ACQUILEX

COMPUTATIONAL MODEL OF THE DICTIONARY ENTRY

Final Report

Carol Peters*, Adriana Roventini**

*Istituto di Elaborazione della Informazione, CNR, Pisa, Italy
**Istituto di Linguistica Computazionale, CNR, Pisa, Italy

ACQUILEX

ESPRIT BASIC RESEARCH ACTION No. 3030
Addendum to Deliverable No.1

WORKING PAPER No. 55

Pisa, Giugno 1992

ILC-ACQ-5-92

2. Encoding Dictionary Data

Printed dictionary entries are highly structured and very complex pieces of text; the information content depends to a large extent on features of layout and the use of a metalanguage, which is generally (but not always completely) explained in the Introduction and User Notes at the beginning of the dictionary. The entry thus contains a mixture of explicit and implicit information interspersed with codes for type-setting and formatting. However, in a computational model of the entry, the lexical information must be extracted from the rest, interpreted and represented explicitly. At the same time, it is useful to maintain a trace of the entry structure, represented in the printed version by means of particular typographic conventions.

2.1 Representing the Project Machine-Readable Dictionaries

The first step towards defining a model for a Common Lexical Entry for ACQUILEX was a detailed analysis of the strucure and content of the separate machine-readable dictionaries to be adopted in the Project. The definition of a uniform and explicit representation language that could be used to describe both the textual and lexical content of these dictionaries was considered essential for the following reasons:

- to permit the exchange of data between project members or with other projects in a common format;
- to facilitate uniform analyses over different dictionaries;
- to write generalized parsers for different dictionaries;
- to facilitate the standardization of the contents of the data fields.

The first part of the preliminary report thus gave a detailed and explicit representation of the machine-readable dictionaries to be used in the project. The dictionaries examined were: *Il Nuovo Dizionario Garzanti*, a monolingual dictionary for Italian, the *Collins Concise English-Italian, Italian-English Dictionary*, the *VanDale Groot Woordenboek Hedendaags Nederlands,* a monolingual Dutch dictionary, the *VanDale Bilingual Dutch/English Dictionary*, the *Longman Dictionary of Contemporary English*, the *Oxford Advanced Learner's Dictionary of Current English,* the *VOX Spanish Monolingual Dictionary*. These dictionaries are referred to

1. Introduction

The scope of the preliminary report on the Computational Model of the Dictionary Entry for the ACQUILEX Project, produced at 6 months, was to produce a working definition of a general representation of a lexical entry, capable of handling not only the dictionaries actually in use in the Project but also the other MRD sources which might be made available during the project life-time. The report thus (i) provided a uniform representation of the content and structure of the mono- and bilingual machine-readable dictionaries to be used within the ACQUILEX Project and (ii) used this first analysis to define a Common Lexical Entry (CLE) onto which all the Project Dictionaries could be mapped; a brief specification of the database model adopted for the project was also given.

The aim of the final report is to evaluate the representation on the basis of the experience acquired during the course of the Project, and to present the definitive version of the ACQUILEX Common Lexical Entry.[1] The report will be brief; it has been found necessary to make only a few significant changes to the first proposal. In fact, after the initial stage in which the project dictionaries were mapped onto the representation structure described and into the project databases (see Carroll, 1990; Marinai et al, 1990), the main activity of the project has been concentrated on the extraction of lexical-semantic information from the lexical databases and the construction of a common Multilingual Lexical Knowledge Base (see Copestake, 1991). The report thus summarizes briefly the motivation and the results of the preliminary study (Section 2), describes the first applications of the Lexical Entry and feedback from other projects (Section 3), illustrates the changes which have been made to the CLE presenting the final proposal (Section 4), and discusses the compatibility of our representation with others, i.e. that of the Text Encoding Initiative (Section 5).

---

[1] The final report is to be considered an addendum to the preliminary study; for full details on the semantics of the representation language adopted and for examples of the possible values that can be taken by the attribute tags of the CLE, the reader is referred back to Calzolari et al., (1990).

as Garzanti, Collins, VanDale Monolingual, VanDale Bilingual, LDOCE, OALD and VOX, respectively.

