Constructing a Glossary for the Electronic *Beowulf*

Michael ELLIS

Abstract. The glossary for the Electronic *Beowulf* differs from previous glossaries for the poem in that it was written using HTML (HyperText Markup Language) and was designed to be used alongside images of the manuscript, or any of several other texts included in the edition. Construction of a large electronic glossary involves all the usual textual and linguistic problems, plus the additional problem of adapting HTML elements to enhance the glossary's usefulness. At present, it incorporates more than 5,000 hyperlinks for cross-referencing and for accessing additional information.

Résumé. Le glossaire pour le *Beowulf* électronique a été réalisé en HTML (Langage hypertexte à ferrets), ce qui le différencie des glossaires précédents, et permet son emploi à côté des images du manuscrit, qu'il s'agisse du *Beowulf* ou des autres textes électroniques compris dans l'édition. La réalisation d'un important glossaire électronique comporte toutes les difficultés habituelles, en matière de texte et de linguistique, auxquelles s'ajoute encore la difficulté d'adapter les éléments du langage HTML afin d'augmenter l'utilité de l'outil. À présent, le glossaire comprend plus de 5 000 liens hypertextes permettant d'établir des liens référentiels et d'accéder à des informations supplémentaires.

Keywords: Electronic glossary, *Beowulf*, HTML.

Mots-clés : Glossaire électronique, *Beowulf*, HTML.

1. Introduction

Constructing a glossary for the Electronic *Beowulf* involved many of the same challenges which have confronted compilers of previous glossaries. The extensive vocabulary of the poem requires entries for more than

✉ Southwest Missouri State; Department of English; 901 South National Avenue; Springfield, Missouri 65804 (U.S.A.).
Fax: +1 417 836 4226

3,200 different words, including many which occur nowhere else in Old English literature. Moreover, the condition of the *Beowulf* manuscript, particularly around its damaged outer margins, often necessitates reconstruction of words or parts of words which are now lost.[1]

However, the glossary for the Electronic *Beowulf* differs from its predecessors in that it began as a computer-generated concordance and was constructed as an electronic text using HTML (HyperText Markup Language), a medium commonly used for creating sites on the World Wide Web. As part of a comprehensive and multidimensional electronic edition edited by Kevin Kiernan at the University of Kentucky, the glossary was designed to run alongside the digital images of the manuscript, or any of several ancillary texts included in the edition. Therefore, in addition to the usual textual, linguistic, and lexicographic challenges, making a glossary for the Electronic *Beowulf* involved the special problem of adapting HTML "hyperlinks" so that the glossary can accomplish what would be difficult or impossible in a printed version. In its present form, the glossary contains more than 5,000 of these links, both for cross-referencing and for layering information in the entries.

2. The Electronic *Beowulf* Project

The compact disk (CD) edition of Electronic *Beowulf* is currently in the final stages of development and testing, and will be available through British Library Publications and the University of Michigan Press. The original Electronic *Beowulf* project, with Kevin Kiernan and Paul Szarmach as co-directors, began in 1993 as part of the British Library's "Initiatives for Access" program, and information about the project, as well as selected images of the manuscript, have been available for some time through the project's ftp and URL sites located at the British Library and at the University of Kentucky.[2]

[1] See KIERNAN (Kevin): 1985, "The state of the *Beowulf* manuscript 1882–1983", *Anglo-Saxon England*, 13, pp. 23–42; 1986, *The Thorkelin Transcripts of Beowulf* (Copenhagen: Rosenkilde and Bagger); 1996, *Beowulf and the Beowulf Manuscript*, revised edition, (Ann Arbor: University of Michigan Press).

[2] For more information on the Electronic *Beowulf* project see Kevin Kiernan's essays at http://www.uky.edu/~kiernan/welcome.html and http://portico.bl.uk/access/Beowulf/ ✂ electronic-Beowulf.html.

The forthcoming CD edition includes full-color digital images of every folio in British Library, Cotton Vitellius A. xv, the manuscript which includes *Beowulf* and several of other Old English texts. The CD also includes several important ancillary documents, including digital images of two eighteenth-century transcriptions of the poem and two nineteenth-century collations of the manuscript with the first printed edition of *Beowulf*. These images, along with a new edition of the poem, a translation, glossary, and hundreds of fiber-optic backlit images of damaged portions of the manuscript, provide users with unprecedented access to the *Beowulf* manuscript.

