Edward Jacobs
Protocols for Transcribing and Formatting Textual Witnesses for Scholarly Editing

These protocols stipulate how to transcribe printed texts into Word so that the resulting Word files will be maximally accurate to the source text or "textual witness," maximally amenable to the procedures of critical editing, and maximally compatible with other digital programs for scholarly editing, such as the collation software Juxta. Aside from stipulating uniform formats for font, paragraphing, and such, these protocols minimize the ways that the various default and automatic settings in Word may interfere with accurate transcription of the source text. They also include instructions for how to enter the source text into Word files that are least likely to distort the source text.

The basic rule for transcribing texts into machine-readable programs like Word is to keep the formatting as simple as possible, avoiding document-wide codes, which can interfere with editorial changes to the file and make its incompatible with other programs.

A. Word Preferences and Format Settings for Files (These Settings Change Default Settings in Word and Must Be Actively Selected)

1. Use Times New Roman 12 font for all text. This includes the title page(s); don't try to reproduce their font sizing.
2. Format all paragraphs to double line spacing. Also, in the format paragraph dialog box, check the box "Don't add a space between paragraphs in the same style." [Otherwise, Word will automatically add paragraph breaks not in the source text].
3. Use left justified text.
4. Number pages at bottom center.
5. Under Word>Preferences>Autocorrect>Autocorrect, uncheck all boxes EXCEPT "replace text." [Otherwise, Word will automatically change what you type in ways that may distort the source text].
6. Under Word>Preferences>Autocorrect>Autoformat>Apply as you type, uncheck all boxes. [Otherwise, Word will automatically change what you type in ways that may distort the source text].
7. Under Word>Preferences>Autocorrect>Autoformat>Replace as you type, uncheck all boxes EXCEPT "straight quotes with smart quotes," "symbol characters with symbols," and "internet and network paths with hyperlinks." [Otherwise, Word will automatically change what you type in ways that may distort the source text.]
8. Under Word>Preferences>Autocorrect>Autoformat>Automatically as you type, uncheck all boxes if they are not already unchecked (which is usually the default setting).
9. Under Word>Preferences>Spelling and Grammar>Spelling, uncheck all boxes EXCEPT "always suggest corrections," "ignore internet addresses," and "flag repeated words." [Otherwise, as you type Word will underline "misspellings" in ways that will distract you from accurate transcription.]
10. Under Word>Preferences>Spelling and Grammar>Grammar, uncheck all boxes. [Otherwise, as you type Word will underline grammatical "errors" in the source text in ways that will distract you from accurate transcription; moreover, when

you spell-check Word will suggest myriad "corrections" for (eccentric or historically/culturally specific) syntax that you have accurately transcribed.]

11. NOTE: THE ABOVE SETTINGS FOR "PREFERENCES" REFER TO WORD FOR MAC 2011. IF YOU USE WORD FOR WINDOWS (WORD 2007 OR 2010), YOU CAN GET TO MOST OF THESE OPTIONS BY CLICKING "WORD OPTIONS" > "PROOFING" UNDER THE WORD ICON. MAKE SURE THAT WITHIN THE "PROOFING" PAGE YOU CLICK "AUTOCORRECT OPTIONS," WHICH WILL TAKE YOU TO THE MOST IMPORTANT SETTING ABOVE. ONCE YOU ARE IN THE "PROOFING" WINDOW, THE CATEGORIES FOR SETTINGS ARE BASICALLY THE SAME AS ABOVE AND YOU CAN FOLLOW THE DIRECTIONS ABOVE.

B. Entering and Formatting Text

1. Use the tab key to indent the first line of paragraphs indented in the source text. DO NOT format paragraphs or use a style sheet to automatically indent paragraphs.

2. Use Word's paragraph formats for any needed block indents, hanging line paragraphs. DO NOT use tabs or spaces to reproduce indentation or other paragraph formats in the source text.

3. Use Word's italics code for italics, underline code for underlined words, etc. Note that punctuation and/or spaces following words in special fonts should not be formatted in special font. That is, select ONLY the characters in special font in the source for special font formatting in Word.

4. Use Word codes for all possible accents or special characters. If there is no Word code for the special character, then replace the source text character with "[SC]" for SPECIAL CHARACTER. You will later have to decide how to reproduce the special character.

5. If the source text contains illustrations or non-alphanumerical text, insert "[ILLUSTRATION X]" where the illustration occurs, substituting for "X" the illustration caption if it has one or otherwise its sequential number among other non-alphanumerical text within the source text.

6. If the text is eighteenth-century or before and contains the "long s," which looks much like the letter "f" but lacks the cross-bar on the f, transcribe that "long s" as "s." For example, if the text reads "trespafs," transcribe that as "trespass."

7. Do note reproduce original page breaks or numbers, except for the break between the title page and the text and ones between prefaces, dedications, etc., and the main text.

