Bruce Fraser, 2005. ***Author-tagging: a dictionary DTD as part of the writing environment***.

Published online 2007 at *http://www.chlt.org/lexicon/papers/Author-Tagging.pdf*

Because markup languages are capable of mapping every detail of a text, they are especially useful for tagging dictionaries, which are among the most densely-structured of all texts. However, DTDs which have been developed for encoding a range of dictionaries must have a wide scope, with a consequent loss of descriptive power. Even a DTD which has been tailored for a specific dictionary must still allow a great level of flexibility to capture the variations in its structure.

This paper describes how greater precision can be attained if the dictionary editors themselves tag as they compose. Author-tagging not only enables the markup to be designed according to a specific lexicographic approach, but also has a number of significant advantages for the editors: by applying constraints to the writing process, it facilitates the development of a consistent style, and it constitutes a lexicographic tool, by allowing the editors to search the XML documents during composition, and to include annotations and other revision material.

The discussion is illustrated with examples taken from a DTD and its associated XSL stylesheets which have been developed for a Greek-English lexicon currently being written at Cambridge University.

Dictionary-writing in an XML environment

The tension between generality and descriptive power in a widely-applicable dictionary DTD, such as that of the Text Encoding Initiative (TEI),[1] has been described by Ide and Véronis (1995), Tutin and Véronis (1998) and others. Although much work has been done on customising that DTD (see Kang 1997, Tutin and Véronis 1998), it tends to be done by relaxing constraints, as by the use of the TEI's generalised structure 'entryFree', instead of the standard 'entry'. It is argued here that, on the contrary, more rigidity is needed.

That can be achieved only by designing a specific system for each dictionary. A notable example is the SGML tagging developed for the second edition of the *OED* (described by Simpson and Weiner 1989, Berg 1992, Tompa 1993), which provides a clear mapping of Murray's structure, because it was designed specifically for the task. However, even such a DTD which maps to an existing structure must incorporate considerable flexibility.

More complex variations can be achieved by the use of typesetting programs, such as the TEX program which has been used for a series of Icelandic dictionaries (described by Pind 1990). The adoption of XML and XSL-FO as publishing standards raises the possibility that fairly simple structures can adequately map lexicographic composition style.

The key seems to be to develop the lexicographic details and tagging structure simultaneously, with the writers involved in the development of the DTD, by composing articles in the emerging digital environment. Such 'author tagging' also brings major benefits to the writers, which offset the extra work involved in following the discipline of tagging, because the DTD and the styling can be configured to help the writers maintain their chosen lexicographic methodology. Precision can be combined with flexibility by making the DTD hierarchically-constrained (that is, progressively more flexible down the levels of structure) and by using XSLT styling rules to supplement the structure of mixed-content elements, as described below.

XML tagging

User-directed benefits

The user-directed advantages of tagging are well-known. From the publishing viewpoint, articles can be produced with precisely-designed formats and consistently-organised information, so minimal typesetting is required, and the structure can be easily adapted for revised and adapted editions. For the reader, tagging allows the electronic editions to be searched. For the research linguist, tagging maximises information, even when there is no space in the print editions: a dictionary can even be linked to a textual databank.

Author-directed benefits

By contrast, the author-directed benefits can easily be overlooked, because tagging has usually been treated as a very different operation from composition. This may be appropriate for literary works, but is less relevant to dictionaries, where the writing is unusually condensed and structured. The advantages of author-tagging outweigh the disadvantages, for four major reasons:

1) A tagged structure helps the writers maintain a consistent style. The digital environment actually constitutes a teaching tool for the authors, who are constrained (within tolerable limits) to follow a template for each entry. This is particularly useful for dictionaries which have a collaborative authorship, and so it is especially difficult to maintain a 'house-style'.

2) Citation forms. If the work is tagged from the start, the editors can search the documents as they are being written. This helps cross-referencing and consistency. By contrast, in the large Liddell and Scott Greek lexicon (LSJ), a single passage from Euripides *Antiope* is cited in five different ways: '*Antiop*.iv B, 46 A, *Antiop*.iv B line 46 Arn., *Antiop*.iv B line 46 Arnim, *Antiop*.p.21 A, and *Antiop*. B 58 p.21 A'. Other inconsistencies include the citation forms of the Homeric Hymns (sometimes cited by number, sometimes by their traditional Latin titles). Consistency could have been maintained if the authors of LSJ had been able to compare all their citations easily; tagging the articles makes this a simple task.

