12 Print Dictionaries

This chapter defines a base tag set for encoding human-oriented monolingual and polyglot dictionaries (as opposed to computational lexica, which are intended for use by language-processing software). Dictionaries are most familiar in their printed form; however, increasing numbers of dictionaries exist also in electronic forms which are independent of any particular printed form, but from which various displays can be produced — e.g. CD-ROM dictionaries.

Both typographically and structurally, dictionaries are extremely complex. In addition, dictionaries interest many communities with different and sometimes conflicting goals. As a result, many general problems of text encoding are particularly pronounced here, and more compromises and alternatives within the encoding scheme may be required.[96] Two problems are particularly prominent.

First, because the structure of dictionary entries varies widely both among and within dictionaries, the simplest way for an encoding scheme to accommodate the entire range of structures actually encountered is to allow virtually any element to appear virtually anywhere in a dictionary entry. It is clear, however, that strong and consistent structural principles do govern the vast majority of conventional dictionaries, as well as many or most entries even in more 'exotic' dictionaries; ideally, a set of encoding guidelines should capture these structural principles. We therefore define two distinct elements for dictionary entries, one (`<entry>`) which captures the regularities of most conventional dictionary entries, and a second (`<entryFree>`) which uses the same elements, but allows them to combine much more freely. It is recommended that `<entry>` be used in preference to `<entryFree>` wherever the structure of the entry allows it. These elements and their contents are described in sections 12.2 *The Structure of Dictionary Entries*, 12.6 *Unstructured Entries*, and 12.4 *Headword and Pronunciation References*.

Second, since so much of the information in printed dictionaries is implicit or highly compressed, their encoding requires clear thought about whether it is to capture the precise typographic form of the source text or the underlying structure of the information it presents. Since both of these views of the dictionary may be of interest, it proves necessary to develop methods of recording both, and of recording the interrelationship between them as well. Users interested mainly in the printed format of the dictionary will require an encoding to be faithful to an original printed version. However, other users will be interested primarily in capturing the lexical information in a dictionary in a form suitable for further processing, which may demand the expansion or rearrangement of the information contained in the printed form. Further, some users wish to encode *both* of these views of the data, and retain the links between related elements of the two encodings. Problems of recording these two different views of dictionary data are discussed in section 12.5 *Typographic and Lexical Information in Dictionary Data*, together with mechanisms for retaining both views when this is desired.

Whichever view is adopted, a parameter entity `TEI.dictionaries` must be declared within the document type subset of any document using this base tag set. This should have the value `INCLUDE`, as further described in section 3.3 *Invocation of the TEI DTD*. A document using this base tag set and no other additional tag sets will thus begin as follows:

```
<?xml version="1.0" encoding="UTF-8" ?>
<!DOCTYPE TEI.2 PUBLIC "-//TEI P4//DTD Main Document Type//EN" "tei2.dtd" [
  <!ENTITY % TEI.XML          'INCLUDE' >
```

[96] We refer the reader to previous and current discussions of a common format for encoding dictionaries. For example, Robert A. Amsler and Frank W. Tompa, *An SGML-Based Standard for English Monolingual Dictionaries*, in *Information in Text: Fourth Annual Conference of the U[niversity of] W[aterloo] Centre for the New Oxford English Dictionary* October 26–28, 1988, Waterloo, Canada, pp. 61–79; Nicoletta Calzolari et al., *Computational Model of the Dictionary Entry: Preliminary Report*, Acquilex: Esprit Basic Research Action No. 3030, Six-Month Deliverable, Pisa, April 1990; John Fought and Carol Van Ess-Dykema, *Toward an SGML Document Type Definition for Bilingual Dictionaries*, TEI working paper TEI AIW20 (available from the TEI); Nancy Ide and Jean Veronis, *Encoding Print Dictionaries*, *Computers and the Humanities* 29: 167–195, 1995; Nancy Ide, Jacques Le Maitre, and Jean Veronis, *Outline of a Model for Lexical Databases*, (Information Processing and Management, 29, 2, 159–186, 1993); Nancy Ide, Jean Veronis, Susan Warwick- Armstrong, Nicoletta Calzolari, *Principles for Encoding machine readable dictionaries*, *Proceedings of the Fifth EURALEX International Congress, EURALEX'92*, University of Tampere, Finland; The DANLEX Group, *Descriptive tools for electronic processing of dictionary data*, in *Lexicographica, Series Maior* (Tübingen: Niemeyer, 1987); and A. Tutin and Jean Véronis, J. (1998). *Electronic dictionary encoding: customizing the TEI Guidelines*, in *Proceedings of the Eighth Euralex International Congress*, 1998.

```
<!ENTITY % TEI.dictionaries 'INCLUDE' >
]>
```

12.1 Dictionary Body and Overall Structure

Overall, dictionaries have the same structure of front matter, body, and back matter familiar from other texts; the base tag set for dictionaries uses the same front-matter and back-matter elements as other TEI base tag sets; these are documented in chapter 7 *Default Text Structure*. In addition, dictionaries define the elements <entry>, <entryFree>, and <superEntry> as component-level elements which can occur directly within a text division or the text body.

The following tags should be used to mark the gross structure of a printed dictionary; the dictionary-specific tags are discussed further in the following section.

<text> contains a single text of any kind, whether unitary or composite, for example a poem or drama, a collection of essays, a novel, a dictionary, or a corpus sample.

<front> contains any prefatory matter (headers, title page, prefaces, dedications, etc.) found at the start of a document, before the main body.

<body> contains the whole body of a single unitary text, excluding any front or back matter.

<back> contains any appendixes, etc. following the main part of a text.

** ** contains a subdivision of the front, body, or back of a text.

<div0> contains the largest possible subdivision of the body of a text.

<div1> contains a first-level subdivision of the front, body, or back of a text (the largest, if <div0> is not used, the second largest if it is).

<entry> contains a reasonably well-structured dictionary entry.

<entryFree> contains a dictionary entry which does not necessarily conform to the constraints imposed by the <entry> element.

<superEntry> groups successive entries for a set of homographs.

The text-division elements <div2> through <div7> may also be used, as described in chapter 7 *Default Text Structure*.

As members of the class entries, <entry> and <entryFree> share the following attributes:

type indicates type of entry, in dictionaries with multiple types. Suggested values are:
- `main` a main entry (default).
- `hom` a homograph with a separate entry.
- `xref` a reduced entry whose only function is to point to another main entry (e.g. for forms of an irregular verb or for variant spellings: was pointing to be, or esthete to aesthete).
- `affix` an entry for a prefix, infix, or suffix.
- `abbr` an entry for an abbreviation.
- `supplemental` a supplemental entry (for use in dictionaries which issue supplements to their main work in which they include updated information about entries).
- `foreign` an entry for a foreign word in a monolingual dictionary.

key contains a (sortable) character sequence reflecting the entry's alphabetical position in the printed dictionary.

The front and back matter of a dictionary may well contain specialized material like lists of common and proper nouns, grammatical tables, gazetteers, a 'guide to the use of the dictionary', etc. These may be tagged as elements defined in the core tag set (chapter 6 *Elements Available in All TEI Documents*) or as specialized dictionary elements as defined in this chapter.

The <body> element consists of a set of *entries*, optionally grouped into one or several , <div0>, or <div1> elements. These text divisions might correspond, for example, to sections for different languages in bilingual dictionaries, sections for different letters of the alphabet, etc.[97] In print dictionaries, entries are typically typographically distinct entities, each headed by some morphological form of the lexical item described (the *headword*), and sorted in alphabetical order or (especially for non-alphabetic scripts)

[97] It is unlikely that many conventional dictionaries will require smaller divisions, but all the usual division elements <div2> through <div7> may be used.

in some other conventional sequence. Dictionary entries should be encoded as distinct successive items, each marked as an <entry> element. The type attribute may be used to distinguish different types of entries, for example main entries, related entries, run-on entries, or entries for cross-references, etc.

Some dictionaries provide distinct entries for homographs, on the basis of etymology, part-of-speech, or both, and typically provide a numeric superscript on the headword identifying the homograph number. In these cases each homograph should be encoded as a separate entry; the <superEntry> element may optionally be used to group such successive homograph entries. In addition to a series of <entry> elements, the <superEntry> may contain a preliminary <form> group (see section 12.3.1 *Information on Written and Spoken Forms*) when information about hyphenation, pronunciation, etc., is given only once for two or more homograph entries. If the homograph number is to be recorded, the global attribute n should be used for this purpose. In some dictionaries, homographs are treated in distinct parts of the same entry; in these cases, they may be separated by use of the <hom> element, for which see section 12.2.1 *Hierarchical Levels*.

A sort key, given in the key attribute, is often required for superentries and entries, especially in cases where the order of entries does not follow the local character-set collating sequence (as, for example, when an entry for “3D” appears at the place where “three-D” would appear).

The body of a bilingual dictionary with two parts will thus have an overall structure resembling the following:

```
<body>
  <div0 type='dictionary'>
    <!-- English-French -->
    <entry><!--...--></entry>
    <entry><!--...--></entry>
    <entry><!--...--></entry>
    <!-- ... -->
  </div0>
  <div0 type="dictionary">
    <!-- French-English -->
    <entry><!--...--></entry>
    <entry><!--...--></entry>
    <entry><!--...--></entry>
    <!-- ... -->
  </div0>
</body>
```

A dictionary with no internal divisions might have a structure like the following; a <superEntry> is shown grouping two homograph entries.

```
<body>
  <entry><!--...--></entry>
  <entry><!--...--></entry>
  <!-- ... -->
  <superEntry>
    <entry type='hom' n='1'><!--...--></entry>
    <entry type='hom' n='2'><!--...--></entry>
  </superEntry>
  <!-- ... -->
</body>
```

The base tag set for dictionaries is contained in the files teidict2.ent and teidict2.dtd. The first of these defines the class comp.dictionaries, so that the generic text-division elements , <div0>, <div1>, etc. can contain <entry> elements:

```
<!-- 12.1: Element classes for dictionary base-->
<!--Text Encoding Initiative Consortium:
Guidelines for Electronic Text Encoding and Interchange.
Document TEI P4, 2002.
Copyright (c) 2002 TEI Consortium. Permission to copy in any form
is granted, provided this notice is included in all copies.
These materials may not be altered; modifications to these DTDs should
be performed only as specified by the Guidelines, for example in the
```

```
chapter entitled 'Modifying the TEI DTD'
These materials are subject to revision by the TEI Consortium. Current versions
are available from the Consortium website at http://www.tei-c.org-->
<!--First we define attributes available on all the elements in
this tag set.-->
<!--declarations from 12.5.4: Attributes for dictionary work inserted here -->
<!--Next we define comp.dictionaries, which will be used in the
declaration of component, within file TEI2.DTD.-->
<!ENTITY % x.comp.dictionaries "" >
<!ENTITY % m.comp.dictionaries "%x.comp.dictionaries; %n.entry; |
%n.entryFree; | %n.superEntry;">
<!ENTITY % mix.dictionaries '| %m.comp.dictionaries;' >
<!--Next, we declare some specialized element
classes, used in various content models in the dictionary
tag set.-->
<!ENTITY % a.entries '
      type CDATA "main"
      key CDATA #IMPLIED'>
<!--declarations from 12.2.2: Class for top-
level structure of dictionary entries inserted here -->
<!--declarations from 12.3.1: Classes for morphological and form information inserted here -->
<!--declarations from 12.3.2: Elements for grammatical information inserted here -->
<!--declarations from 12.4: Classes for headword references inserted here -->
<!--declarations from 12.6: Model class for unstructured dictionary entries inserted here -->
<!-- end of 12.1-->
```

The dictionary-specific elements are all declared in the file teidict2.dtd, which has the following overall structure.

```
<!-- 12.1: Base tag set for printed dictionaries-->
<!--Text Encoding Initiative Consortium:
Guidelines for Electronic Text Encoding and Interchange.
Document TEI P4, 2002.
Copyright (c) 2002 TEI Consortium. Permission to copy in any form
is granted, provided this notice is included in all copies.
These materials may not be altered; modifications to these DTDs should
be performed only as specified by the Guidelines, for example in the
chapter entitled 'Modifying the TEI DTD'
These materials are subject to revision by the TEI Consortium. Current versions
are available from the Consortium website at http://www.tei-c.org-->
<!--First we embed the default text structure.-->
<![%TEI.singleBase;[
<!ENTITY % TEI.structure.dtd PUBLIC '-//TEI P4//ELEMENTS Default Text
Structure//EN' 'teistr2.dtd' >
%TEI.structure.dtd;
]]>
<!--Now we define the dictionary-specific material.-->
<!--declarations from 12.2.1: Dictionary entries and their structure inserted here -->
<!--declarations from 12.3.1: The form group inserted here -->
<!--declarations from 12.3.2: The gram group inserted here -->
<!--declarations from 12.3.3.1: Definition text inserted here -->
<!--declarations from 12.3.3.2: Translation information inserted here -->
<!--declarations from 12.3.4: Etymologies inserted here -->
<!--declarations from 12.3.5.1: Examples and citations inserted here -->
<!--declarations from 12.3.5.2: Usage information inserted here -->
<!--declarations from 12.3.5.3: Cross References inserted here -->
<!--declarations from 12.3.6: Related entries inserted here -->
<!--declarations from 12.4: Headword references inserted here -->
<!-- end of 12.1-->
```

12.2 The Structure of Dictionary Entries

A simple dictionary entry may contain information about the form of the word treated, its grammatical characterization, its definition, synonyms, or translation equivalents, its etymology, cross-references to other entries, usage information, and examples. These we refer to as the *constituent parts* or *constituents* of the entry; some dictionary constituents possess no internal structure, while others are most naturally

viewed as groups of smaller elements, which may be marked in their own right. In some styles of markup, tags will be applied only to the low-level items, leaving the constituent groups which contain them untagged. We distinguish the class of *top-level constituents* of dictionary entries, which can occur directly within entries, from the class of *phrase-level* constituents, which can normally occur only within top-level constituents. The top-level constituents of dictionary entries are described in section 12.2.2 *Groups and Constituents*, and documented more fully, together with their phrase-level sub-constituents, in section 12.3 *Top-level Constituents of Entries*.

