Encoding a Historical Dictionary with the TEI (With reference to the Electronic Scottish National Dictionary Project)

Susan RENNIE, Edinburgh, UK

Abstract

This paper examines the practicalities of applying the Text Encoding Initiative (TEI) guidelines to the encoding of historical dictionaries, with particular reference to the *Electronic Scottish National Dictionary* (*eSND*) project. It discusses in detail the features of the *eSND* which are not specifically covered by the current TEI guidelines, and offers practical and simple solutions which historical lexicographers can use to adapt and extend the current TEI coding scheme. The paper will assume the audience has a basic awareness of SGML or XML mark-up, but does not require a detailed knowledge of the TEI encoding scheme.

1 The eSND Project

The *Electronic Scottish National Dictionary* (*e-SND)* project aims to produce an electronic, Internet version of the *Scottish National Dictionary* (*SND*). The *SND* is a 10-volume, historical dictionary which covers the modern Scots language from 1700 to the present[1]. Compilation of the *SND* began in 1929, under the editorship of William Grant, and was continued on Grants death in 1946 by David Murison, who edited the bulk of the dictionary (roughly from the letter D onwards). Publication was in fascicles, with the last sections (comprising a Supplement to the earlier volumes) being published in 1976. The first stage in the *eSND* project has been to digitise the 40 fascicles comprising the main text and original Supplement of the *SND*, together with the original Prelims and Appendices. The unbound fascicles have been sent to the Centre for Data Digitisation and Analysis (CDDA) at Queen's University in Belfast (http://wwwparent.qub.ac.uk/ss/csr/cdda), where they are being scanned and passed through Optical Character Recognition (OCR) software. The data is returned in the form of MS Word documents which are subsequently proofread and then converted to XML format (this 'up-conversion' is achieved in stages by running a series of computer programs followed by close editorial checking). The original *SND* Supplement entries will be incorporated into the main text, but will retain information about their source (see 3. below). Although we will also be including a certain amount of updatings – especially more recent citations – in the *eSND*, these will echo the style of the original Supplement and be designed to fit into the existing *SND* structure. We are therefore digitising an existing 'legacy' text, which we do not want to change substantially; any adaptation or compromises to make our text 'fit' the TEI scheme had to take place in the mark-up, not in the dictionary text itself.

2 The TEI Encoding Scheme

The TEI guidelines (http://www.tei-c.org) are now the recognised international standard for encoding scholarly texts in SGML (and now also XML) format, and several major historical dictionary projects – notably the electronic *Middle English Dictionary* (http://www.hti.umich.edu/dict/med) and the *Electronic Johnson* (http://www.hti.umich.edu/english/johnson) – have now used the TEI to encode their texts.

The *eSND* has followed this example. We decided at the outset of the project to encode our data in XML, the emerging standard for encoding structured text. Although the TEI was originally conceived as an SGML implementation, most of the recommendations apply equally well to XML documents and we were not put off by the task of converting our subset of theTEI to make it XML-compliant[2].

In the course of our project we have come across features of the *SND* (some of which are common to many historical or scholarly dictionaries) which do not easily fit into the existing TEI guidelines for encoding dictionaries. This situation was to some extent anticipated by the TEI editors; the TEI has a built-in procedure for modifying its rules within certain parameters[3]. There are detailed procedures for incorporating such changes and it is possible to follow the recommendations below and still produce a TEI-conformant text.

In encoding the *eSND*, we have tried to maintain the overall structure and spirit of the TEI coding scheme, and have only deviated from the recommended guidelines if there was no other sensible way to represent the information in our dictionary. We therefore used two basic strategies to adapt the TEI (both of which are allowed for in the TEI modification scheme):

1) We extended the contexts where existing TEI elements or attributes could occur.

2) We created new elements or attributes where nothing appropriate was found in the existing TEI tagsets. Wherever possible for these, we used an existing TEI name and only created a new name when there was no other sensible option.