3) Authorial searches also allow the writers to take account of the relationship between meaning and word formation, which is especially important in a highly-inflected language like Ancient Greek, where compounds are common. It is useful to be able to compare the article for (say) any of the 18 compounds of the verb βαίνω, go (ἀναβαίνω, ἀντιβαίνω, and so on, down to ὑποβαίνω) as a group, during the process of composition. And if entries are tagged with a semantic 'key word', the vocabulary can be collated in semantic groupings, so the writers can easily compare their entries for semantically-related words, such as those describing shapes, colours, or sounds.

4) It is simple to add authorial annotations. In a tagged system, notes and queries on semantic and textual points can be filed with each entry in the same document, and so be easily accessible to the other editors and reviewers. And their comments can accompany the notes in the revised drafts. We find this feature particularly useful in our Greek-English lexicon, because dictionary-writing is so condensed that much information gathered during the research is usually lost.[2]

Tagging and semantics

Our Ancient Greek lexicon was originally planned to be an updated version of the intermediate-size student lexicon of Liddell and Scott. However, the example of the Oxford Latin Dictionary (*OLD*) suggested a new lexicographic approach, similar to the *OED* in concentrating on semantic principles, rather than syntactic forms.[3] Consequently, the entry sections describe semantic groupings, which are (as in *OLD* and *OED*) illustrated by explanatory definitions as well as single-word translations, and accompanied by examples of their textual contexts.

A DTD was developed to map the structure, and has itself led us to modify the lexicographic methodology, in a beneficially-circular process. Great attention has been paid to creating a structure which is maximally constrained, yet flexible enough to accommodate variations in word form, meaning, and function. Feedback is provided through XSL-FO output, so we can view the entries in print-ready form, and, as we have a dedicated citation database, we can instantly refer to the textual citations. The tagging also allows us to make authorial annotations as we write.

Examples

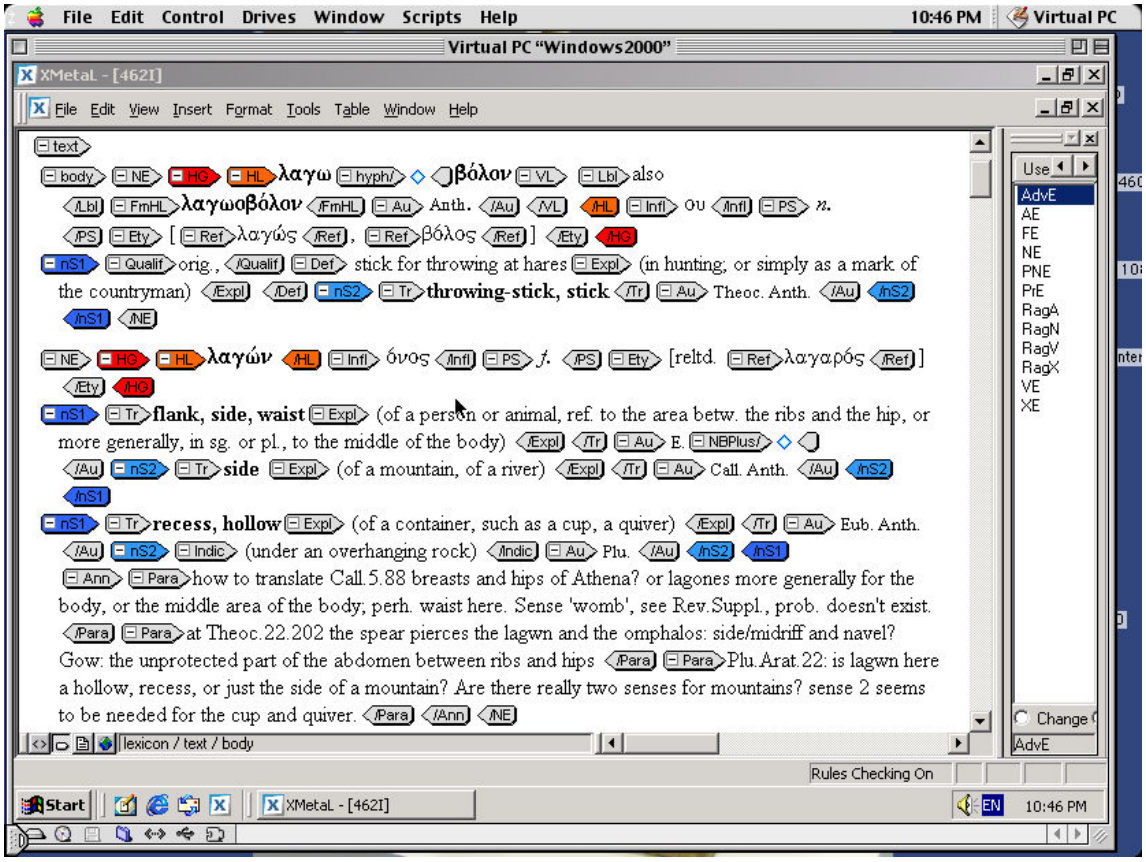
The formatting is illustrated in Fig. 1, which shows two articles, for the nouns λαγωβόλον (throwing-stick) and λαγών (flank):