In addition, however, dictionary entries often have a complex hierarchical structure. For example, an entry may consist of two or more sub-parts, each corresponding to information for a different part-of-speech homograph of the headword. The entry (or part-of-speech homographs, if the entry is split this way) may also consist of senses, each of which may in turn be composed of two or more sub-senses, etc. Each sub-part, homograph entry, sense, or sub-sense we call a *level*; at any level in an entry, any or all of the constituent parts of dictionary entries may appear. The hierarchical levels of dictionary entries are documented in section 12.2.1 *Hierarchical Levels*.

12.2.1 Hierarchical Levels

The outermost structural level of an entry is marked with the elements `<entry>` or `<entryFree>`. The `<hom>` element marks the subdivision of entries into homographs differing in their part-of-speech. The `<sense>` element marks the subdivision of entries and part-of-speech homographs into senses; this element nests recursively in order to provide for a hierarchy of sub-senses of any depth. All of these levels may each contain any of the constituent parts of an entry. A special case of hierarchical structure is represented by the `<re>` (related entry) element, which is discussed in section 12.3.6 *Related Entries*. Finally, the element `<dictScrap>` may be used at any point in the hierarchy to delimit parts of the dictionary entry which are structurally anomalous, as further discussed in section 12.6 *Unstructured Entries*.

<entry> contains a reasonably well-structured dictionary entry.

<entryFree> contains a dictionary entry which does not necessarily conform to the constraints imposed by the `<entry>` element.

<hom> groups information relating to one homograph within an `<entry>`.

<sense> groups together all information relating to one word sense in a dictionary `<entry>` (definitions, examples, translation equivalents, etc.) Attributes include:

level gives the nesting depth of this sense.

Values any string of digits

<dictScrap> encloses a part of a dictionary entry in which other phrase-level dictionary elements are freely combined.

For example, an entry with two senses will have the following structure:

```
<entry>
   <!-- ... information common to both senses -->
   <sense n="1"> <!-- ... sense number 1 --> </sense>
   <sense n="2"> <!-- ... sense number 2 --> </sense>
</entry>
```

An entry with two homographs, the first with two senses and the second with three (one of which has two sub-senses), may have a structure like this:

```
<entry>
   <!-- ... information common to both homographs, if any ... -->
   <hom n="1">
      <sense n="1"> ... </sense>
      <sense n="2"> ... </sense>
   </hom>
   <hom n="2">
      <sense n="1">
         <sense n="a"> ... </sense>
         <sense n="b"> ... </sense>
      </sense>
      <sense n="2"> ... </sense>
```

```
      <sense n="3"> ... </sense>
   </hom>
</entry>
```

In some dictionaries, homographs typically receive separate entries; in such a case, as noted in section 12.1 *Dictionary Body and Overall Structure*, the two homographs may be treated as entries, optionally grouped by a <superEntry>:

```
<superEntry>
   <!-- ... form information common to both homographs, if any ... -->
   <entry n="1">
      <sense n="1"> ... </sense>
      <sense n="2"> ... </sense>
   </entry>
   <entry n="2">
      <sense n="1">
         <sense n="a"> ... </sense>
         <sense n="b"> ... </sense>
      </sense>
      <sense n="2"> ... </sense>
      <sense n="3"> ... </sense>
   </entry>
</superEntry>
```

The hierarchical levels of dictionary entries are declared as shown in the following DTD fragment. As may be seen, the content model for <entry> specifies that entries do not nest, that homographs nest within entries, and that senses nest within entries, homographs, or senses, and may be nested to any depth to reflect the embedding of sub-senses. Any of the top-level constituents (<def>, <usg>, <form>, etc.) can appear at any level (i.e., within entries, homographs, or senses).

```
<!-- 12.2.1: Dictionary entries and their structure-->
<!ELEMENT superEntry %om.RO;  ((form?, entry+)| dictScrap)>
<!ATTLIST superEntry
      %a.global;
      %a.entries;
      TEIform CDATA 'superEntry'  >
<!ELEMENT entry %om.RO; ( hom | sense | %m.dictionaryTopLevel;
                        | %m.Incl; )+>
<!ATTLIST entry
      %a.global;
      %a.entries;
      TEIform CDATA 'entry'  >
<!ELEMENT entryFree %om.RO; ( #PCDATA | %m.dictionaryParts; |
                           %m.phrase; | %m.inter; | %m.Incl; )* >
<!ATTLIST entryFree
      %a.global;
      %a.entries;
      %a.dictionaries;
      TEIform CDATA 'entryFree'  >
<!ELEMENT hom %om.RO; ( sense | %m.dictionaryTopLevel; )* >
<!ATTLIST hom
      %a.global;
      %a.dictionaries;
      TEIform CDATA 'hom'  >
<!ELEMENT sense %om.RR; ( #PCDATA | sense | %m.dictionaryTopLevel; |
                         %m.phrase; | %m.Incl; )* >
<!ATTLIST sense
      %a.global;
      %a.dictionaries;
      level CDATA #IMPLIED
      TEIform CDATA 'sense'  >
<!ELEMENT dictScrap %om.RO; ( #PCDATA | %m.dictionaryParts; |
                           %m.phrase; | %m.inter; | %m.Incl; )* >
<!ATTLIST dictScrap
      %a.global;
```

```
      TEIform CDATA 'dictScrap'  >
<!-- end of 12.2.1-->
```

12.2.2 Groups and Constituents

As noted above, dictionary entries, and subordinate levels within dictionary entries, may comprise several constituent parts, each providing a different type of information about the word treated. The *top-level constituents* of dictionary entries are:

- information about the form of the word treated (orthography, pronunciation, hyphenation, etc.)
- grammatical information (part of speech, grammatical sub-categorization, etc.)
- definitions or translations into another language
- etymology
- examples
- usage information
- cross-references to other entries
- notes
- entries (often of reduced form) for related words, typically called *related entries*

Any of the hierarchical levels (<entry>, <entryFree>, <hom>, <sense>) may contain any of these top-level constituents, since information about word form, particular grammatical information, special pronunciation, usage information, etc., may apply to an entire entry, or to only one homograph, or only to a particular sense. The examples below illustrate this point.

The following elements are used to encode these top-level constituents:

<form> groups all the information on the written and spoken forms of one headword. Attributes include:

type classifies form as simple, compound, etc.

Suggested values include:

`simple` single free lexical item
`lemma` the headword itself
`variant` a variant form
`compound` word formed from simple lexical items
`derivative` word derived from headword
`inflected` word in other than usual dictionary form
`phrase` multiple-word lexical item

<gramGrp> groups morpho-syntactic information about a lexical item, e.g. <pos>, <gen>, <number>, <case>, or <itype> (inflectional class).

<def> contains definition text in a dictionary entry.

<trans> contains translation text and related information (within an entry in a multilingual dictionary).

<eg> (in a dictionary) contains an example text containing at least one occurrence of the word form, used in the sense being described; examples may be quoted from (named) authors or contrived.

<usg> contains usage information in a dictionary entry. Attributes include:

type classifies the usage information using any convenient typology.

Sample values include:

`geo` geographic area
`time` temporal, historical era (archaic, old, etc.)
`dom` domain
`reg` register
`style` style (figurative, literal, etc.)
`plev` preference level (chiefly, usually, etc.)
`lang` lang (language for foreign words, spellings pronunciations, etc.)
`gram` grammatical usage
`syn` synonym given to show use

`hyper` hypernym given to show usage
`colloc` collocation given to show usage
`comp` typical complement
`obj` typical object
`subj` typical subject
`verb` typical verb
`hint` unclassifiable piece of information to guide sense choice

<xr> contains a phrase, sentence, or icon referring the reader to some other location in this or another text. Attributes include:

type indicates the type of cross reference, using any convenient typology.

Sample values include:

`syn` cross reference for synonym information
`etym` etymological information
`cf` related or similar term
`illus` illustration of an object

<etym> encloses the etymological information in a dictionary entry.

<re> contains a dictionary entry for a lexical item related to the headword, such as a compound phrase or derived form, embedded inside a larger entry. Attributes include:

type classifies the related entry according to any convenient typology.

Values any string of characters

<note> contains a note or annotation. Attributes include:

type describes the type of note.

Values Values can be taken from any convenient typology of annotation suitable to the work in hand; e.g. annotation, gloss, citation, digression, preliminary, temporary

resp (responsible) indicates who is responsible for the annotation: author, editor, translator, etc.

Sample values include:

`auth[or]` note originated with the author of the text.
`ed[itor]` note added by the editor of the text.
`comp[iler]` note added by the compiler of a collection.
`tr[anslator]` note added by the translator of a text.
`transcr[iber]` note added by the transcriber of a text into electronic form.
`(initials)` note added by the individual indicated by the initials.

place indicates where the note appears in the source text.

Sample values include:

`foot` note appears at foot of page.
`end` note appears at end of chapter or volume.
`inline` note appears as a marked paragraph in the body of the text.
`left` note appears in left margin.
`right` note appears in right margin.
`interlinear` note appears between lines of the text.
`app[aratus]` note appears in the apparatus at the foot of the page.

anchored indicates whether the copy text shows the exact place of reference for the note.

Legal values are:

`yes` copy text indicates the place of attachment for the note.
`no` copy text indicates no place of attachment for the note.

target indicates the point of attachment of a note, or the beginning of the span to which the note is attached.

Values reference to the ids of element(s) which begin at the location in question (e.g. the id of an <anchor> element).

targetEnd points to the end of the span to which the note is attached, if the note is not embedded in the text at that point.

Values reference to the id(s) of element(s) which *end* at the location(s) in question, or to an empty element at the point in question.

In a simple entry with no internal hierarchy, all top-level constituents appear at the <entry> level.[98]

com.peti.tor /k@m"petit@(r)/ n person who competes. [OALD]

```
<entry>
   <form>
      <orth>competitor</orth>
      <hyph>com|peti|tor</hyph>
      <pron>k@m"petit@(r)</pron>
   </form>
   <gramGrp>
      <pos>n</pos>
   </gramGrp>
   <def>person who competes.</def>
</entry>
```

For the elements which appear within the <form> and <gramGrp> elements of this example, see below, section 12.3.1 *Information on Written and Spoken Forms*, and section 12.3.2 *Grammatical Information*.

As mentioned above, any top-level constituent can appear at any level when the hierarchical structure of the entry is more complex. The most obvious examples are <def> and <trans>, which appear at the <sense> level when several senses or translations exist:

disproof (dIs"pru:f) n. 1. facts that disprove something. 2. the act of disproving. [CED]

```
<entry>
   <form>
      <orth>disproof</orth>
      <pron>dIs"pru:f</pron>
   </form>
   <gramGrp> <pos>n</pos> </gramGrp>
   <sense n="1"> <def>facts that disprove something.</def> </sense>
   <sense n="2"> <def>the act of disproving.</def> </sense>
</entry>
```

In the following example, <gramGrp> is used to distinguish two homographs:

bray /breI/ n cry of an ass; sound of a trumpet. • vt [VP2A] make a cry or sound of this kind. [OALD]

```
<entry>
   <form>
```

[98] Each example taken from a real dictionary indicates its source using the following abbreviations for dictionary names:

C/R Beryl T. Atkins et al., *Collins Robert French-English English-French Dictionary* (London: Collins, 1978, rpt. 1983)

CED Collins English Dictionary

CP Collins Pocket

DNT *Le Dictionnaire de Notre Temps*, ed. Françoise Guerard (Paris: Hachette, 1990).

LDOCE Longman Dictionary of Contemporary English

NPEG *The New Penguin English Dictionary* (London: Penguin, 1986, rpt. 1987).

OALD *Oxford Advanced Learner's Dictionary of Current English*, ed. A. S. Hornby with A. P. Cowie and A. C. Gimson (Oxford University Press, 1974).

PLC *Petit Larousse en Couleurs* (Paris: Larousse, 1990).

PLI *Pequeño Larousse Illustrado* por Ramón García-Pelayo y Gross (Buenos Aires, Mexico, Paris: Ediciones Larousse, 1964).

PR *Le Petit Robert*

SSSE *Simon and Schuster's International Dictionary English/Spanish Spanish/English* ed. Tana de Gómez (New York: Simon and Schuster, 1973).

W7 Webster's 7th Collegiate

WNC *Webster's New Collegiate Dictionary* (Springfield, Mass.: G. & C. Merriam Co., 1975).