3 The TEI Dictionary Tagset

The TEI is a modular system, comprising a core tagset, with guidelines for its appropriate use, and several so-called ‘base’ tagsets for encoding particular types of text[4]. There is a specific base tagset for Print Dictionaries, which is fully documented in section 12 of the TEI Guidelines. Although the dictionary set was the starting-point for encoding the *eSND*, we have also made extensive use of certain ‘core’ TEI tags. The TEI <corr> tag, for example, gave us an off-the-shelf system for describing editorial changes to the printed *SND* text, from correcting simple typos or deleting ghost cross-references to correcting erroneous information. The original, erroneous text, is retained as an attribute of the <corr> element, thus:

. . . Robert <corr orig=‘Burs’> Burns </corr>

Although this might seem rather over-the-top for trivial corrections, it allows us a fast and efficient means to check that all necessary corrections to the *SND* (currently noted on a 'master copy' of the text) have been carried out. Further, the TEI <name> tag allows us to mark up references to significant names that occur outwith citations. It is helpful for electronic searching, for example, if any references to Burns the poet in etymologies or definitions are distinguished from references to 'burns' in its Scots meaning of 'streams'.

In addition to allowing <corr> and <name> tags to appear at any point in the text, we have added a global attribute called 'source', which can be attached to any single *eSND* element. This has stated values of 'SND', 'Supplement' or 'New Data' and thus allows us to identify any element as coming from the *SND* main text or Supplement, or from the New Data files. Although the Supplement and New Data information will be merged editorially with the main *SND* text, the changes can therefore still be identified and searched for.

The following specific features of the *SND* will be discussed with reference to the shortcomings of the TEI. In each case, specific examples will be discussed, and details will be given of the solutions adopted for the *e-SND*:

1) usage information
2) pronunciations
3) citations
4) etymologies

4 Usage Information

The recommended TEI encoding for any kind of usage information is to use the generic <usg> element [tei p3 12.3.5.2]; particular kinds of usage information can be distinguished by the use of a 'type' attribute, thus:

<usg type='geo'> Scot. </usg> or <usg type='domain'> law </usg>

However, two categories of usage information in the *SND* seemed to require more emphasis than this generic treatment allowed: firstly, the ubiquitous regional labels (abbreviations of the old Scottish county names, for example Abd for Aberdeenshire), used in the *SND* to indicate any of the many dialects of Scots; and secondly, typographic symbols which were used throughout to indicate obsolete, obsolescent or nonce usage. Both of these categories of usage information can occur anywhere in an *SND* entry, and are sometimes combined with each other to indicate that a word is now obsolete or obsolescent in a particular region.

4.1 Regional Labels

The Scots language is rich in regional dialects and, in compiling the *SND*, information on current usage was collected from every Scots-speaking region, from Orkney and Shetland, the North-East and Central Lowlands to the Borders and the South-West of Scotland, and also from Northern Ireland for information on Ulster Scots usage. Any single piece of lexicographical information in the SND can potentially be assigned a regional label to indicate its restriction to that area.

The prevalence and relative importance of these regional labels, and the special way in which they behaved in the *SND*, led us to decide give them an element name of their own. We therefore upgraded ‘geo’ from being merely an allowable value of the TEI ‘usg’ attribute to being a full element name, and allowed it to legally appear at any point in an entry. The flexibility which this solution gave us can be seen in the sample encodings of regional citations and pronunciations given below.

4.2 Usage Symbols

Like many historical dictionaries, the *SND* uses typographical markers for usage information: specifically, a dagger (†) to indicate ‘obsolete’ usage, a double dagger (‡) for ‘obsolescent’, a paragraph symbol (t’, a paragraph symbol (¶) for ‘nonce’, and an asterisk (*) for ‘hypothetical’ (as in unattested forms in etymologies, or for reconstructed historical pronunciations). This usage information can occur at any point within an entry: qualifying a headword or variant orthographic form, a sense definition, a pronunciation or regional label. More unusually, they can also qualify a bibliographic reference, or even a single word in a definition.