Users will have several viewing options available, including full-screen images of the manuscript and ancillary texts, as well as a split screen for collating different manuscript images. The glossary and the fiber-optic backlit images can be viewed in vertical frame to the left of the manuscript images, the edition, or other optional texts. Users access the glossary by clicking on one of the Old English letters listed vertically on the right side of the screen, or they can access backlit images by simply clicking on the damaged portions in the manuscript. The glossary also has a full screen option, and a search interface is being developed to facilitate various kinds of searches. An important feature of the glossary is that it includes line references for both the manuscript and the edition, so that users will have the option of translating directly from the manuscript images or from the edition.

3. Lemmatization and Entry-Writing

There were three distinct stages in the construction of the Electronic *Beowulf* glossary, and after each stage I sent draft versions of the glossary files over the Internet to Kevin Kiernan at the University of Kentucky so that the files could be tested and incorporated into preliminary versions of the CD edition. The first stage consisted of transforming computer-generated concordance files, sent to me over the Internet from the University of Kentucky, into glossary files. The concordance files, one for each of twenty-one Old English letters, consist of alphabetized lists of every form as it occurs in the poem, along with manuscript and edition line references for each occurrence. Lemmatization of the concordance files involved the sorting and listing of all occurrences of each word under a headword or *lemma*. This process was followed by the actual writing of the entries, in which words were identified by part of speech, supplied with a definition, and the various occurrences of the word arranged and given abbreviated grammatical designations. Certainly, this stage in glossary construction most closely followed the methods used in traditional lexicography, with frequent reference to standard Old English grammars, concordances, and dictionaries, and most particularly to the new Toronto *Dictionary of Old English* for the letters A through E.[3] I also consulted previous editions of the poem, including their glossaries, and these proved valuable as models, as well as resources for checking details included in the entries.[4] However, the most important reference sources for construction of the glossary were Kevin Kiernan's edition-translation of the poem and the manuscript images. Like the concordance files, the edition-translation was written in HTML format and was received by way of the Internet from the University of Kentucky. HTML documents have the advantage of being readable on all platforms, so that all the Electronic *Beowulf* resources, including a preliminary CD which included the digital images of the manuscript, could be easily accessed through an Internet browser. The ease of moving back and forth between the browser and a word processor became increasingly important for testing the many hyperlinks that were added to the glossary.

Transforming a concordance file into a glossary file can be a formidable task, depending on the size of the original letter file and a variety of other common lexicographic problems. For example, the various grammatical forms of some words, especially forms of Old English strong verbs, may be widely scattered in a concordance file. Some words also occur in a number of variant spellings and these, too, may be widely scattered in a concordance file, and in some cases, in more than one file. In *Beowulf*, for example, the various grammatical forms and spellings of the verb *cuman*, "to come," include *cuman*, *cymast*, *cymeð*, *cume*, *cymen*, *cwom*,

[3] CAMERON (Angus), AMOS (Ashley Crandell), and HEALEY (Antonette diPaolo), eds.: 1986–1996, *Dictionary of Old English, A–E* (Toronto: Pontifical Institute of Medieval Studies).

[4] Previous editions consulted include DOBBIE (E.V.K.), ed.: 1953, *Beowulf and Judith* (New York: Columbia University Press); HEYNE (M.), SCHÜCKING (L.) and VON SCHAUBERT (E.), eds.: 1963, *Beowulf* (Paderborn: Verlag Ferdinand Schöningh); JACK (George), ed.: 1994, *Beowulf, a Student Edition* (Oxford: Oxford University Press); KLAEBER (Fr.), ed: 1950, *Beowulf and the Fight at Finnsburg*, 3rd edition (Boston: D.C. Heath and Co.); WRENN (C.L.) and BOLTON (W.F.), eds.: 1973, *Beowulf with the Finnesburg Fragment*, 3rd edition (New York: St. Martin's Press); Wyatt (A.J.) and CHAMBERS (R.W.), eds.: 1914, *Beowulf and the Finnsburg Fragment*, new edition (Cambridge: Cambridge University Press).

com, *cwomon*, *comon*, *cwome*, *come*, *cumen*, and *cumene*, all of which must be gathered together under the headword *cuman*. Before a form was moved, a copy of the original concordance entry was left in place so that it could later serve as the basis for a cross-reference.

