

Correcting the online Bosworth/Toller and Cleasby/Vigfusson

Revised 7 October 2004

There is a bit of a learning curve to correcting these dictionaries. Please don't let this scare you off: it should only take ten minutes to understand the essentials. The main things you need to learn about are *tags* (to indicate boldface and italic) and *entities* (to indicate special characters).

The dictionaries contain many special characters (á, Æ, etc.) and also use **bold** and *italic* text styles. This is a real problem because different volunteers prefer to work on different systems (Windows, Linux, Macintosh) using many different programs (Microsoft Word, Emacs, TextEdit, etc.). Unfortunately, there is no single standard way of encoding these text styles and characters which is currently supported by every program.

The solution we're using in this project is to use plain text (ASCII) as the format for the editable file which you'll be correcting and submitting, because this is the one format which all programs can read and write. But don't worry; we're not losing the special characters or the boldface/italics. We simply have to represent these things in a special way within the plain text file. The scheme we're using here is a simple one which shouldn't take long to learn, and which should already be familiar to anyone who works with HTML or XML.

Fortunately, when you're doing the initial proofreading, all of the special characters and boldface/italics will look as they should. The PDF file has everything properly typeset. The only time you need to worry about how to enter special characters and boldface/italics in a plain text file is when you're actually typing in your corrections.

How special characters are represented:

Original page: under his cláðum
PDF file: under his cláðum
txt file: under his cláðum

When you're typing corrections into the editable .txt file, please use entities like **á** instead of special characters like á, even if your text editor allows you to enter á.

The last three pages of this manual contain a list of the most common special characters and their corresponding entities. It would be a good idea to print those three pages and keep them close at hand when you type in your corrections.

How boldface and italics are represented:

Original page: **fyrwit** *curiosity*, Salm. Kmbl. 117; Sal. 58. v. fyrwet.
PDF file: **fyrwit** *curiosity*. Salm. Kmbl. 117; Sal. 58. v. fyrwet.
txt file: **fyrwit** *<I>curiosity.</I>* Salm. Kmbl. 117; Sal. 58. v. fyrwet.

Notice that the boldface and italics are represented in the editable .txt file with HTML-style tags **...** and *<I>...</I>*.

The tag **** means “start boldface,” and the tag **** means “end boldface.” If you need to make something bold or italic, please use these tags.

Please *don't* use the bold or italic commands in your word processor. The .txt file is a plain text file, and these text attributes won't be saved.

There is also a third tag, **<HEADER>**, which is used to indicate the special line at the top of the page which contains the page number and the first and last headword. **<I>**, ****, and **<HEADER>** are the only valid tags at this stage of the project.

Typical errors and corrections

Error: wrong letter or letters

Original page: Bd. de nat. rerum;
PDF file: Bd. de nat. rerurn;
txt file: Bd. de nat rerurn;
Corrected: Bd. de nat rerum;

Error: one word split into two; extra characters

Original page: some caşes,
PDF file: some ca. ses,
txt file: some ca. ses,
Corrected: some cases,

Error: missing diacritics (acute accent mark, umlaut mark, etc.)

Original page: Hé oubeád đæt hē of Róme côme and
PDF file: He onbeád ŏæt hé of Rome come and
txt file: He onbe´d ð&elig;t h´e; of
Rome come and
Corrected: H´e; onbe´d ð&elig;t h´e;e;
of R´ome c´ome and

Error: extra italics

Original text: wisan ne can, Ælfc. Gen.
PDF file: wisan ne *can*, Ælfc. Gen.
txt file: wísan ne <I>can,</I> Ælfc. Gen.
Corrected: wísan ne *can*, Ælfc. Gen.

Error: a word should be bold, but it isn't

Original text: Grett. 121. all-gildir, adj.
PDF file: Grett. 121. all-gildir, adj.
txt file: Grett. 121. all-gildir, adj.
Corrected: Grett. 121. all-gildir, adj.

This is a very common problem within entries in Cleasby-Vigfusson. The OCR process did very bad job of distinguishing boldface from plain text, so boldface tags are usually missing altogether except where they were automatically added to the first word of each entry.

Error: whole word should be bold, but part of it isn't

Original text: læce-seax, es; *n. A surgeon's knife*
PDF file: læ ce-seax, es; *n. A surgeon's knife*
txt file: læce-seax, es; <I>n. A surgeon's knife</I>
Corrected: læce-seax, es; <I>n. A surgeon's knife</I>

Please correct all errors in boldfacing and italics, even if it is a seemingly minor error such as just one punctuation mark being incorrectly marked (or not marked) as bold or italic. The reason is that we will later write a program which will use this information to figure out the structure of the entry (we'll be using a markup standard known as the *TEI Guidelines* to make the full structure of the dictionary explicit).