The entry for **laich**, for example (see figure 1) has 3 nonce forms, and 1 obsolete and 1 obsolescent form in the section for orthographic variants alone; the pronunciation section gives one obsolete and two obsolescent pronunciations. Later on in the entry, the double dagger is used to qualify a bibliographic reference (in this case, George Watson’s Roxburghshire *Word Book*), indicating that the relevant form was described as being obsolescent in that source.

There were two possible, conservative encodings that would not require any modification to the TEI: 1) to treat the symbols typographically and code them simply as special characters; and 2) to surround the symbols themselves with <usg> tags. The first of these approaches ignored the significance of the symbol, and the second failed to express any relation between the symbol and the information to which it was attached. We decided, rather, to take a more radical approach. The *SND*’s use of these symbols is more akin to an SGML or XML attribute than to an element. We therefore simply created a new attribute, called **usg**, which would be a global attribute: that is, it could be attached to any element from an entry downwards. This approach (of repurposing an existing TEI element name as an attribute) is analogous to the TEI use of <lang> as an element when the information is explicit in the dictionary, and as an attribute when it is not (for example, in cited forms in etymologies).

For the *eSND*, we allow the following values for the **usg** attribute:

1) obs (for obsolete words)
2) obsol (for obsolescent words)
3) nonce (for some other dictionaries, this could be ‘dubious’ or ‘erroneous’)
4) hyp (for hypothetical forms in etymologies, or for reconstructed historical pronunciations)

The various forms and variant pronunciations in LAICH can therefore be neatly encoded using the new, global usage attribute:

<entry> <form type="lemma"> laich </form> ... Also <form type="variant"> laigh </form> , <form type="variant"> leagh </form> , <form type="variant" usg="nonce"> laihh- </form> ... <form

type="variant" usg="obs"> lay- </form> ; <form type="variant" usg="obsol"> lyaach </form> ... <pron> <geo> Sc. </geo> <phon> le:ç </phon> , <geo> Abd. </geo> + <phon usg="obs"> ljα:x </phon> , <geo> s.Sc. </geo> <phon usg="obsol"> ljux&turnw; </phon> , <phon usg="obsol"> lα: </phon> ... </pron>

In the case of single words in definitions being ascribed an obsolete marker, we simply allowed for a split in the <def> tags. The entry for OST, for example, includes an obsolescent part-definition, which we coded thus:

<pos> vbl.n. </pos> <form type ="derivative"> ostin </form> ... <def> curdled milk </def> ..., <def usg="obsol"> sometimes used as a poultice </def> .

5 Pronunciations

The current TEI guidelines for dictionaries assume that pronunciations will a) follow closely upon the corresponding orthographic form; and b) contain only phonetic information [tei p3 12.3.1]. The recommended TEI markup therefore is to place the <pron> element immediately after the corresponding <orth> element, which is always nested within a <form> element.

However, in the *SND*, the pronunciation of the main form, and of all variants, follows the list of orthographic variants in a separate section (see figure 2); further, not all given pronunciations have a corresponding orthographic form. Our <pron> elements, therefore, have to be allowed to exist outwith the <form> s. This obviously negates the advantages of the TEI nesting system in that it provides no link from form to pronunciation. We intend in the next phase of mark-up to insert such links (and thus to add information that cannot easily be derived from the current *SND*), by giving individual pronunciations an id number that can be related to the appropriate form. A simpler alternative might also be to reorder the entries to a more TEI-conformant model.

Also, in common with other historical dictionaries, the *SND* often qualifies individual pronunciations, for example by regional labels or by usage symbols (see, for example, the entry for LAIRD, figure 3 above); even bibliographic information can occur within the pronunciation section.

One approach we considered was to treat any qualifying information as an attribute of the <pron> element, eg <pron usg=Abd> </pron> . However, this was quite clumsy, especially when we came to treating bibliographic information with many sub-elements.

