

# WordStat v5.1 - Addendum

WordStat is a text mining and content analysis module specifically designed to study textual information such as responses to open-ended questions, interviews, titles, journal articles, public speeches, and electronic communications. Version 5.1 extend the capabilities of that program by the addition of new options and new features in many areas. This document updates the information in the WordStat v5.0 manual by presenting these new features as well as an updated description of the existing ones.

## Dictionaries Page

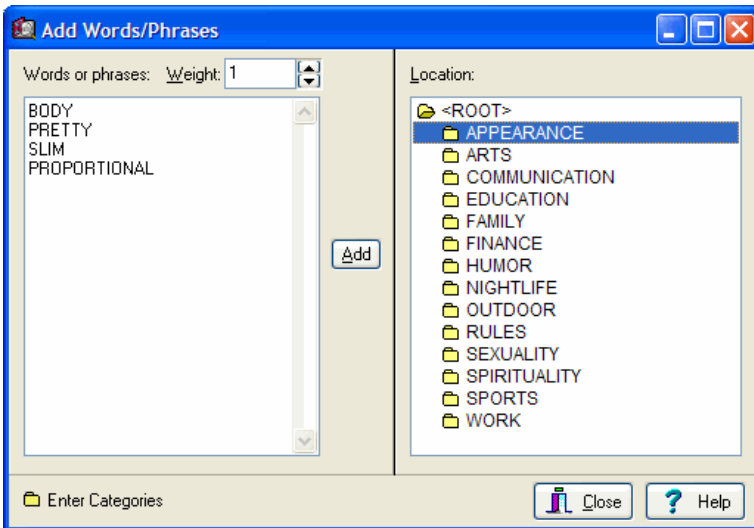
**SHOW WARNINGS** - Some items in an exclusion list or categorization dictionary may remain undetected in documents because of their incompatibility with some analysis options. This occurs, for example, when an item is found both in the categorization dictionary and the exclusion list, or when this item includes non-alphabetic characters that have not been specified as valid. The following table displays the various types of problems that may be identified by WordStat:

<b>TYPE</b>	<b>DESCRIPTION</b>
Item includes invalid characters	WordStat identifies individual words using alphabetic characters and other special characters specified by the user in the Valid Characters option. So, to make sure any item containing non-alphabetic characters is properly recognized, this special character must be added to the list of valid characters.
Item includes numeric character	An item in the categorization dictionary or the exclusion list that includes numeric characters cannot be recognized since the Accept Numeric Characters option is currently disabled.
Item also in the exclusion list	An item found in a categorization dictionary cannot be recognized if it matches an item found in the exclusion list.
Phrase starts with an excluded word	In order to be recognized, a phrase cannot start with a word found in the exclusion list. Therefore, this excluded word should preferably be removed from the exclusion list in order for the phrase to be recognized.

Enabling the **Show Warnings** option instructs WordStat to identify potential compatibility problems affecting items in a dictionary, and it displays a list of those problems in a special dialog box. This dialog is displayed prior to the application of dictionaries for a content analysis.

## To add new words to an existing dictionary:

If you choose to add a word to the exclusion list, the word will automatically be stored in this file without any dialog box. If the Inclusion dictionary is selected, the program will display a dialog box similar to the following:



- Type the words or phrases you would like to add in the edit box, one item per line. Spaces are automatically replaced by an underscore character.
- Select the proper category from the Location box.
- Click on the Add button.


Wildcards such as \* and ? are supported.

## To search for an entry in a dictionary:

- Right-click anywhere in the categorization dictionary.
- Select the **FIND** command from the pop-up menu. A search dialog box will appear.
- Type the desired search string in the **Find What** edit box. To restrict the search to items starting with the search string, enable the **Match Beginning of Item** option and to restrict it to whole words matching the search string, enable the **Match Whole Word Only**.
- Click on the **Find** button to search the first item matching the entry. Clicking on this button again finds the next occurrence of the search string, starting at the currently selected item.

# Frequencies Page

## To export the frequency table to disk:

- Click on the  button. A Save File dialog box will appear.
- In the **Save as Type** list box select the file format in which to save the table. The following formats are supported: ASCII file (\*.TXT), Tab delimited file (\*.TAB), Comma delimited file (\*.CSV), HTML file (\*.HTM;\*.HTML), and Excel spreadsheet file (\*.XLS).
- Type a valid file name with the proper file extension.
- Click on the **Save** button.