It is well known that different dictionaries do not always present the same type of lexical information, neither do they necessarily present the information in the same way. A typical lexical entry, in its most basic configuration, is formed of a lemma with which phonetic, syntactic and semantic information is associated according to a widely varying pattern. In fact, the lexical entry is a particularly complex and difficult "object" to represent by means of a model because of the many different ways in which the information it contains can be combined.

In our preliminary proposal of a Computational Lexical Entry, our principal goal was to identify and represent all the possible types of linguistic information which could be embedded within each field in the entry. We thus had to define a language which was not only explicit but could also represent the hierarchical structure of the entry and the potential relations among its various components. The representation consisted of three parts: specific lexical entry templates for each dictionary; a description of the semantics of the tags used in the templates; an exemplification of the set of possible values for each Attribute.

The Lexical Entry Templates were intended to permit a representation of the internal organization of the "maximal" entry in each dictionary, i.e. all possible fields should be foreseen in all possible positions, no information should be excluded. The hierarchical structure of the dictionary entry was evidenced by the use of Group_tags, gathering semantically and logically connected constituents (Attribute_tags). Tokens were used to show the obligatority vs. optionality of the elements within the entry. In order to represent dependencies and scope in this first representation, we used indentation in the Template structure. The Template thus made transparent the hierarchical structure of the entry and the relations among its components and provided a means to represent all the information contained in the text of the entry.

A considerable effort was necessary by all the project partners in order to analyse exhaustively the structure and information content

of their dictionaries so that the standard representation language could be defined and templates could be produced for each dictionary using this language. This first "lower" level of standardization imposed sufficient uniformity to be a useful basis for the exchange of MR dictionary documents and was an essential basis for the subsequent development of the Common Entry into which the data from any dictionary could be mapped.

For the list of all the Tags used, a detailed description of their semantics, and a list of the possible values for the Attribute Tags of each different dictionary, the reader is referred to Part One of the preliminary report (Calzolari et al, 1990).

During this stage one of the problems was to reconcile two, to some extent conflicting, goals: to preserve the original document in its integrity while, at the same time, analyzing its contents in all possible detail. For this reason, the template included two special Attributes: '_text' and '_type' attributes. Attributes whose names end with the string '_text' are used to preserve the data as it is represented in the source text, when the process of parsing the dictionary into the different fields otherwise results in the source text being no longer recoverable, or when a portion of the source text has not yet been analyzed. (The maintenance of the source text was considered important as, for example, the requirements of further and more refined analyses may necessitate going back to the source.) Whereas, attributes whose names end with the string '_type' were used for information which was not explicitly present in the printed dictionaries, but which could be derived at an early stage by relatively simple automatic or semi-automatic procedures.

2.2 The Project Database Model

The underlying computational model used to structure dictionary data in lexical database form is crucial as it conditions the kinds of access and query functions which are possible on the data. The two-level hierarchical-tagged model, proposed by Boguraev et al (1990) and adopted by ACQUILEX, combines the perspectives of both the tagged and hierarchical representations. In fact, although the tagged dictionary model preserves all of the information contained in the entry, explicitly labelling each information field to facilitate

successive retrieval, it does not satisfactorily represent the structural relationships which are implicit in the source. The hierarchical model is considered far more appropriate to represent the lexical information encoded in the typical dictionary entry as it permits the clustering and nesting of data at different levels; lexical entries can be characterised as shallow hierarchies of attribute-value pairs with a variable number of instances of specific nodes at each level, e.g. multiple homographs within an entry or multiple senses within a homograph (see Neff et al., 1987). In the two-level model of the lexicon, the source dictionary is seen as the primary repository of lexical data and the database system automatically generates from it subsidiary sets of indices which encode all the information that it has been decided to represent in the LDB (see Carroll, 1990).

2.3 The Common Lexical Entry

Once the different MRDs used in the Project had been analysed and described using a common and simple representation language, the next step was to design a Common Lexical Entry (CLE) onto which the different dictionaries could be mapped. The Common Lexical Entry has the following characteristics:
- it no longer reflects the linear order of the text of the single project dictionaries but can still contain all the information found in the source, even if this has been rearranged or is represented in a new way;
- it contains more types of "derived" information, i.e. information which is not explicitly present in the source dictionary, but which can be extracted by processing the data;
- it is not bound and constrained by the necessity of compacting the data and saving space as is the printed medium;
- it can be easily expanded and updated.