λαγω-βόλον (also λαγωοβόλον Anth.) ou *n*. [λαγώς, βόλος] orig., stick for throwing at hares (in hunting; or simply as mark of the countryman); **throwing-stick**, **stick** Theoc. Anth.

λαγών ῶνος *f*. [reltd. λαγαρός] **1 flank**, **side**, **waist** (of a person or animal, ref. to the area betw. the ribs and the hip, or more generally, in sg. or pl., to the middle of the body) E. +; **side** (of a mountain, of a river) Call. Anth. **2 recess**, **hollow** (of a container, such as a cup, a quiver) Eub. Anth.; (under an overhanging rock) Plu.

The semantic approach can be seen in the use of a definition ('stick for throwing at hares') as well as translation equivalents ('throwing-stick, stick') in the first article, where the connection between the translation and the word form (whose stem is derived from the Greek for *hare*) is made in the (plain-text) definition. And explanatory parentheses are used in both articles.

The structure underlying this text is shown in Fig. 2, which shows the same articles, as they appear in the editing software:



This is the view which the writers see as they compose the articles. As well as typing the text, they insert the appropriate tags at each point, starting with the top-level tags for each part of speech, which can be seen in the right-hand column.

Five key features of the XML structure

1: Article structure and part of speech

As noted above, the structure of each article is specific to the part of speech of the lemma. There are dedicated structures for nouns (NE), proper names (PNE), adjectives (AE), adverbs (AdvE), verbs (VE), prepositions (PrE), other function words (FE), and XE for cross-reference entries. The use of provisional 'Rag' entries even allows us to write experimental entries that may well be omitted in the final version. This specificity enables us to impose very precise constraints on our writing style, and helps the writers as they work through an article, because at each point they are presented with a quite narrow choice out of the total of 136 elements. All are configured with a minimal number of child elements, and are designed to realise a specific configuration.

2: Contextual information

This tagging system includes the provision of contextual information. Because they organise senses in groupings, dictionary definitions are generalised, but they ultimately depend on how the words are used in context.

We define meaning, not by copying previous dictionaries, but by looking again at the original textual sources. Because Ancient Greek texts are almost all digitised, we have an excellent opportunity to reassess the contexts. We have collected the sources for every dictionary

headword in a large digital database, giving us an access to a great range of passages from the major Greek authors which is practically instantaneous, because all the passages have been identified and collated in advance, using the Perseus Greek morphological analyser.[4] Because the number of textual passages for each word can run into the hundreds, the database is configured to collect the citations used by the previous LSJ lexicon (and still the textual basis for subsequent dictionaries) separately, so we can examine them first, and then refer to other passages and other authors.

Using specific structures for different parts of speech also addresses an unevenness in the lexicographic approaches to different parts of speech. In lexicographic theory, 'meaning' concentrates almost entirely on referring words, and especially nouns and adjectives. This is not perhaps surprising, since they are in the majority. Of the 36,467 headwords in the Intermediate Greek Lexicon of 1889, 25,373 (=70%) are nouns and adjectives (with perhaps 3,000 = 8% of the whole, also having adverbial forms), while 10,672 = 29% are verbs in -ω, -μαι and -μι. Grammatical words including prepositions,' improper' adverbials, subordinating conjunctions and particles comprising the remaining 1%. Yet the frequency with which these words occur gives them an importance beyond their variety.