To simplify the electronic presentation of this document on systems with limited character sets, most of the pronunciations are presented using the transliteration found in the electronic edition of the *Oxford Advanced Learner's Dictionary*. Also, the middle dot in quoted entries is rendered with a full stop, while within the sample transcriptions hyphenation and syllabification points are indicated with |, regardless of their rendition in the source text.

```
      <orth>bray</orth>
      <pron>breI</pron>
    </form>
    <hom>
      <gramGrp> <pos>n</pos> </gramGrp>
      <def>cry of an ass; sound of a trumpet.</def>
    </hom>
    <hom>
      <gramGrp>
        <pos>vt</pos>
        <subc>VP2A</subc>
      </gramGrp>
      <def>make a cry or sound of this kind.</def>
    </hom>
  </entry>
```

Information of the same kind can appear at different levels within the same entry; here, grammatical information occurs both at entry and homograph level.

ca.reen /k@"ri:n/ vt,vi 1 [VP6A] turn (a ship) on one side for cleaning, repairing, etc. 2 [VP6A, 2A] (cause to) tilt, lean over to one side. [OALD]

```
<entry>
  <form>
    <orth>careen</orth>
    <hyph>ca|reen</hyph>
    <pron>k@"ri:n</pron>
  </form>
  <gramGrp>
    <pos>vt</pos>
    <pos>vi</pos>
  </gramGrp>
  <sense n="1">
    <gramGrp> <subc>VP6A</subc> </gramGrp>
    <def>turn (a ship) on one side for cleaning,
    repairing, etc.</def>
  </sense>
  <sense n="2">
    <gramGrp> <subc>VP6A</subc> <subc>VP2A</subc> </gramGrp>
    <def>(cause to) tilt, lean over to one side.</def>
  </sense>
</entry>
```

Alone among the constituent groups, <form> can appear at the <superEntry> level as well as at the <entry>, <hom>, and <sense> levels:

a.ban.don 1 /@"b&nd@n/ v [T1] 1 to leave completely and for ever; desert: The sailors abandoned the burning ship. 2 ... **abandon** 2 n [U] the state when one's feelings and actions are uncontrolled; freedom from control: The people were so excited that they jumped and shouted with abandon / in gay abandon. [LDOCE]

```
<superEntry>
  <form>
    <orth>abandon</orth>
    <hyph>a|ban|don</hyph>
    <pron>@"b&amp;nd@n</pron>
  </form>
  <entry n="1">
    <gramGrp>
      <pos>v</pos>
      <subc>T1</subc>
    </gramGrp>
    <sense n="1">
      <def>to leave completely and for ever ... </def>
      <!-- ... -->
    </sense>
    <sense n="2"> <!-- ... --> </sense>
  </entry>
  <entry n="2">
    <gramGrp>
```

```
      <pos>n</pos>
      <subc>U</subc>
    </gramGrp>
    <def>the state when one's feelings and actions are
      uncontrolled; freedom from control</def>
    <!-- ... -->
  </entry>
</superEntry>
```

The class of top-level constituents for dictionary entries is defined by the following DTD fragment:

```
<!-- 12.2.2: Class for top-level structure of dictionary entries-->
<!ENTITY % x.dictionaryTopLevel "" >
<!ENTITY % m.dictionaryTopLevel "%x.dictionaryTopLevel; %n.def; |
%n.dictScrap; | %n.eg; | %n.etym; | %n.form; | %n.gramGrp; | %n.note; | %n.re; |
%n.trans; | %n.usg; | %n.xr;">
<!-- end of 12.2.2-->
```

The individual constituents are declared below, each in the section which documents it in more detail.

12.3 Top-level Constituents of Entries

This section describes the top-level constituents of dictionary entries, together with the phrase-level constituents peculiar to each.

- the <form> element, which groups orthographic information and pronunciations, is described in section 12.3.1 *Information on Written and Spoken Forms*
- the <gramGrp> element, which groups elements for the grammatical characterization of the headword, is described in section 12.3.2 *Grammatical Information*
- the <def> and <trans> elements, which describe the meaning of the headword, are described in section 12.3.3 *Sense Information*
- the <etym> element and its special phrase-level elements are documented in section 12.3.4 *Etymological Information*
- the <eg>, <usg>, <lbl>, <xr>, and <note> elements are described in section 12.3.5 *Other Information*
- the <re> element, which marks nested entries for related words, is described in section 12.3.1 *Information on Written and Spoken Forms*

12.3.1 Information on Written and Spoken Forms

Dictionary entries most often begin with information about the form of the word to which the entry applies. Typically, the orthographic form of the word, sometimes marked for syllabification or hyphenation, is the first item in an entry. Other information about the word, including variant or alternate forms, inflected forms, pronunciation, etc., is also often given.

The following elements should be used to encode this information: the <form> element groups one or more occurrences of any of the others; it can also be recursively nested to reflect more complex subgrouping of information about word form(s), as shown in the examples.

<form> groups all the information on the written and spoken forms of one headword. Attributes include:

type classifies form as simple, compound, etc.

Suggested values include:

`simple` single free lexical item
`lemma` the headword itself
`variant` a variant form
`compound` word formed from simple lexical items
`derivative` word derived from headword
`inflected` word in other than usual dictionary form
`phrase` multiple-word lexical item

<orth> gives the orthographic form of a dictionary headword. Attributes include:

type gives the type of spelling.
Values Any convenient word or phrase, e.g. 'lat' (latinate), 'std' (standard), 'trans' (transliterated), etc.
extent gives the extent of the orthographic information provided.
Sample values include:
`full` full form
`pref` prefix
`suff` suffix
`part` partial

<pron> contains the pronunciation(s) of the word. Attributes include:
extent indicates whether the pronunciation is for whole word or part.
Sample values include:
`full` full form
`pref` prefix
`suff` suffix
`part` partial

<hyph> contains a hyphenated form of a dictionary headword, or hyphenation information in some other form.

<syll> contains the syllabification of the headword.

<stress> contains the stress pattern for a dictionary headword, if given separately.

<lbl> in dictionaries, contains a label for a form, example, translation, or other piece of information, e.g. 'abbreviation for', 'contraction of', 'literally', 'approximately', 'synonyms:', etc. Attributes include:
type classifies the label using any convenient typology.
Values any string of characters, such as 'usage', 'sense restriction', etc.

In addition to those listed above, the following elements, which encode morphological details of the form, may also occur within `<form>` elements:

<gram> within an entry in a dictionary or a terminological data file, contains grammatical information relating to a term, word, or form. Attributes include:
type classifies the grammatical information given according to some convenient typology — in the case of terminological information, preferably the dictionary of data element types specified in ISO WD 12 620.
Sample values include:
`pos` part of speech (any of the word classes to which a word may be assigned in a given language, based on form, meaning, or a combination of features, e.g. noun, verb, adjective, etc.)
`gen` gender (formal classification by which nouns and pronouns, and often accompanying modifiers, are grouped and inflected, or changed in form, so as to control certain syntactic relationships)
`num` number (e.g. singular, plural, dual, ...)
`animate` animate or inanimate
`proper` proper noun or common noun

<gen> identifies the morphological gender of a lexical item, as given in the dictionary.

<number> indicates grammatical number associated with a form, as given in a dictionary.

<case> contains grammatical case information given by a dictionary for a given form.

<per> contains an indication of the grammatical person (1st, 2nd, 3rd, etc.) associated with a given inflected form in a dictionary.

<tns> indicates the grammatical tense associated with a given inflected form in a dictionary.

<mood> contains information about the grammatical mood of verbs (e.g. "indicative", "subjunctive", "imperative")

<itype> indicates the inflectional class associated with a lexical item. Attributes include:

type indicates the type of indicator used to specify the inflection class, when it is necessary to distinguish between the usual abbreviated indications (e.g. 'inv') and other kinds of indicators, such as special codes referring to conjugation patterns, etc.
Sample values include:
`abbrev` abbreviated indicator
`verb table` coded reference to a table of verbs

Of these, the <gram> element is most general, and all of the others are synonymous with <gram> elements with appropriate values (`gen, number, case`, etc.) for the type attribute.

Different dictionaries use different means to mark hyphenation, syllabification, and stress, and they often use some unusual glyphs (e.g., the 'middle dot' for hyphenation). All of these glyphs should however be available in the Unicode character set, and may be represented using either the standard ISO name or an appropriate character entity reference, as discussed in 2.7.3 *Character references*. When transcribing representations of pronunciation the International Phonetic Alphabet should be used. It may be convenient (as has been done in the text of this chapter) to use a simple transliteration scheme for the phonetic transcription scheme employed; such a scheme should however be properly documented, either informally in the header, or formally using a Writing System Declaration (37 *Obtaining TEI WSDs*).

In the simplest case, nothing is given but the orthography:

```
<form>
   <orth>doom-laden</orth>
</form>
<!-- [CED] -->
```

Often, however, pronunciation is given.

soucoupe [sukup] ... [DNT]

```
<form>
   <orth>soucoupe</orth>
   <pron>sukup</pron>
</form>
```

For a variety of reasons including ease of processing, it may be desired to split into separate elements information which is collapsed into a single element in the source text; orthography and hyphenation may for example be transcribed as separate elements, although given together in the source text. For a discussion of the issues involved, and of methods for retaining both the presentation form and the interpreted form, see section 12.5 *Typographic and Lexical Information in Dictionary Data*.

This example splits orthography and hyphenation, and adds syllabification because it differs from hyphenation:

ar.ea ... [W7]

```
<form>
   <orth>area</orth>
   <hyph>ar|ea</hyph>
   <syll>ar|e|a</syll>
</form>
```

Multiple orthographic forms may be given, e.g. to illustrate a word's inflectional pattern:

brag ... vb. brags, bragging, bragged ... [CED]

```
<form>
   <orth>brag</orth>
   <!-- ... -->
</form>
<gramGrp>
   <pos>vb</pos>
</gramGrp>
<form type="infl">
   <orth>brags</orth>
   <orth>bragging</orth>
   <orth>bragged</orth>
</form>
<!-- ... -->
```

Or the inflectional pattern may be indicated by reference to a table of paradigms, as here:

horrifier [ORifje] (7) vt ... [C/R]

```
<form>
   <orth>horrifier</orth>
   <pron>ORifje</pron>
   <itype type="vbtable">7</itype>
</form>
```

As noted, <itype> etc. are synonymous with appropriately typed instances of the general <gram> element; the last example might equally be tagged thus:

```
<form>
   <orth>horrifier</orth>
   <pron>ORifje</pron>
   <gram type="itype / vbtable">7</gram>
</form>
```

Explanatory labels may be attached to alternate forms:

MTBF *abbrev. for* mean time between failures. [CED]

```
<entry>
   <form type="abbrev">
      <orth>MTBF</orth>
   </form>
   <form type="full">
      <lbl>abbrev. for</lbl>
      <orth>mean time between failures</orth>
   </form>
</entry>
```

When multiple orthographic forms are given, a pronunciation may be associated with all of them, as here:

biryani or **biriani** (%bIrI"A:nI) ... [CED]

```
<!-- The pronunciation is associated with both forms. -->
<form>
   <orth>biryani</orth>
   <orth>biriani</orth>
   <pron>%bIrI"A:nI</pron>
</form>
```

In other cases, different pronunciations are provided for different orthographic forms; here, the <form> element is repeated to associate the first orthographic form explicitly with the first pronunciation, and the second orthographic form with the second pronunciation:

mackle ("m&k^@l) or macule ("m&kju:l) ... [CED]

```
<!-- &supschwa; is a small superscript schwa -->
<form>
   <orth>mackle</orth>
   <pron>"m&amp;k&supschwa;l</pron>
</form>
<form>
   <orth>macule</orth>
   <pron>"m&amp;kju:l</pron>
</form>
```

Recursive nesting of the <form> element can preserve relations among elements that are implicit in the text. For example, in the CED entry for “hospitaller”, it is clear that “U.S.” is associated only with “hospitaler”, but that the pronunciation applies to both forms. The following encoding preserves these relations:

hospitaller or U.S. hospitaler ("hQspIt@l@) ... [CED]

```
<form>
   <orth>hospitaller</orth>
   <form>
      <usg type="geo">U.S.</usg>
      <orth>hospitaler</orth>
   </form>
```

```
   <pron>"hQspIt@l@</pron>
</form>
```

The formal declarations for the elements of the <form> group are these:

```
<!-- 12.3.1: The form group-->
<!ELEMENT form %om.RR;  ( #PCDATA | %m.phrase; | %m.inter;
                    | %m.formInfo; | %m.Incl; )* >
<!ATTLIST form
      %a.global;
      %a.dictionaries;
      type CDATA #IMPLIED
      TEIform CDATA 'form'  >
<!ELEMENT orth %om.RO;  %paraContent;>
<!ATTLIST orth
      %a.global;
      %a.dictionaries;
      type CDATA #IMPLIED
      extent CDATA "full"
      TEIform CDATA 'orth'  >
<!ELEMENT pron %om.RO;  %paraContent;>
<!ATTLIST pron
      %a.global;
      %a.dictionaries;
      extent CDATA "full"
      notation CDATA #IMPLIED
      TEIform CDATA 'pron'  >
<!ELEMENT hyph %om.RO;  %paraContent;>
<!ATTLIST hyph
      %a.global;
      %a.dictionaries;
      TEIform CDATA 'hyph'  >
<!ELEMENT syll %om.RO;  %paraContent;>
<!ATTLIST syll
      %a.global;
      %a.dictionaries;
      TEIform CDATA 'syll'  >
<!ELEMENT stress %om.RO;  %paraContent;>
<!ATTLIST stress
      %a.global;
      TEIform CDATA 'stress'  >
<!--(LBL is declared with USG, elsewhere.)-->
<!--Elements for morphological information:-->
<!ELEMENT gram %om.RO;  %paraContent;>
<!ATTLIST gram
      %a.global;
      %a.dictionaries;
      type CDATA #IMPLIED
      TEIform CDATA 'gram'  >
<!ELEMENT gen %om.RR;  %paraContent;>
<!ATTLIST gen
      %a.global;
      %a.dictionaries;
      TEIform CDATA 'gen'  >
<!ELEMENT number %om.RR;  %paraContent;>
<!ATTLIST number
      %a.global;
      %a.dictionaries;
      TEIform CDATA 'number'  >
<!ELEMENT case %om.RR;  %paraContent;>
<!ATTLIST case
      %a.global;
      %a.dictionaries;
      TEIform CDATA 'case'  >
<!ELEMENT per %om.RO;  %paraContent;>
<!ATTLIST per
      %a.global;
```

```
      %a.dictionaries;
      TEIform CDATA 'per'  >
<!ELEMENT tns %om.RO;  %paraContent;>
<!ATTLIST tns
      %a.global;
      %a.dictionaries;
      TEIform CDATA 'tns'  >
<!ELEMENT mood %om.RO;  %paraContent;>
<!ATTLIST mood
      %a.global;
      %a.dictionaries;
      TEIform CDATA 'mood'  >
<!ELEMENT itype %om.RR;  %paraContent;>
<!ATTLIST itype
      %a.global;
      %a.dictionaries;
      type CDATA #IMPLIED
      TEIform CDATA 'itype'  >
<!-- end of 12.3.1-->
```

The classes of morphological elements, and of elements allowed within the <form> group, are declared thus:

```
<!-- 12.3.1: Classes for morphological and form information-->
<!ENTITY % x.morphInfo "" >
<!ENTITY % m.morphInfo "%x.morphInfo; %n.case; | %n.gen; | %n.gram; |
%n.itype; | %n.mood; | %n.number; | %n.per; | %n.tns;">
<!ENTITY % x.formInfo "" >
<!ENTITY % m.formInfo "%x.formInfo; %n.form; | %n.hyph; | %n.lbl; |
%m.morphInfo; | %n.orth; | %n.pron; | %n.syll; | %n.usg;">
<!-- end of 12.3.1-->
```

12.3.2 Grammatical Information

The <gramGrp> element groups grammatical information, such as part of speech, subcategorization information (e.g., syntactic patterns for verbs, count/mass distinctions for nouns), etc. It can contain any of the following elements:

<pos> indicates the part of speech assigned to a dictionary headword (noun, verb, adjective, etc.)