Our solution was therefore to create a new element, called <phon> , which would surround only the actual phonetic transcription. This freed the <pron> element to behave in the same way as the TEI <etym> element for etymologies (see below), enclosing *all* the given pronunciation information, not just individual transcriptions. It also allowed us to apply our new usage attribute (see 2. above) to the <phon> element only, so that hypothetical pronunciations could be easily encoded.

The obsolete Aberdonian pronunciation of LAIRD can now be coded thus:

<pron> <phon> le:rd </phon> , <geo> Abd. </geo>
+ <phon usg="obs"> lja:rd </phon> ... </pron>

Figure 1: Dictionary entries 1, 2 and 3

A further logical stage would be to group the individual pronunciations with all accompanying information, for example by creating a new element called <pronGrp> (by analogy with the TEI <gramGrp> which is used to enclose any kind of grammatical or morpho-syntactic information [see tei p3 12.3.2]):

<pronGrp> <pron> <phon> le:rd </phon> </pron> , <pron> <geo> Abd. </geo> + <phon usg=obs> lja:rd </phon> </pron> ; <pron> in sense 1.
<phon> l&ah;rd </phon> </pron> </pronGrp>

This second encoding would require considerably more editorial checking and is being considered for future phases of the project.

6 Citations

Because of the central importance of dialect information in the *SND*, all citations are assigned a particular regional label; if the citation is from a source considered to be non-dialectal, the 'regional' label is given as 'general Scots' (abbreviated to 'Sc.'). The entry for **malagrugrous**, for example (see below, figure 4), has one 'general Scots' citation (from *Blackwood's Magazine*), and two citations assigned to regional dialects of Scots (an Ayrshire example from a John Galt novel, and an Aberdonian one from a local periodical).

The works of some well-known authors are ascribed to a particular region, if it was felt that their work was predominantly in that dialect; citations from Burns and Galt, therefore, are typically labelled as Ayrshire usage, even if the word they illustrate has a wider provenance, though citations from Stevenson will typically have a 'general Scots' designation.

Although the TEI does not allow for this syntax (the TEI <cit> element was intended to contain only bibliographic information), this potential violation had already been defused by the solutions put in place to deal with usage information. As we had allowed the new <geo> element free passage within any *eSND* entry, there was no difficulty in allowing it to appear within citations.

However, the *SND* also includes many oral, as well as written, citations from attributed sources. These comprise transcriptions from speech, either overheard by, or spoken to, one of the many 'contributors' to the SND. Oral sources are identified in the *SND* by a superscript number assigned to their respective regional area. The entry for **malafooster**, for example (see below, figure 5), includes an oral citation from a Perthshire contributor.

The superscript number identifies a particular contributor to the *SND*, whose full name is given in a list in the *SND* Prelims (in the *eSND*, the oral-source references will eventually be hyperlinked to their expansions in this list). Some of the anonymous identity numbers mask quite famous contributors; the eminent lexicographer, William Craigie, for example (one of the original editors of the *OED*, who also initiated the compilation of *DOST*), is labelled as simply Ags.[6], and the contributions of the first *SND* editor, William Grant, are modestly ascribed to Mry.[2] and Kcb.[5].

Oral-source citations of this kind are not specifically dealt with in the TEI, which allows for only printed (or manuscript) citations, or contrived examples (which are rare in historical dictionaries) [see tei p3 12.3.5.1]. In a sense, these citations comprise a mini spoken corpus of Scots

(with admittedly very small samples). Using the TEI-recommended tags for corpora, however, was not a serious option: we had no idea how accurately the quotes had been transcribed, nor who the original speakers were, and our only honest recourse, therefore, was to define them as quotes ascribed to the contributor.

