Nancy Ide, Jean Véronis
Vassar College (Poughkeepsie, NY)
and
GRTC, CNRS (Marseille)
Susan Warwick-Armstrong
ISSCO (Geneva)
Nicoletta Calzolari
ILC, CNR (Pisa)

Principles for encoding machine readable dictionaries

Fifth Euralex International Congress,
University of Tampere, Finland, August 4-9, 1992

Abstract

We provide an outline of a standard format for encoding machine readable dictionaries, based on work which is ongoing within the dictionary work group of the Text Encoding Initiative. The format is suitable for encoding a wide range of dictionaries, and is flexible enough to accomodate many esoteric dictionaries as well. It is also suitable for encoding different "views" of a dictionary simultaneously in the same document, specifically, a view which sees the dictionary in its textual format, and a view which sees the information in the dictionary without concern for its physical rendering.

1. Introduction

The Text Encoding Initiative (TEI) is an international project to develop a common format for the encoding of machine readable literary and linguistic data, using the international standard markup language SGML (ISO, 1986). The primary goal of the TEI is to support data interchange and compatibility.

A working group of the TEI is developing an encoding format suitable for mono- and bilingual dictionaries.[1] Previous work (for example, The DANLEX Group, 1987; Amsler and Tompa, 1988; Fought and Van Ess-Dykema, 1990; Calzolari et al., 1990; Ide, Le Maitre, and Véronis, 1991) has made it clear that the development of a common dictionary encoding format is extremely difficult, due to the complexities and variations in dictionary structure as well as the varying and often conflicting goals of those who want to encode dictionaries.

We first discuss the background and context of the work of the TEI committee, in particular, the user communities the TEI is attempting to serve and their goals in encoding dictionaries. We then outine a set of guidelines for encoding dictionaries.[2]

2. Context and goals

2.1 Different communities of users

The dictionary encoding format being developed within the TEI is intended for use primarily by the following general groups:

(a) PUBLISHERS AND LEXICOGRAPHERS, who are developing databases of lexical information to enable the manipulation, presentation, and use of this information in various ways, and to provide the potential to produce different types of dictionaries (for example, a full version, a concise version, and a pocket version) from the same data. A common format for dictionary data would enable them to check coherency across related dictionaries and exchange lexical data among different dictionaries, potentially by automatic means.

(b) COMPUTATIONAL LINGUISTS, who use printed dictionaries as a rich source of ready-made linguistic data, from which computational lexicons for natural language processing systems can be constructed. In the past decade, computational linguists have commonly analyzed typesetter's tapes for printed dictionaries to identify and extract different fields of information. Their goal is typically to represent this data in a LEXICAL DATABASE, which contains the same kinds of information found in printed dictionaries as well as additional linguistic information. A common encoding format would enable computational linguists to exchange data, in particular translated typesetter's tapes, and to more easily merge information from different sources.

(c) PHILOLOGISTS AND PRINT HISTORIANS, who want to study and compare historical dictionaries. They are potentially interested in all aspects of physical layout of dictionaries, including page breaks, hyphenation, etc. However, philologists are at the same time interested in the content, and may in fact be interested in the relations between content and printed rendering. They need a common encoding format to enable data sharing among researchers and the use of common software to process dictionaries.

(d) DICTIONARY USERS, who want to be able to retrieve lexical information as they would from a database, but want the results to appear as in a printed book. The advantage of a common format for dictionary users is the potential for common software for processing dictionaries distributed in electronic form.

2.2 Multiple views

There are at least three different views of dictionaries:

(a) the TYPOGRAPHIC VIEW, which is concerned with the two-dimensional, printed page, including information about line and page breaks and other features of layout. This view is effectively the output of the typesetting process, and represents the exact form of a given printing. For example, a domain indication in a dictionary entry may be broken over a line and therefore hyphenated (e.g., "naut-" "ical"); the typographic view of the dictionary preserves this information.

