An SGML-based Standard for English Monolingual Dictionaries

Robert A. Amsler

Bellcore
Morristown, NJ, USA

and

Frank Wm. Tompa

Bellcore
and
The Centre for the New Oxford English Dictionary
University of Waterloo
Waterloo, Ontario, Canada

ABSTRACT

This paper summarizes the progress to date on the development of an interchange standard for English monolingual machine-readable dictionaries. The need for such a standard has become urgent in the fields of computational lexicology, lexicography, and linguistics with the rapid dispersion of copies of several different machine-readable dictionaries and extensive research on the applications of such dictionaries.

1. The Diversity of Existing Machine-Readable Dictionaries

In addition to raw typesetting tapes, which presumably have existed since computers were used to typeset dictionaries, there are several formatted dictionaries in fairly common use in the computational lexicology community. The first dictionaries to be prepared for computational applications include the *Merriam-Webster Seventh Collegiate* and the *Longman Dictionary of Contemporary English*, both of which are available in record-based formats as described by [Peterson 86] and [Michiels 82]. More recently, the *Oxford English Dictionary* and the *Oxford Advanced Learner's Dictionary of Current English* have become available in formats based on "descriptive document markup," which do not segment the text into more conventional records. Unfortunately, the variation in formats among these dictionaries is so great, that it has been virtually impossible to implement dictionary-independent application programs, much less programs that access more than one dictionary (e.g., to support comparative lexicology). To complicate matters, multiple machine-readable versions of each dictionary have been developed as researchers convert the dictionaries into alternative forms more suited to their own applications. Thus, corrections applied to the text of a dictionary by one researcher have typically been unavailable to others who supposedly use the same dictionary.

In an attempt to correct these problems, work began on developing a standard at the ‘‘Workshop on the Lexicon in a Computational and Theoretical Perspective’’ held during the Summer Linguistics Institute hosted by Stanford University in the summer of 1987. Because most computational lexicologists have used *W7* or *LDOCE*, the initial goal of a dictionary standard was approached by assuming that the existing *ad hoc* conventions used for the two record-based formats could be modified to reach a new compatible standard which would then be extendible to all machine-readable dictionaries [Slocum 87]. However, subsequent to that meeting, it became clear that basing the standard on an existing internationally endorsed standard language for textual information would better serve the needs of the research community and perhaps influence the publishing community as well. That standard language, of course, is ISO-8879, the Standard Generalized Markup Language [SGML 86].

2. Standard Generalized Markup Language

SGML basically is a language in which to write standards for text (called *applications*). SGML itself does not require any specific delimiters or tags be used; it merely sets forth a method by which an application of SGML can itself describe its particular choice of delimiters and terminology in a formal manner. Therefore, the decision to use SGML in no way prescribes what the representation of text must look like and any decision to endorse *this* application of SGML should be made independently of the decision to adopt *some* SGML-based application for representing machine-readable dictionaries.

Following the conventions used by most SGML applications, the contents of a document are tagged as follows:

$$< \textbf{tid} \ \mathbf{a_1 = v_1} \ \mathbf{a_2 = v_2} \ \ldots \ \mathbf{a_n = v_n} \ > \quad \text{content} \quad < / \ \textbf{tid} \ >$$

Each unit of text to be marked is surrounded by a pair of tags having an identifier (represented here by *tid*) indicative of the role of the unit in the document (e.g., ‘‘paragraph’’, ‘‘headword’’). Opening tags may contain attributes-value pairs, represented here by $a_i = v_i$.

A second markup feature in SGML is the availability of ‘‘entity references,’’ which are used to represent encoded units stored outside the document (e.g., graphics or separately described documents). The most frequent use of entity references, in practice, is for representing non-ASCII characters. In fact, Appendix D in the published SGML standard includes declarations of several ‘‘public character sets’’ for non-English alphabets, special mathematical symbols, standard characters used in publishing, etc. Entity references are typically encoded in SGML applications using the syntax &**ref**; in which *ref* is the name of the encoded unit.