In addition to variant forms of words, another common problem for lemmatization is the necessity of distinguishing between homographs, forms which may appear identical, but which represent items belonging under separate headwords. The spelling *for* may represent the preposition *for*, or it may represent the preterit singular of the verb *faran*, ("to go, travel"). Since the program which generates the concordance files cannot distinguish homographs, it is necessary to consult the edition-translation to determine what word the form represents. Especially difficult are the many high-frequency words which are homographs or which may belong to different grammatical categories of a single word. These occurrences require a great deal of sorting, again with constant reference to the edition-translation to insure that all line references are put in under the appropriate headword. For example, there are 319 occurrences of the form *þa*. This form may represent adverb *þa* ("then") or it may be conjunction *þa* ("when"), while in other cases it may be the demonstrative pronoun *þa*. As a demonstrative pronoun, *þa* may be nominative plural, accusative plural, or feminine accusative singular, and each occurrence must be sorted into the appropriate grammatical category in their respective entries. Demonstratives also must be sorted by usage, since they may function variously as demonstrative adjectives, as relatives, or as substantives. Finally, all occurrences must also be sorted depending on whether the initial consonant is *þ* or *ð*, since distinguishing between these two letters may provide useful evidence, particularly concerning differences between the two scribes in the manuscript. The adverb and conjunction forms of *þa* remain in the TH letter file, but the occurrences of demonstrative *þa* must be moved to a special file which contains the many different demonstrative pronouns which occur in *Beowulf*. There are numerous other cases in which a word generated in one concordance file belongs under a headword in another file. Many of these cases involve variant spellings of initial vowels, but the most common examples of this difficulty, by far, are the many words which carry a *ge-* prefix. These forms are generated in the letter G concordance file, but the conventional practice in Old English dictionaries and glossaries (e.g., the *Dictionary of Old English*) is to ignore the *ge-* prefix for purposes of alphabetization. Consequently, a very large number of words and their accompanying line references had to be moved out of the letter G file and into other letter files. In these cases, the words and line references to be moved were simply copied and appended to the letter file in which they should appear so that they could be sorted later.

Appearance of the entries was, of course, an important concern during the process of writing the entries. HTML elements can be used to create a variety of formats and typographical styles, and can be used to create special characters used in the Old English alphabet. However, a shortcoming of HTML is that it does not recognize more than one space, so that in order to maintain boundaries between the different parts of an entry it was necessary to punctuate the glossary more heavily than is the common practice in printed glossaries and dictionaries. Also, in order to facilitate recognition of forms and to enhance the separation between different grammatical forms or spellings, each of these variants was spelled out completely in bold within the entry. The example below shows how an entry appears in the browser:

> **sittan**, st. 5, *to sit*; 141r8:492, 144v11:640; pres. 3s. **siteð** 192v19:2907; imp. s. **site** 141r3:488; pret. 3s. **sæt** 132v17:130, 136r12:286, 137v17:356, 141r16:499, 155v9:1166, 156r15:1192, 191v17:2853, 192v10:2895; pret. 3p. **sæton** 155v6:1163.
>
> (abbreviations: st. 5 = strong verb, class 5; pres. = present tense; pret. = preterit; imp. = imperative; s. = singular; p. = plural)

In this example, the line reference immediately following the definition are occurrences of the infinitive.

4. Linking

With lemmatization and entry writing completed, glossary construction entered its second stage with the incorporation of HTML hyperlinks for cross-referencing and for accessing additional information. While HTML elements which control format and appearance are certainly important, elements which are used to create hyperlinks are particularly important for enhancing the capabilities of an electronic glossary. These links are of two basic kinds, those which are used simply to access another HTML file, and those which are used to move to specific points within a file or specific points between files. In order to create abbreviated entries for high-frequency words such as pronouns, prepositions, and conjunctions, special files were made to contain the line references for these words, and these secondary files linked from a brief entry in one of the main letter files.