The Common Lexical Entry was represented using the same formalism as that adopted for the individual dictionary templates; the CLE can, in fact, be considered to some extent as the "union" of the idiosyncratic entries since one of the guiding criterion was that no information from any dictionary should be excluded. However, as the structure chosen for the CLE is different from that of the first templates and as much more "derived" information is now represented, some new Tags were necessary. Particular attention was

given to the representation of semantic information, e.g. data on taxonomic and other semantic relations that have been extracted by analysing the definitions, and new fields were introduced in our model so that particular lexical features (such as, for example, superordinates, or roles such as agent, patient, or other semantic features such as quantity, quality, action, place, etc.) could be made accessible; in this way the CLE responds to issues which have been raised or results which have emerged from recent linguistic studies.

It was decided to structure the entries in the CLE on a "one-entry one-major-part-of-speech" basis in which the particular relations between lexical items given implicitly in the source dictionaries (such as lexical or grammatical homograph, variant or derivative form) are indicated in a special field. This means that, where in the source dictionaries homographs representing more than one part of speech were compacted together in a single entry, in the CLE they were split to create one DB entry for each separate part of speech while, however, maintaining a link between them. In the same way, much other information which was previously found in a single entry in many dictionaries, e.g. variant forms, derivatives formed from the headword by adding a suffix, phrasal verbs, etc., is given in the LDB in separate entries. It was thus essential to represent explicitly relations between lexical items which were given implicitly in the source dictionary, but which risk being lost once the splitting of source entries has been completed. Thus, for each LDB entry, the entry type is specified and all the entries that are indicated in the source dictionary as being related to this entry are listed in successive Related_Entry fields, each one followed by indication of the particular relationship involved in a Related_entry_type field; for example, the homographic relationship between DB entries with the same form but different major POSs, previously compacted under the same headword in the source dictionary, is identified by the value "Gram_Hom" entered in the Related_entry_type field. In the same way, values will be entered in this field for all those entries which are shown in the source dictionary as being related in a dependent fashion to a given entry, e.g. variant forms, run-ons, etc.. The relationship between entries with the same graphical form (homographs and homonyms) which appear in the source dictionary as separate entries has been maintained. The number which appeared in the source dictionary indicating the occurrence of more than one

entry for the same form is entered as value in the Hdwd_Hom_No. field.[2]

3. Using the Common Lexical Entry

The work done by the different partners in defining a standard structure for dictionary data representation was very important at the beginning of the project as it facilitated mapping of the different dictionaries into the project databases[3] using common data fields. During this phase the Common Lexical Entry (or a subset of it) represented a reference point for each partner. However, since then, the main activity of the project has been concentrated on the extraction of information from the different LDBs and the construction of a common Multilingual Lexical Knowledge Base. Attention has thus been focused on the definition of a common type system and feature structures which could be used to represent the lexical information in the LKB and away from work on the lexical database. For this reason, after the initial stage, not much attention has been given to the CLE within the project. The ACQUILEX Common Lexical Entry has, however, already been adopted with success in other projects, giving us a further chance for experimentation, and interest in the proposal has been shown from outside the project and in particular by the Text Encoding Initiative (see below).

3.1 Applications

We can cite here two applications of the CLE in Pisa, external to ACQUILEX.

---

[2] The Template for the Common Lexical Entry, the new tags and the values they can take are specified in the second part of the Preliminary Report.

[3] Two LDB systems were developed: the Cambridge system implemented on Apple Macintosh machines (see Carroll, 1990) and the Pisa MLDB developed on personal computers running the MS/DOS operating system (see Marinai et al, 1990). In this way, the maximum diffusion of the project results was guaranteed. The Pisa lexical databases can also be loaded onto the Cambridge system.

A procedure that has been studied to permit semi-automatic mapping between the monolingual and bilingual LDBs contained in the Pisa integrated MLDB system, developed within the project, is described in Marinai et al (1991). Equivalent entries from separate electronic dictionaries are combined and links are created between their senses, mainly on the basis of the information which can be extracted from definitions, examples and semantic labels, in order to create a new more complete composite entry which represents the sum of the information contained in the individual dictionaries. The system has been designed to permit the linguist or lexicographer to compare and study the lexical information given for the same item in different sources. It has been experimented on the DMI[4] and Garzanti Italian monolingual dictionaries, and on the Collins Italian/English bilingual LDB.