(b) the TEXTUAL VIEW--the one-dimensional sequence of tokens which can be seen as the input to the typesetting process. Here, the particular form in which the domain name is given in a particular dictionary (e.g., as "nautical", "naut.", "Naut.", etc.) would be preserved, but not necessarily its printed rendering.

(c) the LEXICAL VIEW--this view includes the information represented in a dictionary, without concern for its exact textual form. Thus the only information preserved concerning domain may be "nautical", whatever the form in which it appears.

Publishers begin with the lexical view--i.e., lexical data as it might appear in a database--and generate first the textual view (i.e., information reflecting editorial choices for a particular dictionary, such as the use of the abbreviation "naut." for "nautical", etc.), and finally the typographic view representing a particular printed rendering. Ideally, this translation is automatic, and therefore publishers need to retain only the lexical view.

Computational linguists and philologists often begin with the typographic view and analyze it to obtain the textual and/or lexical views. Computational linguists may ultimately be concerned with retaining only the lexical view, or they may wish to preserve the typographic or textual views as a reference text, since information can be lost or misinterpreted in the translation process. Philologists potentially want to see the three views simultaneously, since they may well be interested in questions which span all of them. For instance, they may want to determine all the (potentially inconsistent) variant forms in which the domain "nautical" is used in a given edition of a dictionary. Thus they need to access the lexical and the textual views simultaneously.

The need to access more than one view of the dictionary at the same time necessitates the development of not only a means to encode each view, but also a mapping among them that preserves their relations.

2.3 Variations in structure

Dictionaries present special difficulties because they are not linear texts, but are instead highly structured. In particular, information in dictionary entries is typically FACTORED so that common information is not re-specified. For example, information such as pronunciation, orthographic form, part of speech, etc. is "factored out" at the head of an entry in order to make it clear that it applies to a number of senses. Coupled with this, there is a well-developed "override" system in dictionary entries: it is very common to give exceptions for a specific sense when factored information does not apply. For example, part of speech often appears at the entry level, but can be overriden for a particular sense.

Moreover, variations in dictionary structure present difficulties for the development of an encoding format:

(a) variations in structure may occur WITHIN dictionaries, especially when entry formats are inconsistent or elaborate. In some dictionaries such as the OED, wide variations in structure exist due to the complexity of entries: for example, in one case it may be advantageous to give an etymology at the beginning rather than the end of an entry, or, rather than giving a pronunciation within parentheses, providing an elaborate explanation of pronunciation variants. In such cases, it may be impossible to define a fixed structure that accomodates the original structure of every dictionary.

(b) variations in structure AMONG dictionaries. A gross example of this is the difference in the way dictionaries factor information: in one dictionary, all senses of a given orthographic form with the same etymology will be grouped in a single entry, regardless of part of speech, whereas in another, different entries for the same orthographic form are given if the part of speech is different. Even if we exclude exotic dictionaries such as the OED, such variations make it difficult to develop a common format which can be applied across dictionaries.

3. Recommendations for encoding

3.1 Tags

We propose two sets of tags:

(a) ATOMIC TAGS that mark elementary fields of information in a dictionary such as

```
<orth>     : orthographic representation of the word
<pron>     : pronunciation
<pos>      : part of speech
<usg>      : usage note (register, geography, etc.)
<def>      : definition text
<hn>       : homograph number
<sn>       : sense number
```

etc.

(b) BRACKETTING TAGS to group together related elements (such as orthographic form and pronunciation, or part of speech and gender), such as:

```
<entry>    : groups all information in an entry
<homograph>: groups all information in a homograph
<sense>    : groups all information in a sense
<form>     : groups graphic and phonetic form of a word
<gram>     : groups grammatical information
<etym>     : groups etymological information
<xr>       : groups cross reference information
<eg>       : groups example information (text, author,
             date, etc.)
```

3.2 Document Type Definitions

The development of an encoding format involves not only defining the base elements of a text, but also, more importantly, describing the structure of the text in terms of relations among these elements. In SGML, the structure of a class of texts is specified in a DOCUMENT TYPE DEFINITION (DTD), which is a context-free grammar describing the nesting of elements.