How to handle various things

When you're not sure

If you are uncertain how to handle something, type the word UNCERTAIN in all capital letters right after the word(s) in question. This signals for us to have a look at the problem.

It's a really good idea to ask questions if you're not sure about something. Don't worry at all about pestering. Send questions to kurisuto@unagi.cis.upenn.edu. Or, better yet, post them to the message board.

Illegible text in the original

Please type the word ILLEGIBLE in place of any words which are illegible in the scanned image of the original page. Occasionally there is an ink blob or other problem.

Errors in the original

If you see something which you think is an error in the original Bosworth and Toller, type ERROR after the word or words in question. It's OK to do this even if you just suspect an error but are not sure.

Greek, Hebrew, and Runic letters

If you encounter text in Greek, Hebrew, or Runic orthography, please replace the garbage produced by the OCR program with the word GREEK, HEBREW, or RUNE, respectively, all in capital letters. Later, someone else will go through and add this text by hand.

Even if you know Greek or Hebrew, please don't transliterate the text into Roman letters. We will later search for the exact words GREEK, etc., and encode these non-Roman characters in a uniform way throughout the dictionary. If you transliterate these words, we won't know where to find them.

Examples:

Original text: **The Anglo-Saxon Rune X not only stands for**

Wrong: The Anglo-Saxon Rune X not only stands for

Right: The Anglo-Saxon Rune **RUNE** not only stands for

Original text: *named λύχνις στεφανική and*

PDF file: *named vrvts arf\aviK\at and*

Wrong: named **luchnis stefanike** and

Right: named **GREEK** and

Separating lines

There should be one empty line between dictionary entries. (In other words, there should be two newline characters between entries.)

This is all one entry. It's OK
that it's broken into four lines,
because there's no empty line
between the full lines.

This is the next entry. It's
separated from the last entry by
an empty line.

Sometimes one entry is incorrectly broken into two. Sometimes two entries are incorrectly run together. Please correct such errors. You'll often have to add or remove boldfacing in such cases.

Sometimes an entry is broken between two columns; it starts in the left column of the original page and ends in the right column. When this happens, join the pieces together into one entry. (Don't worry about entries which start on one page and end on another; just leave these as they are. We'll fix those later.)

Mixed up order

Occasionally, the OCR program gets confused when it's figuring out how the text is laid out on the page, and it puts some of the text in the wrong order. If this happens, please use the cut-and-paste commands in your text editor to fix the order.

Hyphens at the end of a line

Hyphens serve two functions: 1) to show that a word is broken between two lines, and 2) to show the internal divisions within a word (or compound word).

Unfortunately, this double use of hyphens creates an ambiguity, since it's not always obvious whether the hyphen should be kept or discarded when you're putting back together a word which has been broken across two lines.

Michelle has a healthy self-confidence when she talks about physics. Her understanding of the subject matter is obviously good.

When the lines are rejoined, the word *self-confidence* should keep the hyphen (we don't write *selfconfidence*), but *understanding* should not (we don't write *under-standing*). When you find a hyphen at the end of a line, there's no simple way to tell which kind it is; you just have to know.

If you have a clear judgment as to whether a hyphen should be kept or discarded in a particular case, follow your judgment (this will be easiest in the case of modern English words). If you aren't sure, though, then please type **&dash-uncertain;** in place of the problematic hyphen. This is little awkward, but it allows the problem to be marked for later consideration rather than forcing an immediate decision.

Example:

If the word *Man-drihtne* is broken across two lines and you're not sure whether to keep the dash, type `Man&dash-uncertain;drihtne.`

The :— symbol

The symbol sequence :— is frequently found in both dictionaries. Please type this as :-- which is a colon followed by two hyphens. Please don't use a one-em dash (—) even if your word processor allows you to type it. The reason is that the one-em dash is not available in all text editors, and it's much easier to manage things if everybody enters the data the same way.

Tables and graphics

Occasionally, an entry contains a formatted table with rows and columns. There are also rare entries which contain some kind of graphical diagram. Please type UNCERTAIN in these cases. We'll fix these later. Typing UNCERTAIN shows us where the problems are.

Page header

The page number and page header (e.g. ANDETTING - ANDRED) should be on a single line together, enclosed with the tags <HEADER>...</HEADER>.