The first step in the procedure was to map the lexical data from the separate dictionaries onto the CLE. In this way, the different criteria adopted by the separate dictionaries to distinguish between homographs and homonyms and to record derivative and variant forms can be eliminated. For example, in these dictionaries, the DMI gives a separate entry for each lexical and grammatical homograph while Collins and Garzanti both usually, but not always, list lexical homographs separately but put grammatical homographs into a single entry, subdivided for the different grammatical categories; Garzanti lists many derivatives as run-ons to the main entry while Collins gives them separate headword status. As the CLE provides a separate entry for each major part of speech, for homonyms and derivatives, these differences were eliminated. Two important modifications were made to the CLE in this application. An additional field which contained a standardised value for the equivalent grammatical categories of the different dictionaries was introduced in order to facilitate the combining of all the information available for a given lexical item from each dictionary, grammatical category by grammatical category. Tables setting standard values for the semantic labels in the different dictionaries (subject, semantic, register, usage and geographic codes) were implemented so that the procedure which proposes the mapping between senses could match the same information, which is presented in different ways in

---

[4] The Italian Machine Dictionary (DMI), constructed at the ILC-CNR, Pisa, is based to a large extent on the Zingarelli Italian dictionary.

different dictionaries. (Thus, for example, the semantic labels *ant, antiq, rar, arc,* from Garzanti, *ant* from Collins, and *1* and *3* from the DMI, all translated into a standard label **OLD** denoting old or archaic form.)

Similarly, the Lexicographic Workstation (also implemented in Pisa, see for example Picchi et al. 1992) which provides a series of tools useful for the creation and revision of monolingual and bilingual dictionaries has adopted a slightly modified version of the CLE for the mono and bilingual LDB systems it maintains, and for the dictionary editor developed to assist the lexicographer compiling the entry. When the editor environment is entered, the user is presented with a default (mono- or bilingual) entry template based on the CLE, that lists explicitly all the phonetic, syntactic and semantic information to be contained in the printed dictionary. However, the lexicographer can modify the structure of the template proposed by the system, to some extent, to meet the particular characteristics of the dictionary or lexicon on which he is working.

Both these applications use simplified versions of the CLE which was designed to represent the totality of the information contained in all the project databases.

3.2 Feedback

The recent interest in the most suitable structure to represent the information contained in the dictionary entry has not been confined to ACQUILEX. Studies have included those of Amsler and Tompa (1988) for the monolingual dictionary and Fought and Van Ess-Dykema (1990) for the bilingual entry. The most important coordinated international activity currently under way in this field is the Text Encoding Initiative, co-sponsored by the Association for Computational Linguistics (ACL), Association for Literary and Linguistic Computing (ALLC), Association for Computing in the Humanities (ACH) and by the European Community. The intention of TEI is to provide guidelines and standards for the representation and exchange of texts in machine-readable format using the SGML mark-up language (see TEI, 1990) and the general programme of activity includes a study on the definition of a common representation for the dictionary entry (see TEI 1990: 183-187 for a

first analysis of the structural components of dictionary entries, and see Ide et al., 1991, for a more detailed description).

Our preliminary report on the Acquilex dictionary model was used as input by the TEI group working on the dictionary. The first results of this group have been extremely useful in evaluating the CLE. It is our intention that the final proposal for the Acquilex Computational Model of the Dictionary Entry will be compatible with the work of the Text Encoding Initiative and that it can be formulated using SGML[5].

4. Final Proposal

The initial study of the lexical entry was very thorough. All the information which could be stored explicitly or implicitly was analysed in great detail and care was taken to make the preliminary proposal for the computational model as complete as possible. For this reason, it has not been necessary to make many significant alterations to the preliminary proposal. Perhaps the principal defect was the lack of a formal mechanism to represent the hierarchical structure of the entry and to clearly define the scope and dependency of embedded information.