<subc> contains subcategorization information (transitive/intransitive, countable/non-countable, etc.)

<colloc> contains a collocate of the headword. Attributes include:

type classifies the collocation, using any convenient typology.
Values any string of characters, e.g. ‘preposition’.

In addition, <gramGrp> can contain any of the morphological elements defined in section 12.3.1 *Information on Written and Spoken Forms* for <form>:

<gram> within an entry in a dictionary or a terminological data file, contains grammatical information relating to a term, word, or form. Attributes include:

type classifies the grammatical information given according to some convenient typology — in the case of terminological information, preferably the dictionary of data element types specified in ISO WD 12 620.
Sample values include:

- `pos` part of speech (any of the word classes to which a word may be assigned in a given language, based on form, meaning, or a combination of features, e.g. noun, verb, adjective, etc.)
- `gen` gender (formal classification by which nouns and pronouns, and often accompanying modifiers, are grouped and inflected, or changed in form, so as to control certain syntactic relationships)
- `num` number (e.g. singular, plural, dual, ...)
- `animate` animate or inanimate
- `proper` proper noun or common noun

<itype> indicates the inflectional class associated with a lexical item. Attributes include:

type indicates the type of indicator used to specify the inflection class, when it is necessary to distinguish between the usual abbreviated indications (e.g. 'inv') and other kinds of indicators, such as special codes referring to conjugation patterns, etc.
Sample values include:
`abbrev` abbreviated indicator
`verb table` coded reference to a table of verbs

<gen> identifies the morphological gender of a lexical item, as given in the dictionary.

<number> indicates grammatical number associated with a form, as given in a dictionary.

<case> contains grammatical case information given by a dictionary for a given form.

<per> contains an indication of the grammatical person (1st, 2nd, 3rd, etc.) associated with a given inflected form in a dictionary.

<tns> indicates the grammatical tense associated with a given inflected form in a dictionary.

<mood> contains information about the grammatical mood of verbs (e.g. "indicative", "subjunctive", "imperative")

Elements conveying morphological information bear different interpretations within `<gramGrp>` and `<form>` groups, the difference being that in the `<form>` group, the morphological information specified pertains to the specific alternate form in question, while within `<gramGrp>` it applies to the headword form. For example, in the entry "**pinna** ('pIn@) n., pl. -nae (-ni:) or -nas" [CED], the word defined can be either singular or plural; the "pl." specification applies only to the inflected forms provided. Compare this with "pants (paents) pl. n.", where "pl." applies to the headword itself.

As noted above in section 12.3.1 *Information on Written and Spoken Forms*, the elements for morphological information are simply shorthand for the general purpose `<gram>` element. Consider this entry for the French word 'médire':

médire v.t. ind. (de) ... [PLC]

This entry can be tagged using specialized grammatical elements:

```
<form> <orth>m&eacute;dire</orth> </form>
<gramGrp>
   <pos>v</pos>
   <subc>t ind</subc>
   <colloc type="prep">de</colloc>
</gramGrp>
```

Or using the `<gram>` element:

```
<form> <orth>m&eacute;dire</orth> </form>
<gramGrp>
   <gram type="pos">v</gram>
   <gram type="subc">t ind</gram>
   <gram type="colloc / prep">de</gram>
</gramGrp>
```

Like `<form>`, `<gramGrp>` can be repeated, recursively nested, or used at the `<sense>` level to show relations among elements.

isotope adj. et n. m. ... [DNT]

```
<form> <orth>isotope</orth> </form>
<gramGrp>
   <pos>adj</pos>
</gramGrp>
<gramGrp>
   <pos>n</pos>
   <gen>m</gen>
</gramGrp>
```

wits (wIts) pl. n. 1. (sometimes sing.) the ability to reason and act, esp. quickly ... [CED]

```
<entry>
   <form>
      <orth>wits</orth>
      <pron>wIts</pron>
```

```
   </form>
   <gramGrp>
      <number>pl</number>
      <pos>n</pos>
   </gramGrp>
   <sense n="1">
      <gramGrp>
         <number>sometimes sing.</number>
      </gramGrp>
      <def>the ability to reason and act, esp. quickly ...</def>
      <!-- ... -->
   </sense>
</entry>
```

The following gives the formal declarations for elements in the grammatical-information group.

```
<!-- 12.3.2: The gram group-->
<!ELEMENT gramGrp %om.RR; (#PCDATA | %m.phrase; | %m.inter; | %m.gramInfo;
| %m.Incl;)*>
<!ATTLIST gramGrp
      %a.global;
      %a.dictionaries;
      TEIform CDATA 'gramGrp'  >
<!ELEMENT pos %om.RO;  %paraContent;>
<!ATTLIST pos
      %a.global;
      %a.dictionaries;
      TEIform CDATA 'pos'  >
<!ELEMENT subc %om.RO;  %paraContent;>
<!ATTLIST subc
      %a.global;
      %a.dictionaries;
      TEIform CDATA 'subc'  >
<!ELEMENT colloc %om.RO;  %paraContent;>
<!ATTLIST colloc
      %a.global;
      %a.dictionaries;
      type CDATA #IMPLIED
      TEIform CDATA 'colloc'  >
<!-- end of 12.3.2-->
```

The class of elements allowed within the <gramGrp> element is declared thus. The class *morphInfo* is defined above in section 12.3.1 *Information on Written and Spoken Forms*.

```
<!-- 12.3.2: Elements for grammatical information-->
<!ENTITY % x.gramInfo "" >
<!ENTITY % m.gramInfo "%x.gramInfo; %n.colloc; | %n.gramGrp; | %n.lbl; |
%m.morphInfo; | %n.pos; | %n.subc; | %n.usg;">
<!-- end of 12.3.2-->
```

12.3.3 Sense Information

Dictionaries may describe the meanings of words in a wide variety of different ways — by means of synonyms, paraphrases, translations into other languages, formal definitions in various highly stylized forms, etc. No attempt is made here to distinguish all the different forms which sense information may take; all alike may be tagged using the <def> element described in section 12.3.3.1 *Definitions*.

Because as a special case it is frequently desired to distinguish the provision of translation equivalents in other languages from other forms of sense information, however, the specialized elements <tr> (translation equivalent) and <trans> (which groups a translation equivalent with related information such as its grammatical description) are defined for this purpose in section 12.3.3.2 *Translation Equivalents*.

Whether sense information in multilingual dictionaries is consistently tagged using <tr> or <def> is a matter of the encoder's choice; no blanket recommendation is made here.

12.3.3.1 Definitions

Dictionary definitions are those pieces of prose in a dictionary entry that describe the meaning of some lexical item. Most often, definitions describe the headword of the entry; in some cases, they describe

translated texts, examples, etc.; see <tr>, section 12.3.3.2 *Translation Equivalents*, and <eg>, section 12.3.5.1 *Examples*. The <def> element directly contains the text of the definition; unlike <form> and <gramGrp>, that is, it does not serve solely to group a set of smaller elements. The close analysis of definition text, such as the tagging of hypernyms, typical objects, etc., is not covered by these Guidelines.

Definitions may occur directly within an entry; when multiple definitions are given, they typically are identified as belonging to distinct senses, as here:

demigod (...) n. 1.a. a being who is part mortal, part god. b. a lesser deity. 2. a godlike person. [CP]

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

In multilingual dictionaries, it is sometimes possible to distinguish translation equivalents from definitions proper; here a <def> element is distinguished from the translation information within which it appears.

rémoulade [Remulad] nf remoulade, rémoulade (*dressing containing mustard and herbs*). [C/R]

```
<entry>
   <form>
      <orth>r&eacute;moulade</orth>
      <pron>Remulad</pron>
   </form>
   <gramGrp>
      <pos>n</pos>
      <gen>f</gen>
   </gramGrp>
   <trans>
      <tr>remoulade</tr>
      <tr>r&eacute;moulade</tr>
      <def>dressing containing mustard and herbs</def>
   </trans>
</entry>
```

The following gives the formal definition of <def>:

```
<!-- 12.3.3.1: Definition text-->
<!ELEMENT def %om.RO;  %paraContent;>
<!ATTLIST def
      %a.global;
      %a.dictionaries;
      TEIform CDATA 'def'  >
<!-- end of 12.3.3.1-->
```

12.3.3.2 Translation Equivalents

Multilingual dictionaries contain information about translations of a given word in some source language for one or more target languages. Minimally, the dictionary provides the corresponding translation in the target language; other information, such as morphological information (gender, case), various kinds of usage restrictions, etc., may also be given. If translation equivalents are to be distinguished from other kinds of sense information, they may be encoded using the <tr> element.

As in monolingual dictionaries, the <sense> element is used in multilingual dictionaries to group information (forms, grammatical information, usage, translation(s), etc.) about a given sense of a word where necessary, as in monolingual dictionaries. Information about the individual translation equivalents within a sense is grouped using <trans> element. This information may include the translation

text (tagged <tr> or <def>), morphological information (<gen>, <case>, etc.), usage notes (<usg>), translation labels (<lbl>), and definitions (<def>).

<trans> contains translation text and related information (within an entry in a multilingual dictionary).

<tr> contains a translation of the headword or an example.

<lbl> in dictionaries, contains a label for a form, example, translation, or other piece of information, e.g. 'abbreviation for', 'contraction of', 'literally', 'approximately', 'synonyms:', etc. Attributes include:

type classifies the label using any convenient typology.

Values any string of characters, such as 'usage', 'sense restriction', etc.

Note how in the following example, different translation equivalents are grouped into the same or different senses, following the punctuation of the source and the usage labels:

dresser ... (a) (Theat) habilleur m, -euse f; (Comm: window ~) étalagiste mf. she's a stylish ~ elle s'habille avec chic; V hair. (b) (tool) (for wood) raboteuse f; (for stone) rabotin m. [C/R]

```
<entry n="1">
   <form>
      <orth>dresser</orth>
      <!-- ... -->
   </form>
   <!-- ... -->
   <sense n="a">
      <sense>
         <usg type="dom">Theat</usg>
         <trans>
            <tr>habilleur</tr>
            <gen>m</gen>
         </trans>
         <trans>
            <tr>-euse</tr>
            <gen>f</gen>
         </trans>
      </sense>
      <sense>
         <usg type="dom">Comm</usg>
         <form type="compound"> <orth>window <oRef/> </orth> </form>
         <trans>
            <tr>&eacute;talagiste</tr>
            <gen>mf</gen>
         </trans>
      </sense>
      <eg>
         <q>she's a stylish <oRef/></q>
         <trans>
            <tr>elle s'habille avec chic</tr>
         </trans>
      </eg>
      <xr type="see">V. <ref target="hair">hair</ref>
      </xr>
   </sense>
   <sense n="b">
      <usg type="category">tool</usg>
      <sense>
         <usg type="hint">for wood</usg>
         <trans>
            <tr>raboteuse</tr>
            <gen>f</gen>
         </trans>
      </sense>
      <sense>
         <usg type="hint">for stone</usg>
         <trans>
            <tr>rabotin</tr>
```

```
            <gen>m</gen>
          </trans>
        </sense>
      </sense>
    </entry>
```

In this encoding, a distinction is made between the translation equivalent (“OAS”) and a descriptive phrase providing further information for the user of the dictionary.

O.A.S. ... nf (abrév de **Organisation de l’Armée secrète**) OAS (*illegal military organization supporting French rule of Algeria*). [C/R]

```
<entry>
   <!-- ... -->
   <trans>
      <tr>OAS</tr>
      <def>illegal military organization supporting French
      rule of Algeria</def>
   </trans>
</entry>
```

Note that <tr> may also be used in monolingual dictionaries when a translation is given for a foreign word:

havdalah or **havdoloh** Hebrew. (Hebrew hAvdA"lA; Yiddish hAv"dOl@) n. Judaism. the ceremony marking the end of the sabbath or of a festival, including the blessings over wine, candles and spices. [literally: separation] [CED]

```
<entry type="foreign">
   <form>
      <orth>havdalah</orth>
      <orth>havdoloh</orth>
   </form>
   <!-- ... -->
   <usg type="dom">Judaism</usg>
   <def>the ceremony marking the end of the sabbath or of a festival,
    including the blessings over wine, candles and spices.</def>
   <trans>
      <lbl>literally</lbl>
      <tr>separation</tr>
   </trans>
</entry>
```

The formal definition of these elements is as follows:

```
<!-- 12.3.3.2: Translation information-->
<!ELEMENT trans %om.RO;  ( #PCDATA | %m.phrase; | %m.inter; |
                     %m.dictionaryParts; | %m.Incl; )* >
<!ATTLIST trans
      %a.global;
      %a.dictionaries;
      TEIform CDATA 'trans'  >
<!ELEMENT tr %om.RO;  %paraContent;>
<!ATTLIST tr
      %a.global;
      %a.dictionaries;
      TEIform CDATA 'tr'  >
<!-- end of 12.3.3.2-->
```

12.3.4 Etymological Information

The element <etym> marks a block of etymological information. Etymologies may contain highly structured lists of words in an order indicating their descent from each other, but often also include related words and forms outside the direct line of descent, for comparison. Not infrequently, etymologies include commentary of various sorts, and can grow into short (or long!) essays with prose-like structure. This variation in structure makes it impracticable to define tags which capture the entire intellectual structure of the etymology or record the precise interrelation of all the words mentioned. It is, however, feasible to mark some of the more obvious phrase-level elements frequently found in etymologies, using tags defined in the core tag set or elsewhere in this chapter. Of particular relevance for the markup of etymologies are:
<etym> encloses the etymological information in a dictionary entry.