Adding comments

Please don't add comments or questions to the text (for example, don't add comments like "I'M NOT SURE ABOUT THE LAST WORD.") If the notations described above aren't adequate (i.e., UNCERTAIN, ERROR, ILLEGIBLE, GREEK, HEBREW, RUNE), then please send a separate email describing the problem to kurisuto@unagi.cis.upenn.edu.

Please avoid inventing new tags. The three valid ones are <I> for italics, for bold, and <HEADER> for the header line.

Occasionally, you might find a special character which isn't in the list of entities. If you think you've found such a case, please send an email to kurisuto@unagi.cis.upenn.edu describing where you found the character, so that it can be added to the list of characters which we're counting as valid.

Thank you!

Thank you for your valuable help with this project!

Proofreader symbols

In terms of the final product, it doesn't make any difference at all what proofreader symbols you use while marking up the errors on paper. Use whatever works best for you.

If you'd like some suggestions, the following are a some symbols which you might find handy. Some of these are standard proofreader symbols, but others are ones have evolved for the unusual needs of this particular type of project.

Change the underlined character(s)
into the character(s) written above.
Non-standard.

a-^celeopiaⁿ;

Here is an example where a period
should be changed to a comma.
Non-standard.

terror_!

Here is an example where a
comma should be changed to a
period.
Non-standard.

Exon,_.

Insert a space.
Standard.

To cram, [#]fill;

Here is another way to show that a space should be inserted. Some people might find it a little easier to read, even though it is a non-standard proofreader symbol.
Non-standard.



hieĎá

Close up the space (delete the space character).
Standard.



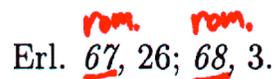
form.

Delete this character.
Standard.



Ác-leá

Set this italic text in roman (plain) text. (Use “ital” to indicate that something should be set in italics.)
Standard.



Erl. 67, 26; 68, 3.

Add an acute accent to this character. *Non-standard.*



móde,

Table of special characters

The following table includes the most frequently occurring special characters. These entity names are standard, except in cases where there is no standard entity name for the particular character.

The full list of entities currently recognized as valid is at:

http://www.ling.upenn.edu/~kurisuto/germanic/aa_character_encoding.html

If you find a character which isn't on the list below, please consult the full list, or just type UNCERTAIN after the word to have someone else take care of it.

General characters

þ	þ	Þ	Þ
ð	ð	Ð	Ð
æ	æ	Æ	Æ (i.e., AE ligature)
œ	œ	Œ	Œ
å	å	Å	Å
ø	ø	Ø	Ø
ǣ	&yogh;	Ƿ	þ-bar; (abbreviation for þæt)
		ł	&l-bar;

Characters with acute accent

á	á	Á	Á
é	é	É	É
í	í	Í	Í
ó	ó	Ó	Ó
ú	ú	Ú	Ú
ý	ý	Ý	Ý
ǽ	æ-acute;	Ǽ	Æ-acute;

Characters with umlaut (diaeresis)

ä	ä	Ä	Ä
ë	ë	Ë	Ë
ï	ï	Ï	Ï
ö	ö	Ö	Ö
ü	ü	Ü	Ü

Characters with circumflex

â	â	Â	Â
ê	ê	Ê	Ê
î	î	Î	Î
ô	ô	Ô	Ô
û	û	Û	Û

Characters with short sign (found in Latin words)

ă	&a-short;	Ă	&A-short;
ě	&e-short;	Ě	&E-short;
ĩ	&i-short;	Ĩ	&I-short;
ö	&o-short;	Ö	&O-short;
ű	&u-short;	Ű	&U-short;

Characters with long sign (found in Latin words)

ā	&a-long;	Ā	&A-long;
ē	&e-long;	Ē	&E-long;
ī	&i-long;	Ī	&I-long;
ō	&o-long;	Ō	&O-long;
ū	&u-long;	Ū	&U-long;

Greek letters

Please use these *only* when a single Greek letter is used to indicate an item on an ordered list. Please type GREEK in place of a full word in Greek.

α	α	β	β
γ	γ	δ	δ
ζ	ζ	ε	ε

Other characters

&	&	(Use & instead of the bare & character, because the & character itself begins an entity.)
<	<	(Use > and < instead of < and > because the < and > characters are used for tags like <I> or .)
>	>	
§	§	
¶	¶	
•	•	
~	˜	

The **˜** entity is not standard. The reason we use it is that the OCR program uses ~ for any character which it can't recognize, so it's easier if the checking system can treat all instances of ~ as errors. We use **˜** for those rare cases where the ~ character really does occur in the text, to distinguish it the errors.