Intelligent Archive

Budgerigar Version

Developed at the Centre for Literary and Linguistic Computing, University of Newcastle, Australia

Director: Hugh Craig
Software developers: R Whipp, Michael Ralston

Introduction

What is the Intelligent Archive?

The Intelligent Archive program is a Java based piece of software used for text analysis within the University of Newcastle's Centre for Literary and Linguistic Computing (CLLC). The software is used in various different ways by the Centre researchers who are focusing on different aspects of text analysis. Professor Hugh Craig is the chief architect of the IA's development; much of the core functionality that is required for Prof. Craig's work also applies to others working within the CLLC.

The typical CLLC project involves preparing a set of texts for computational stylistics operations, with the ultimate purpose of determining authorship of a disputed literary work, or analysing the style of a work or group of works. The IA serves these projects by organising sets of texts and making word counts which can be exported for analysis in an external spreadsheet or statistics program. It is an interface to an archive of texts, and incorporates a range of counting functionalities which can be determined by the user, hence is an 'intelligent archive'. While most text-processing programs focus on more linguistic outputs, such as concordances, or lists of the commonest collocates of a given word, the IA's primary function is more statistical, centred on producing frequency counts of words.

What does it do?

The Intelligent Archive Budgerigar software currently provides the following core facilities:

- Management of individual texts of different formats within a virtual library or repository
- Management of text sets, which are user-created groups of these texts
- Word frequency analysis on individual texts, tagged sections within texts, text sets, contiguous block segments of a specified size within texts, etc.

System Requirements

The Intelligent Archive is written using the Java platform. As such it is able to run on any operating system which supports Java and therefore will require the installation of a Java Runtime Environment. This can be downloaded, free of charge, from http://www.java.com

The specifications of the computer used will vary according to which features of the software you wish to use. The core functionality only requires a very basic system with at least 512MB of memory.

The software does not require a fast CPU, however, it will be able to provide its results quicker if equipped with a quicker CPU. The software does not currently benefit from being used on a system with multiple CPUs or CPU cores.

The software itself uses less than 1MB of disk space. You will also require enough disk space to store all texts added to the text repository.

Installation and Running

You will have been provided with a ZIP archive containing all of the files and directories necessary to use the Intelligent Archive. You will need to extract this archive to a location of your choice, for example you may simply put the "Intelligent Archive" directory contained in the archive directly on your desktop.

There are two ways to start the Intelligent Archive. Inside the Intelligent Archive directory there will be two files: "Intelligent_Archive.jar" and "Intelligent_Archive.bat". Double clicking on the ".jar" file will start the program with Java's default memory allocation. This is generally sufficient for all core functions of the software. With exceptionally large or numerous text files, you may need to use the ".bat" file, which will requisition additional memory. Double clicking on the ".bat" file will allocate the program this memory.

The file named "teilite.dtd" is used by the Intelligent Archive to validate tei files. You should not open, modify or delete this file.

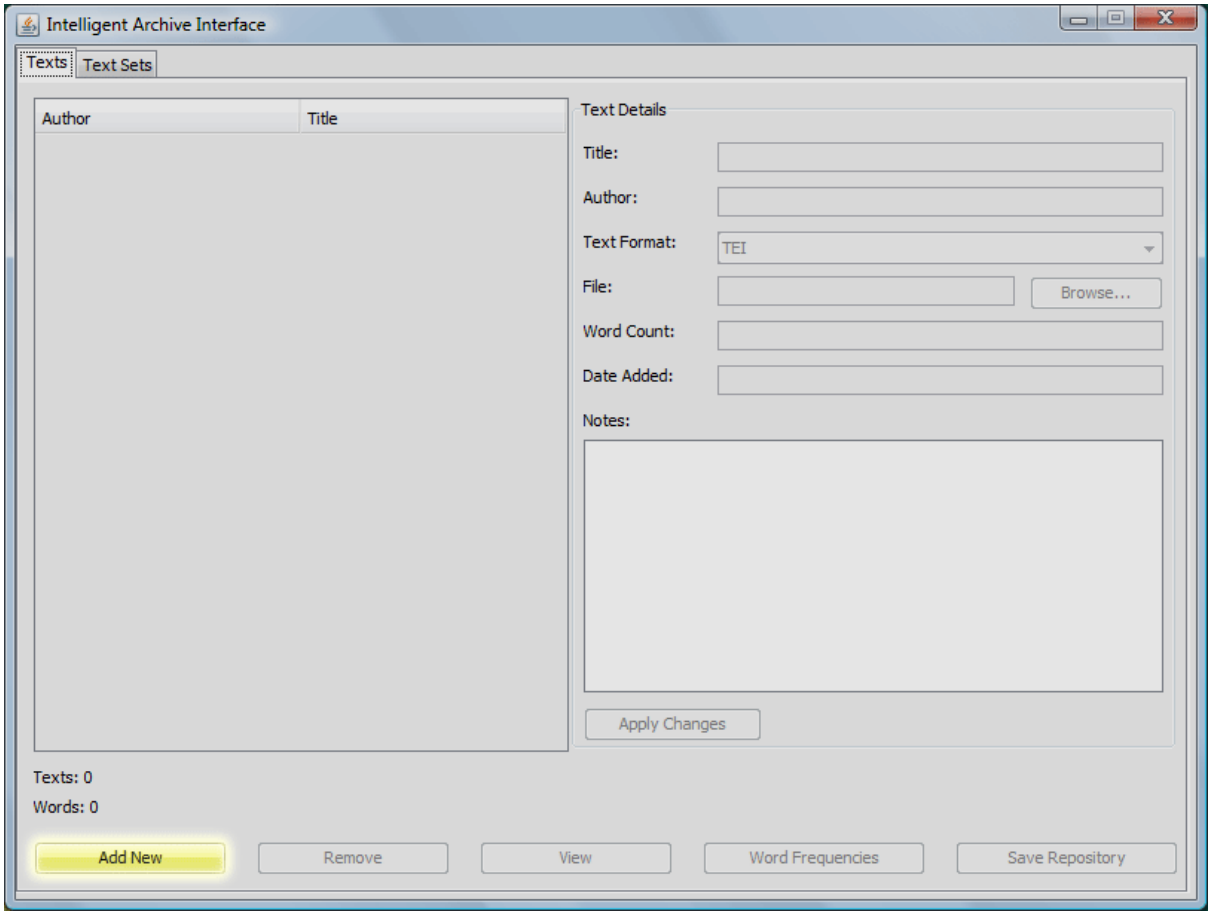
The file named "RepositoryData.ser" is created by the Intelligent Archive, and contains information about the texts you have added to the software. This is a background file used by the software which you should not open, modify or delete this file.

Core Functionality

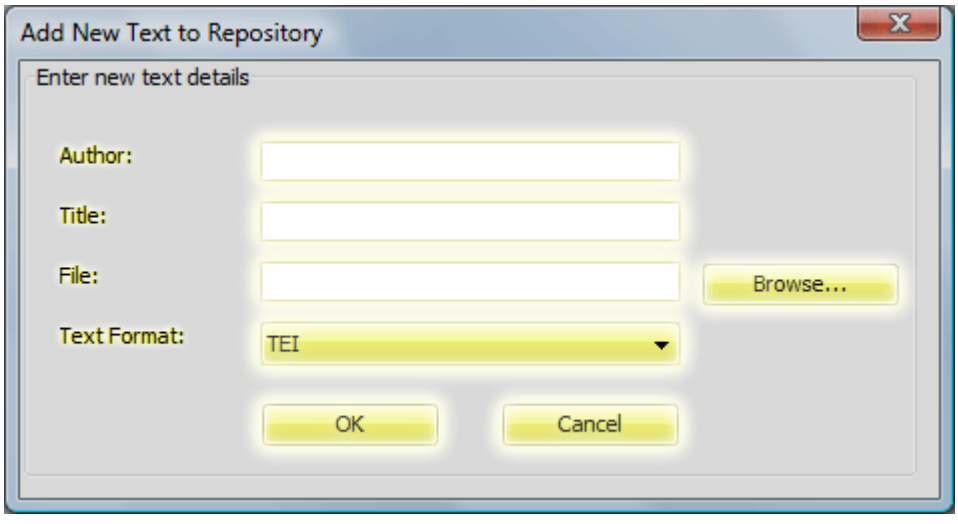
The Text Repository

The main purpose of the Intelligent Archive is to calculate the frequencies of words contained in a text. In order to do this, the software must first be provided with the text(s) you wish to examine. This process stores the text and various information about the text in a database, the Intelligent Archive's "Repository". This allows reuse of the text at a later time.