A major reason for treating the parts of speech independently is to help specify the contextual information, which varies for each part of speech: for adjectives the most important companion word is the noun it qualifies (given in our element QualN), while for adverbs it is usually the verb modified (ModVb). The meaning of verbs is especially connected with other sentence constituents, including their subject, object, and other types of complement (including dependent genitives and datives, dependent prepositional phrases and infinitival, participial and finite clauses). We give these in dedicated elements. Function words like particles have elements to mark their prosodic hosts and their textual context. The meaning of nouns is probably least dependent on their syntactic function, but we do sometimes give dependent genitives (Gntv), and other words which appear regularly as collocations (Cllc).

We do not, of course, give every collocation, but we generalise the contexts as much as possible, in order to focus on their contribution to word sense. An example may be seen in the entry for the adjective ῥᾳδινός, where it can be seen that the senses are organised in terms of the qualified nouns. Each translation, in bold, is preceded by parentheses, which give semantic as well as syntactic information (because the nouns are grouped by semantic field):

> **ῥᾳδινός** ή όν, Aeol. **βρᾰ́δινος** ᾱ ον *adj.* **1** (of a horse-whip, javelin, flower, sapling, cypress) app. **slender, slim, delicate, shapely** (sts. perh. w.connot. of flexibility or suppleness) Il. Lyr. Theoc.
> **2** (of feet, hands or arms, thighs) **slender, shapely** Hes. hHom. Anacr. Thgn. AR. Theoc.; (of a cheek) A.(dub.); (of the body) **slim** (through healthy diet) X.; (of a boy) **slim, shapely** Sappho (cj.); (of a girl) Theoc.

Further parentheses add detail to the translations, and author abbreviations give an indication of date and genre. Even though we do not include the full numbered citations in our printed lexicon, we can link to them in the electronic edition. This can be done for every entry, via the elements which cite the author names. For example, the article for λαγωβόλον (Figs. 1 and 2) notes that the word appears in Theocritus and the Anthology. A link to the electronic slips for that headword will enable the reader to access the relevant passages.[5] Consequently, although the printed edition of the lexicon will be quite small (about 1,100 pages), the combination of articles and source material in an electronic edition can be very extensive.[6]

This multi-level approach to semantic explanation may eventually have wider applications for the sense-tagging of classical corpora, which has so far been carried out very informally, although it ultimately depends for its criteria on dictionary definitions (see Véronis 2000).

3: Targets for electronic searches

Some elements will not appear in the printed lexicon, though they will be searchable in the electronic editions:

a) Elements giving information on word formation, such as word stems and affixes, which are especially productive in Ancient Greek. We can enter the word stem (or stems of a compound), root, prefix, and suffix in dedicated elements, which will be available only in the electronic editions, as constraints of space preclude their inclusion in the printed edition.

b) A 'keyword' element, in which we can note a semantic grouping like 'colour term' or 'botanical word', as targets for searches.

c) The 'annotation' element contains lexicographic remarks and queries that we wish to note in the XML, but which are not rendered in the print-ready copy. We render them separately as PDFs for our own reference and for our reviewers, and their responses can be entered alongside the primary annotations. Our discussions are recorded and preserved, up to three stages of dialogue. In Fig. 2 above, the second entry, for λαγών, includes such an annotation which discusses three textual difficulties. We make great use of this feature, as it enables us to add notes for our reviewers, as well as providing an archive of the research undertaken for the articles, which will be permanently available for future reference.

4: Simple micro-structure

For practical reasons, the structure needs to be easy to learn. Rather than using a schema, which defines structure contextually, we define XML structure by a DTD which gives each component a fixed definition, and we do not use variable attributes (arrays of values, which can be associated with any particular XML element). This is simply a matter of convenience: either system is equally powerful, but it is much easier for the writer to add an element by clicking on a list, than by setting a particular attribute.

Secondly, element names are chosen not only to be transparent, but to be as brief as possible. This is because each element tag takes up a finite space in the editing window, and the larger the name, the less space there is for the typed text. Although provision can be made for a separate 'page view', it is essential for writers to have a clear view of the text in the editing window. Consequently, the brevity of our element names is much closer to the style of the *OED* than to the *TEI*.

Thirdly, our elements are as simple as possible, and relate to the lexicographic structure: there are almost no purely formatting elements (only 4 out of 100).