8. When formatting the title page reproduce letter case and font qualities (like bold or italics) as well as special fonts (such as Gothic). However, even if you switch from Times New Roman font, do NOT vary from the 12-point font size.

9. When transcribing the title page, reproduce line-breaks in the main title (such as "Memoirs of the Right Villanous John Hall") but do NOT reproduce line-breaks in longer paragraphs (such as the ones describing the nature of the edition and why it has been published in HUE and MJH). Reproduce centering or justification using Word's paragraph codes.

10. When you cannot clearly read the source text (for example, the source contains smudges, ink bleed-through, or it has been scanned in ways that cut off margins), make your best guess but put all characters about which you are visually unsure in brackets, as for example "oth[er] hand[,] it was…" If you are reasonably sure about characters that have been cut off by scanning, just enter those without brackets. For example if one line the source text reads "they retired each to" and the next one begins "eir own houses," but you can see that scanning has cut off the margin, it is reasonable to transcribe this as "they retired each to their own houses" rather than "they retired each to [th]eir own houses." If you have doubts, however, use brackets.
11. Do NOT reproduced text (such as lists) set in multiple columns in the source text, UNLESS the columns correlate substantively, with for example items in one column being a version of or response to parallel items in the other column(s). If there is no substantive correlation between columns, assume that the text was set this way to save space and transcribe them as one column, starting with the leftmost column and then continuing to the top of the next column to the right and so on. (Note: if multiple column text in the source text continues over a page break, transcribe items left to right column by column across page breaks UNLESS there is another obvious logic of organization in the source text, such as alphabetization, dating, or numbering. In such cases, transcribe the items into one column based on the substantive logic of organization in the source text.
12. If the source text contains bold lines indicating section breaks (called "leads" typographically), transcribe those "leads" by entering [leaded break] in a separate paragraph.
13. Do not reproduce the gathering signatures (notations such as "A" or "B2") that appear in the lower right corner of some pages in early hand-press texts.
14. Do not reproduce "catchwords," the words in the lower-right corner of each page that anticipate the first word of the following page. Catchwords were used to help printers and binders put pages in order and can be useful during collation and emendation, but they are not technically part of the text's substance and hence should not be transcribed.

C. Spell-Checking Files
Note: Word Spell-Checking is NOT a substitute for scholarly proofreading of transcribed texts against source texts as stipulated in Jacobs, "Protocols for Scholarly Proofreading." However, if done according to the protocols below, Word's Spell-Check can make scholarly proofreading of transcribed texts easier, by highlighting and enabling you to correct keyboarding errors.
1. ALWAYS check the spellings that Word suggests against the source text before you accept Word's suggestion, because otherwise you may inadvertently distort the source text. Alternatively, under Word>Preferences>Spelling and Grammar>Spelling, you can uncheck "always make suggestions" and directly check all flagged "errors" against the source text.
2. NEVER use Word's "Change All" Spell-Check option, because doing so may invisibly and inaccurately "regularize" variant spellings in the source text of the same word.

D. Saving Files as Rich Text (.docx)
1. Simply save the file as normal

E. Saving Files as UTF-8 Plain Text (.txt)
Note: Saving documents both as Word Rich Text files and as UTF-8 Plain Text files is necessary because most digital programs for scholarly editing (such as the collation software Juxta) cannot accept Rich Text files because Rich Text contains proprietary format coding that interferes with the transfer of files across program platforms. Basically, saving a Word Rich Text file as a UTF-8 Plain Text file strips the Rich Text file of all paragraph formatting and all font formatting except for case formatting, but otherwise preserves substance and paragraph breaks. Consequently, any Rich Text transcription that has been saved as a UTF-8 Plain Text File will need to be visually check for font and paragraph formatting against the source text.
1. AFTER you save your Word document as a Rich Text (.docx) file, go to File>Save As and select Plain Text (.txt) from the Format drop down options. CRUCIAL NOTE: IF YOU SAVE YOUR WORD DOCUMENT (BY DEFAULT CODED IN RICH TEXT) AS A UTF-8 PLAIN TEXT FILE BEFORE YOU SAVE IT AS A (DEFAULT) RICH TEXT FILE, YOU WILL LOSE ALL OF THE FONT AND PARAGRAPH FORMATTING ENCODED IN YOUR ORIGINAL WORD DOCUMENT.
2. Click Save. (Doing so doesn't yet save the file in Plain Text UTF-8 coding).
3. In the dialogue box that appears, under "Text Encoding," select "Other encoding."
4. In this same dialogue box, select Unicode 5.1 UTF-8 from the options for encoding. NOTE: Depending upon your version of Word, the Unicode number may differ from that cited above, but you must select whatever Unicode option your Word version offers that includes a "UTF-8" encoding.
5. Click OK. Your file will then be saved with the same filename and in the same location as your original Rich Text file, but with the format suffix .txt rather than .docx.