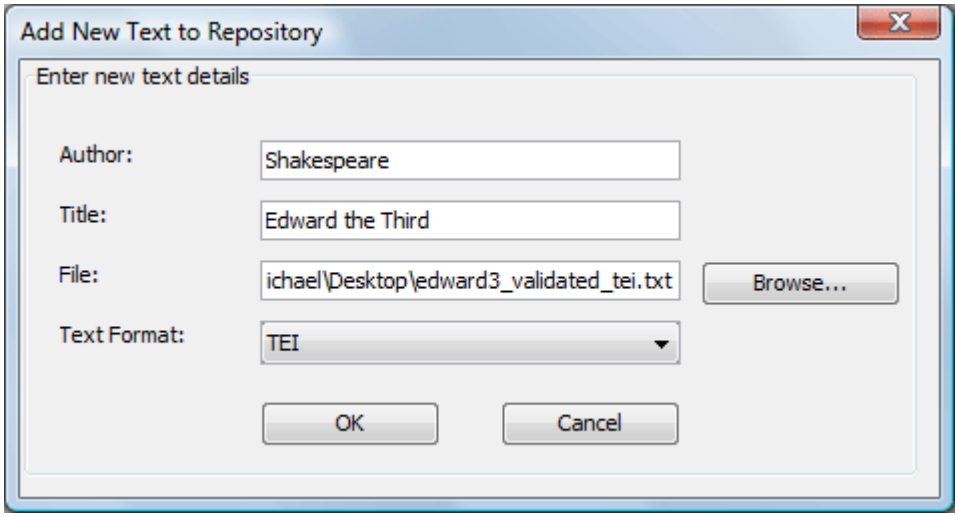
Upon first starting the Intelligent Archive, you will have a window which contains no texts. To add a text, click the "Add New" button, in the bottom left of the window.



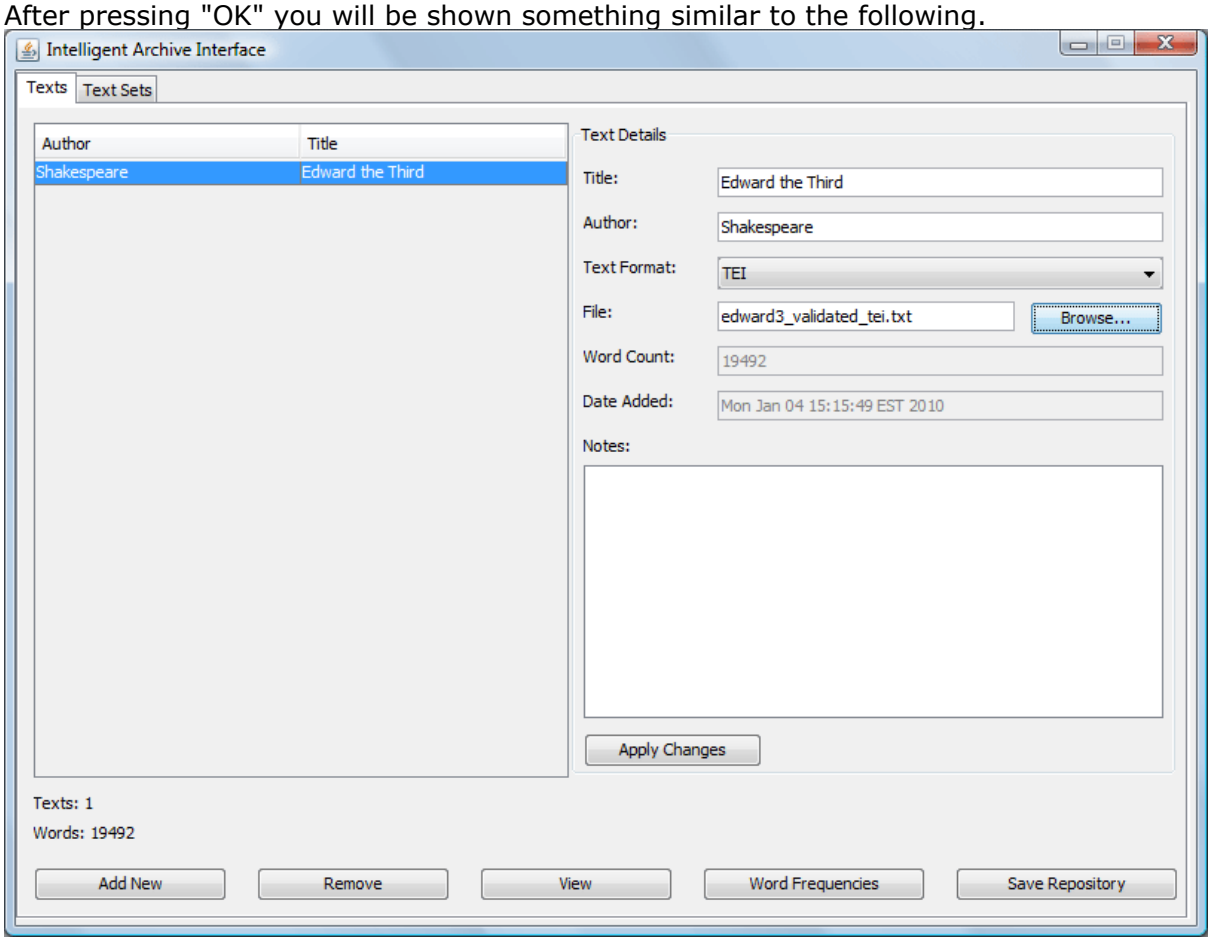
You will then be presented with the following window. Enter information in all fields. The information stored here will be used to reference the text inside the Intelligent Archive's repository. This information will not affect the results of Word Frequency calculation. Click the "Browse button" and locate the file on your computer that you wish to add to the repository. In the Budgerigar version two texts formats are supported: 'TEI' and 'Hybrid'. The TEI format requires a text valid according to any version of the Text Encoding Initiative protocols. The Hybrid version accepts plain ASCII text.