## To copy the entire table to the clipboard

- Right-click anywhere in the frequency table.
- Select the COPY TO CLIPBOARD | TABLE command from the pop-up menu.


## To copy selected rows to the clipboard


- Select the rows you would like to copy (multiple but separate rows can be selected by clicking while holding down the CTRL key).
- Right-click anywhere in the frequency table.
- Select the COPY TO CLIPBOARD | SELECTED ROWS command from the popup menu.

## To search for a specific item

- Right-click anywhere in the frequency table.
- Select the FIND command from the popup menu. A search dialog box will appear.
- Type the desired search string in the **Find What** edit box. To restrict the search to items starting with the search string, enable the **Match Beginning of Item** option. To restrict the search to whole words matching the search string, enable the **Match Whole Word Only** option.
- Click on the **Find** button to search the first item matching the typed string. Clicking on this button again finds the next occurrence of the search string, starting at the currently selected item.

## Creating barcharts or line charts

The  button allows one to produce barcharts or pie charts to visually display the distribution of specific keywords or categories. To produce such charts:

- Set the **Sort By** option to the order in which you wish the values be shown graphically.
- Select the rows you would like to plot (multiple but separate rows can be selected by clicking while holding down the CTRL key)
- Click on the  button.

## Exportation of Data files

DATA TO SAVE - This option allows one to specify the data to be saved. Four different kinds of data may be saved:


- Keyword frequency
- Case Occurrence (i.e. a dummy variable with 0 when absent, 1 when present)
- Percentage of words (i.e the frequency of the keyword divided by the total number of words in the case)
- TF\*IDF (i.e. the keyword frequency weighted by inverse document frequency).

VARIABLE NAMES - This option sets what method should be used by WordStat to create new variable names. When set to KEYWORD, the program will attempt to use each keyword as the name of a new variable. Illegal characters are automatically removed and long names are truncated to the first 10 characters. Duplicated variable names are distinguished by the substitution of numerical digits at the end of the name. When this option is set to PREFIX, variable names are created by adding successive numeric values to a user-defined prefix. For example, if the edit box at the right of the prefix option is set to "WORD\_", the variable names will be WORD\_1, WORD\_2, WORD\_3, etc.. The order of creation of the variables correspond to the sort order used in the FREQUENCY page.

SAVE TOTAL NUMBER OF WORDS - This option appends a numeric variable named TOTWORDS that contains the total number of words processed in each case.

# Barchart and Pie Chart

WordStat allows one to produce barcharts or pie charts to display visually the distribution of specific keywords or categories. To produce such charts:

- Move to the Frequencies page.
- Set the **Sort By** option to the desired graphic order of the values.
- Select the rows you would like to plot (multiple but separate rows can be selected by clicking while holding down the CTRL key)
- Click on the  button.

Three types of charts may be used to depict the distribution of keywords or content categories:



The vertical bar chart is the default chart used to display absolute or relative frequencies of keywords or content categories.



The horizontal bar chart displays the same information as the vertical one but is especially useful when the number of keywords is high and their labels cannot be displayed entirely on the bottom axis.



The pie chart is useful to display the relative frequency of each keyword and compare individual values to other values and to the whole. Numerical values displayed in pie charts are always expressed in percentages of either the total frequency or case occurrences.

The **Plot** option allows one to select the values that will be used as the scale for the length of bars in barcharts or as the percentage base for pie charts. For barchart the options are:

FREQUENCY	Number of occurrences of the keyword
% SHOWN	Percentage based on the total number of keywords displayed in the table
% PROCESSED	Percentage based on the total number of words encountered during the analysis
% TOTAL	Percent based on the total number of words that have not been excluded
NB OF CASES	Number of cases where this keyword appears
% CASES	Percentage of cases where this keyword appears


For pie charts, two options are available to specify how percentages will be computed:

FREQUENCY	Percentage based on the total frequency the keywords
NB OF CASES	Percentage based on the total number of case occurrences

The **View Others** option displays an additional bar or slice representing all items in the frequency table that have not been selected.