5: Hierarchically-constrained configuration

It is also important to control the degree of homogeneity in the tagging structure. The *OED* tagging system, described in Berg (1993), has a structure which is similar at all levels, as shown in the simplified view given in Fig. 7:

```
 <HG> HEADWORD GROUP
        <HL>  Headword Lemma ,
                Lookup Form (LF), Stressed Form (SF), Murray Form (MF)  </HL>
        <MPR> Murray Pronunciation
        <PS> Part of Speech
        <HO> Homonym Number
</HG>

<VL>  VARIANT FORM LIST
        Variant Date <VD>, Variant Form <VF>
</VL>

<ET> ETYMOLOGY
</ET>

<S0><S1>...<S8> SENSE(S)

        <#>Sense Number
        <DEF> Definition
        <QP>Quotation Paragraph
                <EQ>Earliest Quote
                        <Q>    Quote
                                Date <D> Author <A> Work  <W>, Text <T>...</T></Q>
                </EQ>
                <Q>...</Q> Quote(s)
                <LQ><Q>...</Q></LQ> Latest Quote (Obsolete Entries Only)
        </QP>

        <SE>  Sub-Entry (Preceded by"Hence")
                Bold Lemma ( <BL>...</BL>,  and similar tags
                        to those following Headword Lemma)
        </SE>
</S0></S1>...</S8>
```

This has a clear structure, but it does not allow much variation in the ordering of the structural elements (though a great variety of purely formatting elements are used at the lower levels). This is not a great problem providing the elements are not too detailed, so only simple searches are possible, and if word-forms are regular and so easy to list, as in English. We needed greater granularity, because we wanted to maximise the possible types of searches that can be performed, and also because Greek word structure is exceptionally complex, and so it is often necessary to give a large number of variant and inflectional forms, and yet to allow considerable variety in their formatting (as in the Form Group shown below).

Our approach has been to configure the elements according to their level in the whole structure, using mixed-content elements only at the lower levels. In our DTD, the top two levels (the entry, and its child elements, including the head group and S1 sense groupings) are all standard-content elements, as shown in the entry structure for adjectives, AE:

<!ELEMENT AE (HG , HG2? , Summ? , S0?, aS1+ , (Adv | RelW | Ethn | NPS)* , Keywd?, Ann? , Ed?)>

Here, the formal information is given in the head group (HG). The main sense groupings (aS1) can be preceded by an introductory summary written in continuous prose (Summ, S0), and can be followed by sub-entries for related adverbs, ethnic nouns, or other parts of speech (Adv, RelW, Ethn, NPS). The entry closes with a keyword giving the semantic grouping, lexicographic annotations, and editorial notes.

Within the aS1 sense groups, sub-sections (sS2 elements) are also standard-content, and may themselves contain aS2 elements down to any depth.

<!ELEMENT aS1 (Nm? , Ety? , GLbl? , UsgLbl? , Indic? , Qualif? , Def? , Tr? , Au? , SeqLbl? , (Cmpl | QualN)* , Phr* , Extra? , aS2* , SGrm* , XR?)>

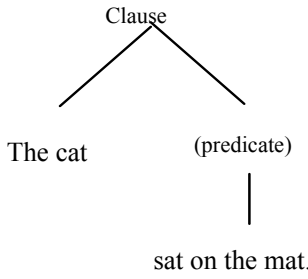
The sense sections can start with information on etymology (Ety), syntax (GLbl), linguistic genre or register (UsgLbl), and introductory semantic material (Indic). Then descriptive definitions and translations may be followed by contextual examples (QualN and Cmpl), specific quotations (Phr), possible sub-sections aS2 and SGrm, and a possible cross-reference.

Only below this level do mixed-content elements appear. For these, a constrained structure is maintained partly by keeping these elements simple by having as few child elements as possible: for example, the element corresponding to the TEI 'tr' has just five children:

<!ELEMENT Tr (#PCDATA | or | rom | ital | Prnth | Expl)*>

The link between XML and XSL

We also add constraints at the lower levels by means of the XSL styling. This is necessary because XML has a specific structural weakness: the mixed-content element. If we consider that a sentence is not just a linear sequence, but also a hierarchy of dependent elements, then, in the clause 'The cat sat on the mat', the subject ('the cat') is at a different logical level from the predicate, because the subject governs the predicate, as shown in Fig. 3:



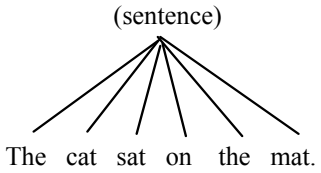
The predicate itself includes a verb governing a prepositional phrase, and so a similar hierarchy holds at every level of detail. Now, XML is a 'language' (or at least the deep-structure of a language), and reflects natural languages in having a hierarchical structure. However, it is a rather poor language, in that its hierarchies are structured very loosely, and do not give much information about the governing relations. This is because XML hierarchies are built using 'mixed-content' elements, which contain both words and also other 'child' elements. But, unlike the units of natural languages, the XML elements give little information about the organisation of their 'children'.