4.1 Modifying the preliminary CLE

We have maintained the basic philosophy of the CLE which is organised on a "one-entry one-major-part-of-speech" basis. However, as lexical data was to be stored in the LKB on a "word sense" basis, it was originally decided to repeat the Grammatical Information node for each Sense_Group in order to facilitate the later mapping of entries to the Knowledge Base. On the basis of the experience acquired applying the CLE, this structure has been found

---

[5] It is perhaps still early to expect much feedback from other projects. However, interest has been shown by researchers working in the Research Institute of Computing Machinery, Prague, who used our model to represent a bilingual English/Czech dictionary, introducing a number of new attributes and dependencies. They have communicated their results to us making a number of useful observations. Their intention is to build a lexical database.

unnecessarily redundant. In fact, in both the applications described in 3.1, the entire Gram_Inf_Group has been implemented at the headword node level. It has thus been decided to adopt the same solution in the Final Proposal. The Gram_Inf_Group is no longer repeated for each sense but has been raised to the headword level. This is the most important change which has been made to the original CLE. If separate entries are to be created for each sense of any lexical item, then the grammatical information can be inherited appropriately. In most cases, the different senses of an entry will inherit all the grammatical features stored at the headword level, without any change. If, for a given sense, the grammatical information should vary then this is shown by repeating the relevant fields in the sense group and thus overriding the grammatical information appearing at Headword_Group level. For example, some words can assume a different sense when used in the plural. In this case, the Number tag of the Gram_Inf_Group will be given explicitly within the Sense_Group, and the new value will override the value given at headword level. It is evident that an overriding of an attribute value inherited from an upper level is limited in scope to that particular sense.

A particular problem we encountered when applying the computational model to our dictionaries was that the subset of functional or grammatical words caused difficulties as they have no true semantic value and thus, instead of a definition, we found either an equivalent or nearly equivalent functional word or some indications of their grammatical and syntactical use followed by a number of examples. For such cases, we feel that the Gram_Inf_Text field can be used at the Sense_Group level to cater with this type of information.

The other significant change made to the original proposal is an attempt to evidence in a more precise and formal mode, particular phenomena that are typical of lexical entries such as inheritance, scoping and overriding of information. We have thus provided an example of a TEI-style Document Type Definition, adopting the following tags (based on SGML, see Camuglia, 1992):

Delimiters

[	DTD open delimiter
]	DTD close delimiter
<!	element definition open
>	element definition close
<	start tag open
>	tag close
</	end tag close
(	group open delimiter
)	group close delimiter

Connectors:

,	sequence connector
|	or connector
&	and connector

Occurrence indicators:

?	optional
+	required and repeatable
*	optional and repeatable

The final proposal for the project Common Lexical Entry is given in the following pages. The first four pages give the expanded Template showing all the possible information fields in all the possible positions for both monolingual (M) and bilingual (B) dictionary entries. The structure has been slightly modified with respect to the original proposal but no new information fields have been added. The expanded template is followed by Abbreviated Schema for both Monolingual and Bilingual Lexical Entries and then by an example of a Document Type Definition which formally specifies the structure of the CLE.

One thing which is perhaps still lacking is the definition of standard values for many of the tags (such as those adopted for the POS and Semantic label fields by Marinai et al. (1991)) in order to facilitate the matching and comparing of information between LDBs. This problem has not received high priority within the project as it becomes irrelevant once the lexical information from the LDBs is mapped onto the LKB typed feature structures.

COMMON LEXICAL ENTRY TEMPLATE

FINAL PROPOSAL

Tags at Dictionary Level:

```
DICT_SOURCE
   Dict_name:
   Dict_type:

   LANGUAGE
M     Lang:
B       L1:
B       L2:
B     Metalanguage:

   PHONETIC TRANSCRIPTION
        IPA:
        Notes:
```