# Crosstab Page

## To export the frequency table to disk:

- Click on the  button. A **Save File** dialog box will appear.
- In the **Save as Type** list box select the file format in which to save the table. The following formats are supported: ASCII file (\*.TXT), Tab delimited file (\*.TAB), Comma delimited file (\*.CSV), HTML file (\*.HTM;\*.HTML), and Excel spreadsheet file (\*.XLS).
- Type a valid file name with the proper file extension.
- Click on the **Save** button.

## To copy the table to the clipboard

- Right-click anywhere in the frequency table.
- Select the COPY TO CLIPBOARD | TABLE command from the pop-up menu.

## To copy selected rows to the clipboard

- Select the rows you would like to copy (multiple but separate rows can be selected by clicking while holding down the CTRL key).
- Right-click anywhere in the table.
- Select the COPY TO CLIPBOARD | SELECTED ROWS command from the pop-up menu.

## To search for a specific item

- Right-click anywhere in the frequency table.
- Select the FIND command from the popup menu. A search dialog box will appear.
- Type the search string in the **Find What** edit box. To restrict the search to items starting with the search string, enable the **Match Beginning of Item** option. To restrict the search to whole words matching the search string, enable the **Match Whole Word Only** option.
- Click the **Find** button to search the first item matching the typed string. Clicking on this button again finds the next occurrence of the search string, starting at the currently selected item.

# Keyword-In-Context Page

When displaying rules, only the keywords or key phrases associated with the first item of those rules are displayed. For example, in a rule like:

#SATISFACTION before #TEACHER and not near #NEGATION

the KWIC list will contain only items in the SATISFACTION category meeting the conditions specified by this rule.

Once an inconsistency has been detected, it becomes possible to reduce it by making changes to the textual data or to the dictionaries. For example, the researcher may change all occurrences of the word KILL in the original text for either KILL1 or KILL2 in order to differentiate the different meanings and then add only one of these modified words (say KILL1) to the categorization dictionary. The word KILLY may also be added to the dictionary of excluded words. The categorisation of phrases may also be used to distinguish various meanings of a word. For example, the use of KIND to refer to the adjective ("considerate and helpful nature") may be reliably differentiated from the use of KIND as a noun ("category of things") or as an adverb by categorizing the phrase "KIND OF" as instances of this word used as a noun or as an adverb and by categorizing the remaining instances of KIND as the adjective. Disambiguation may also be performed by identifying words in close proximity that are associated with specific meanings and by creating categorization rules.

## Cluster Analysis

### 2D and 3D Map controls

**Nb Clusters** - This option allows the setting of the number of clusters that the clustering solution should have. Different colors are used both in the dendrogram and in the 2D and 3D maps to indicate membership of specific items to different clusters. The slide ruler located on the top toolbar of the dialog box may also be used to quickly change the number of clusters. Please note that when the **Remove single word clusters** option is enabled, changing the number of clusters in either way often causes the program to recompute MDS maps to take into account the different number of items displayed.

## Correspondence Analysis




Items distributed in a very similar way among subgroups of the categorical variables may be plotted on top of each other, making them hard to differentiate. Clicking down this button adds some random noises to the location of individual words or keywords, allowing one to clearly identify those that overlap. To remove the random noises, click on the button again.

# Feature Extraction - Phrases Finder

The **Min. Frequency** or **Min. Cases** options allow one to eliminate from the list phrases that appear only a few times by setting a minimum frequency criteria. When set to **Min. Frequency**, the criteria specifies the minimum number of times a phrase must appear regardless of whether it comes from a single document or from multiple documents. Setting it to **Min. Cases** allows one to require those occurrences to appear in a minimum specified number of cases.

## Finding overlaps

While WordStat tries to reduce redundancy in the list of phrases by automatically removing short phrases that are part of larger ones, the resulting list may still contain items that are not independent of each other such as phrases that sometimes overlap. In order to allow users to take into account potential overlaps when selecting phrases, WordStat provides a display option that allows one to see when a selected phrase includes a shorter one, is part of a larger one, or sometimes overlaps other phrases.