The scheme we devised to encode oral-source citations was as follows: 1) we added a new 'type' attribute with two posible values of 'oral' or 'written' to the TEI <cit> element (so that all citations of either kind could be easily isolated for searching); and 2) we gave our new <geo> tag an 'id' attribute[5] , which would repeat the contributor's superscript number (and could therefore form the basis of future hyperlinks). In this encoding scheme, the citation describing the unfortunate snowman becomes:

<cit> <geo id="4"> Per. </geo> <date> 1950 </date> :

<q> The big laddie's malafoostered oor snowman. </q>

As well as representing the lexicographic structure of the *SND*, this solution made it possible to use the same query syntax to find regional labels occurring at any point in an *eSND* entry.

One other slight modification we allowed ourselves was to relax the TEI rules for where citation and quotation elements could occur within an entry. As noted above, bibliographic references can sometimes occur after individual pronunciations; more often, they are assigned to orthographic variants (see below, figure 6). And occasionally, especially for nonce words which are only evidenced in a glossary or other reference source, the *SND* definition is itself a citation[6] (see below, figure 7).

7 Etymologies

The TEI dictionary guidelines recommend tagging separate elements of etymologies, using multipurpose TEI tags. In particular, the guidelines state that the 'variation in etymological structure makes it impractical to define tags which capture the entire intellectual structure of the etymology or record the precise interrelation of all the words mentioned' [tei p3 12.3.4]. This means that each gloss is tagged as an individual and not attributed to its own target. For example, the etymology for KRINGLE (see below, figure 8) would be encoded thus:

<etym> <lang> Norw. </lang> <mentioned> kringel </mentioned> , <gloss> a circle, ring </gloss> , <mentioned> kringla </mentioned> , <gloss> to lay or place in a circle </gloss> , <lang> O.N. </lang> <mentioned> kringla </mentioned> , <gloss> a circle </gloss> ... </etym>

We felt that this was an unsatisfactory way to treat such a crucial feature of historical dictionaries; it also precluded us from allowing users sophisticated searching within etymologies, eg for etyma with specific glosses or source languages.

Our interim solution has been to make use of the existing TEI global attribute for 'lang' (from the core tagset) to link an etymon to its source language by repeating the content of the <lang> element as an attribute of <mentioned> . For example:

<etym> <lang> Norw. </lang> <mentioned lang="norw"> kringel </mentioned> . . . etc

However, a further logical step (by analogy with the proposed system for pronunciations above) would be to create a 'group' element, called possibly <etGrp> , which would enclose an etymon along with its source language and gloss. The above etymology would thereby become:

Figure 2: Dictionary entries 4 – 8

<etym> <etGrp> <lang> Norw. </lang> <mentioned> kringel </mentioned> , <gloss> a circle, ring </gloss> </etGrp> , <etGrp> <mentioned> kringla </mentioned> , <gloss> to lay or place in a circle </gloss> </etGrp> , <etGrp> <lang> O.N. </lang>
<mentioned> kringla </mentioned> , <gloss> a circle </gloss> </etGrp> </etym>

For the initial phase of the *eSND*, we have followed the TEI recommendation to tag all cited forms in etymologies – including etyma and cognates – with the <mentioned> element. Although we were not happy with the use of the multi-purpose <mentioned> element, as opposed to a unique element for cited forms in etymologies, we felt there were not sufficient grounds to change the element name, given that its use is recommended in the TEI guidelines for this purpose. However, our eventual aim is to sub-divide cited forms to identify etyma and cognates (possibly also to assign chronological values to the etyma, identifying the most recent and the earliest). At that stage, we will need to either replace the <mentioned> element in these cases with an <etymon> or <cognate> element[7], or (less radically) add a 'type' attribute to the <mentioned> element with a value of 'etymon' or 'cognate'. However, these distinctions are not practical for automatic tagging and will need to be done editorially at a future revision stage.