Tags at Entry Level:

```
ENTRY_GROUP
   DB_Entry_Id.:
   Source_Entry_Id:
   Entry_type:

   HEADWORD_GROUP
    ?Hdwd_text:
     Hdwd_form:
     Hdwd_type:
     Hdwd_POS:
    ?Hdwd_label:
    ?Hdwd_Hom_No.:
    ?Hdwd_freq_inf:

      ?PHONETIC_GROUP
         ?Pronunc_text:
         +Pronunciation:
         ?Primary_stress:
         ?Secondary_stress:

      ?ETYMOLOGY_GROUP
          Etymology_text:
```

```
        ?GRAM_INF_GROUP
           ?Gram_Inf_text:
                 ?subcat:
                 ?subtype:
                 ?gender:
                 ?number:
                 ?various:
                 ?g_code:
                 ?aux_form:
              ?INFLECTION_GROUP
                    ?Infl_text:
                    ?Infl_label:
                    ?Infl_stem:
                    +Infl_form:


      *SENSE_GROUP
           Sense_no.:
          *GRAM_INF_GROUP...


           *CROSS_REFERENCE_GROUP
              X-ref_type:
              Related_entry:
              Related_entry_id:

           *SEMANTIC_LABEL_GROUP
              ?Semantic_label_text:
              ?Subject_code:
              ?Semantic_code:
              ?Register_code:
              ?Usage_code:
              ?Geographic_code:
              ?Country_code:

M           +DEFINITION_GROUP
               Def_no:
              ?CROSS_REFERENCE_GROUP..
              ?SEMANTIC_LABEL_GROUP...
              ?GRAM_INF_GROUP...
               Def_text:
              ?TAXONOMY_GROUP
                 ?Taxon_label:
                  Taxon_text:
```

```
B          +TRANSLATION_GROUP
            Trans_no:
           ?Trans_type:
           ?SEMANTIC_LABEL_GROUP...
           ?SEMANTIC_INDICATOR_GROUP:
             ?Semantic_Ind_type:
              Semantic_Ind_text:
            Trans_text:
           ?GRAM_INF_GROUP...

           *EXAMPLE_GROUP
               ?Ex_type:
               ?SEMANTIC_LABEL_GROUP...
               ?SEMANTIC_INDICATOR_GROUP:...
                Example:
               ?Ex_explanation:
B              ?COLLOCATION_GROUP
                   Coll_POS:
                   Coll_word:

B              +EXAMPLE_TRANS_GROUP
                   ?Ex_Trans_type:
                   ?Ex_Trans_label:
                    Ex_Trans_text:
                   ?GRAM_INF_GROUP...
                   ?SEMANTIC_LABEL_GROUP...
                   ?SEMANTIC_INDICATOR_GROUP:...

M              *MULTIWORD_GROUP
                   ?Mwd_type:
                   ?SEMANTIC_LABEL_GROUP...
                    Mwd_form:
                   ?Mwd_explanation:

               *PROVERB_GROUP
                    Prov_text:
                   ?Prov_explan:
```

```
M                SEMANTIC_RELATION_GROUP

                   *SEMANTIC_FEATURES_GROUP
                      Sem_Features:
                      Sem_Roles:

                   *SUPERORDINATE_GROUP
                       Genus_term:
                       Genus_term_id.:

                   *SYNONYM_GROUP
                       Synonym:
                       Syn_entry_id:

                   *ANTONYM_GROUP
                       Antonym:
                       Ant_entry_id:

                   *ALTERATE_GROUP
                       Alterate:

      *RELATED_ENTRY_GROUP
          Related_entry:
          Related_entry_id:
          Related_entry_type:
```

ABBREVIATED SCHEMA OF COMMON LEXICAL ENTRY FOR MONOLINGUAL DICTIONARIES

```
<ENTRY_GROUP>

   <HEADWORD_GROUP>
         <PHONETIC_GROUP>              </PHONETIC_GROUP>
         <ETYMOLOGY_GROUP>             </ETYMOLOGY_GROUP>
         <GRAM_INF_GROUP>
             <INFLECTION_GROUP>            </INFLECTION_GROUP>
         </GRAM_INF_GROUP>

      <SENSE_GROUP>
         <CROSS_REFERENCE_GROUP>       </CROSS_REFERENCE_GROUP>
         <SEMANTIC_LABEL_GROUP>        </SEMANTIC_LABEL_GROUP>
         <DEFINITION_GROUP>
            <TAXONOMY_GROUP>             </TAXONOMY_GROUP>
            <EXAMPLE_GROUP>
               <PROVERB_GROUP>              </PROVERB_GROUP>
               <MULTIWORD_GROUP>            </MULTIWORD_GROUP>
            </EXAMPLE_GROUP>
            <SEMANTIC_RELATION_GROUP>
               <SEMANTIC_FEATURES_GROUP>    </SEMANTIC_FEATURES_GROUP>
               <SUPERORDINATE_GROUP>        </SUPERORDINATE_GROUP>
               <SYNONYM_GROUP>              </SYNONYM_GROUP>
               <ANTONYM_GROUP>              </ANTONYM_GROUP>
               <ALTERATE_GROUP>             </ALTERATE_GROUP>
            </SEMANTIC_RELATION_GROUP>
         </DEFINITION_GROUP>
      </SENSE_GROUP>

   </HEADWORD_GROUP>
        <RELATED_ENTRY_GROUP>        </RELATED_ENTRY_GROUP>
</ENTRY_GROUP>
```