To enable the display of information regarding overlaps, simply click on the  button. A window appears on the right of the table. Selecting a phrase in the table automatically shows all other items that overlap this selected item. Each phrase is accompanied by a ratio indicating the total number of times this other phrase occurs and how many times it overlaps with the selected item. For example, if one selects the phrase I'M LOOKING FOR in the table showing it occurs 26 times in a document collection, one may notice that it overlaps with another phrase, LOOKING FOR SOMEONE, with a ratio of 11 out of 12. This suggests that LOOKING FOR SOMEONE occurs 12 times, but on 11 occasions, both phrases overlap (I'M LOOKING FOR SOMEONE). This ratio also indicates that on one other occasion, this second phrase occurs without overlapping the first one. It is also useful to compare the total number of overlaps with the total frequency of the target phrase. In the above example, we can conclude that the phrase I'M LOOKING FOR - occurring 26 times - is followed by SOMEONE on 11 occasions. Thus, on 15 other occasions, it is followed by something else.

To hide information about overlaps, click on the button again.

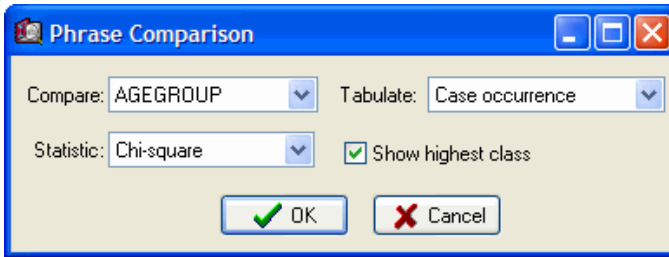
## Comparing frequencies or case occurrences of phrases

The distribution of a phrase among classes of a categorical variable may be quite useful when choosing whether or not to include it in a categorization dictionary. For example, one may want to identify phrases that are typical of some topics in order to better describe them or to differentiate them from other topics. While the Crosstab page in WordStat is normally used for such a purpose, it can only be used for items already included in a categorization dictionary or selected by the content analysis process. In other words, one way to compare the frequency of phrases identified by the phrase finder among classes of a categorical variable is to move all those phrases to a categorization dictionary, and then use this dictionary to obtain the cross frequency of those phrases. However, the phrase finder page offers a convenient way to obtain such information without the need to move those phrases to the categorization dictionary.



## To compare frequencies or case occurrences of phrases:

- Once phrases have been extracted, click on the  button. A dialog box similar to the following one will appear:




- The **Compare** list box shows all categorical variables that are available for comparison. Select the variable on which the comparison will be performed.
- Use the **Tabulate** list box to specify whether the comparison will be based on the frequency or case occurrence of those phrases and to specify whether data will be presented using absolute or relative frequencies. Four options are currently available:

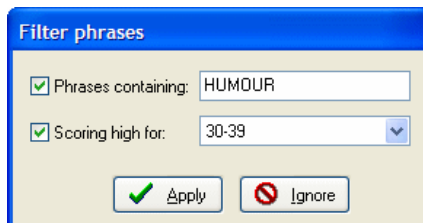
FREQUENCY	Total number of times this phrase occurs
CASE OCCURRENCE	Total number of cases in which this phrase occurs
FREQUENCY PER 100 cases	Number of times this phrases occurs per 100 cases
% OF CASES	Percentage of cases in which this phrase occurs

- Choose the **Statistic** that should be used to assess the relationship between the frequency or case occurrence of the phrases and classes of the categorical variable. The **Chi-square** is the overall chi-square value computed on all classes of the categorical variable, while the **Max Chi<sup>2</sup>** option is the chi-square value computed on the class with the highest case occurrence or frequency against all the other classes. Select **None** if you don't want to display any comparison statistic.
- Check the **Show highest class** option to display a column indicating the label of the class with the highest relative frequency or case occurrence. In the event that two or more classes obtain the same high percentage, the cell will list all the labels associated with each of those classes.
- Click on the **OK** button to perform the computation.

Once the computation is completed, several additional columns are added to the right side of the table. To sort rows based on values in any of the newly created columns, click on the appropriate column heading. Clicking several times on the same column heading toggles between ascending and descending order.