One important *application* of SGML is that produced by The Association of American Publishers (AAP) for a few forms of published text (specifically, books, articles, and serials) [AAP 87]. Because dictionaries can be viewed as another form of published text, it would seem worthwhile to define a dictionary standard as an extension of the AAP standard. However, in the AAP’s proposed standard there are certain assumptions about the nature of tags and tag names for several document classes, not all of which seem appropriate for the computational needs of lexicologists. Thus, where possible we have followed the AAP's guidelines and directions, but in some areas, the nature of the text being represented and the anticipated uses for that text preclude our proposed dictionary standard from being a simple extension of the AAP standard.

3. The Dictionary Encoding Standard

To accommodate the variety of content and presentation methods used in dictionaries, we have based the standard on many sources. The principal sources include two SGML-based dictionaries, the *OED* as documented in [NewOED 87] and the undocumented *OALD*, each in an original and a new version; two record-based dictionaries, *W7* as documented in [Peterson 86] and *LDOCE* as documented in [Michiels 82]; and a new record-based format that emerged from the Stanford Workshop [Slocum 87].

In addition to serving as evidence for consensus or diversity in developing the dictionary encoding standard, we have also attempted to accommodate these dictionaries within the standard whenever they contained unique features. This has been done because the obvious second step to creating a dictionary encoding standard is to convert existing dictionary formats into their representations in the standard format. It is the intention of Bellcore to complete this rewrite process for all the dictionaries with which we are working. This includes, in addition to the previously mentioned sources, a few dictionaries in their original typesetting formats.*

Figure 3-1 shows the entry for *apple* from the *Merriam-Webster Seventh Collegiate Dictionary*. Within the **<ME>** (main entry) tag there are a number of major *sections*, each introduced by a capitalized tag name. These include, in this simple entry, **<E>** (Etymology) and **<M>** (Meaning) sections. Other aspects of the encoding will be discussed below.

In the remainder of this paper, we will describe the major problems that arose in designing a dictionary standard, and we will present the solutions adopted. Where possible, we will relate the discussion to the encoding of the example entry for ‘‘apple.’’

3.1. Content vs. Layout

While the SGML standard provides the means to encode the content of a machine-readable dictionary, it is not intended to describe the layout of the printed book. This initially seems fine, as the layout is one of the most artificial aspects of a dictionary, largely just concerned with where the white space is placed around the content objects. However, there is a shadowy middle-ground between pure layout issues and content issues that is most easily understood in the context of specific editions of historically important dictionaries. Scholars studying such works might, for example, be concerned with where breaks occurred across pages or columns. If there are several editions of such a dictionary, these characteristics might provide insights into the lexicographic decision-making of the editors and publishers or otherwise be needed to work with the data as an acceptable proxy for the printed pages of the original book. Thus, the ability to record in the

* Bellcore has obtained copies of the *Macquarie Dictionary*, the *Collins English Dictionary*, and the *McGraw-Hill Dictionary of Scientific and Technical Terms*, and we hope to obtain additional machine-readable dictionaries with which to experiment in the future, producing standard formatted versions of them as a by-product.

```
<ME pos=n-a>
 <F>
  <orth hyph=2>apple</orth>
  <pron stress=PU syl=2>ap&schwa;l</pron></F>
 <E>
  <es>
   <etymon lang=ME>appel</etymon></es>
  <es>
   <rel>fr.</rel>
   <etymon lang=OE>&aelig;ppel</etymon></es>
  <es>
   <rel>akin to</rel>
   <eu>
    <etymon lang=OHG>apful</etymon>
    <deftext>apple</deftext></eu>
   <eu>
    <etymon lang=OSlav>abl&breve;ko</etymon></eu></es></E>
 <M>
  <S sn=1>
   <deftext>the fleshy usu. rounded and red or yellow edible pome fruit
    of a tree <taxon lev=G>Malus</taxon> of the rose family</deftext>
   <deftext type=also>an apple tree</deftext></S>
  <S sn=2>
   <deftext>a fruit or other vegetable production suggestive of an apple
    </deftext></S></M></ME>
```

Figure 3-1: W7 Entry for *apple* in Dictionary Standard Notation

dictionary encoding standard where line, column or page breaks occurred has been anticipated.