<lang> name of a language mentioned in etymological or other linguistic discussion.

<date> contains a date in any format. Attributes include:

calendar indicates the system or calendar to which the date belongs.

Values Recommended values include: *Gregorian*, *Julian*, *Roman*, *Mosaic*, *Revolutionary*, *Islamic*.

value gives the value of the date in some standard form, usually yyyy-mm-dd.

Values Any string representing a date in standard format; recommended form is ISO 8601:2000 5.2.1.1 Complete representation, extended format (yyyy-mm-dd)

certainty indicates the degree of precision to be attributed to the date.

Values Any appropriate value, e.g. *ca.*, *approx*, *after*, *before*.

<mentioned> marks words or phrases mentioned, not used.

<gloss> identifies a phrase or word used to provide a gloss or definition for some other word or phrase. Attributes include:

target identifies the associated term element

Values must be a valid identifier for some `<term>` element in the current document

<pron> contains the pronunciation(s) of the word. Attributes include:

extent indicates whether the pronunciation is for whole word or part.

Sample values include:

`full` full form
`pref` prefix
`suff` suffix
`part` partial

notation indicates what notation is used for the pronunciation, if more than one occurs in the machine-readable dictionary.

Values Sample values: IPA, Murray, ...

<usg> contains usage information in a dictionary entry. Attributes include:

type classifies the usage information using any convenient typology.

Sample values include:

`geo` geographic area
`time` temporal, historical era (archaic, old, etc.)
`dom` domain
`reg` register
`style` style (figurative, literal, etc.)
`plev` preference level (chiefly, usually, etc.)
`lang` lang (language for foreign words, spellings pronunciations, etc.)
`gram` grammatical usage
`syn` synonym given to show use
`hyper` hypernym given to show usage
`colloc` collocation given to show usage
`comp` typical complement
`obj` typical object
`subj` typical subject
`verb` typical verb
`hint` unclassifiable piece of information to guide sense choice

<lbl> in dictionaries, contains a label for a form, example, translation, or other piece of information, e.g. ‘abbreviation for’, ‘contraction of’, ‘literally’, ‘approximately’, ‘synonyms:’, etc. Attributes include:

type classifies the label using any convenient typology.

Values any string of characters, such as ‘usage’, ‘sense restriction’, etc.

As in other prose, individual word forms mentioned in an etymological description are tagged with `<mentioned>` elements. Pronunciations, usage labels, and glosses can be tagged using the `<pron>`, `<usg>`, and `<gloss>` elements defined elsewhere in these Guidelines. In addition, the `<lang>` element

may be used to identify a particular language name where it appears, in addition to using the lang attribute of the <mentioned> element.

Examples:

abismo m. (del gr. a priv. y byssos, fondo). Sima, gran profundidad. ...

```
<entry>
   <form><orth>abismo</orth></form>
   <!-- ... -->
   <etym>del <lang>gr.</lang> <mentioned>a</mentioned> priv. y
         <mentioned>byssos</mentioned>, <gloss>fondo</gloss> </etym>
   <!-- ... -->
</entry>
```

neume \'n(y)üm\ n [F, fr. ML pneuma, neuma, fr. Gk pneuma breath — more at **pneumatic**]: any of various symbols used in the notation of Gregorian chant ... [WNC]

```
<entry>
   <!-- ... -->
   <etym>
      <lang>F</lang> fr. <lang>ML</lang>
      <mentioned>pneuma</mentioned>
      <mentioned>neuma</mentioned> fr. <lang>Gk</lang>
      <mentioned>pneuma</mentioned>
      <gloss>breath</gloss>
      <xr type="etym">more at <ptr target="pneumatic"/> </xr>
   </etym>
   <def>any of various symbols ... </def>
   <!-- ... -->
</entry>
```

The formal definition for the elements described in this section and not declared elsewhere is:

```
<!-- 12.3.4: Etymologies-->
<!ELEMENT etym %om.RO;  ( #PCDATA | %m.phrase; | %m.inter; | usg | lbl
                      | def | trans | tr | %m.morphInfo; | eg
                      | xr | %m.Incl; )* >
<!ATTLIST etym
      %a.global;
      %a.dictionaries;
      TEIform CDATA 'etym'  >
<!ELEMENT lang %om.RR;  %paraContent;>
<!ATTLIST lang
      %a.global;
      %a.dictionaries;
      TEIform CDATA 'lang'  >
<!-- end of 12.3.4-->
```

12.3.5 Other Information

12.3.5.1 Examples

Dictionaries typically include examples of word use, usually accompanying definitions or translations. In some cases, the examples are quotations from another source, and are occasionally followed by a citation to the author.

The <eg> element contains usage examples and associated information; the example text itself should be enclosed in a <cit> element, if attributed, or a <q> or <quote> element otherwise. The <cit> element associates a quotation with a bibliographic reference to its source.

<eg> (in a dictionary) contains an example text containing at least one occurrence of the word form, used in the sense being described; examples may be quoted from (named) authors or contrived.

<q> contains a quotation or apparent quotation — a representation of speech or thought marked as being quoted from someone else (whether in fact quoted or not); in narrative, the words are usually those of of a character or speaker; in dictionaries, <q> may be used to mark real or contrived examples of usage. Attributes include:

type may be used to indicate whether the quoted matter is spoken or thought, or to characterize it more finely.

Sample values include:

`spoken` representation of direct speech, usually marked by quotation marks.

`thought` representation of thought, e.g. internal monologue.

direct may be used to indicate whether the quoted matter is regarded as direct or indirect speech.

Legal values are:

`y` speech or thought is represented directly.

`n` speech or thought is represented indirectly, e.g. by use of a marked verbal aspect.

`unspecified` no claim is made.

who identifies the speaker of a piece of direct speech.

Values may be an idref

<quote> contains a phrase or passage attributed by the narrator or author to some agency external to the text.

<cit> A quotation from some other document, together with a bibliographic reference to its source.

Examples frequently abbreviate the headword, and so their transcription will frequently make use of the <oRef> or <oVar> elements described below in section 12.4 *Headword and Pronunciation References*.

Examples:

multiplex /.../ adj tech having many parts: the multiplex eye of the fly. [LDOCE]

```
<eg>
   <q>the multiplex eye of the fly.</q>
</eg>
```

As the following example shows, <eg> can also contain elements such as <pron>, <def>, etc.

some ... 4. (*S*~ and *any* are used with *more*): Give me ~ more /s@'mO:(r)/ [OALD]

```
<sense n="4">
   <usg type="colloc">
      <oRef type="cap"/> and <mentioned>any</mentioned>
      are used with <mentioned>more</mentioned>
   </usg>
   <eg>
      <q>Give me <oRef/> more</q>
      <pron extent="part">s@'mO:(r)</pron>
   </eg>
</sense>
```

In multilingual dictionaries, examples may also be accompanied by translations:

horrifier ... vt to horrify. **elle était horrifiée par la dépense** she was horrified at the expense. [C/R]

```
<entry>
   <!-- ... -->
   <trans> <tr>to horrify</tr> </trans>
   <eg>
      <q>elle &eacute;tait horrifi&eacute;e par la d&eacute;pense</q>
      <trans> <tr>she was horrified at the expense.</tr> </trans>
   </eg>
</entry>
```

When a source is indicated, the example should be marked with a <cit> element:

valeur ... n. f. ... 2. Vx. Vaillance, bravoure (spécial., au combat). “La valeur n’attend pas le nombre des années” (Corneille). ... [DNT]

```
<sense n="2">
   <usg type="time">Vx.</usg>
   <def>Vaillance, bravoure (sp&eacute;cial., au combat)</def>
   <eg>
      <cit>
         <q>La valeur n'attend pas le
             nombre des ann&eacute;es</q>
         <bibl> <author>Corneille</author> </bibl>
      </cit>
```

```
   </eg>
</sense>
```

The formal definition of <eg> is:

```
<!-- 12.3.5.1: Examples and citations-->
<!ELEMENT eg %om.RO;  ( q | quote | cit | %m.dictionaryParts;
                        | %m.formPointers; )+ >
<!ATTLIST eg
      %a.global;
      %a.dictionaries;
      TEIform CDATA 'eg'  >
<!-- end of 12.3.5.1-->
```

12.3.5.2 Usage Information and Other Labels

Most dictionaries provide restrictive labels and phrases indicating the usage of given words or particular senses. Other labels, not necessarily related to usage, may be attached to forms, translations, cross references, and examples. Usage and other labels should be marked with the following elements:

<usg> contains usage information in a dictionary entry. Attributes include:

type classifies the usage information using any convenient typology.

Sample values include:

- `geo` geographic area
- `time` temporal, historical era (archaic, old, etc.)
- `dom` domain
- `reg` register
- `style` style (figurative, literal, etc.)
- `plev` preference level (chiefly, usually, etc.)
- `lang` lang (language for foreign words, spellings pronunciations, etc.)
- `gram` grammatical usage
- `syn` synonym given to show use
- `hyper` hypernym given to show usage
- `colloc` collocation given to show usage
- `comp` typical complement
- `obj` typical object
- `subj` typical subject
- `verb` typical verb
- `hint` unclassifiable piece of information to guide sense choice

<lbl> in dictionaries, contains a label for a form, example, translation, or other piece of information, e.g. ‘abbreviation for’, ‘contraction of’, ‘literally’, ‘approximately’, ‘synonyms:’, etc. Attributes include:

type classifies the label using any convenient typology.

Values any string of characters, such as ‘usage’, ‘sense restriction’, etc.

Typical usage labels mark

- temporal use (archaic, obsolete, etc.)
- register (slang, formal, taboo, ironic, facetious, etc.)
- style (literal, figurative, etc.)
- connotative effect (e.g. derogatory, offensive)
- subject field (Astronomy, Philosophy, etc.)
- national or regional use (Australian, U.S., Midland dialect, etc.)

Many dictionaries provide an explanation and/or a list of such usage labels in a preface or appendix. The type of the usage information may be indicated in the `type` attribute on the <usg> element. Some typical values are:

geo geographic area

time temporal, historical era (“archaic”, “old”, etc.)

dom domain

reg register

style style (figurative, literal, etc.)

plev preference level (“chiefly”, “usually”, etc.)

acc acceptability

lang language for foreign words, spellings pronunciations, etc.

gram grammatical usage

In addition to this kind of information, multilingual dictionaries often provide ‘semantic cues’ to help the user determine the right sense of a word in the source language (and hence the correct translation). These include synonyms, concept subdivisions, typical subjects and objects, typical verb complements, etc. These labels are also marked with the `<usg>` element; sample values for the `type` attribute in these cases include:

syn synonym given to show use

hyper hypernym given to show usage

colloc collocation given to show usage

comp typical complement

obj typical object

subj typical subject

verb typical verb

hint unclassifiable piece of information to guide sense choice

In this entry, one spelling is marked as geographically restricted:

colour or U.S. **color** ... [CED]

```
<form>
  <orth>colour</orth>
  <form>
    <usg type="geo">U.S.</usg>
    <orth>color</orth>
  </form>
</form>
```

In this example, usage labels are used to indicate domains, register, and synonyms associated with different senses:

palette [palEt] nf (a) (Peinture: lit, fig) palette. (b) (Boucherie) shoulder. (c) (aube de roue) paddle; (battoir à linge) beetle; (Manutention, Constr) pallet. [C/R]

```
<!-- ... -->
<sense n="a">
   <usg type="dom">Peinture</usg>
   <usg type="style">lit</usg>
   <usg type="style">fig</usg>
   <trans> <tr>palette</tr> </trans>
</sense>
<sense n="b">
   <usg type="dom">Boucherie</usg>
   <trans> <tr>shoulder</tr> </trans>
</sense>
<sense n="c">
   <sense>
      <usg type="syn">aube de roue</usg>
      <trans> <tr>paddle</tr> </trans>
   </sense>
   <sense>
      <usg type="syn">battoir &agrave; linge</usg>
      <trans> <tr>beetle</tr> </trans>
```

```
      </sense>
      <sense>
         <usg type="dom">Manutention</usg>
         <usg type="dom">Constr</usg>
         <trans> <tr>pallet</tr> </trans>
      </sense>
   </sense>
   <!-- ... -->
```

When the usage label is hard to classify, it may be described as a “hint”:

rempaillage [...] nm reseating, rebottoming (*with straw*). [C/R]

```
<entry>
   <!-- ... -->
   <trans>
      <tr>reseating</tr>
      <tr>rebottoming</tr>
      <usg type="hint">with straw</usg>
   </trans>
</entry>
```

The following gives the formal definition of <usg> and <lbl>:

```
<!-- 12.3.5.2: Usage information-->
<!ELEMENT usg %om.RO;  %paraContent;>
<!ATTLIST usg
      %a.global;
      %a.dictionaries;
      type CDATA #IMPLIED
      TEIform CDATA 'usg'  >
<!ELEMENT lbl %om.RO;  %paraContent;>
<!ATTLIST lbl
      %a.global;
      %a.dictionaries;
      type CDATA #IMPLIED
      TEIform CDATA 'lbl'  >
<!-- end of 12.3.5.2-->
```

12.3.5.3 Cross References to Other Entries

Dictionary entries frequently refer to information in other entries, often using extremely dense notations to convey the headword of the entry to be sought, the particular part of the entry being referred to, and the nature of the information to be sought there (synonyms, antonyms, usage notes, etymology, an illustration, etc.)