Fig. 4 shows the coding for the *TEI* Dictionary DTD, for the element "tr", which is designed to contain a translation:

<!ELEMENT tr (#PCDATA | abbr | address | date | dateRange | expan | lang | measure | name | num | rs | time | timeRange | add | corr | del | orig | reg | sic | unclear | oRef | oVar | pRef | pVar | distinct | emph | foreign | gloss | hi | mentioned | soCalled | term | title | ptr | ref | xptr | xref | seg | bibl | biblFull | biblStruct | cit | q | quote | label | list | listBibl | note | stage | text | anchor | gap | alt | altGrp | index | join | joinGrp | link | linkGrp | timeline | cb | lb | milestone | pb)* >

This coding signifies that the element "tr" may contain any amount of text inside it (PCDATA) and also 63 child elements, which may appear any number of times, in any order, interspersed in any way with the text. This very loose configuration is typical of most TEI elements.

And yet, without mixed-content elements, we restrict the hierarchy. A 'flat' structure would be equally uninformative, and perhaps equivalent to the diagram in Fig. 5:



This just reveals that all the words belong to the same clause. The only way of mapping detailed governing relationships in XML is to use mixed-content elements.

The problem would remain, even if we were to use a (context-sensitive) schema rather than a DTD. That would enable us to define elements contextually. But that is unlike natural languages (where, say, a noun phrase behaves the same way anywhere in a sentence), and it does not overcome the weakness of mixed-content elements.

Our solution is to create a maximally-constrained structure by configuring the XML elements according to their level in the overall hierarchy, and also to use the styling rules to capture part of the underlying structure.

Structural styling: standard-content elements[7]

In both the higher-level, standard-content, elements, and the lower, mixed-content, ones, we use the XSL styling to help us attain a high level of precision. In this approach, the formatting can be seen as the written equivalent of prosody, because it reflects the underlying structure.

The mix of constraint and flexibility can be seen in the head group (HG), where the head lemma (HL) is optionally followed by a range of configurations of dialectal lemmas (DL) and variant lemmas (VL), together with their gender inflections (Infl and its variants):

<!ELEMENT HG (HL , ((DL? , DInfl?) | (VL? , VInfl?) | Infl)* , BrVL? , PS , (Ety | InvEty)? , Morph?, Prsd? , FG?)>

A sample of output is shown from the article for the adjective ῥᾴδιος (easy):

ῥᾴδιος α ον (also ος ον E. Pl. D.), ep.Ion. ῥηίδιος η ον (ῥῃδιος Thgn.) *adj*. [ϟραδιος; compar. and superl. suffixes added directly to stem ῥᾶ] | COMPAR.: ῥᾴων, ῥᾶον (sts. in later Att. ῥᾳδιέστερος), Ion. ῥηίων, ῥηίον, ep.Ion. ῥηίτερος η ον (ῥῄτερος Thgn.), Dor. ῥᾴτερος (Pi.) || SUPERL.: ῥᾶστος η ον, ep.Ion. ῥήιστος |

Here, the head lemma (HL: ῥᾴδιος) is followed by two gender inflections (Infl: α ον). Then an alternative inflectional paradigm (VInfl: ος ον) is given, together with abbreviations for the authors who use it, in brackets. This grouping is followed by a comma, and then a dialectal lemma form (DL: ῥηίδιος) together with its inflections, followed by a variant lemma form (VL: ῥῃδιος) used by one author (Theognis).

Both sets of brackets are inserted automatically, yet at the discretion of the author, who simply places the VInfl and VL elements as children of the Infl and DL elements, rather than following them, which is the default (that would automatically change the formatting, with the elements then automatically separated by a comma, as can be seen above, immediately before 'ep.Ion.').

The XSL rule which achieves this particular configuration is a 'when-otherwise' conditional, shown in Fig. 8:

```
<xsl:choose>
    <xsl:when test="parent::lex:DL or parent::lex:HL">
        <fo:inline xsl:use-attribute-sets="lex:normal-font">
            <xsl:text> (</xsl:text>
        </fo:inline>
            <xsl:copy-of select="$content" />
        <fo:inline xsl:use-attribute-sets="lex:normal-font">
            <xsl:text>)</xsl:text>
        </fo:inline>

    </xsl:when>

    <xsl:otherwise>
        <xsl:copy-of select="$content" />
    </xsl:otherwise>
</xsl:choose>
```

This rule links the parentheses and preceding whitespace to the structure, enabling it to be controlled by the editor, in quite an intuitive way. The parentheses appear only when the element is embedded. We use a considerable amount of such automatic yet optional bracketting.