Within SGML there is provision for multiple hierarchies to describe a single document. Layout and content are good examples of two hierarchies for a document. That is, the beginnings and ends of the content items in a document do not typically coincide with the beginnings and ends of the pages, columns, paragraphs and lines. These two hierarchies are thus *not* nested. This is a very nasty representational problem as it complicates all computer processing of text. To handle this problem we are proposing the following convenient SGML "hack".

SGML allows comments to be present in a text if appropriately tagged. Such comments begin and end with a special symbol, which following the AAP standard, we have selected to be '--'. If layout tags are created for pages, columns, lines, etc., then these can be inserted within a content tag set as comments that are harmlessly discarded when processing the content hierarchy of the dictionary entries. Likewise, if one simply reverses the interpretation of the non-comment (content) tags and these layout-comment tags, one can "see" the alternate layout hierarchy of the text with the content tags appearing as mere comments within that data.

3.2. Accommodating Print-Based Conventions

Publishers and editors have adopted various conventions to represent common features in their dictionaries. In order to standardize the machine-readable versions, information idiosyncratically represented in dictionaries (especially as values taken from a closed set of abbreviations) will be represented via attributes. For example, the **<ME pos=n-a>** tag introducing the example entry for apple, includes a part-of-speech attribute with the value *n-a*, the abbreviation for *noun, often attributive* which will be taken from a standard table of designated abbreviations. Use of such tables will minimize cross-dictionary variation, while nevertheless allowing individual dictionaries to retain their own part of speech labels when formatting text based on the SGML codes.

A related problem, and one which arises even among multiple editions of a single dictionary, is that of representing information which is associated with lexical objects but without intrinsic meaning apart from the particular dictionary in which it appears (e.g. homograph or sense numbers). Again, attribute-value pairs will be used to record such information in such a way that it can be ignored when dealing solely with content yet accessed when required, e.g., for cross-reference resolution.

The machine-readable versions of dictionaries can also use attribute-values to achieve a separation of information that is typically interleaved in printed editions. For example, the **<orth hyph=2>** tag contains the orthographic form (i.e., spelling and permissible hyphenation points). Most printed dictionaries record hyphenation (or occasionally syllabification) points intermixed with the characters of the form. Thus, if *apple* were written as *ap˙ple* anyone searching for *apple* in the data would have to know what its hyphenation pattern was in order to find it. To avoid the complication of suppressing such marks when scanning the data, and to provide a more consistent recording convention among dictionaries, we adopt the technique used in the record-based form of *W7*: encode such points in terms of *character distance encoding*.

Character distance encoding is a general method used whenever there is information to be recorded about the position of marks inside content objects. Position information is recorded as a series of alphanumeric *character distances* (1-9,A-Z). The individual digits in this form of encoding detail the number of characters to skip before the next hyphenation (or syllable) boundary. Thus, the attribute-value **hyph=2** indicates that *apple* has a hyphenation point after the first two characters, as *ap˙ple*. For *applesauce*, the code would have been **hyph=23** which represents *ap˙ple˙sauce*.

A similar solution is provided for recording information about pronunciation. The **<pron stress=PU syl=2>** tag gives the sole pronunciation, the stress, and syllable boundaries. Stress is recorded as another type of distance encoding, this time for syllables instead of characters. One letter will appear for each syllable indicating whether it had **P**rimary, **S**econdary, or **U**nstressed status in the dictionary. Thus once again, although dictionaries denote stress with stress marks, these are idiosyncratic and we have chosen to regard them as solely a display characteristic of the published dictionary which is generated from the *syllable distance encoding*.

3.3. Logical Grouping

The tags of the dictionary standard can be divided into three broad classes: first, those that delimit specific content categories of dictionaries, such as orthographies, pronunciations, etymologically related words (etymons), and definition texts; second, those that are common to all forms of unrestricted text, such as italicized words used for biological nomenclature, foreign expressions, or to delimit the names of people or places;

and third, those which provide information about the structure of the dictionary entry by grouping together related content and text tags into units such as etymologies, meanings, or synonymy subsections of a dictionary entry.