Cross references may be tagged in dictionaries using the simple <ref> and <ptr> elements defined in the core tag set (section 6.6 *Simple Links and Cross References*), or the ‘extended’ pointing elements <xref> and <xptr> defined in the additional tag set for linking, segmentation, and alignment (section 14.2 *Extended Pointers*). In addition, the <xr> element may be used to group all the information relating to a cross reference. The following elements may be used for tagging cross references within dictionaries:

<xr> contains a phrase, sentence, or icon referring the reader to some other location in this or another text. Attributes include:

type indicates the type of cross reference, using any convenient typology.

Sample values include:

- `syn` cross reference for synonym information
- `etym` etymological information
- `cf` related or similar term
- `illus` illustration of an object

<ref> defines a reference to another location in the current document, in terms of one or more identifiable elements, possibly modified by additional text or comment. Attributes include:

target specifies the destination of the reference by supplying the value of the id attribute on one or more other elements in the current document.

Values One or more valid identifiers, separated by white space.

<ptr> defines a pointer to another location in the current document in terms of one or more identifiable elements. Attributes include:

target specifies the destination of the pointer by supplying the values used on the id attribute of one or more other elements in the current document
Values One or more valid identifiers, separated by white space.

<xref> defines a reference to another location in the current document, or an external document, using an extended pointer notation, possibly modified by additional text or comment.

<xptr> defines a pointer to another location in the current document or an external document.

<lbl> in dictionaries, contains a label for a form, example, translation, or other piece of information, e.g. 'abbreviation for', 'contraction of', 'literally', 'approximately', 'synonyms:', etc. Attributes include:

type classifies the label using any convenient typology.
Values any string of characters, such as 'usage', 'sense restriction', etc.

As in other types of text, the actual pointing element (e.g. <ref> or <ptr>) is used to tag the cross-reference target proper (in dictionaries, usually the headword, possibly accompanyied by a homograph number, a sense number, or other further restriction specifiying what portion of the target entry is being referred to); as usual, a <ptr> element may be used when the actual text of the target description can be reconstructed automatically, and a <ref> element is typically to be preferred when it cannot be reconstructed automatically. The <xr> element is used to group the target with any accompanying phrases or symbols used to label the cross reference; the cross reference label itself may be tagged as a <lbl> or may remain untagged. Both of the following are thus legitimate:

glee ... Compare **madrigal** (sense 1) [CED]

```
<entry>
   <form> <orth>glee</orth> </form>
   <!-- ... -->
   <xr>Compare <ptr target="madrigal.1"/> </xr>
</entry>
```

hostellerie Syn. de hôtellerie (sens 1). [DNT]

```
<xr type="syn">
   <lbl>Syn. de</lbl> <ref>h&ocirc;tellerie (sens 1)</ref>.
</xr>
```

In addition to using, or not using, <lbl> to mark the cross-reference label, the two examples differ in another way. The former assumes that the first sense of 'madrigal' has the identifier "madrigal.1", and that the specific form of the reference in the source volume can be reconstructed, if needed, from that information. The latter does not require the first sense of "hôtellerie" to have an identifier, and retains the print form of the cross reference; by omitting the target attribute of the <ref> element, however, the second example does assume implicitly either that some software could usefully parse the phrase tagged as a <ref> and find the location referred to, or else that such processing will not be necessary.

The type attribute on the pointing element or on the <xr> element may be used to indicate what kind of cross reference is being made, using any convenient typology. Since different dictionaries may label the same kind of cross reference in different ways, it may be useful to give normalized indications in the type attribute, enabling the encoder to distinguish irregular forms of cross reference more reliably:

rose2 ... vb. the past tense of **rise**. [CED]

```
<entry n="2" type="xr, hom">
   <form> <orth>rose</orth><!-- ... --> </form>
   <!-- ... -->
   <xr type="pastof">
      <lbl>the past tense of</lbl>
      <ref target="rise">rise</ref>
   </xr>
</entry>
```

from cross-references for synonyms and the like:

antagonist ... syn see **adverse** [W7]

```
<xr type="syn">
   <lbl>syn see</lbl>
   <ref target="adverse">adverse</ref>
</xr>
```

Strictly speaking, the reference above is not to the entry for 'adverse', but to the list of synonyms found at that entry. Slightly more complicated is the following reference to an illustration accompanying another entry:

ax, axe ... ⟶ see the illus at **tool** [OALD]

This entry refers to the illustration at the entry for 'tool', not the entry itself. The target attribute might give the identifier of the illustration itself, or of the enclosing entry (in which case the type attribute might be used to infer that the reference is actually to the illustration, not the entry as a whole).

```
<xr type="see illustration">
   <lbl>see the illus at</lbl>
   <ptr target="tool.illus"/>
</xr>
```

In some cases, the cross reference is to a particular subset of the meanings of the entry in question:

globe ...V. **armillaire** (sphère) [PR]

```
<xr>V. <ref target="armillaire">armillaire</ref>
   <lbl type="sense-restriction">sph&egrave;re</lbl>
</xr>
```

Cross-references occasionally occur in definition texts, example texts, etc., or may be free-standing within an entry. These may typically be encoded using <ref> or <ptr>, without an enclosing <xr>. For example:

entacher ... *Acte entaché de nullité*, contenant un vice de forme ou passé par un incapable*. [DNT]

The asterisk signals a reference to the entry for 'incapable'.

```
<def>contenant un vice de forme ou pass&eacute;
     par un <ptr target="incapable"/>.</def>
```

In some cases, the form in the definition is inflected, and thus <ref> must be used, as here:

justifier ...4. IMPRIM Donner a (une ligne) une longeur convenable au moyen de blancs (2, sens 1, 3). [DNT]

```
<sense n="4">
   <usg type="dom">imprim</usg>
   <def>Donner a (une ligne) une longeur convenable au moyen de
        <ref target="blanc-2.1 blanc-2.3">blancs (2, sens 1, 3)</ref>
   </def>
</sense>
```

The formal definition for <xr> is as follows:

```
<!-- 12.3.5.3: Cross References-->
<!ELEMENT xr %om.RO;  ( #PCDATA | %m.phrase; | %m.inter; | usg
                      | lbl | %m.Incl; )* >
<!ATTLIST xr
      %a.global;
      %a.dictionaries;
      type CDATA #IMPLIED
      TEIform CDATA 'xr'  >
<!-- end of 12.3.5.3-->
```

12.3.5.4 Notes within Entries

Dictionaries may include extensive explanatory notes about usage, grammar, context, etc. within entries. Very often, such notes appear as a separate section at the end of an entry. The <note> element should be used for such material.

<note> contains a note or annotation. Attributes include:

type describes the type of note.

Values Values can be taken from any convenient typology of annotation suitable to the work in hand; e.g. annotation, gloss, citation, digression, preliminary, temporary

resp (responsible) indicates who is responsible for the annotation: author, editor, translator, etc.
Sample values include:
`auth[or]` note originated with the author of the text.
`ed[itor]` note added by the editor of the text.
`comp[iler]` note added by the compiler of a collection.
`tr[anslator]` note added by the translator of a text.
`transcr[iber]` note added by the transcriber of a text into electronic form.
`(initials)` note added by the individual indicated by the initials.

place indicates where the note appears in the source text.
Sample values include:
`foot` note appears at foot of page.
`end` note appears at end of chapter or volume.
`inline` note appears as a marked paragraph in the body of the text.
`left` note appears in left margin.
`right` note appears in right margin.
`interlinear` note appears between lines of the text.
`app[aratus]` note appears in the apparatus at the foot of the page.

anchored indicates whether the copy text shows the exact place of reference for the note.
Legal values are:
`yes` copy text indicates the place of attachment for the note.
`no` copy text indicates no place of attachment for the note.

target indicates the point of attachment of a note, or the beginning of the span to which the note is attached.
Values reference to the ids of element(s) which begin at the location in question (e.g. the id of an <anchor> element).

targetEnd points to the end of the span to which the note is attached, if the note is not embedded in the text at that point.
Values reference to the id(s) of element(s) which *end* at the location(s) in question, or to an empty element at the point in question.

For example:

ain't (eInt) *Not standard. contraction of* am not, is not, are not, have not *or* has not: *I ain't seen it.* ....
▲**Usage.** Although the interrogative form *ain't I?* would be a natural contraction of *am I not?*, it is generally avoided in spoken English and never used in formal English. [CED]

```
<entry>
   <form type="contr">
      <orth>ain't</orth>
      <pron>eInt</pron>
   </form>
   <usg type="reg">Not standard</usg>
   <form type="full">
      <lbl>contraction of</lbl>
      <orth>am not</orth>
      <orth>is not</orth>
      <orth>are not</orth>
      <orth>have not</orth>
      <orth>has not</orth>
   </form>
   <eg>
      <q>I ain't seen it.</q>
   </eg>
   <!-- ... -->
   <note type="usage">Although the interrogative form <mentioned>ain't
     I?</mentioned> would be a natural contraction of <mentioned>am I
     not?</mentioned>, it is generally avoided in spoken English and
     never used in formal English.</note>
</entry>
```

The formal declaration for <note> is given in section 6.8 *Notes, Annotation, and Indexing*. It has this form:

```
<!-- 12.3.5.4: [Note]-->
<!ELEMENT note %om.RO;  %specialPara;>
<!ATTLIST note
      %a.global;
      type CDATA #IMPLIED
      resp CDATA #IMPLIED
      place CDATA 'unspecified'
      anchored (yes | no) "yes"
      target IDREFS #IMPLIED
      targetEnd IDREFS #IMPLIED
      TEIform CDATA 'note'  >
<!-- end of 12.3.5.4-->
```

12.3.6 Related Entries

The <re> element encloses a degenerate entry which appears in the body of another entry for some purpose. Many dictionaries include related entries for direct derivatives or inflected forms of the entry word, or for compound words, phrases, collocations, and idioms containing the entry word.

Related entries can be complex, and may in fact include any of the information to be found in a regular entry. Therefore, the <re> element is defined to contain the same elements as an <entry> element, with the exception that it may not contain any nested <re> elements.

Examples:

bevvy ("bEvI) Dialect. ~ n., pl. -vies. 1. a drink, esp. an alcoholic one: we had a few bevvies last night. 2. a night of drinking. ~ vb. - vies, -vying, -vied (intr.) 3. to drink alcohol [probably from Old French bevee, buvee, drinking] —'bevvied adj. [CED]

```
<entry>
   <form>
      <orth>bevvy</orth>
      <pron>"bEvI</pron>
   </form>
   <usg type="reg">Dialect</usg>
   <hom>
      <gramGrp> <pos>n</pos> </gramGrp>
      <!-- ... -->
      <sense n="1">
         <def>a drink, esp. an alcoholic one</def>
         <!-- ... -->
      </sense>
      <!-- ... -->
   </hom>
   <hom>
      <gramGrp> <pos>vb</pos> </gramGrp>
      <!-- ... -->
      <sense n="3"> <def>to drink alcohol</def> </sense>
   </hom>
   <etym>probably from <lang>Old French</lang>
      <mentioned>bevee</mentioned>, <mentioned>buvee</mentioned>
      <gloss>drinking</gloss>
   </etym>
   <re type="derived">
      <form> <orth>bevvied</orth> </form>
      <gramGrp> <pos>adj</pos> </gramGrp>
   </re>
</entry>
```

The formal definition of <re> is

```
<!-- 12.3.6: Related entries-->
<!ELEMENT re %om.RO;  ( #PCDATA | sense | %m.dictionaryTopLevel;
                        | %m.phrase; | %m.Incl; )* >
<!ATTLIST re
      %a.global;
```

```
      %a.dictionaries;
      type CDATA #IMPLIED
      TEIform CDATA 're'  >
<!-- end of 12.3.6-->
```

12.4 Headword and Pronunciation References

Examples, definitions, etymologies, and occasionally other elements such as cross references, orthographic forms, etc., often contain a shortened or iconic reference to the headword, rather than repeating the headword itself. The references may be to the orthographic form or to the pronunciation, to the form given or to a variant of that form. The following elements are used to encode such iconic references to a headword:

<oRef> in a dictionary example, indicates a reference to the orthographic form(s) of the headword. Attributes include:

type indicates the kind of typographic modification made to the headword in the reference.

Sample values include:

`cap` indicates first letter is given as capital

`nohyph` indicates that the headword, though a prefix or suffix, loses its hyphen

<pRef> in a dictionary example, indicates a reference to the pronunciation(s) of the headword.

<oVar> in a dictionary example, indicates a reference to variant orthographic form(s) of the headword. Attributes include:

type indicates the kind of variant involved.

Sample values include:

`pt` past tense

`pp` past participle

`prp` present participle

`f` feminine

`pl` plural

<pVar> in a dictionary example, indicates a reference to variant pronunciation(s) of the headword.

As members of the class formPointers, all these elements share a target attribute, which may optionally be used to resolve any ambiguity about the headword form being referred to.

target identifies the orthographic form referred to.

Headword references come in a variety of formats:

~ indicates a reference to the full form of the headword

pref~ gives a prefix to be affixed to the headword

~suf gives a suffix to be affixed to the headword

A~ gives the first letter in upper case, indicating that the headword is capitalized

pref~suf gives a prefix and a suffix to be affixed to the headword

a. gives the initial of the word followed by a full stop, to indicate reference to the full form of the headword

A. refers to a capitalized form of the headword

The <oRef> element should be used for iconic or shortened references to the orthographic form(s) of the headword itself. It is an empty element and replaces, rather than enclosing, the reference. Note that the reference to a headword is not necessarily a simple string replacement. In the example "**colour**1, (US = color) ...~ films; ~ TV; Red, blue and yellow are ~s." [OALD], the tilde stands for either headword form ('colour', 'color').