8 Conclusion

If the TEI is to be successful as a standard, it is the responsibility of lexicographic projects such as ours to identify shortcomings and suggest revisions of the guidelines. Our experience has been that the TEI is flexible enough to allow for many of the structural idiosyncrasies of the *SND*. Where problems have occurred, we have found it possible to adapt the TEI minimally, to create new coding syntax which draws on analogies elsewhere in the TEI to form natural extensions of the scheme. As mentioned above, the TEI already allows for a certain degree of individual user modification. However, most of the *eSND* modifications have been necessary for features that are not unique to that dictionary, but which are found in many historical or scholarly dictionaries. The *English Dialect Dictionary* [8], for example, includes regional attributions within its citations; the *Dictionary of American Regional English*[9] includes many oral as well as written citations, and often qualifies individual pronunciations with regional labels; and the use of typographical markers to indicate usage information is commonplace in historical lexicography. Plans have been mooted to revive the TEI dictionary group and if such were to happen, a special set of recommendations could be included next time around to cover the needs of scholarly dictionaries.We hope, then, that the experience of the *eSND* and the solutions we have arrived at for our own project can help to stimulate discussions among historical dictionary projects so that common difficulties with the TEI can be established and some possible solutions agreed, making it more likely that such modifications may be incorporated in the future into this international standard.

Notes

[1]The earlier history of the Scots language is covered by the *SND*'s sister dictionary, the *Dictionary of the Older Scots Tongue* (*DOST*), which is currently nearing completion. Plans are in now hand to digitise *DOST* using the same XML/TEI mark-up scheme as the *eSND*, and so to create a single electronic

resource, covering the Scots language from the earliest records to the present, which will share the same search tools and interface.

[2]Since the start of the eSND project, XML-compliant versions of the TEI DTDs have been published on the Internet: see www.tei-c.org. The *eSND* has also been fortunate to have on our backdoorstep the Language Technology Group (LTG) of Edinburgh University (www.ltg.ed.ac.uk), who are involved in XML research and the creation of XML editing and search tools. The LTG are currently working with us on the prototype search mechanisms and web interface for the *eSND*.

[3]Projects which have used and published such modifications to the TEI scheme include the Women Writers Project at Brown University (http://www.brown.edu) and the Corpus Encoding Standard (http://www.cs.vassar.edu/CES).

[4]This modular approach has led to the analogy of the TEI 'pizza chef' system, which allows users to choose combinations of TEI tagsets: see http://www.hcu.ox.ac.uk/TEI/newpizza.html.

[5]id is an existing TEI global attribute (see tei p3 3.5)

[6]This is also a feature of Johnson's *Dictionary* and a similar modification of the TEI was needed to encode the Electronic Johnson. See user booklet accompanying Anne MacDermott ed., *Johnson's Dictionary of the English Language on CD-ROM* (Cambridge University Press, Cambridge, 1996), p. 49.

[7]This approach is also being considered by the *Electronic Middle English Dictionary* project. At the time of writing this paper, the mark-up scheme for the*eMED* had not been finalised. I am grateful to the *eMED* editors for supplying me with information on their mark-up and methodology.

[8]Joseph Wright ed., *The English Dialect Dictionary* (Henry Frowde, London, 1898-1905). Although a project was recently mooted to digitise the *EDD*, at the time of writing, this has been postponed due to lack of funding.

[9]Cassidy, Frederic and Hall, Joan Houston eds., *Dictionary of American Regional English.* Harvard University Press, Cambridge, Massachusetts, 1985-. See also http://polyglot.lss.wisc.edu/dare/dare.html

References

[Grant *et al.* 1976] Grant, William and Murison, David, eds., *The Scottish National Dictionary.* Scottish National Dictionary Association, Edinburgh, 1931-1976.

[Sperberg-McQueen] Sperberg-McQueen, C.M. and Bernard, Lou, eds., *Guidelines for Electronic Text Encoding and Interchange* (TEI P3). Full documentation and electronic copies of the TEI Guidelines are available on the TEI website: http://www.tei-c.org. An excellent guide to the TEI is also available from the University of Virginia's Electronic Text Center at http://etext.lib.virginia.edu/tei/uvatei.html.