ABBREVIATED SCHEMA OF COMMON LEXICAL ENTRY FOR BILINGUAL DICTIONARIES

```
<ENTRY_GROUP>

   <HEADWORD_GROUP>
         <PHONETIC_GROUP>            </PHONETIC_GROUP>
         <ETYMOLOGY_GROUP>           </ETYMOLOGY_GROUP>
         <GRAM_INF_GROUP>
             <POS_GROUP>                 </POS_GROUP>
             <INFLECTION_GROUP>          </INFLECTION_GROUP>
         </GRAM_INF_GROUP>

      <SENSE_GROUP>
         <CROSS_REFERENCE_GROUP>     </CROSS_REFERENCE_GROUP>
         <SENSE_LABEL_GROUP>
            <SEMANTIC_LABEL_GROUP>       </SEMANTIC_LABEL_GROUP>
            <SEMANTIC_INDICATOR_GROUP>  </SEMANTIC_INDICATOR_GROUP>
         </SENSE_LABEL_GROUP>
         <TRANSLATION_GROUP>
            <EXAMPLE_GROUP>
               <EXAMPLE_TRANS_GROUP>       </EXAMPLE_TRANS_GROUP>
            </EXAMPLE_GROUP>
        </TRANSLATION_GROUP>
      </SENSE_GROUP>

     </HEADWORD_GROUP>
        <RELATED_ENTRY_GROUP>       </RELATED_ENTRY_GROUP>
</ENTRY_GROUP>
```

5. Compatibility with the Text Encoding Initiative

In accordance with the stated aim of making our proposal compatible with the work of the Text Encoding Initiative, here below we show how a Document Type Definition (DTD) can be implemented for the abbreviated schema of our monolingual CLE template, following the model given in the most recent report by the TEI Dictionary Working Group.

```
[
<!Element Dictionary...(Entry_Group+)>
<!Element Entry...((Headword_Group),(Related_Entry_Group*)>
<!Element Headword...((Phonetic_Group?),(Etymology_Group?),
                      (Gram_Inf_Group?),(Sense_Group*))>
]
```

If we decompose the most complex group, the Sense_Group, into all its subgroups, we make transparent the numerous levels of embedding given the wide range of semantic information which is made explicit by analyzing the definitions.

```
<!Element Sense...((Cross_Reference_Group*),(Semantic_Label_Group*),
                   (Definition_Group+))>
<!Element Definition...((Taxonomy_Group?),(Example_Group*), (Semantic_Relation_Group))>
<!Element Example...((Multiword_Group*),(Proverb_Group*))>
<!Element Semantic_Relation...((Semantic_Features_Group),(Superordinate_Group*),
                               (Synonym_Group*),(Antonym_Group*),(Alterate_Group*))>
```

This is the outermost level in our definition. If we move in deeper we can consider the atomic elements or leaves within each of these complex groups.

Starting with the Headword_Group we find:

```
Headword_Group:
<!Element Group...((Hdwd_text?),(Hdwd_form),(Hdwd_PoS),(Hdwd_type),
                   (Hdwd_label?),(Hdwd_Hom_No?),(Hdwd_freq_inf?))>
Phonetic_Group:
<!Element Group...((Pronunc_text?),(Pronunciation+),(Primary_stress?),
                   (Secondary_stress?))>
Etymology_Group:
<!Element Group...((Etymology_text),(Etymology_form+))>
```

```
Gram_Inf_Group:
<!Element Group...(((Gram_inf_text?),(subcat?),(subtype?),
                    (gender?),(number?),(various?),(g_code?),
                    (aux_form))((inflection_text?),(inflection_label?),
                    (inflection_stem?),(inflection_form)))>
```

For brevity, we omit here the definition of the atomic elements which are embedded in all the sub-groups of the Sense_Group as the procedure is exactly the same.