The usefulness of optional punctuation rules can be seen further on in the figure, after the part of speech (PS: *adj*.) and etymological information (Ety, surrounded by square brackets). The form group (FG) which gives a number of comparative and superlative inflectional forms, is automatically formatted with vertical bars, small caps and commas. In addition, optional bracketting is added for two variant forms and one author abbreviation (Pi. for Pindar). The result is that even a complex piece of formatting is extremely easy to control.

Structural styling: mixed-content elements

Whitespace is also added in the mixed-content elements, where it depends automatically on the presence of child elements. Though this is unconventional practice, it is extremely helpful for the writers, and it is possible because we have applied pre-set constraints to mixed-content organisation. As noted above, the translation element 'tr' can have five child elements (or, rom, ital, Prnth, Expl) as well as text. Of these children, the first three add optional formatting, while Prnth and Expl contain explanatory text, bracketted. Examples of these can be seen in one of the sense sections for the verb γράφω (write), which includes the translation:
'**propose** (sthg.) **in writing** (before the Assembly)'.

Here, not only the parentheses, but also the external whitespace, are added through the stylesheets, so the Prnth element (containing the text 'sthg.') is bracketted with whitespace preceding and following, while for Expl (containing the text 'before the Assembly'), whitespace is added before the element, but not after:

```
<xsl:template match="lex:Expl" mode="fo:inline">
        <fo:inline xsl:use-attribute-sets="lex:normal-font">
                <xsl:text> (</xsl:text>
        <xsl:apply-templates mode="fo:inline" />
                <xsl:text>)</xsl:text>
        </fo:inline>
</xsl:template>
```

It may seem unnecessary to use stylesheets to achieve formatting which could be keyed-in. It has, however, proved invaluable, for two reasons. First, it ties the styling to the structure, providing more constraints and ensuring that the final documents are consistent.

Secondly, the writers need to concentrate on the lexicographic issues, not the formatting. Working in even the best text editor adds an extra burden, and so it is vital to balance this by making the software take as much of the load as possible. Every extra keystroke is a distraction from the task of composition. The automatic addition of whitespace has proved to be particularly useful, as the editing displays do not give a perfect rendition of the formatting. And although it is unorthodox, we have found that it increases efficiency and reduces writer error.

Because the formatting and the structure are matched with our lexicographic methodology, the variations can easily be controlled by the writer. This is very useful in the writing process, because the number of variables can be immense. Clearly, we cannot predict just how complex unwritten articles may need to be, yet we have been able to develop the tagging structure alongside the lexicographic details, at the beginning of the project, and maintain it over a period of years.

It was only by considerable trial and error that we have been able to combine the necessary constraints with the required flexibility. Although the styling rules are simple, they are very extensive, constituting a large part of the coding: while the DTD can be printed out on twelve A4 pages, the XSLT stylesheet fills over 60. However, the resulting working environment has proved highly compatible with the tasks of composing and editing entries. The XML is validated in the normal way, and our stylesheets are used by the typesetters in preparing the final copy.

Conclusion

This paper has described how the design of dictionary articles can best utilise tagging, by integrating it in the writing environment. Our experience of this approach has been positive, and we anticipate that the ability to search the completed documents will prove increasingly more useful over the next years, as the lexicon approaches completion. The DTD is published online, and so can be examined by interested scholars.[8]

This 'customised' approach is not intended to neglect the potential for translation and interchange of information between dictionaries. The precision afforded by a structure which is specific, not just to each dictionary but to each part of speech, leads to greater consistency, and ultimately to a greater ability to exchange information.

Acknowledgements

The work described here was partly funded by a grant from the European Commission Information Society. Dr Anne Thompson developed the lexicographic methodology which the DTD was designed to realise. Thanks are due to Chris Hamilton-Emery, then of Cambridge University Press, for introducing me to XML methodologies including the TEI DTD, and to Dr Jeni Tennison for writing the first draft of the XSL stylesheets. The citation database was designed and programmed by Professor Jeffrey Rydberg-Cox of the University of Missouri at Kansas City.