Examples:

colonel ... army officer above a lieutenant-~. [OALD]

```
<def>army officer above a lieutenant-<oRef/></def>
```

academy ... The Royal A~ of Arts [OALD]

```
<q>The Royal <oRef type="cap"/> of Arts</q>
```

The following example demonstrates the use of the target attribute to refer to a specific form of the headword:

vag- or **vago-** comb form ... : vagus nerve < **vag**al > < **vago**tomy > [W7]

```
<entry>
   <form>
      <orth id="o1">vag-</orth>
      <orth id="o2">vago-</orth>
   </form>
   <!-- ... -->
   <def>vagus nerve</def>
   <eg>
      <q><oRef target="o1" type="nohyph"/>al</q>
      <q><oRef target="o2" type="nohyph"/>tomy</q>
   </eg>
</entry>
```

In many cases the reference is not to the orthographic form of the headword, but rather to another form of the headword — usually to an inflected form. In these cases, the element <oVar> should be used; this element takes as its content the string as it appears in the text.

take ... < Mr Burton **took** us for French > [NPEG]

```
<eg>
   <q>Mr Burton <oVar type="pt">took</oVar> us for French</q>
</eg>
```

take ... < was quite **~n** with him > [NPEG]

```
<eg>
   <q>was quite <oVar type="pp"><oRef/>n</oVar> with him</q>
</eg>
```

The next example shows a discontinuous reference, using the attributes next and prev, which are defined in the additional tag set for linking, segmentation, and alignment (see chapter 14 *Linking, Segmentation, and Alignment*) and therefore require that that tag set be selected in addition to that for dictionaries.

mix up... < it's easy to **mix** her **up** with her sister > [NPEG]

```
<eg>
  <q>it's easy to <oVar next="ov2" id="ov1">mix</oVar> her
     <oVar prev="ov1" id="ov2">up</oVar> with her sister</q>
</eg>
```

In addition, some dictionaries make reference to the pronunciation of the headword in the pronunciation of related entries, variants, or examples. The <pRef> and <pVar> elements should be used for such references.

hors d'oeuvre /ˌaw'duhv (Fr O:r dœvr)/ n, pl hors d'oeuvres also hors d'oeuvre /'duhv(z) (Fr ~)/ [NPEG]

```
<form>
   <orth>hors d'oeuvre</orth>
   <pron>%aU"dUv</pron>
   <form>
      <usg type="lang">Fr</usg>
      <pron id="p2">OR dOvR</pron>
   </form>
</form>
<!-- ... -->
<form type="infl">
   <number>pl</number>
   <orth>hors d'oeuvres</orth>
   <orth>hors d'oeuvre</orth>
   <pron extent="part">"dUv(z)</pron>
   <form>
      <usg type="lang">Fr</usg>
      <pron> <pRef target="p2"/> </pron>
   </form>
</form>
```

Because headword and pronunciaton references can occur virtually anywhere in an entry, the <oRef>, <oVar>, <pRef>, and <pVar> elements can appear within any other element defined for dictionary entries.

Since existing printed dictionaries use different conventions for headword references (swung dash, first letter abbreviated form, capitalization or italicization of the word, etc.) the exact method used should be documented in the header.

The class of headword references is defined thus:

```
<!-- 12.4: Classes for headword references-->
<!ENTITY % a.formPointers '
      target IDREF #IMPLIED'>
<!-- end of 12.4-->
```

The formal declaration for headword reference elements is:

```
<!-- 12.4: Headword references-->
<!ELEMENT oRef %om.RO;  EMPTY>
<!ATTLIST oRef
      %a.global;
      %a.formPointers;
      %a.dictionaries;
      type CDATA #IMPLIED
      TEIform CDATA 'oRef'  >
<!ELEMENT oVar %om.RR;  (#PCDATA | oRef)*>
<!ATTLIST oVar
      %a.global;
      %a.formPointers;
      %a.dictionaries;
      type CDATA #IMPLIED
      TEIform CDATA 'oVar'  >
<!ELEMENT pRef %om.RO;  EMPTY>
<!ATTLIST pRef
      %a.global;
      %a.formPointers;
      %a.dictionaries;
      TEIform CDATA 'pRef'  >
<!ELEMENT pVar %om.RR;  (#PCDATA | pRef)*>
<!ATTLIST pVar
      %a.global;
      %a.formPointers;
      %a.dictionaries;
      TEIform CDATA 'pVar'  >
<!-- end of 12.4-->
```

12.5 Typographic and Lexical Information in Dictionary Data

Among the many possible views of dictionaries, it is useful to distinguish at least the following three, which help to clarify some issues raised with particular urgency by dictionaries, on account of the complexity of both their typography and their information structure.

- (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
- (b) the *editorial view* — the one-dimensional sequence of tokens which can be seen as the input to the typesetting process; the wording and punctuation of the text and the sequencing of items are visible in this view, but specifics of the typographic realization are not
- (c) the *lexical view* — this view includes the underlying information represented in a dictionary, without concern for its exact textual form

For example, a domain indication in a dictionary entry might be broken over a line and therefore hyphenated (“naut-” “ical”); the typographic view of the dictionary preserves this information. In a purely editorial view, the particular form in which the domain name is given in the particular dictionary (as “nautical”, rather than “naut.”, “Naut.”, etc.) would be preserved, but the fact of the line break would

not. Font shifts might plausibly be included in either a strictly typographic or an editorial view. In the lexical view, the only information preserved concerning domain would be some standard symbol or string representing the nautical domain (e.g. “naut.”) regardless of the form in which it appears in the printed dictionary.

In practice, publishers begin with the lexical view — i.e., lexical data as it might appear in a database — and generate first the editorial view, which reflects editorial choices for a particular dictionary (such as the use of the abbreviation “Naut.” for “nautical”, the fonts in which different types of information are to be rendered, etc.), and then the typographic view, which is tied to a specific printed rendering. Computational linguists and philologists often begin with the typographic view and analyse it to obtain the editorial and/or lexical views. Some users may ultimately be concerned with retaining only the lexical view, or they may wish to preserve the typographic or editorial views as a reference text, perhaps as a guard against the loss or misinterpretation of information in the translation process. Some researchers may wish to retain all three views, and study their interrelations, since research questions may well span all three views.

In general, an electronic encoding of a text will allow the recovery of at least one view of that text (the one which guided the encoding); if editorial and typographic practices are consistently applied in the production of a printed dictionary, or if exceptions to the rules are consistently recorded in the electronic encoding, then it is *in principle* possible to recover the editorial view from an encoding of the lexical view, and the typographic view from an encoding of the editorial view. In practice, of course, the severe compression of information in dictionaries, the variety of methods by which this compression is achieved, the complexity of formulating completely explicit rules for editorial and typographic practice, and the relative rarity of complete consistency in the application of such rules, all make the mechanical transformation of information from one view into another something of a vexed question.

This section describes some principles which may be useful in capturing one or the other of these views as consistently and completely as possible, and describes some methods of attempting to capture more than one view in a single encoding. Only the editorial and lexical views are explicitly treated here; for methods of recording the physical or typographic details of a text, see chapter 18 *Transcription of Primary Sources*. Other approaches to these problems, such as the use of repetitive encoding and links to show their correspondences, or the use of feature structures to capture the information structure, and of the ana and inst attributes to link feature structures to a transcription of the editorial view of a dictionary, are not discussed here. (For feature structures, see chapter 16 *Feature Structures*. For linkage of textual form and underlying information, see chapter 15 *Simple Analytic Mechanisms*.)

12.5.1 Editorial View

Common practice in encoding texts of all sorts relies on principles such as the following, which can be used successfully to capture the editorial view when encoding a dictionary:

1. All characters of the source text should be retained, with the possible exception of *rendition text* (for which see further below).
2. Characters appearing in the source text should typically be given as character data content in the document, rather than as the value of an attribute; again, rendition text may optionally be excepted from this rule.
3. Apart from the characters or graphics in the source text, nothing else should appear as content in the document, although it may be given in attribute values.
4. The material in the source text should appear in the encoding in the same order. Complications of the character sequence by footnotes, marginal notes, etc., text wrapping around illustrations, etc., may be dealt with by the usual means (for notes, see section 6.8 *Notes, Annotation, and Indexing*).[99]

[99] Complications of sequence caused by marginal or interlinear insertions and deletions, which are frequent in manuscripts, or by unconventional page layouts, as in concrete poetry, magazines with imaginative graphic designers, and texts about the nature of typography as a medium, typically do not occur in dictionaries, and so are not discussed here.

In a very conservative transcription of the editorial view of a text, *rendition characters* (e.g. the commas, parentheses, etc., used in dictionary entries to signal boundaries among parts of the entry) and *rendition text* (for example, conjunctions joining alternate headwords, etc.) are typically retained. Removing the tags from such a transcription will leave all and only the characters of the source text, in their original sequence.[100]

Consider, for example, the following entry:

> **pinna** ('pIn@) n., pl. -nae (-ni:) *or* -nas. 1. any leaflet of a pinnate compound leaf. 2. *Zoology*. a feather, wing, fin, or similarly shaped part. 3. another name for **auricle** (sense 2). [C18: via New Latin from Latin: wing, feather, fin] [CED]

A conservative encoding of the editorial view of this entry, which retains all rendition text, might resemble the following:

```
<entry>
  <form>
    <orth>pinna</orth>
    <pron>("pIn@)</pron>
  </form>
  <gramGrp> <pos>n.</pos>, </gramGrp>
  <form type="infl">
    <number>pl.</number>
    <form>
      <orth type="lat" extent="part">-nae</orth>
      <pron extent="part">(-ni:)</pron>
    </form>
    or
    <orth type="std" extent="part">-nas</orth>
  </form>
  <sense n="1">1. <def>any leaflet of a pinnate compound leaf.</def> </sense>
  <sense n="2">2.
    <usg type="dom">Zoology</usg>
    <def>a feather, wing, fin, or similarly shaped part.</def>
  </sense>
  <sense n="3">3.
    <xr type="syn">
      <lbl>another name for</lbl>
      <ref target="auricle.2">auricle (sense 2).</ref>
    </xr>
  </sense>
  <etym>[<date>C18</date>: via <lang>New Latin</lang> from
    <lang>Latin</lang>: <gloss>wing</gloss>, <gloss>feather</gloss>,
    <gloss>fin</gloss>]</etym>
</entry>
```

A somewhat simplified encoding of the editorial view of this entry might exploit the fact that rendition text is often systematically recoverable. For example, parentheses consistently appear around pronunciation in this dictionary, and thus are effectively implied by the start- and end-tags for `<pron>`.[101] In such an encoding, removing the tags should exactly reproduce the sequence of characters in the source, minus rendition text. The original character sequence can be recovered fully by replacing tags with any rendition text they imply.

Encoding in this way, the example given above might resemble the following. The `<tagUsage>` element in the header would be used to record the following patterns of rendition text:

- parentheses appear around `<pron>` elements
- commas appear before inflected forms

[100] This is a slight oversimplification. Even in conservative transcriptions, it is common to omit page numbers, signatures of gatherings, running titles and the like. The simple description above also elides, for the sake of simplicity, the difficulties of assigning a meaning to the phrase "original sequence" when it is applied to the printed characters of a source text; the "original sequence" retained or recovered from a conservative transcription of the editorial view is, of course, the one established during the transcription by the encoder.

[101] The omission of rendition text is particularly common in systems for document production; it is considered good practice there, since automatic generation of rendition text is more reliable and more consistent than attempting to maintain it manually in the electronic text.

- the word “or” appears before alternate forms
- brackets appear around the etymology
- full stops appear after <pos>, inflection information, and sense numbers
- senses are numbered in sequence unless otherwise specified using the global n attribute

```
<entry>
  <form>
    <orth>pinna</orth>
    <pron>"pIn@</pron>
  </form>
  <gramGrp> <pos>n</pos> </gramGrp>
  <form type="infl">
    <number>pl</number>
    <form>
      <orth type="lat" extent="part">-nae</orth>
      <pron extent="part">-ni:</pron>
    </form>
    <orth type="std" extent="part">-nas</orth>
  </form>
  <sense n="1"> <def>any leaflet of a pinnate compound leaf.</def> </sense>
  <sense n="2">
    <usg type="dom">Zoology</usg>
    <def>a feather, wing, fin, or similarly shaped part.</def>
  </sense>
  <sense n="3">
    <xr type="syn">
      <lbl>another name for</lbl>
      <ref>auricle (sense 2).</ref>
    </xr>
  </sense>
  <etym>
    <date>C18</date>: via <lang>New Latin</lang> from
    <lang>Latin</lang>: <gloss>wing</gloss>, <gloss>feather</gloss>,
    <gloss>fin</gloss>
  </etym>
</entry>
```

When rendition text is omitted, it is recommended that the means to regenerate it be fully documented, using the <tagUsage> element of the TEI header.

If rendition text is used systematically in a dictionary, with only a few mistakes or exceptions, the global attribute rend may be used on any tag to flag exceptions to the normal treatment. The values of the rend attribute are not prescribed, but it can be used with values such as no-comma, no-left-paren, etc. Specific values can be documented using the <rendition> element in the TEI header.

In the following (imaginary) example, no left parenthesis precedes the pronunciation:

biryani or **biriani** %bIrI"A:nI) any of a variety of Indian dishes ... [from Urdu]

This irregularity can be recorded thus:

```
<entry>
   <form>
      <orth>biryani</orth>
      <orth>biriani</orth>
      <pron rend="noleftparen">%bIrI"A:nI</pron>
   </form>
   <def>any of a variety of Indian dishes ... </def>
   <etym>from <lang>Urdu</lang></etym>
</entry>
```

12.5.2 Lexical View

If the text to be interchanged retains only the lexical view of the text, there may be no concern for the recoverability of the editorial (not to speak of the typographic) view of the text. However, it is strongly recommended that the TEI header be used to document fully the nature of all alterations to the original data, such as normalization of domain names, expansion of inflected forms, etc.

