Kb	25,780 Kb
Time Consumed to Load	60 minutes	40 minutes
Time to Create TEI File	4 minutes	3 minutes
TEI File Size	10,275 Kb	7,386 Kb

Table 1. Processing specifications for 2 dictionary databases organised in a Parse State Structure mounted on a SPARC server 1000.

Although we have tried to make our model independent of a specific dictionary, some attributes with a closed set of values, e.g. part of speech (pos), still need to be mapped to the standard nomenclature of a local dictionary. Therefore a mapping table is necessary for each dictionary so

that comparison of the value sets of the same feature across different dictionaries will be possible. For example, in one dictionary the intransitive verb is abbreviated as 'vi' while in another dictionary it is 'v.i.', hence the condition D1.pos = D2.pos has to be checked based on the mapping tables provided by each dictionary.

The development of this project requires us to complete a federation of a variety of dictionary sources. Currently we have the following list of dictionaries in electronic form, whose structure is represented in a TEI(FS)-like way:

- EDBL (Lexical Database of Basque for NLP applications, stored in a relational database),
- LPPL (Le Plus Petit Larousse, ordinary dictionary of French, stored in a relational database),
- HLEH (Hauta-Lanerako Euskal Hiztegia, dictionary of Basque automatically parsed and encoded following the TEI guidelines for dictionaries).

7. Conclusion

This design accomplishes:

- a higher level of architectural descriptions of a federation of lexical and dictionary databases,
- a unique proposal for parallel retrieval with integrated response to the user,
- and a description of an implementation of a Lexical & Dictionary Definition Language (LDDL) using a Parse State Structure (PSS), that is more useful than a mark-up language or a fixed database schema for cross communicating between different platforms which support a Generalized Lexical & Dictionary Query Language (GLDQL).

Our architecture will support a variety of classes of users including the general public, translators, lexicographers, etc. and it provides for universal access to electronic dictionaries without attempting to enforce the dictionary publisher to convert existing paper or electronic dictionaries into a particular schema structure or DBMS platform, hence lowering co-operation and general accessibility.

Finally, it provides for a database architecture suitable for readily coalescing and cross-referencing a variety of source materials, including mono-, cross- and multi-lingual dictionaries, thesauri, phrase books, encyclopaedic information, etc.

8. References

- Agirre E., Arregi X., Artola X., Diaz de Ilarraza A., Evrard F., Sarasola K. Intelligent Dictionary Help System. Proc. 9th Symposium on Languages for special Purposes. Bergen (Norway), 1993.

- Agirre E., Arregi X., Artola X., Diaz de Ilarraza A., Sarasola, K. Lexical Knowledge Representation in an Intelligent Dictionary Help System. Proc. of COLING'94, 544–550. Kyoto (Japan), 1994.
- Aulestia G. Basque-English Dictionary, University of Nevada Press, Reno, 1990.
- Aulestia G. & White L. English-Basque Dictionary, University of Nevada Press, Reno, 1990.
- Boguraev B., Briscoe T. eds., Computational Lexicography for Natural Language Processing. New York: Longman, 1989.
- Copestake A. The ACQUILEX LKB: representation issues in semi-automatic acquisition of large lexicons, Proceedings 3rd. Conference on Applied Natural Language Processing (Trento, Italia), 88–95. 1992.
- EDR Electronic Dictionary Technical Guide. Japan Electronic Dictionary Research Institute, Ltd. TR-042, 1993.
- Ide N., Le Maitre J., Veronis J. Outline of a Model for Lexical Databases. Information Processing and Management, vol. 29, no. 2, pp. 159–186, 1993.
- Sperberg-McQueen C.M., Burnard L. eds., Guidelines for Electronic Text Encoding and Interchange. Chicago, Oxford: TEI P3 Text Encoding Initiative, 1994.