One such tag is **<form>**, which permits the grouping of related orthographies and pronunciations within an entry. Thus, any given word may have multiple ways of being spelled or pronounced. For each such spelling there may be one or more acceptable associated pronunciations (or, visa versa, for each pronunciation there may be one or more acceptable spellings). The **<form>** tag is used to group the related pronunciations and spellings together. A related grouping function is served by the **<es>** and **<eu>** tags in etymologies. At the highest levels of the dictionary, the **<E>** (Etymology), **<M>** (Meaning), and **<RE>** (Related Entry) tags group together the major components of the entry.

Components of an entry (especially senses) may be organized into hierarchical groups and numbered accordingly (using Roman, Arabic or alphabetic symbols). These groups often reflect grammatical or historical organizing principles deemed important to users' understanding of dictionary content, but grouping is sometimes simply adopted to reduce the repetition of labels in the printed book. It is important that the organizing principles of the dictionaries be preserved. However there is no one organizing principle: some dictionaries group first by part of speech and then by historical development, others group first by historical development, and still others ignore historical development and group by subject only. Imposing one standard hierarchy on all dictionaries would lose the editors' intents, but accommodating all possible organizational structures in a standard results in unnecessarily idiosyncratic representations which will complicate document interchange and processing software.

In place of specific tags that denote hierarchical grouping (e.g., <sense-group>), the standard provides an attribute (**sn**) for the content tags of the elements of the group, having as its value the group number. In the example, the **<S>** (sense) tags record the sense number as the value for attribute **sn**. More generally, the value of the attribute is a ordered list of the group numbers giving the complete path from the root of the hierarchy. Thus, for example, the tag **<S sn="II 4 d">** represents a typical sense number for the *OED*.

As a consequence of adopting this convention, the components of nested structures can be more easily manipulated individually, while the nesting information is still preserved for identifying structurally-related components. Thus, for example, given part-of-speech attributes for the sense tags as well, a program can be written to extract all definitions for adjectival uses of a word from a dictionary regardless of whether part of speech, etymology, or domain is used as the basis of the sense hierarchy.

4. Universality of the Standard

There are three major problems in creating a standard to accommodate all dictionaries, even when restricted to English monolingual ones. The most obvious is to anticipate what might occur in a dictionary that is not among the sources at hand. We are confident that the wide variety of dictionaries at our disposal, in printed and machine-readable form, has broadened our definition of the standard to encompass the dominant characteristics of other dictionaries as well. As with all standards, however, we need to provide for extension mechanisms to allow for unpredicted tags, entity references, or values assumable by attributes.

The second problem is to accommodate the exceptions that lexicographers have included in existing dictionaries (whether intentional or not). For example, we have chosen to represent part of speech as an attribute that takes its value from a (to-be-determined) closed set. However, how should the information ‘‘collective noun taking a singular verb’’ be represented? Presumably as a special value in the set. But what then of ‘‘collective noun *often* taking a singular verb, except when...’’? In fact, dictionary editors rely on the option to use unconstrained text to describe many aspects of our highly flexible language. The standard must adopt a method for balancing the need for diverse modes of explanation against the computational need for consistency.

Thirdly, there is the problem of our own shortcomings in understanding the structure of a dictionary's entry, when the documentation of that structure is so sparse. In our work to date, this problem was apparent when designing the encoding for etymologies. In spite of well-written prefatory material in the several dictionaries (esp. [OED 28, W3I 61, Century 11]), and extensive reference books about lexicography and computational linguistics (e.g. [Crystal 86, Landau 84]), we were unable to uncover a definitive description of the structure within a typical etymology (e.g., the meaning of punctuation symbols and the scope of language names). The solution we adopted is to include tags for *etymons* (the word forms, with language as an attribute), *etymological units* (**eu**, the equivalent of a lexical entry, including form, pronunciation, meaning, and so forth) and *etymological segments* (**es**, branches of a hypothetical universal etymology tree, including information about the relationships **rel** among the components). We must wait for other experts to help us determine whether or not this organization is adequate for the standard.

As a closing note, we observe that the real test of the standard cannot be accomplished until several dictionaries have been encoded successfully. We expect to begin such work in the near future.