6. Conclusions

The definition of a data model for the lexical entry was an important moment at the beginning of the ACQUILEX project as it stimulated a careful and detailed analysis by all the partners of exactly what type of lexical information was stored in their dictionaries (and can be stored in dictionaries in general), how it is structured, and what is the best way to represent it formally. The results of the study should provide a useful starting point for other projects aiming at studying and structuring machine-readable data and constructing lexical databases.

The effort of this final proposal to make the ACQUILEX lexical data model compatible with the work of the Text Encoding dictionary group recognizes the importance of providing a generally accepted standard format for the representation and exchange of dictionary data.

References

Amsler R., Tompa F. (1988), An SGML-based standard for English monolingual dictionaries, in *Proc. Fourth Annual Conf. of the UW Center for the New OED: Information in Text*, Waterloo, 61-79.

Boguraev B., Briscoe E, Carroll J., Copestake A. (1990) Database Models for Computational Lexicography, presented at the EURALEX Conference, Malaga, Spain, 28-31 August 1990.

Calzolari N., Peters C., Roventini A. (1990), Computational Model of the Dictionary Entry: Preliminary Report, ACQUILEX, Esprit BRA 3030, Six Month Deliverable, ILC-ACQ-1-90, Pisa, 90p.

Camuglia , G. (1992), Manuale di Introduzione al SGML e al MARKIT, ILC-CNR, Pisa, Internal Report.

Carroll J. (1990), Lexical Database System: User Manual, ACQUILEX, Esprit BRA 3030,Deliverable No. 2.3.3 (a).

Copestake A. (1991), The LKB: A System for Representing Lexical Information extracted from Machine-Readable Dictionaries, *Proceedings of the Acquilex Workshop on Default Inheritance in the Lexicon,* Cambridge.

Fought J., Van Ess-Dykema C. (1990), Toward an SGML Document Type Definition for Bilingual Dictionaries, TEI Internal Report.

Ide N., Veronis J., Warwick-Armstrong S., Calzolari N. (1991), Principles for Encoding Machine Readable Dictionaries, TEI WP, A15W6.

Marinai E., Peters C., Picchi E. (1990), "The Pisa Multilingual Lexical Database System", Esprit Basic Research Action No. 3030, Twelve Month Deliverable, ILC-ACQ-2-90, Pisa, 61p.

Marinai E., Peters C., Picchi E. (1991), A prototype system for the semi-automatic sense linking and merging of mono- and bilingual LDBs, in N.Ide and S. Hockey (eds.), *Research in Humanities Computing*, OUP, (forthcoming).

Neff M., Byrd R. and Rizk O. (1987), Creating and querying hierarchical lexical data bases, in *Proceedings of the Second ACL Conference on Applied Natural Language Processing*, Austin, Texas, 84-93.

Picchi, E, Peters, C., Marinai, E. (1992), The Pisa Lexicographic Workstation: The Bilingual Components, accepted for EURALEX'92, to be held Tampere, Finland, 4-9 August, 1992.

TEI 1989, Text Encoding Initiative: Proposal for a Second Development Cycle, Technical Report TEI SCG 10, ACH, ACL, ALLC.

Project Dictionaries

Il Nuovo Dizionario Italiano Garzanti, Garzanti, Milano, 1984.

Collins Concise English-Italian, Italian-English Dictionary, Collins, London and Glasgow, 1985.

Van Dale Groot Woordenboek Hedendaags Nederlands, P.G.J. van Sterkenburg and W.J.J. Pijnenburg (eds.), Van Dale Lexicografie, Utrecht/Antwerpen, 1984.

Van Dale Groot Woordenboek Nederlands-Engels, W.Martin and G.A.J.Tops (eds.), Van Dale Lexicografie, Utrecht/Antwerpen, 1986.

Longman Dictionary of Contemporary English (LDOCE), P. Procter et al. (eds.), Longman, Harlow and London, 1978.

Oxford Advanced Learner's Dictionary of Current English (OALD), A.S.Hornby (ed.)., Oxford University Press, Oxford, 1974.

Vox: Diccionario General Ilustrado de la Lengua Española, Bibliograf S.A., 1987.

Zingarelli N. (1970), *Vocabolario della Lingua Italiana,* Zanichelli, Bologna, 1970.