In an encoding of the lexical view of a text, there are degrees of departure from the original data: normalizing inconsistent forms like “nautical”, “naut”., “Naut.”, etc., to “nautical” is a relatively slight alteration; expansion of “delay -ed -ing” to “delay, delayed, delaying” is a more substantial departure. Still more severe is the rearranging of the order of information in entries — for example,

- reorganizing the order of elements in an entry to show their relationship, as in

  clem (klEm) or clam vb. clems, clemming, clemmed or clams, clamming, clammed [CED]

  where in a strictly lexical view one might wish to group “clem” and “clam” with their respective inflected forms.
- splitting an entry into two separate entries, as in

  celi.bacy /"selIb@sI/ n [U] state of living unmarried, esp as a religious obligation. celi.bate /"selIb@t/ n [C] unmarried person (esp a priest who has taken a vow not to marry). [OALD]

  For some purposes, this entry might usefully be split into an entry for “celibacy” and a separate entry for “celibate”.

An encoding which captures the lexical view of the example given in the previous section might look something like the following. In this encoding,

- abbreviated forms have been silently expanded
- some forms have been moved to allow related forms to be grouped together
- the part of speech information has been moved to allow all forms to be given together
- the cross reference to “auricle” has been simplified

```
<entry>
  <form>
    <orth>pinna</orth>
    <pron>"pIn@</pron>
    <form type="infl">
      <number>pl</number>
      <form>
        <orth type="lat">pinnae</orth>
        <pron>'pIni:</pron>
      </form>
      <orth type="std">pinnas</orth>
    </form>
  </form>
  <gramGrp> <pos>n</pos> </gramGrp>
  <sense n="1"> <def>any leaflet of a pinnate compound leaf.</def> </sense>
  <sense n="2">
    <usg type="dom">Zoology</usg>
    <def>a feather, wing, fin, or similarly shaped part.</def>
  </sense>
  <sense n="3">
    <xr type="syn"> <ptr target="auricle.2"/> </xr>
  </sense>
  <etym>
    <date>C18</date>: via <lang>New Latin</lang> from
    <lang>Latin</lang>: <gloss>wing</gloss>, <gloss>feather</gloss>,
    <gloss>fin</gloss>
  </etym>
</entry>
```

12.5.3 Retaining Both Views

It is sometimes desirable to retain both the lexical and the editorial view, in which case a potential conflict exists between the two. When there is a conflict between the encodings for the lexical and editorial views, the principles described in the following sections may be applied.

12.5.3.1 Using Attribute Values to Capture Alternate Views

If the order of the data is the same in both views, then both views may be captured by encoding one 'dominant' view in the character data content of the document, and encoding the other using attribute values on the appropriate elements. If all tags were to be removed, the remaining characters would be those of the dominant view of the text.

The attribute class dictionaries is used to provide attributes for use in encoding multiple views of the same dictionary entry. These attributes are available for use on all elements defined in this chapter when the base tag set for dictionaries is selected.

When the editorial view is dominant, the following attributes may be used to capture the lexical view:

norm gives a normalized form of information given by the source text in a non-normalized form

split gives the list of split values for a merged form

When the lexical view is dominant, the following attributes may be used to record the editorial view:

orig gives the original string or is the empty string when the element does not appear in the source text.

mergedin gives a reference to another element, where the original appears as a merged form.

One attribute is useful in either view:

opt indicates whether the element is optional or not

For example, if the source text had the domain label "naut.", it might be encoded as follows. With the editorial view dominant:

```
<usg norm="nautical" type="dom">naut.</usg>
```

The lexical view of the same label would transcribe the normalized form as content of the <usg> element, the typographic form as an attribute value:

```
<usg orig="naut." type="dom">nautical</usg>
```

If the source text gives inflectional information for the verb 'delay' as "delay, -ed, -ing", it might usefully be expanded to "delayed, delayed, delaying". An encoding of the editorial view might take this form:

```
<form>
   <orth>delay</orth>
   <form type="infl">
      <orth norm="delayed" extent="part">-ed</orth>
      <tns norm="pst,pstp"/>
   </form>
   <form type="infl">
      <orth norm="delaying" extent="part">-ing</orth>
      <tns norm="prsp"/>
   </form>
</form>
```

Note the use of the <tns> tag with null content, to enable the representation of implicit information even though it has no print realization.

The lexical view might be encoded thus:

```
<form>
   <orth>delay</orth>
   <form type="infl">
      <orth orig="-ed">delayed</orth>
      <tns orig="">pst</tns>
      <tns orig="">pstp</tns>
   </form>
   <form type="infl">
      <orth orig="-ing">delaying</orth>
      <tns orig="">prsp</tns>
```

```
   </form>
</form>
```

A particular problem may be posed by the common practice of presenting two alternate forms of a word in a single string, by marking some parts of the word as optional in some forms. The following entry is for a word which can be spelled either “thyrostimuline” or “thyréostimuline”:

thyr(é)ostimuline [tiR(e)ostimylin] ...

With the editorial view dominant, this entry might begin thus:

```
<form>
  <orth split="thyrostimuline, thyr&eacute;ostimuline">thyr(&eacute;)ostimuline</orth>
  <pron split="tiRostimylin, tiReostimylin">tiR(e)ostimylin</pron>
</form>
```

With the lexical view dominant, however, two <orth> and two <pron> elements would be encoded, in order to disentangle the two forms; the orig attribute would be used to record the typographic presentation of the information in the source.

```
<form>
   <orth id="o1" orig="thyr(&eacute;)ostimuline">thyrostimuline</orth>
   <pron id="p1" orig="tiR(e)ostimylin">tiRostimylin</pron>
</form>
<form>
   <orth mergedin="o1">thyr&eacute;ostimuline</orth>
   <pron mergedin="p1">tiReostimylin</pron>
</form>
```

This example might also be encoded using the opt attribute combined with the attributes next and prev defined in chapter 14 *Linking, Segmentation, and Alignment*.

```
<form>
   <orth next="o2" id="o1">thyr</orth>
   <orth next="o3" prev="o1" id="o2" opt="y">&eacute;</orth>
   <orth prev="o2" id="o3">ostimuline</orth>
   <pron next="p2" id="p1">tiR</pron>
   <pron next="p3" prev="p1" id="p2" opt="y">e</pron>
   <pron prev="p2" id="p3">ostimylin</pron>
</form>
```

Note that this transcription preserves both the lexical and editorial views in a single encoding. However, it has the disadvantage that the strings corresponding to entire words do not appear in the encoding uninterrupted, and therefore complex processing is required to retrieve them from the encoded text. The use of the opt attribute is recommended, however, when long spans of text are involved, or when the optional part contains embedded tags.

For example, the following gives two definitions in one text: “picture drawn with coloured chalk made into crayons”, and “coloured chalk made into crayons”:

pas.tel /"p&stl US: p&"stel/ n 1 (picture drawn with) coloured chalk made into crayons. 2... [OALD]

A simple encoding solution would be to leave the definition text unanalysed, but this might be felt inadequate since it does not show that there are two definitions. A possible alternative encoding would be:

```
<sense n="1">
   <def>coloured chalk made into crayons</def>
   <def>picture drawn with coloured chalk made into crayons</def>
</sense>
```

This transcribes some characters of the source text twice, however, which deviates from the usual practice. The following encoding records both the editorial and lexical views:

```
<sense n="1">
   <def next="d2" id="d1" opt="y">picture drawn with</def>
   <def prev="d1" id="d2">coloured chalk made into crayons</def>
</sense>
```

A more complex example is the following, in which the optional element contains additional tags:

canary ...(Geog) Canary Isles, Canaries (îles fpl) Canaries fpl... [C/R]

```
<re type="cmpd">
   <usg type="dom">Geog</usg>
   <form>
      <orth>Canary Isles</orth>
      <orth>Canaries</orth>
   </form>
   <trans next="t2" id="t1" opt="y">
      <tr>&icirc;les</tr>
      <gen>f</gen>
      <number>pl</number>
   </trans>
   <trans prev="t1" id="t2">
      <tr>Canaries</tr>
      <gen>f</gen>
      <number>pl</number>
   </trans>
</re>
```

12.5.3.2 Recording Original Locations of Transposed Elements

The attributes described in the previous section are useful only when the order of material is the same in both the editorial and the lexical view. When the two views impose different orders on the data, the ID/IDREF mechanism may be used to show the original location of material transposed in an encoding of the lexical view.

If the original is only slightly modified, the <anchor> element may be used to mark the original location of the material, and the location attribute may be used on the lexical encoding of that material to indicate its original location(s). Like those in the preceding section, this attribute is defined for the attribute class dictionaries:

opt indicates whether the element is optional or not

For example:

pinna ("pIn@) n., pl. -nae (-ni:) or -nas. [CED]

```
<form>
   <orth>pinna</orth>
   <pron>'pIn@</pron>
   <anchor id="p1"/>
   <form type="infl">
      <number>pl</number>
      <form>
         <orth extent="part">-nae</orth>
         <pron extent="part">-ni:</pron>
      </form>
      <orth extent="part">-nas</orth>
   </form>
</form>
<gramGrp>
   <pos location="p1">n</pos>         <!-- moved -->
</gramGrp>
```

12.5.4 Attributes for Dictionary Elements

The attributes provided for all dictionary-specific elements and documented in this section are defined thus:

```
<!-- 12.5.4: Attributes for dictionary work-->
<!ENTITY % a.dictionaries '
      expand CDATA #IMPLIED
      norm CDATA #IMPLIED
      split CDATA #IMPLIED
      value CDATA #IMPLIED
      orig CDATA #IMPLIED
      location IDREF #IMPLIED
      mergedin IDREF #IMPLIED
```

```
      opt (y | n) "n"'>
<!-- end of 12.5.4-->
```

12.6 Unstructured Entries

The content model for the <entry> element provides an entry structure suitable for many average dictionaries, as well as many regular entries in more exotic dictionaries. However, the structure of some dictionaries does not allow the restrictions imposed by the content model for <entry>. To handle these cases, the <entryFree> and <dictScrap> elements are provided to support much wider variation in entry structure. The <dictScrap> element offers less freedom, in that it can only contain phrase level elements, but it can itself appear at any point within a dictionary entry where any of the structural components of a dictionary entry are permitted. As such, it acts as a container for otherwise anomalous parts of an entry.

The <entryFree> element places no constraints at all upon the entry: any element defined in this chapter, as well as all the phrase-level and inter-level elements defined in the core tag set, can appear anywhere within it. With the <entryFree> element, the encoder is free to use any element anywhere, as well as to use or omit grouping elements such as <form>, <gramGrp>, etc.

The <entryFree> element allows the encoding of entries which violate the structure specified for the <entry> element. For example, in the following entry from a dictionary already in electronic form, it is necessary to include a <pron> element within a <def>. This is not permitted in the content model for <entry>, but it poses no problem in the <entryFree> element.

```
<ent h='demigod'><hwd>demi|god</hwd>
<pr><ph>"demIgQd</ph></pr>
<hps ps='n'>
<hsn><def>one who is partly divine and partly human</def>
<def>(in Gk myth, etc) the son of a god and a mortal woman,
eg<cf>Hercules</cf><pr><ph>"h3:kjUli:z</ph></pr></def>
</hsn></hps></ent>

[OALD electronic]
<entryFree>
  <form>
    <orth>demigod</orth>
    <hyph>demi|god</hyph>
    <pron>"demIgQd</pron>
  </form>
  <gramGrp>
    <pos>n</pos>
  </gramGrp>
  <def>one who is partly divine and partly human</def>
  <def>(in Gk myth, etc) the son of a god and a mortal woman,
       eg <mentioned>Hercules</mentioned>
    <pron>"h3:kjUli:z</pron></def>
</entryFree>
```

The <entryFree> element also makes it possible to transcribe a dictionary using only phrase-level ('atomic') elements—that is, using no grouping elements at all. This can be desirable if the encoder wants a completely 'flat' view, with no indication of or commitment to the association of one element with another. The following encoding uses no grouping elements, and keeps all rendition text:

biryani or **biriani** (%bIrI"A:nI) any of a variety of Indian dishes...[from Urdu] [CED]

```
<entryFree>
   <orth>biryani</orth> or
   <orth>biriani</orth>
   <pron>(%bIrI"A:nI)</pron>
   <def>any of a variety of Indian dishes ...</def>
   <etym>[from <lang>Urdu</lang>]</etym>
</entryFree>
```

Here is an alternative way of representing the same structure, this time using <dictScrap>:

```
<entry>
  <dictScrap>
    <orth>biryani</orth> or
    <orth>biriani</orth>
    <pron>(%bIrI"A:nI)</pron>
    <def>any of a variety of Indian dishes ...</def>
    <etym>[from <lang>Urdu</lang>]</etym>
  </dictScrap>
</entry>
```

Declarations for <dictScrap> and for <entryFree> were given above in section 12.2 *The Structure of Dictionary Entries*.

```
<!-- 12.6: Model class for unstructured dictionary entries-->
<!--This entity declares the class of elements defined
specifically for use in dictionary entries, except those which are
included in the phrase class. This class
is used in defining the 'free'
dictionary entry.-->
<!ENTITY % x.dictionaryParts "" >
<!ENTITY % m.dictionaryParts "%x.dictionaryParts; %n.case; | %n.colloc;
| %n.def; | %n.eg; | %n.etym; | %n.form; | %n.gen; | %n.gramGrp; |
%n.hom; | %n.hyph; | %n.itype; | %n.lbl; | %n.mood; | %n.number; | %n.orth; |
%n.per; | %n.pos; | %n.pron; | %n.re; | %n.sense; | %n.stress; | %n.subc; |
%n.superEntry; | %n.syll; | %n.tns; | %n.tr; | %n.trans; | %n.usg; | %n.xr;">
<!-- end of 12.6-->
```