## Filtering the table:

Extracting phrases from a large collection of documents can result in a very large table containing thousands of phrases. Clicking on the  button brings a dialogbox offering filtering options that allow one to view only phrases containing either a key word or phrases that are characteristic of a specific class. Filtering conditions are specified in a dialog box similar to this one:



Enabling the **Phrase containing** option and entering a string in the edit box allows one to display only phrases containing the specified string. If a comparison has been performed between classes of a categorical variable, one may also view phrases that are characteristic of a class by enabling the **Scoring high for option** and selecting the value associated with this class. In the above example, both filtering options were used, restricting the phrases displayed in the table to those containing the string HUMOUR and found to be characteristic of the 30-39 age group.

To apply the filtering condition, click on the **APPLY** button. To remove those filtering conditions and display all extracted phrases, click on the **IGNORE** button.

## To copy the entire table to the clipboard:

- Right-click anywhere in the frequency table.
- Select the COPY TO CLIPBOARD | TABLE command from the pop-up menu.

## To copy selected rows to the clipboard:

- Select the rows you would like to copy (multiple but separate rows can be selected by clicking while holding down the CTRL key).
- Right-click anywhere in the frequency table.
- Select the COPY TO CLIPBOARD | SELECTED ROWS command from the pop-up menu.

## To search for a specific item:

- Right-click anywhere in the table.
- Select the **FIND** command from the pop-up menu. A search dialog box will appear.
- Type the desired search string in the **Find What** edit box. To restrict the search to items starting with the search string, enable the **Match Beginning of Item** option and to restrict it to whole words matching the search string, enable the **Match Whole Word Only** option.
- Click on the **Find** button to search the first item matching the typed string. Clicking on this button again finds the next occurrence of the search string, starting at the currently selected item.


# Automated Text Classification - Select Features Page

The strength of the relationship between an item and the classes of the categorical variable can be computed either on the occurrence (present or absent), on the frequency of items in each class, or on the percentage of words. To change the base statistic used for assessing differences among classes, set the Compute statistics on list to the proper option.

The discriminative strength of each item is assessed using three statistics and is presented in a table containing the following information:

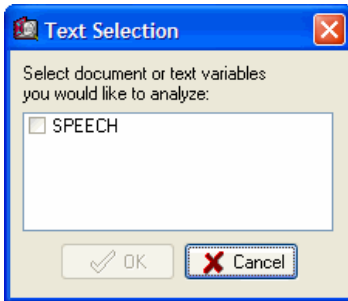
<b>Name</b>	The word, keyword or content category.
<b>Global Chi<sup>2</sup></b>	The overall chi-square value computed on all classes of the categorical variable.
<b>P</b>	The probability of the above chi-square value.
<b>Max Chi<sup>2</sup></b>	The chi-square value computed on the class with the highest case occurrence or frequency against all the other classes.
<b>P</b>	The probability of the Max Chi <sup>2</sup> value.
<b>Biserial</b>	The biserial correlation computed between the class of the categorical variables with the highest case occurrences and the remaining classes. This coefficient assumes that the presence or absence of a class is determined by a trait normally distributed. Contrary to the standard correlation coefficient, this measure of association may yield a value lower than -1.0 or higher than +1.0.
<b>Predict</b>	Indicates the class in which the item most frequently occurs. When the highest case occurrence <b>or frequency</b> appears for more than one class, the column includes the labels of all those classes.


By default, bars in the chart are displayed in descending order of frequency. Moving from one feature to another will cause the order of the values to be rearranged according to the observed frequencies. To display bars

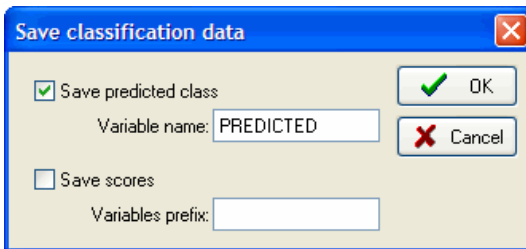
associated with specific values at a fixed location, click on the  button.

## To classify documents in another data file:

- Set the **Text to Classify** list box to **External Data File**.
- Click on the **Open File** button to locate the Simstat/QDA Miner data file containing the documents to be classified. A dialog box similar to the following one will appear.