Allowing for the widest range of variation in structure in a DTD results in a completely general description which effectively says that anything can go anywhere in SOME dictionary. In the TEI, we provide such a "free" DTD in order to accomodate all possible dictionaries, which allows bracketting tags to contain any bracketting tag, atomic tag, or character data, in any order and as many times as necessary:

```
<!DOCTYPE dict [
<!ENTITY % atoms          "orth|pron|usg|pos|hn|sn|def|...">
<!ENTITY % brackets
                 "homograph|sense|form|gram|etym|xr|eg|...">
<!ELEMENT dict         - - (entry)+                        >
<!ELEMENT entry        - - (#PCDATA|%atoms|%brackets;)+    >
<!ELEMENT (%brackets;) - - (#PCDATA|%atoms|%brackets;)+    >
]>
```

However, it is clear that there are some strong and consistent structural principles both within and among the majority of average dictionaries, as well as in many regular entries in more exotic dictionaries, that a set of guidelines should capture. Such a description can provide guidance for those who are both encoding and creating dictionaries. We therefore propose, in addition to the "free" DTD, a "regular" DTD that captures these regularities, to be used whenever the structure of the dictionary allows it:

```
<!DOCTYPE dict [
<!ENTITY % brackets      "form|gram|etym|xr|eg|..."        >
<!ELEMENT dict        - - (entry)+                         >
<!ELEMENT entry       - - (homograph|sense|def|%brackets;)+>
<!ELEMENT homograph   - - (hn? & (sense|def|%brackets;)+)  >
<!ELEMENT sense       - - (sn? & (sense|def|%brackets;)+)  >
<!ELEMENT form        - - (form|orth|pron|infl|...)+       >
<!ELEMENT gram        - - (gram|pos|subcat|morph|...)+     >
...
]>
```

This DTD specifies that homographs can nest within entries, and senses can nest within either entries or homographs. Senses themselves may be nested to any depth to reflect the embedding of sub-senses. Further, the DTD specifies that `<form>` can contain only `<form>`, `<orth>`, `<pron>`, `<infl>`, etc., that `<gram>` can contain only `<gram>`, `<pos>`, `<subcat>`, `<morph>`, etc.

Although this DTD is more restrictive than the free DTD, it is nonetheless flexible enough to represent information factored in very different ways. This flexibility results from the fact that although bracketting tags are restricted in their content, most elements--e.g., `<form>`, `<gram>`, `<def>`, `<eg>`, `<usg>`, `<xr>`, `<etym>`--can appear at any level (i.e., within entries, homographs, or senses).

The following is an example of a simple entry from the Collins Pocket Dictionary, which follows the regular DTD:

> **demigod** ('demɪ,god) *n.* **1.** a. a being who is part mortal, part god. **b.** a lesser deity. **2.** a godlike person.

```
<entry>
  <form>
    <orth>demigod</orth>
    <pron>'demI,god</pron>
  </form>
  <gram>
    <pos>n.</pos>
  </gram>
  <sense>
    <sn>1</sn>
    <sense>
      <sn>a</sn>
      <def>a being who is part mortal, part god.</def>
    </sense>
    <sense>
      <sn>b</sn>
      <def>a lesser deity.</def>
    </sense>
  </sense>
  <sense>
    <sn>2</sn>
    <def>a godlike person.</def>
  </sense>
</entry>
```

3.3 User-modified DTDs

Even the regular DTD provided above is very general. It may be desirable in some cases to more tightly control the structure of a PARTICULAR dictionary for the purposes of validation. Validation of format is often important for publishers, who wish to check the consistency of entry format when a dictionary is created. For example, a publisher may want to check that pronunciation is never inadvertently given at the sense level in a particular dictionary, but both DTDs above allow this.

To enable a more precise description of a particular dictionary, structure can be added to the regular DTD. This DTD can be made more restrictive by eliminating elements from within bracketting tags, or by constraining the order in which, and/or the number of times, tags may appear.