For example, entering the following information for Shakespeare Edward the third:

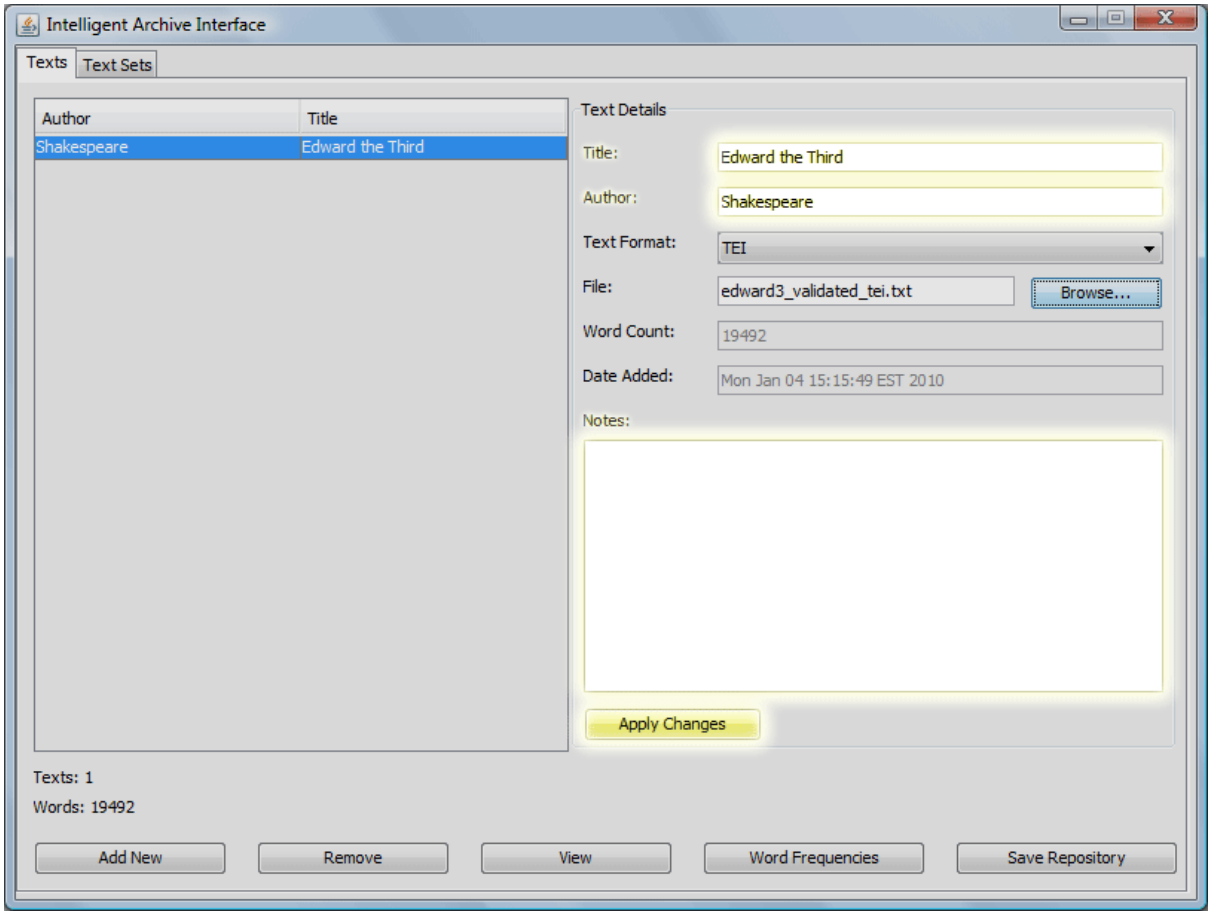


After pressing "OK" you will be shown something similar to the following.



To save the information about the new text in the Repository of the program so that it will be available next time you open it, click Save Repository.

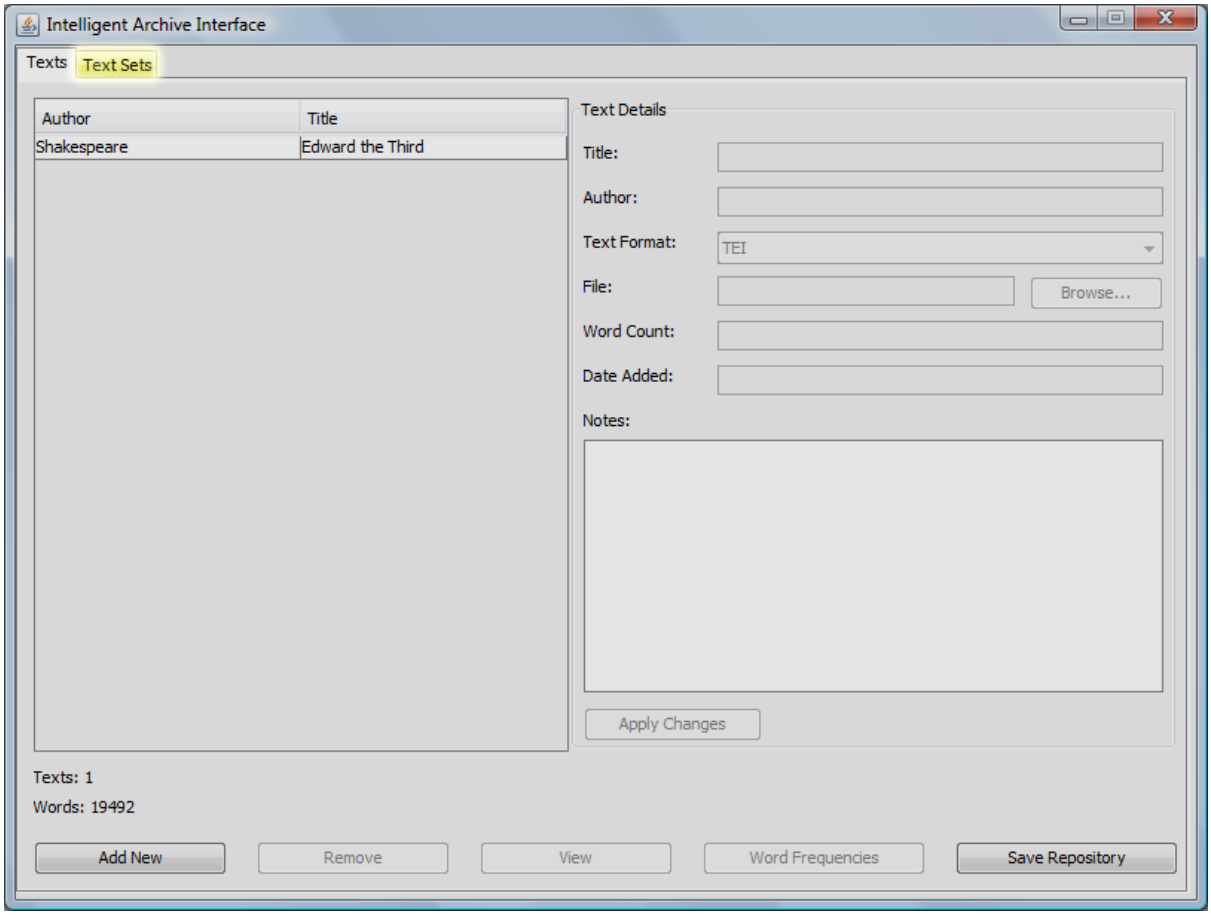
You may now modify the information stored in the repository by click the following highlighted fields, then pressing the "Apply Changes" button.