Bibliography

Berg, D.L. 1993. 'The Research Potential of the Electronic OED2 Database: a Guide for Scholars' (*http://www.chass.utoronto.ca/chass/oed/Berg-1.html*, online ed. W. McCarty, Centre for Computing in the Humanities, University of Toronto, 1993; converted to HTML 1995).

Crane, G. 1991. 'Generating and Parsing Classical Greek', *Literary and Linguistic Computing* 6.4, 243-245.

Ide, N. and J. Véronis, eds. 1998. *Word Sense Disambiguation* (Special issue of *Computational Linguistics*, 24.1).

--------1995. 'Encoding Print Dictionaries', *Computers and the Humanities*, 29, 1-3, 167-95.

Johnson, S. 1755. *Dictionary of the English Language*, London: Knapton.

Kang, B.-m. 1997/8. 'Modifying the *TEI* DTD: The Case of Korean Dictionaries', *Computers and the Humanities*, 31 (5): 433-449.

LSJ 1925-1940. Liddell, H. G. and Scott, R., revised by H.S. Jones, with the assistance of R. McKenzie, *A Greek-English Lexicon: a New Edition*, published in ten Parts, Oxford: Clarendon.

OED 1888-1928. Murray, J. A. H. et al. *A New English Dictionary on Historical Principles* (reprinted 1933 with the title *Oxford English Dictionary*), Oxford: Clarendon.

OLD 1968-82. Wyllie, J. M., Glare, P. G. W. et al. *Oxford Latin Dictionary*, Oxford: Clarendon.

Pind, J. 1990. 'Computers, Typesetting, and Lexicography', in J. Pind and E. Rögnvaldsson, eds. (1990) *Papers from the Seventh Scandinavian Conference of Computational Linguistics, Reykjavík 1989*, Reykjavík: Institute of Lexicography, 308-325.

Rydberg-Cox, J. A. 2005. 'Word Profile Tool.' *Cultural Heritage Language Technologies* (*www.chlt.org/CHLT/Slip_Sample/index.html*).

Silva, P. 2000. 'Time and Meaning: Sense and Definition in the OED', in L. Mugglestone, ed., *Lexicography and the OED: Pioneers in the Untrodden Forest*, Oxford: Oxford University Press.

Simpson, J.A. and E.S.C. Weiner 1989. 'The New Oxford Dictionary Project', in *OED*, 2nd ed., Vol. I, l-lvi, Oxford: Clarendon.
Sperlberg-McQueen, C.M. and L. Burnard 1994. *Guidelines for Electronic Text Encoding and Interchange*, Text Encoding Initiative: Oxford (published online at *www.tei-c.org/P4X/*).
Tompa, F.W. 1992. 'An Overview of Waterloo's Database Software for the *OED'*, *Computing in the Humanities Working Papers* 2; repr. B.13, 1996 (published online at *http://www.chass.utoronto.ca/epc/chwp/tompa/index.html*).
Tutin, A. and J. Véronis 1998. 'Electronic Dictionary Encoding: Customizing the *TEI* Guidelines', *Eighth Euralex International Congress (EURALEX '98)*, 4-8, Liège.
Véronis, J. 2000. 'Sense tagging: Don't look for the meaning but for the use', *Computational Lexicography and Multimedia Dictionaries (COMLEX 2000)*, 1-9. Online at *www.up.univ-mrs.fr/~veronis*.

---

[1]For details of the *TEI* DTD, see Sperlberg-McQueen and Burnard (1994).
[2]Although the card slips containing the textual citations for the *OED* are retained in an archive for consultation by the editors of subsequent editions, those for LSJ and *OLD* have been lost.
[3]John Chadwick, who planned the dictionary, was (in addition to his work on Linear B) a member of the editorial staff of the *OLD* from 1946-52, and was on the Project Committee of the Revised Supplement to LSJ, from 1979 until its completion. For the semantic approach of the *OED*, see Silva (2000).
[4] See Rydberg-Cox (2005). A description of the software can be found in Crane (1991).
[5]Idylls 4.49 and 7.128; and *AP* 6. 152, 177, 188 and *APl* 16.258.
[6]The lexicon and associated database will be published online on the Perseus site.
[7]'Standard-content' describes an element which contains only child elements, as opposed to elements with text-only content, mixed-content, and empty elements.
[8]The DTD is published online at *http://www.chlt.org/lexicon/papers/lexicon-DTD.txt.*