For example, a DTD which restricts etymology to appear at the `<entry>` level and prevents `<form>` and `<gram>` from occurring within senses, would include the following modifications to the regular DTD:

```
<!ELEMENT homograph   - - (hn? & (sense|def|%brackets;)+)
                                                  -(etym) >
<!ELEMENT sense       - - (sn? & (sense|def|%brackets;)+)
                                        -(etym|form|gram) >
```

This DTD is more specific to a given dictionary structure, but loses the generality of the regular DTD. This demonstrates the trade-off between the ability to validate and the generality of the model.

3.4 Encoding simultaneous views

In section 2, different views of the dictionary were described. The DTDs proposed above are intended to apply to both the lexical and textual views, independently.[3] The TEI must also provide a mechanism to enable the simultaneous representation of both views.

There are several means to enable simultaneous views. If the alternatives are LOCAL, that is, if they span the content of only one tag, then we recommend that one view be given as content and the other in an attribute. For example, encoding "spectateur, -trice" and maintaining the textual view as content yields

```
<orth>spectateur</orth>
<orth expand="spectatrice">-trice</orth>
```

Maintaining the lexical view as content yields

```
<orth>spectateur</orth>
<orth original="-trice">spectatrice</orth>
```

If the lexical view demands re-ordering the content, it may be necessary to encode each view separately and utilize an alignment mechanism to associate corresponding elements from each view (Sperberg-McQueen and Burnard, 1990, section 6.2.5). This can be done within a file--for example, for individual entries--or on a larger scale for two separate files, each containing the encoding of one view.

4. Conclusion

We provide an outline of a standard format for encoding machine readable dictionaries, based on work which is ongoing within the dictionary work group of the Text Encoding Initiative. The format is suitable for encoding a wide range of dictionaries, and is flexible enough to accomodate many esoteric dictionaries as well. It is also suitable for encoding different "views" of a dictionary simultaneously in the same document, specifically, a view which sees the dictionary in its textual format, and a view which sees the information in the dictionary without concern for its physical rendering.

Endnotes

[1] The current members of the TEI print dictionary working group are Robert Amsler (co-chair), Nicoletta Calzolari (co-chair), Carol Van Ess-Dykema, John Fought, Nancy Ide, W. Frank Tompa, Jean Veronis, and Susan Warwick-Armstrong. The authors would like to acknowledge the contribution of discussions with other committee members to the ideas in this paper.

[2] It should be noted that the work of the TEI committees is still ongoing at this writing, and therefore the final recommendations made by the TEI may differ from those outlined here.

[3] Typographic views of text are treated within a separate working group of the TEI, and are therefore not discussed here.

Bibliography

AMSLER, R. A., TOMPA, F. W. (1988): An SGML-Based Standard for English Monolingual Dictionaries. *Fourth Annual Conference of the U[niversity of] W[aterloo] Centre for the New Oxford English Dictionary,* Waterloo, Canada, 61-79.

CALZOLARI, N. , PETERS, C., ROVENTINI, A. (1990): *Computational Model of the Dictionary Entry: Preliminary Report,* Acquilex: Esprit Basic Research Action NO. 3030, Six-Month Deliverable, Pisa.

FOUGHT, J., VAN ESS-DYKEMA, C. (1990): *Toward and SGML Document Type Definition for Bilingual Dictionaries,* TEI Working paper TEI AIW20 (available from the TEI).

IDE, N., LE MAITRE, J., VERONIS, J. (1991): *Outline of a Model for Lexical Databases,* GRTC-CNRS Technical Report #496 (to appear in *Information Processing and Management*).

ISO (1986): ISO 8879: *Information processing--Text and office systems--Standard Generalized Markup Language (SGML),* International Organization for Standardization, Geneva.

SPERBERG-McQUEEN, C.M., BURNARD, L. (1990): *Guidelines for the encoding and interchange of machine-readable texts,* Draft, Version 1.0. ACH, ACL, and ALLC.

THE DANLEX GROUP (1987): *Descriptive tools for electronic processing of dictionary data.* Lexicographica, Series Maior, Tubingen: Niemeyer.

Keywords

machine-readable dictionaries, encoding, standards, Text Encoding Initiative