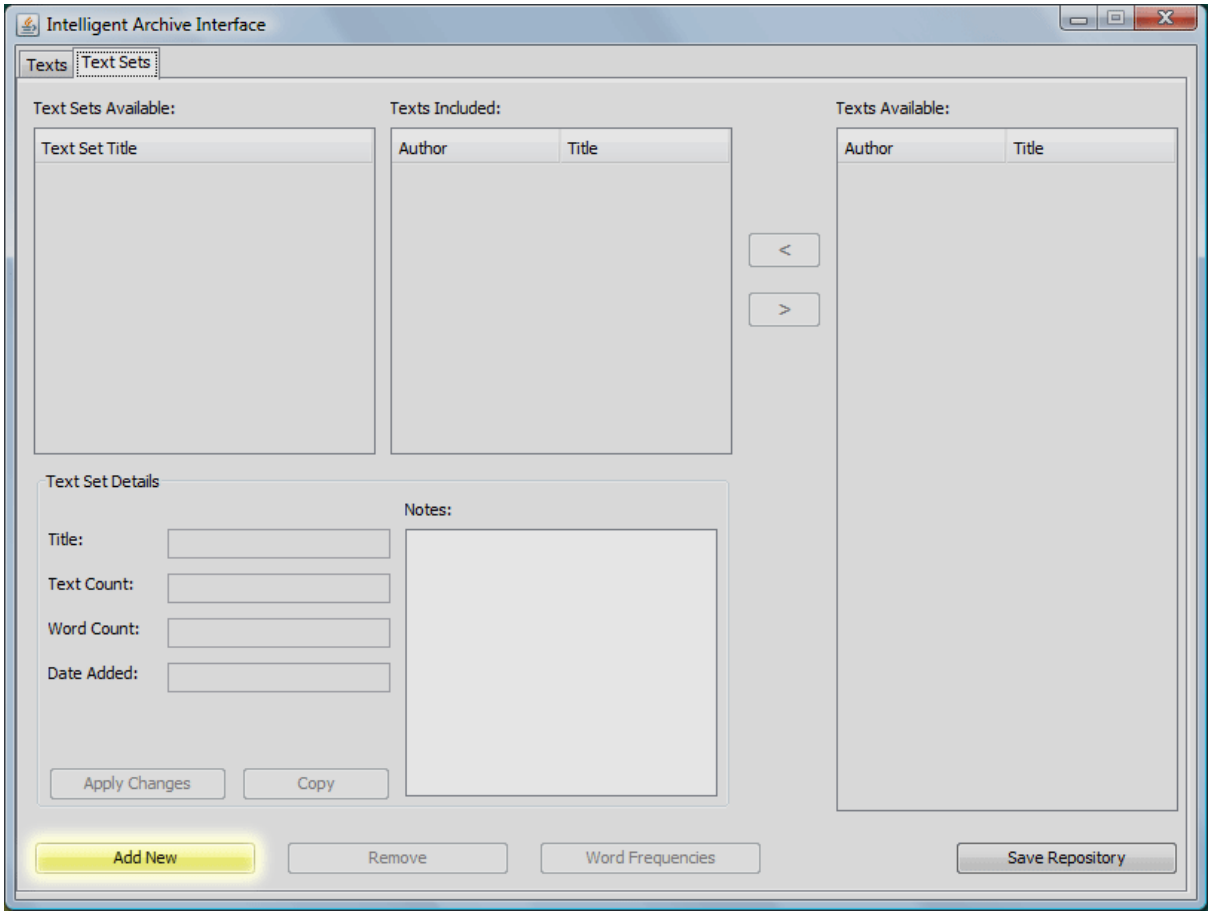
Text Sets

Text sets are groupings of texts. They usually contain more than one text. They can be used to perform frequency analysis on multiple texts at once. This could be used to compare the frequency analysis of several works by the same author, or to collect frequency data at one pass from a large, diverse corpus of texts.

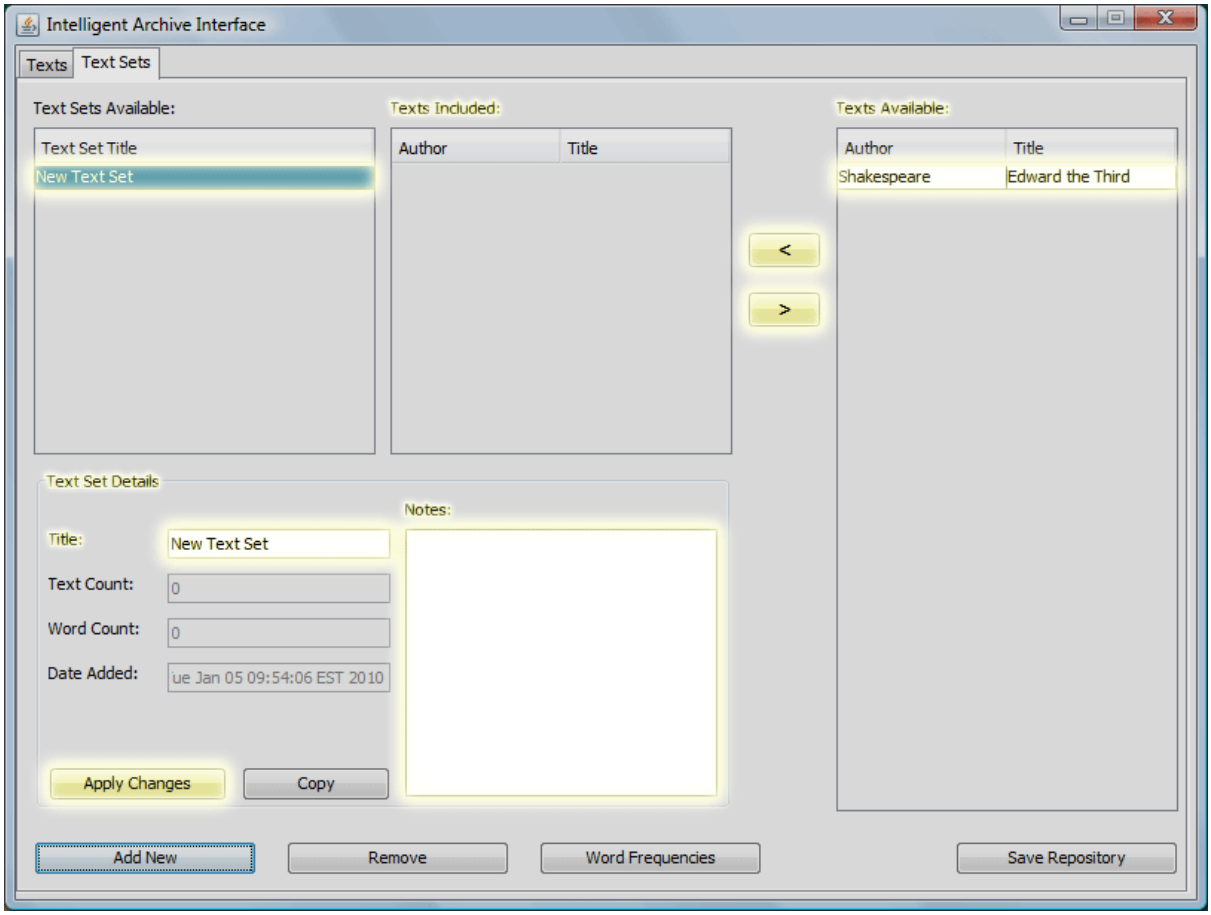
To obtain the Text Sets interface, click the "Text Sets" tab at the top of the window.



To create a Text Set, click the "Add New" button in the bottom left.



You will now see a window similar to below. The new text set will automatically contain the title "New Text Set". To modify this title, select the "Title" field, under the "Text Set Details" heading. You can also enter notes for your own reference in the text box under the "Notes" heading.



