WordStat
Content Analysis Module for SIMSTAT

User’s Guide

Designed and written by Normand Péladeau
Provalis Research

Copyright © 1998-1999 Provalis Research. All right reserved.
DISCLAIMER

This software and the disk on which it is contained are licensed to you, for your own use. This is copyrighted software owned by Provalis Research. By purchasing this software, you are not obtaining title to the software or any copyright rights. You may not sublicense, rent, lease, convey, modify, translate, convert to another programming language, decompile, or disassemble the software for any purpose. You may make as many copies of this software as you need for backup purposes. You may use this software on more than one computer, provided there is no chance it will be used simultaneously on more than one computer. If you need to use the software on more than one computer simultaneously, please contact us for information about site licenses.

WARRANTY

The WORDSTAT product is licensed "as is" without any warranty of merchantability or fitness for a particular purpose, performance, or otherwise. All warranties are expressly disclaimed. By using the WORDSTAT product, you agree that neither Provalis Research nor anyone else who has been involved in the creation, production, or delivery of this software shall be liable to you or any third party for any use of (or inability to use) or performance of this product or for any indirect, consequential, or incidental damages whatsoever, whether based on contract, tort, or otherwise even if we are notified of such possibility in advance. (Some states do not allow the exclusion or limitation of incidental or consequential damages, so the foregoing limitation may not apply to you). In no event shall Provalis Research's liability for any damages ever exceed the price paid for the license to use the software, regardless of the form of claim. This agreement shall be governed by the laws of the province of Quebec (Canada) and shall inure to the benefit of Provalis Research and any successors, administrators, heirs, and assigns. Any action or proceeding brought by either party against the other arising out of or related to this agreement shall be brought only in a PROVINCIAL or FEDERAL COURT of competent jurisdiction located in Montréal, Québec. The parties hereby consent to in personam jurisdiction of said courts.

COPYRIGHT

Copyright © 1998-1999 Provalis Research. All rights reserved. No part of this publication may be reproduced or distributed without the prior written permission of Provalis Research, 2414 Bennett Street, Montreal, QC, CANADA, H1V 3E4.

 TRADEMARK

Microsoft Windows is a registered trademark of Microsoft Corporation.
Excel, MS Access and FoxPro are products of Microsoft Corporation
SPSS/PC+ and SPSS for Windows are a registered trademark of SPSS Inc.
Other product names mentioned in this manual may be trademarks or registered trademarks of their respective companies and are hereby acknowledged.
# TABLE OF CONTENT

<table>
<thead>
<tr>
<th>Section</th>
<th>Page</th>
</tr>
</thead>
<tbody>
<tr>
<td>Introduction to WordStat</td>
<td>4</td>
</tr>
<tr>
<td>Program's capabilities</td>
<td>5</td>
</tr>
<tr>
<td>The content analysis &amp; categorization process</td>
<td>7</td>
</tr>
<tr>
<td>A quick tour: Performing your first content analysis</td>
<td>9</td>
</tr>
<tr>
<td>Preparing and importing data</td>
<td>13</td>
</tr>
<tr>
<td>Preliminary text preparation</td>
<td>13</td>
</tr>
<tr>
<td>Importing spreadsheet files</td>
<td>14</td>
</tr>
<tr>
<td>Importing database files</td>
<td>15</td>
</tr>
<tr>
<td>Importing text files</td>
<td>16</td>
</tr>
<tr>
<td>The working environment</td>
<td></td>
</tr>
<tr>
<td>First screen - Dictionaries</td>
<td>18</td>
</tr>
<tr>
<td>Second screen - Options</td>
<td>23</td>
</tr>
<tr>
<td>Third screen - Word count</td>
<td>27</td>
</tr>
<tr>
<td>Third screen - Crosstab</td>
<td>29</td>
</tr>
<tr>
<td>Fourth screen - Key-Word-In-Context</td>
<td>34</td>
</tr>
<tr>
<td>Common tasks</td>
<td></td>
</tr>
<tr>
<td>Creating, developing, and maintaining dictionaries</td>
<td>38</td>
</tr>
<tr>
<td>Using lexical tools for dictionary-building</td>
<td>42</td>
</tr>
<tr>
<td>Viewing and editing text</td>
<td>50</td>
</tr>
<tr>
<td>Hierarchical clustering and multidimensional scaling</td>
<td>53</td>
</tr>
<tr>
<td>Correspondence analysis</td>
<td>58</td>
</tr>
<tr>
<td>Filtering records</td>
<td>62</td>
</tr>
<tr>
<td>Expression operators and rules</td>
<td>63</td>
</tr>
<tr>
<td>Supported xBase functions</td>
<td>65</td>
</tr>
<tr>
<td>Performing analysis on manually entered codes</td>
<td>71</td>
</tr>
<tr>
<td>Computing inter-raters agreement statistics</td>
<td>73</td>
</tr>
<tr>
<td>Saving numeric and text results into a data file</td>
<td>76</td>
</tr>
<tr>
<td>Performing multivariate analysis on words or categories</td>
<td>79</td>
</tr>
<tr>
<td>References</td>
<td>81</td>
</tr>
<tr>
<td>Technical support</td>
<td>82</td>
</tr>
</tbody>
</table>
Introduction to WordStat

WordStat is a content analysis module specifically designed to process textual information such as responses to open-ended questions, interviews, titles, journal articles, public speeches, electronic communications, etc. WordStat may be used both for automatic categorization of text using a dictionary approach and for manual coding.