A number of generic links are also used to link a large number of entries or points within the glossary to a single HTML file. A generic link at the beginning of each letter file accesses a brief menu of explanatory material which includes information on alphabetization and arrangement of entries, an explanation of line references, an explanation of linking within the glossary, a list of abbreviations, and a list of references consulted. An explanation of the system of line numbering was especially important since many lines in the Electronic *Beowulf* edition diverge from those used in previous editions. This explanatory material includes a table for finding equivalent line numbers in other editions, and also incorporates links to the places in the edition where lines diverge, as well as links to passages from previous editions with examples of how the lines in these passages have been interpreted in previous editions or how various interpolations have been supplied by editors.

A second kind of generic link accesses paradigms for the seven classes of Old English strong verbs, with each strong verb in the glossary linked to one of seven files containing additional information, including examples of each strong verbs for each class. The example below shows the entry for the verb *faran* followed by the linked information about class 6 strong verbs.

> **faran**, st. 6, *to fare, go forth, travel, voyage, raid*; 132v12:124, 148v14:864, 186v9:2552, 193r5:2916, 193v7:2946; pret. 3s. **for** 161r5:1406, 171v15:1911, 181r13:2309, 189A(197)r14:2673; pret. 3p. **foron** 171v3:1898; ger. **([f]ar)enne** *169v3:1808.

The user simply clicks on the st. 6 following the headword to access the file:

> **Strong Verb, Class 6**, Ablaut series *a, o, o, a*.
>
> Examples (inf., pret., pret. plural, past ptc.):
> **faran, for, foron, (ge)faren;**
> **scacan, scoc, scocon, scacen;** pp. also occurs as **sceacen;**
> **standan, stod, stodon, (ge)standen;** variant forms include inf. **stondan** and pret. plural **stodan;**
>
> Examples of Class 6 verbs with e or y in the inf.:
> **hebban, hof, hofon, hafen;** pp. also occurs as **hæfen;**
> **scyppan, scop, scopon, scepen;**
>
> Example of a Class 6 contracted form:
> **slean, sloh, slogon, slægen;** pret. singular also occurs as **slog.**

All examples used in the strong verb files also link to the full entries for these verbs in the glossary, so that by clicking on any of the verb forms for *hebban*, for example, the user will be taken to the full entry in the letter H file. There are also additional files in the form of paradigms for the third-person personal pronouns and for the demonstrative pronouns, and these files distinguish between the forms used by the two scribes in the manuscript, Scribe 1, who copied the portions of the poem from the beginning through folio 172v line three, and Scribe 2, who copied the portions from that point to the end of the poem. As in the strong verb files, all the examples used in the pronoun files are linked to the full entries in the glossary.

In addition to links which access HTML files, the glossary makes considerable use of the HTML *NAME* element, which allows linking to a specific point or *anchor* within an existing file or to a specific entry in another HTML file. The *NAME* element made it possible to create a large number of links for cross-referencing without the necessity of creating additional files. For example, by clicking on the **faran** in the cross reference "**for**, verb, see **faran**" the user is taken to the *faran* entry without the necessity of any additional scrolling. There are more than six hundred cross-reference links in the files to accommodate the various grammatical forms and variant spellings of words, and these links are installed for all cross-references unless the *anchor* is immediately adjacent. There are also more than one thousand additional cross-reference links for compound and prefixed forms of headwords. Unlike the cross-references for grammatical forms and variant spellings, these cross-references are at the ends of entries, as can be seen in the example below. Clicking on any of the five compounds at the end of the entry will the user to the entry for that word.

fæt, n., *vessel, cup; beaker;* ap. **fatu** 190r4:2762. See also **banfæt, dryncfæt, mæþdumfæt, sincfæt, wunderfæt.**

(abbreviations: n. = neuter noun; ap. = accusative plural)

5. Layering

Although cross-reference linking certainly increased the usefulness of early versions of the glossary, the number of relatively large entries remaining in the letter files, including those containing links for compounds and prefixed forms, continued to cause problems for scrolling. Even with the addition of an option which allows the user to go to full screen for viewing the glossary, scrolling can be a time-consuming process for some of the larger letter files. The H and S letter files, for example, each contain more than four hundred entries. However, the creation of additional files to contain line references for the longer entries created problems both for file management and for revision of entries. The problem of entry size and scrolling was complicated further by the desire to add even more information to the entries. The solution to the problem was to maximize the use of linking, but to accomplish this with a relatively small number of files. An early version of the glossary contained nearly three hundred individual HTML files; at present, with a considerable increase in the number of links, the total number of files is less than sixty. Therefore, the third stage in glossary construction was to use the HTML *NAME* element to "layer" the information in the letter files. The present version of the glossary is constructed in two layers with each entry in the primary layer linked to an expanded entry in a secondary layer file. Entries at the primary layer contain essential information, including a definition and a summary of forms in which the word occurs. Cross-referencing of grammatical forms and variant spellings also remains in the primary layer. Secondary layer entries include all manuscript and edition line references, linked cross-references to compound and prefixed words, and listing of forms containing variation between *þ* and *ð*. Additional information at the secondary layer includes significant variation between the two scribes of the manuscript, and information about restorations or emendations used in previous editions of the poem.[5] The example below shows a primary layer entry and the corresponding secondary layer entry.