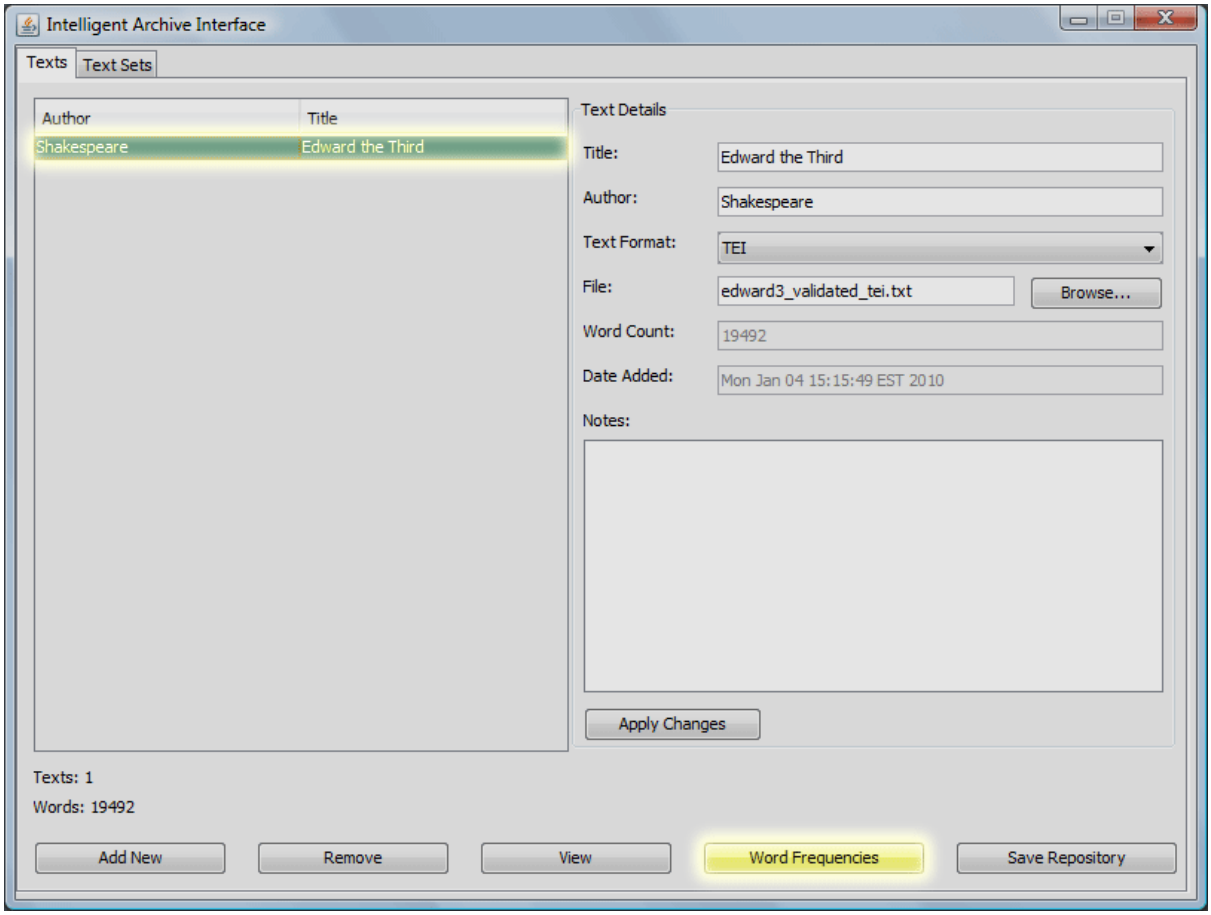
To save this text set for future use, click Save Repository.

To add texts to the set, select the text from the list under the heading "Texts Available", then click the left arrow, "<". The text will the appear in the list under the heading "Texts Included". Likewise to remove a text from the set, select it in the "Texts Included" list and click the right arrow ">".

Word Frequencies

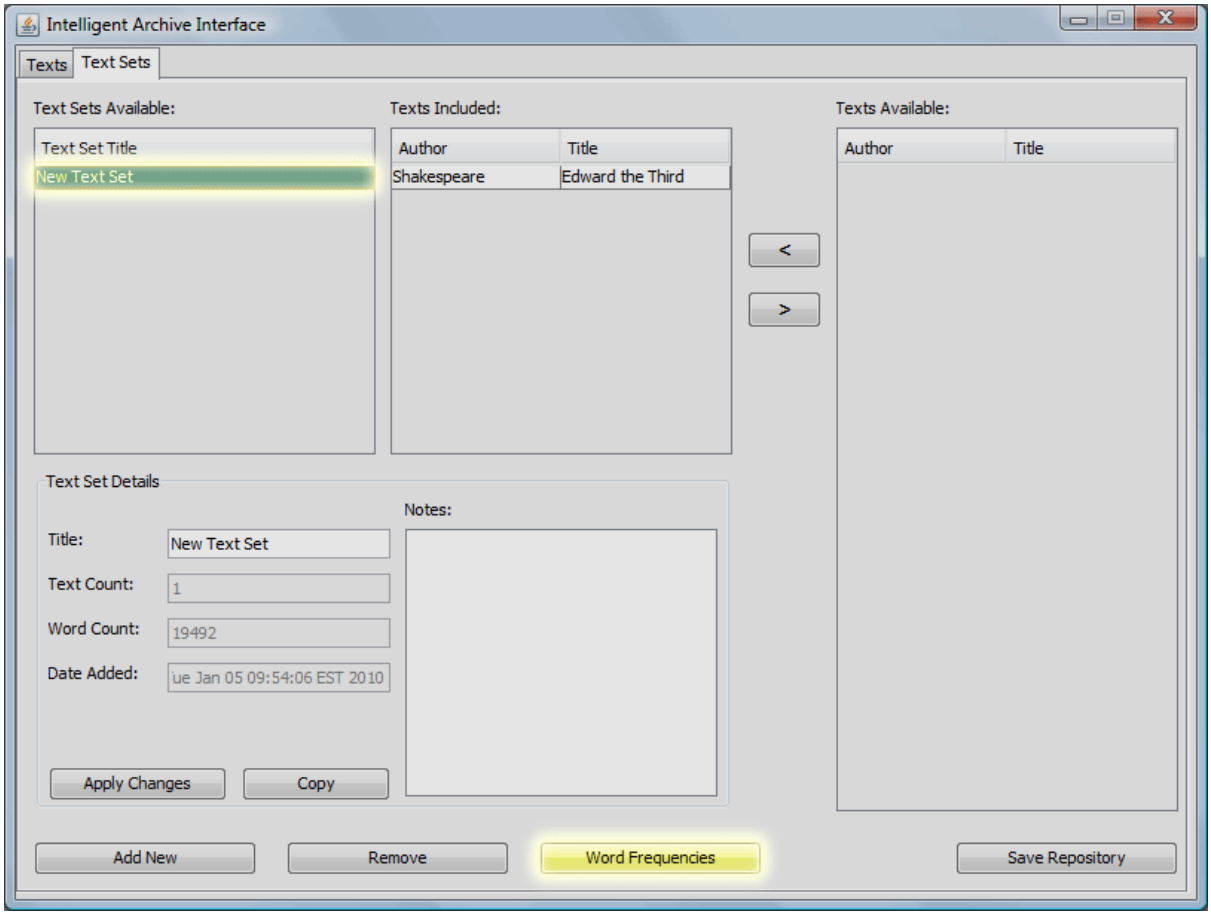
Text

To calculate the word frequencies for a single text, first select the text you wish to use from the list of texts, the press the "Word Frequencies" button.