References

[AAP 87]
Standard for Electronic Manuscript Preparation and Markup
Version 2.0 edition, AAP, Washington, D.C., 1987.
Electronic Manuscript Project.

[Benbow 87]
Timothy Benbow, Peter Carrington, Gayle Johannesen, Frank Wm. Tompa, and Edmund Weiner.
Report on the New Oxford English Dictionary Survey
Technical Report OED-87-05, University of Waterloo Centre for the New Oxford English Dictionary, Waterloo, Ontario, Canada, November, 1987.

[Boguraev and Levin 88]
Boguraev, Branimir and Beth Levin.
Machine-Readable Dictionaries: A Computational Linguistics Perspective.
In *Second Conference on Applied Natural Language Processing.* Association for Computational Linguistics, 9 February, 1988.
Copies of overhead transparencies and bibliography to accompany tutorial presented at the conference.

[CED 79]
Laurence Urdang (editor).
Collins English Dictionary.
Collins, London, 1979.

[Century 11]
William Dwight Whitney (editor).
The Century Dictionary: An Encyclopedic Lexicon of the English Language.
The Century Company, New York, 1911.

[Crystal 86]
Crystal, David.
A Dictionary of Linguistics and Phonetics.
Basil Blackwell, New York, 1986.

[DOSATT 86]
Sybil P. Parker (editor).
McGraw-Hill Dictionary of Scientific and Technical Terms.
McGraw-Hill Book Company, New York, 1986.

[Kazman 86]
Rick Kazman.
Structuring the Text of the Oxford English Dictionary through Finite State Transduction.
Master's thesis, Computer Science Department, University of Waterloo, 1986.
Also Printed as Technical Report CS-86-20 of the Data Structuring Group.

[Landau 84]
Sidney I. Landau.
Dictionaries: The Art and Craft of Lexicography.
Charles Scribner's Sons, New York, 1984.

[LDOCE 78]
Paul Procter (editor).
Longman Dictionary of Contemporary English.
Longman Group UK Limited, Burnt MIll, Harlow, Essex, England, 1978.

[Macquarie 86]
Arthur Delbridge (editor).
The Penguin Macquarie Dictionary.
Penguin Books, New York, 1986.

[Michiels 82]
Archibal Michiels.
Exploiting a Large Dictionary Data Base.
PhD thesis, Universite de Liege, 1982.

[NewOED 87]
Oxford University Press.
New Oxford English Dictionary Project -- Phase 1: Description of the Processed Text of the New OED.
Technical Report, Oxford University Press, Oxford, England, July, 1987.

[OALDCE 77]
Hornby, A.S. (editor).
Oxford Advanced Learner's Dictionary of Current English.
Oxford University Press, Oxford, England, 1977.

[OED 28]
Murray, James A.H.
The Oxford English Dictionary.
Oxford at the Clarendon Press, Oxford, England, 1928.

[Peterson 86]
James L. Peterson.
Webster's Seventh New Collegiate Dictionary: A Computer-Readable File Format.
Technical Report, Microelectronics and Computer Technology Corporation (MCC), Austin, Texas, August 8th, 1986.

[Pullum&Ladusaw 86]
Geoffrey Pullum and William A. Ladusaw.
Phonetic Symbol Guide.
University of Chicago Press, Chicago, 1986.

[SGML 86]
International Standard ISO 8879: Information Processing -- Text and Office Systems -- Standard Generalized Markup Language (SGML)
first edition -- 1986-10-15 edition, ANSI, New York, NY, 1986.
Ref. No. ISO 8879-1986 (E).

[Slocum 87]
Jonathan Slocum.
Dictionary Exchange Format: re-revised "record oriented" proposal.
October, 1987.
Personal communication to the authors as a follow-up to a 1-day dictionary standards meeting at MCC in Austin, Texas.

[W3I 61]
Philip Babcock Gove (editor).
Webster's Third New International Dictionary of the English Language.
G. & C. Merriam Company, Springfield, Massachusetts, 1961.

[W7 67]
Philip Babcock Gove (editor).
Webster's Seventh New Collegiate Dictionary.
G. & C. Merriam Company, Springfield, Massachusetts, 1967.