sweord, n., *sword*; ns. **sweord, swurd**; as. **sweord, swurd, swyrd**; gs. **swcordes**; ds. **sweorde**; np. **swyrd**; ap. **sweord, swurd**; gp. **sweorda**; dp. **sweordum.**

sweord, n.,
sword; 158r20:1288, 158v3:1291, 164v3:1571, 165r13:1607, 165r21:1617,

[5] KELLY (Birte): 1983, "The formative stages of *Beowulf* textual scholarship: part 1", *Anglo-Saxon England*, 11, pp. 247–274; "The formative stages of *Beowulf* textual scholarship: part 2", *Anglo-Saxon England*, 12, pp. 239–275.

> *167r10:1698, 185v4:2500, 189A(197)r4:2660, 189A(197)r21:2682, 189A(197)v17:2701, **swurd** 149r16:889; as. **sweord** 139v13:436, 145r18:671, *169v6:1811, 180r1:2253, 185v21:2519, 186v19:2563, **swurd** 171v8:1904, **swyrd** 187v20:2611, 194r19:2988; gs. **sweordes** 154r9:1105, 178v8:2196, *183r2-3:2387; ds. **sweorde** 142v16-17:560, 143r8:573, 145v4:678, 185r19:2493, 192r19:2881, 192v18:2905; np. **swyrd** 195v6:3050; gp. **sweorda** 152v7:1039, 193r21:2937; dp. **sweordum** *143r1:566, 143r18:585, 149r10:883, 193v20:2962; ap. **sweord** 188v5:2639, **swurd** 142r15:538. **swe(ordum)** 143r1:566 usu. em. **sweo[r]dum**. **sweordum** 193v20:2962 usu. em. **sweorda**. See also **ealdsweord**, **gomelsweord**, **guðsweord**, **maþþumsweord**, **wægsweord**.
>
> (Abbreviation: usu. em. = as usually emended in previous editions)

In order to create the primary and secondary layers it was first necessary to duplicate each of the letter files, and rename the files which would form the secondary layer (e.g., the original `A.htm` file became `A.htm` and `A2.htm`). Then, each primary and secondary layer file was edited so that files contain only the information appropriate to that layer. The final step was to add HTML codes for each entry in both layers so that the primary layer would link to the secondary layer. Although writing in codes for more than 3,200 entries is a time-consuming process, the layering of information in the entries improved the overall appearance of entries at the primary level and had a significant impact on the problem of scrolling. Finally, given the large number of links in the final version of the glossary, and since these links occur consistently, it was no longer necessary to highlight the links.

6. Conclusions

HTML may not be the ideal medium for constructing an electronic glossary, but despite its shortcomings it does have the advantage of being fairly easy to learn and to use. Recently, SGML (Standard Generalized Markup Language), the format from which HTML is derived, has received considerable acceptance as a standard for electronic texts, and SGML guidelines have been promulgated through the Text Encoding Initiative (TEI). SGML is also the format used by Humanities Text Initiative (HTI) at the University of Michigan, which includes among its resources an on-going Glossarial DataBase of Middle English. It will be very interesting to see how projects using SGML develop in near future, but at present HTML remains a viable alternative. Expansion of features for an electronic glossary would appear to be limited only by the amount of time it takes to incorporate them. Some additional features anticipated for the Electronic *Beowulf* include hyperlinks between the edition and the glossary, particularly at damaged or otherwise problematic passages, and an expanded treatment of the grammar of the poem with generic links included for all parts of speech. The search engine being developed for the Electronic *Beowulf* will certainly prove invaluable for various kinds of investigations and will add to the overall usefulness of the glossary.