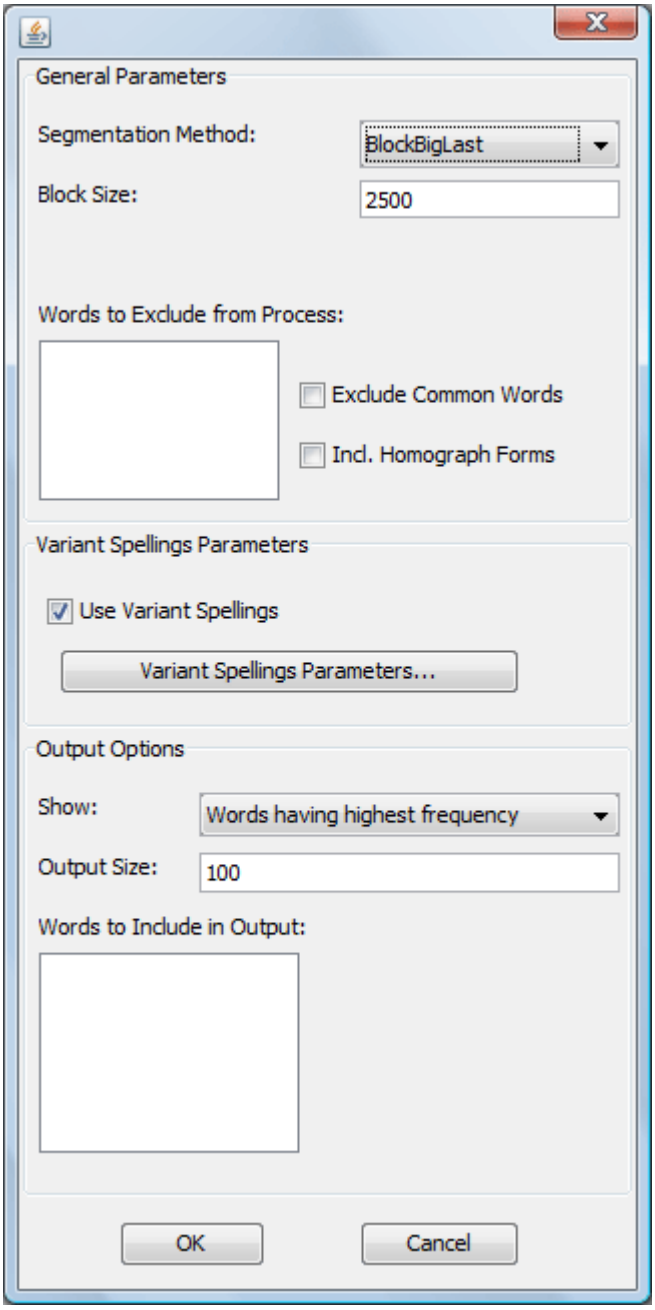
Text Set

To calculate the word frequencies for a text set, first select the "Text Sets" tab, then select the text set from the list of "Text Sets Available", then press the "Word Frequencies" button.



Parameters Window

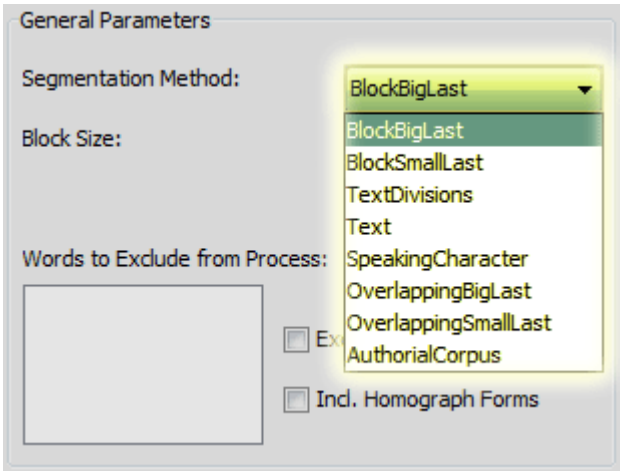
You will then be displayed a window similar to below. There are several options here so we will go through them one at a time. Clicking the "OK" button will cause the Intelligent Archive to process the select Text(s) and display the Output Window.



General Parameters

Segmentation Method

There are several different methods available for cutting up the text (or text set) into segments. The following table describes them and their parameters.

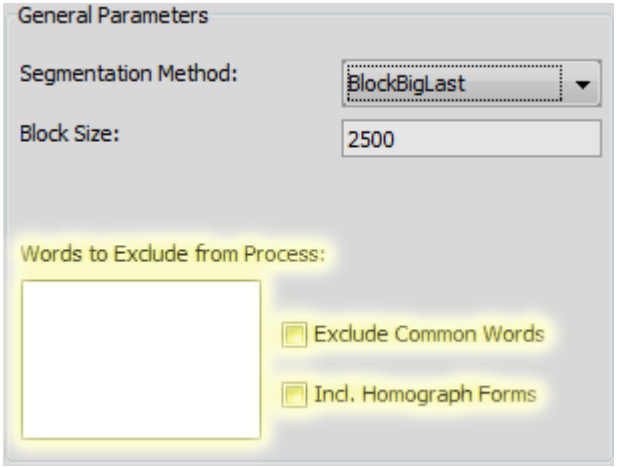


Segmentation Method	Description	Parameters
BlockBigLast	Divides the text up into blocks of size **Block Size**. If the final block is smaller than **Block Size** then it is appended to the previous block. Eg. Text with 5550 words, block size 1000. Results in 4 x 1000 word blocks + 1 x 1550 word final block.	**Block Size** - How many words you want in each block
BlockSmallLast	Divides the text up into blocks of size **Block Size**. If the final block is smaller than **Block Size** then it is left alone. Eg. Text with 5550 words, block size 1000. Results in 5 x 1000 word blocks + 1 x 550 word final block.	
TextDivisions	Only text inside certain 'div' tags will be processed.	**Text Marker** - What tag you wish to use to identify text to be processed, e.g. 'div1', 'div2', etc.
Text	Entire text will be used as a segment. When used on a single text, will result in a single segment. When used on a text set, will result in one segment per text.	None
SpeakingCharacter	Will segment according to the who="" attribute of TEI speech tags. Eg... <sp who="John">	None

OverlappingBigLast	Divides the text up into blocks of size **Block Size**. The start of each block is **Advancement** words further into the text than the previous one. If the final block is smaller than **Block Size** then it is appended to the previous block. Eg. Text with 5550 words, block size 1000, and advancement size 200. Results in 22 x 1000 word blocks, word 1-1000, 201-1200, 401-1400 and so on. The final block will be 1150 words, from word 4401-5550.	**Block Size** - How many words you want in each block **Advancement** - How many words along in the text you want the next block to start
OverlappingSmallLast	Divides the text up into blocks of size **Block Size**. The start of each block is **Advancement** words further into the text than the previous one. If the final block is smaller than **Block Size** then it is appended to the previous block. Eg. Text with 5550 words, block size 1000, and advancement size 200. Results in 23 x 1000 word blocks, word 1-1000, 201-1200, 401-1400 and so on. The final block will be 950 words, from word 4601-5550.	
AuthorialCorpus	Only available when calculating the word frequencies of a text set, and only useful for a text set with multiple authors. Segments by author.	None

With TextDivisions you can define your own segments within the text (e.g. to divide a play by acts or scenes, or a novel by volumes or chapters). OverlappingSmallLast and OverLappingBigLast, also known as 'rolling segments', are useful where you are unsure where the 'natural' divisions within a text are, and want multiple segments which only differ by a small amount of text, while maintaining a segment size minimum. (An instance would be a text begun by one author but completed by another, where the handover point is unknown.)

Words to Exclude from Process



Words entered here will not be counted or output. Words can be one per line or space separated.

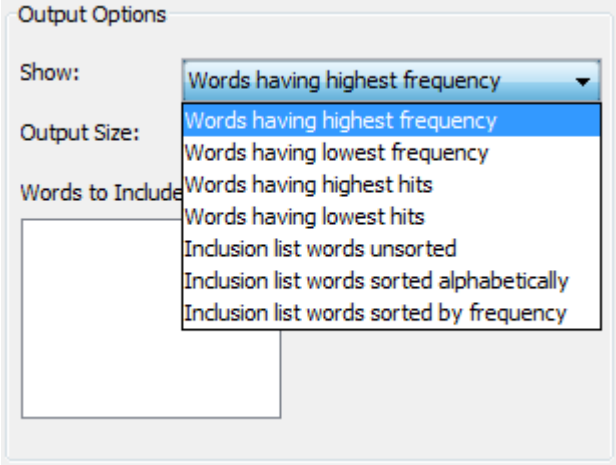
Homograph Forms

The TEI defines tags attached to words as 'entities'. The beginning of an entity is marked with an ampersand and the end with a semi-colon. The TEI Document Type Definition which is included as a separate file in your Intelligent Archive folder defines a set of entities which refer to grammatical function, 'noun', 'preposition' etc. The IA refers to words with these entities as 'Homograph Forms,' i.e. identical word-forms that have distinct grammatical functions. By default, the Intelligent Archive ignores the homograph part of words. For example the words to&H4; and to&H9; in a TEI document will both be counted as the word "to". Checking the box labelled "Incl. Homograph Forms" will change this behaviour. It will result in words with homograph tags being counted as separate words.

