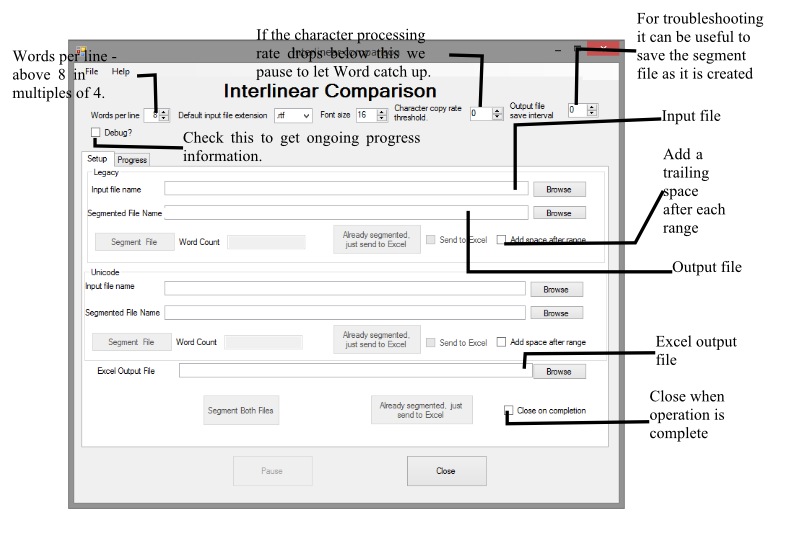
Stephen Palmstrom 17 May 2017

# Introduction

After documents have been converted from legacy fonts to Unicode, it is useful to compare the legacy and Unicode documents to see how successful the conversion has been. This can be done by looking at the documents side-by-side, but this can become difficult, particularly as there may be subtle differences between the fonts that may change the pagination. It is therefore useful to be able to see the two documents in one place, line-by-line. This program does this.

This document describes how to use Interlinear, a program that takes two Microsoft Word documents, generates segmented files (i.e. files with a fixed number of words per line) and copies them into an Excel spreadsheet that gives a line-by-line comparison of the two documents.

# Licensing and Copyright

The program was developed for MissionAssist, so it is copyright to that organisation, but distributed under the GNU General Public License (<http://www.gnu.org/licenses/gpl.txt>).

# Installation

Open the Interlinear folder in the zip file and execute setup.exe. Until the code can be digitally signed, Windows 8 may throw up security warnings that you will have to ignore.

# Execution

After installation, you will see this:

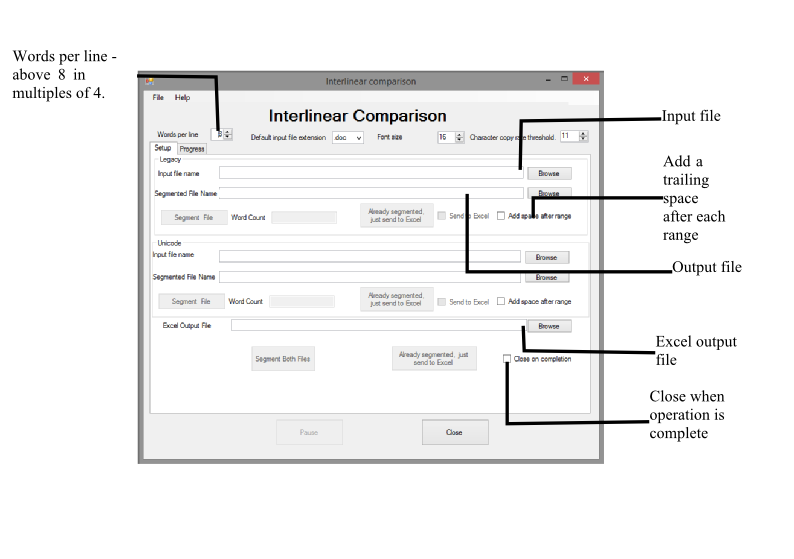
Click the Browse buttons to insert the file names for legacy, Unicode and the Excel spreadsheet. The buttons allowing you to segment the files and create the Excel output will be enabled as you add the file names. The installation process is supposed to put this in your start menu. This means that the *Already segmented, just send to Excel* buttons and the *Send to Excel* tick boxes will only become enabled when you give the program an Excel file name to which to send the data and the former when the segmented files exist.

Figure 1 Explaining input and output files.

Once you have entered the file names you can click on the Segment buttons. If the Excel file name has been entered, the default is that the segmented data are sent to Excel. If you have all the file names completed, the Segment Both button will do segment and create the Excel file in one operation. The progress tab reports progress after you have started the process. Unfortunately, because the program is doing much copying and pasting, it runs very slowly at present, though we hope to improve performance in future.

The program now remembers the last folders you used for input and output.

# The Excel File

The Excel file consists of one worksheet with the legacy and Unicode text interleaved.

# A glimpse under the bonnet

This application creates hidden instances of Microsoft Word and Excel. If instances are already open, Interlinear will connect to them and remember that it has done so. It will make them invisible while it runs to improve performance, and make them visible again when it finishes. If the application {Word or Excel) is not running before you start, it should shut down when Interlinear exits, but in the event of a program crash, they may remain running and have to be shut down manually using Task Manager.

# Modus Operandum

The program first opens the input document read-only, and saves it as the output document. Because we are trying to compare all the text in the document, it copies headers, footers, footnotes etc. to the main part of the document. In Word-speak, it copies the text content of all the stories[[1]](#footnote-1) that are not the main story to the main story. This means we have to copy much less than if we copied the whole content to a new document.

The program then cleans up the text, removing text boxes (there may not be any at this point in any case), converting tables to text, also removing various control characters, graphics, figures etc. etc., but it copies any text associated with them. It removes multiple spaces, before replacing spaces with paragraph markers so we are left with one word per paragraph. It then merges these paragraphs into segments of words, the number selected by you, but the default is eight preceding them with a single quote[[2]](#footnote-2). It finally selects and copies and pastes the paragraphs into Excel, the legacy text in odd rows and Unicode text in even rows, line by line to preserve the various fonts that may change mid-word.

# Copying to Excel

This version does a copy and paste to Excel, line by line. Unfortunately, it is the slowest part of the Interlinear operation now that copying is much faster than in previous versions. For the exercise to be useful, we need, however, to preserve formatting. It is possible that future versions will be able to copy the information more efficiently.

1. A Word document has a number of so-called *stories*, such as headers, footers, footnotes, endnotes etc. as well as the main story which is the major part of the document. [↑](#footnote-ref-1)
2. This is necessary because Excel treats any text starting with an equals sign as a formula, not text. [↑](#footnote-ref-2)