- Select one or several text or document variables that will be used for classification purpose and click OK. The content of the data file is displayed in a table while the text to be classified is displayed on its right. You can resize this text window by dragging its left border.
- Click on the **Classify** button to apply the current classifier to all documents contained in the selected text variables.
- To store the predicted class or the computed score obtained for every class, click on the  button. A dialog box similar to the following one will appear:



- To save the predicted class, put a check mark beside **Save predicted class** and enter a variable name.
- To save the scores associated with each class and upon which the classification has been made, put a check mark beside **Save scores** and enter a variable prefix (up to 7 characters). Variable names are created by adding successive numeric values to this prefix. For example, if the edit box at the right of the Variable Prefix option is set to "CLASS", the variable names will be CLASS1, CLASS2, CLASS3, etc.
- If any one of the specified variables does not exist, WordStat will create new ones and store the numerical values associated with either the predicted class or the class scores. A confirmation dialog box will ask to confirm the creation of those new variables as well as the overwriting of any existing ones.

# WordStat Document Classifier

The WordStat Document Classifier utility program is a stand-alone application that may be used to perform content analysis and automatic text classification on a text pasted from the clipboard or stored in a file. It may also be used to analyze a collection of documents stored in a Simstat or QDA Miner data file.

Performing a content analysis or a text classification on an existing document, or collection of documents is quite simple and involves three easy steps: 1) loading the document or data file into the main editing window, 2) opening the classification or categorization model previously saved on disk and 3) applying the model. Detailed results of the content analysis or classification are displayed in tables at the bottom of the dialog box.


## Analyzing a single document

The document classifier supports several document file formats such as ASCII text files, HTML, Rich Text, MS Word, WordPerfect and Acrobat PDF files.

## Analyzing a collection of documents


The Document Classifier can perform content analysis and automatic text classification on a collection of documents stored in a Simstat or QDA Miner data file. Categorization and classification results as well as scores per class may then be stored back into the data file or exported to another file.

### Step #1 - Opening the collection of documents

- Select the OPEN DATA FILE command from the DOCUMENT menu or on click the  button. An Open File dialog box will be displayed. Locate the data file containing the documents to analyze, select it and click the on **Open** button.

The content of the data file is displayed in a table while the text to be categorized or classified is displayed on its right.

### Step #2 - Opening the model

- To open the categorization or the text classification model, select the OPEN command from the MODEL menu or click the  button. Content analysis models are stored in files with a .wcat file extension, while document classification models are stored in files with a .wclas file extension.
- Select the model you would like to use and click the **Open** button.


### Step #3 - Applying the model

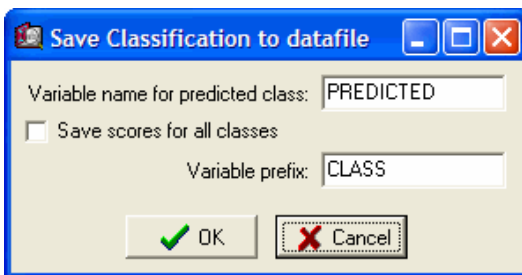
- Select the APPLY command from the MODEL menu or click on the  button.

When a categorization dictionary is applied, a single frequency table is displayed at the bottom of the screen with the number of occurrences of each keyword included in the model as well as the total number of words.

When a classifier is used, a second table is shown, allowing one to examine the classification decision made by the classifier as well as the computed values associated with each class of the categorical variable. This table is synchronized with the database shown at the top of the screen so that moving from one row to another either in the database or this classification table moves to the corresponding row in the other table.

If the k-Nearest Neighbors algorithm is used for classification and the database containing the training set can be located, a third table is shown, displaying the "k" most similar documents, their ranking and their similarity scores.

- To store in the opened data file the predicted class or the computed score obtained for every class, click on the  button. The following dialog box will appear:



- Enter the variable name that will contain the predicted class.
- To save the scores associated with each class and upon which the classification has been made, put a check mark beside **Save Scores for all classes** and enter a variable prefix (up to 7 characters). Variable names are created by adding successive numeric values to this prefix. For example, if the edit box at the right of the Variable Prefix option is set to "CLASS", the variable names will be CLASS1, CLASS2, CLASS3, etc.

If any one of the specified variables does not exist, WordStat will create new ones and store the numerical values associated with either the predicted class or the class scores. A confirmation dialog box will ask to confirm the creation of those new variables, as well as to overwrite any existing variables.