Homograph tags will be translated into human-readable text according to the definitions in the "homographs.txt" file contained in the Config directory, e.g. to&H4; will be translated to to[preposition] and to&H9; will be translated to to[infinitive].

Output Options

The report produced by the Intelligent Archive has some options available to filter the words which appear in the output. The following options are available:



Words having highest frequency

This option will display words sorted by the highest frequency. Frequency is the number of times a word appears in a block. The number of words displayed is determined by the output size field. Any words entered into the field Words to Include in Output will be shown in addition to the words selected by highest frequency.

Words having lowest frequency

This option will display words sorted by the lowest frequency. The number of words displayed is determined by the output size field. Any words entered into the field Words to Include in Output will be shown in addition to the words selected by lowest frequency.

Words having highest hits

This option will display words sorted by the highest hits, hits being the number of blocks a word appears in. The number of words displayed is determined by the output size field.

Any words entered into the field Words to Include in Output will be shown in addition to the words selected by highest hits.

Words having lowest hits

This option will display words sorted by the lowest hits. The number of words displayed is determined by the output size field. Any words entered into the field Words to Include in Output will be shown in addition to the words selected by lowest hits.

Inclusion list words unsorted

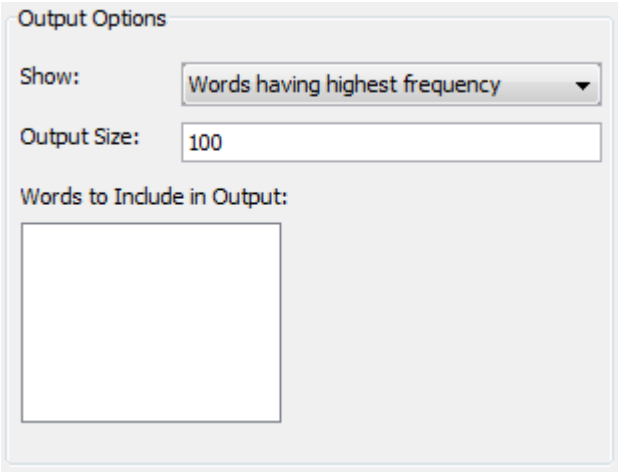
This option will display the words which are listed in the "Words to Include in Output" field. The words will appear in the order in which they are entered into that field.

Inclusion list words sorted alphabetically

This option will display the words which are listed in the "Words to Include in Output" field. The words will appear in alphabetical order.

Inclusion list words sorted by frequency

This option will display the words which are listed in the "Words to Include in Output" field. The words will appear in order of highest to lowest frequency.



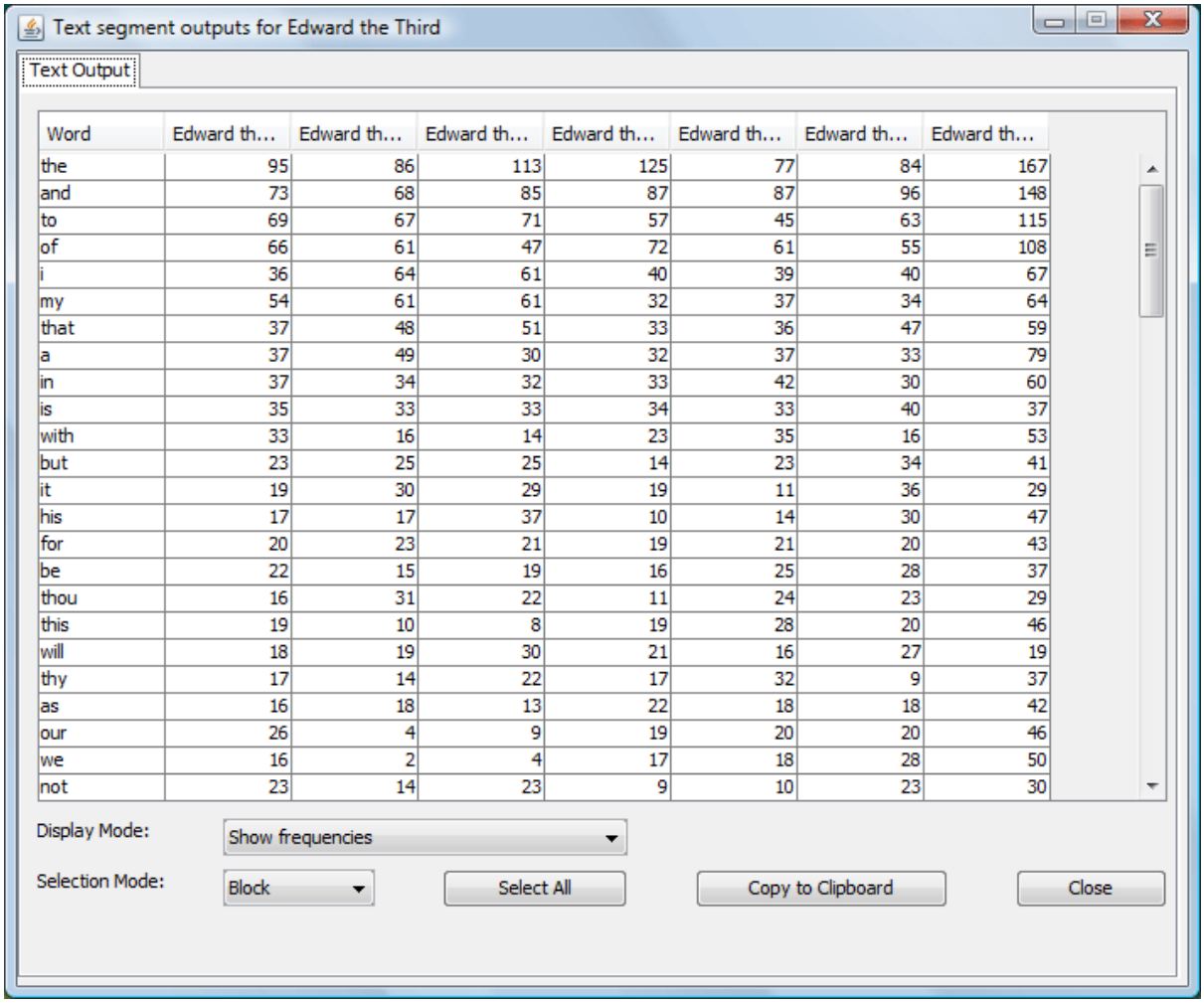
Output Size

This specifies how many different words will appear in the output. This is only relevant for the first four Output types. The inclusion list output types will ignore this field.

Words to Include in Output

For the "Words having" output types, the words listed here will appear in addition to the words selected by frequency or hits. For the "Inclusion list" output types, the words listed here will be the only words output.

Output Window

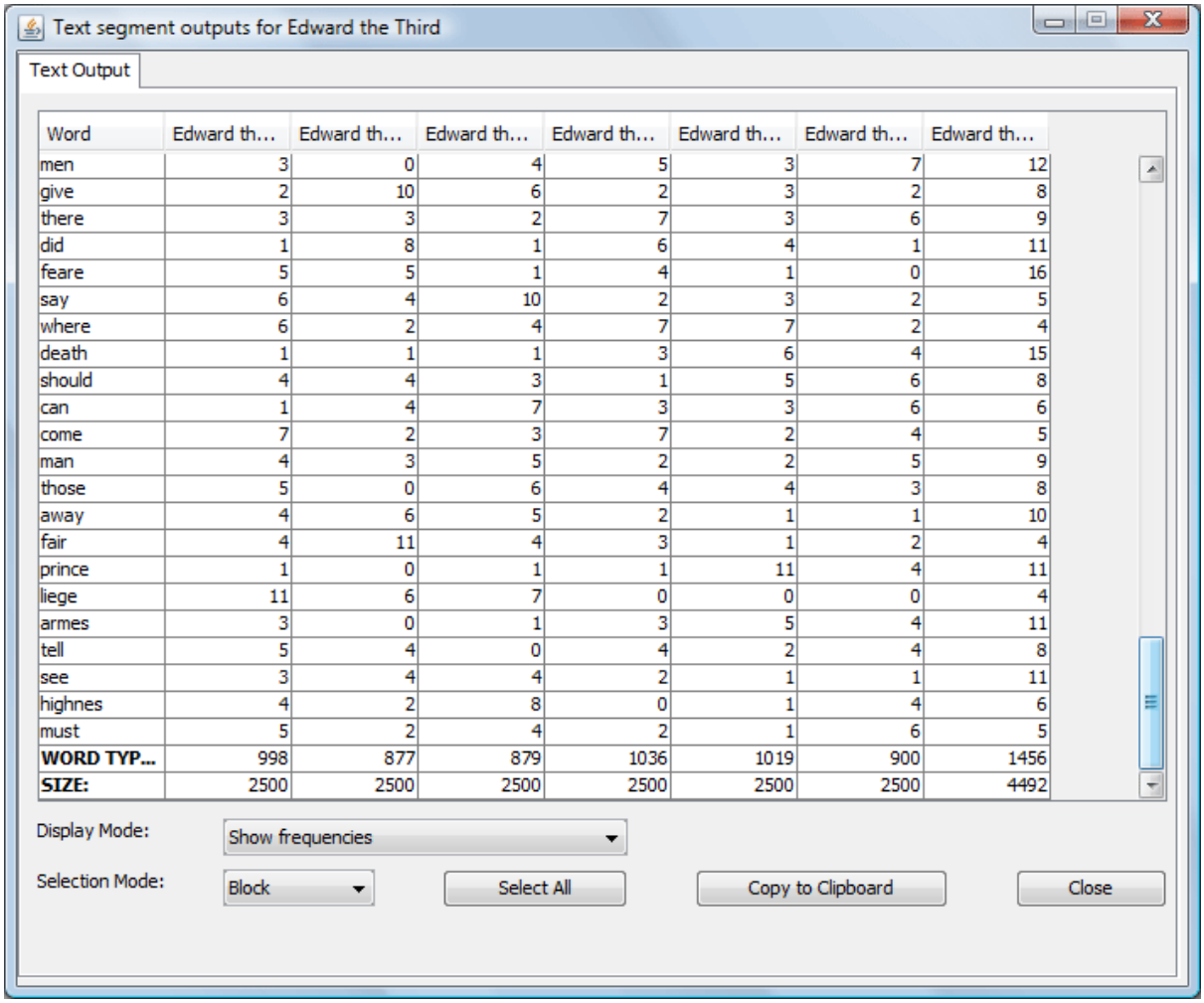


The Output Window will look similar to the above. The words are listed down the left, and the frequencies (or hits, if you selected hits in the output options) are shown for each segment.

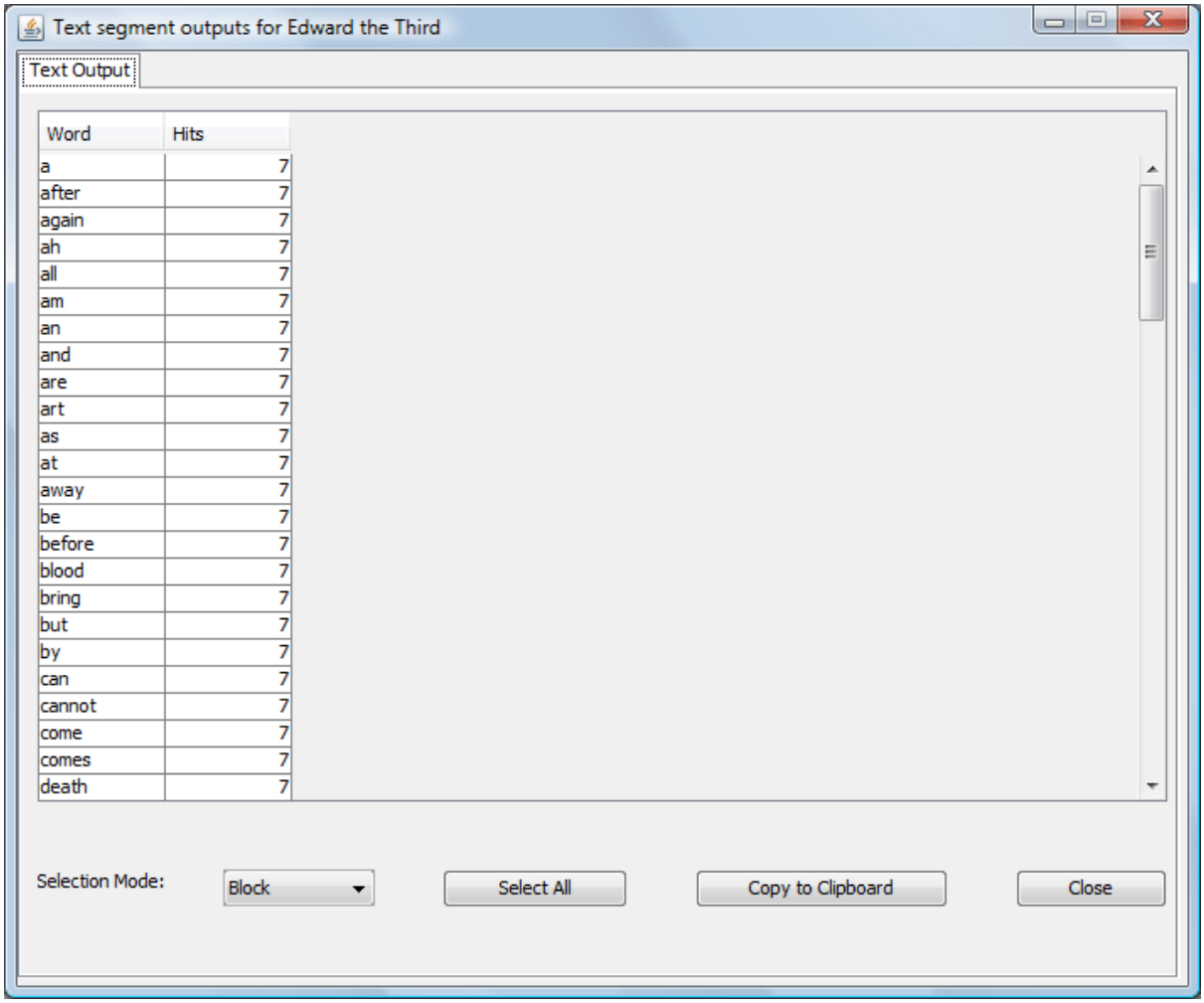
For frequencies output, there are two Display Modes available, "Show Frequencies" is selected by default, also available is "Show Proportions". These are the frequencies divided by the 'Size' of the text, i.e. the total number of words in it.

You can select cells in the output window by clicking and dragging your mouse. This will select the cells from the table in blocks by default. You can also change Selection Mode to column or row to more easily select the cells you require. The Select All button will select the entire table. Once cells are selected you can copy them to the clipboard with the Copy to Clipboard button, or you could also use the standard ctrl+c key combination. You can then paste the data into a spreadsheet or other statistical package.

Some users have experienced problems copying very large tables of data from the Intelligent Archive. If you have this problem it is suggested that you select half of the data and copy it, then select the other half and copy it.



This shows the output window, scrolled all the way to the bottom. You can see the final two rows display "Word Types" - the number of different words in the block and "Size" - the total number of words in the block.



Finally, this is a window showing what the output will look like if you selected "hits" from the output options.