The program can apply existing dictionaries or categorization schema to a new text corpus. It also facilitates the development and validation of new dictionaries.

When used in conjunction with manual coding, the program can provide assistance for a more systematic application of coding rules, help uncover differences in word usage between subgroups of individuals, assist in the revision of existing coding using KWIC (Key-Word-In-Context) tables, and assess the reliability of coding by the computation of inter-raters agreement statistics.

The numerical results of a categorization or word count may be combined with information stored in other numeric fields, allowing the user to compare how different subgroups differs in vocabulary, topic, or category. They also may be used to uncover existing relationships between specific categories or words and numeric variables such as age of the respondent, year of publication, etc. In addition, the program allows comparison of word, keyword or category occurrences between different text fields and the assessment of word co-occurrence by the display of word by word frequency matrix. Relationships among words or categories may also be examined using hierarchical clustering and multidimensional scaling analysis.

Its close integration with SIMSTAT, a comprehensive statistical program, facilitates further quantitative analysis on numerical results obtained from the content analysis (ex.: factor analysis or correspondence analysis on word frequencies, multiple regression, etc.).
Program's Capabilities

TEXT PROCESSING CAPABILITIES

- Analyses text stored in several records (or cases) of a data file.
- Performs analyses on alphanumeric fields containing short textual information (up to 255 characters per field) such as responses to open ended questions, titles, descriptions, etc. as well as on longer texts stored in memo fields (a single memo field can contain up to 64K of text).
- Optional exclusion of pronouns, conjunctions, expressions, etc. by the use of user-defined exclusion lists.
- Categorization of words or expression using existing or user-defined dictionaries.
- Word or expression substitution and scoring using wildcards and integer weighting.
- Frequency analysis of words, derived categories or concepts, or user-defined codes entered manually within a text.
- Interactive development and validation of multi-level dictionaries or categorization schema.
- Ability to restrict an analysis to specific portions of a text or to exclude comments and annotations.
- Ability to perform a content analysis on a random sample of records.
- Integrated spell-checking with support for different languages such as English, French, Spanish, etc.
- Integrated thesaurus (English only) to assist the creation of categorization schema.
- Powerful record filtering on any numeric or alphanumeric field and on code occurrence (with AND, OR, and NOT boolean operators)
- Ability to define a memo field linked to individual records to either write annotations, side notes, raters comments, or codes.
- Print presentation quality tables (word count, crosstab or KWIC lists)
- Save any table to HTML, ASCII, Tab separated or comma separated value files.

UNIVARIATE WORD FREQUENCY ANALYSIS

- Univariate word frequency analysis (word count and record occurrence).
- Word co-occurrence matrix (within a record)
- Integrated clustering and dendrogram display of word similarities
- 2D and 3D multidimensional scaling on either joint frequency or co-occurrence of words or categories.

MULTIPLE RESPONSES AND COMPARISONS

- Can perform a single frequency analysis on information stored in several alphanumeric fields (memo or string variables).
- Comparison of word or keyword occurrence between different fields.
- Compute inter-raters agreement measures on codes manually entered in different fields (pct. of agreement, Cohen's Kappa, Scott's Pi, Krippendorf's R and r-bar, free marginal, and intraclass correlation).
BIVARIATE COMPARISONS BETWEEN SUBGROUPS

- Bivariate comparison between any textual field and any nominal or ordinal variable (such as sex of the respondent, specific subgroups, years of publication, etc.).
- Choice between 12 different association measures to assess the relationship between word occurrence and nominal or ordinal variables (Chi-square, Likelihood ratio, Student’s F, Tau-a, Tau-b, Tau-c, symmetric Somers’ D, asymmetric Somers’ Dxy and Dyx, Gamma, Pearson’s R, and Spearman’s Rho)
- Correspondence analysis allows examination of relationships between words or categories and other nominal or ordinal variables.
- Ability to sort word matrix in alphabetic order, by word frequency or word occurrence, on the obtained statistics or on its probability.

KEY-WORD-IN-CONTEXT

- Ability to display a Key-Word-In-Context (KWIC) table of any included, not included or user defined word or word pattern.
- KWIC tables may be sorted in ascending order of record number, words with context, or on values of independent variables.
- Ability to jump from a specific occurrence in the KWIC table to the original text field in order to view or edit the selected word.
- KWIC tables may be saved in data files for further processing.
- Customizable KWIC and report function to display all hits as lists of paragraphs, sentences or user defined segments.

FULL INTEGRATION WITH A STATISTICAL PROGRAM

- Alphanumeric variables are stored in the same file as all other numeric variables.
- Variable selection and analysis are performed within the main statistical program using a simple 3-step operation:
  - Open the existing data file.
  - Select one or several alphanumeric fields as dependent variables and, optionally, other nominal or ordinal variables to be treated as independent.
  - Execute the CONTENT ANALYSIS command from the STATISTICS drop-down menu.
- Final numerical results and KWIC tables are automatically added to existing statistical outputs.
- New variables representing frequency or occurrence of words, keywords or concepts can be added to the existing data file or exported to a new data file in order to be submitted to more advanced analysis (such as cluster analysis, correspondence analysis, multiple regression, etc.).
- Data can be imported from and exported to different file format including dBase, Paradox, Excel, Quattro Pro, Lotus 1-2-3, SPSS for DOS, SPSS for Windows, comma or tab separated text files, etc.
- Ability to perform numeric and alphanumeric transformation or to apply filters on records of the data file to restrict the analysis to specific subgroups.
The Content Analysis & Categorization Process

The most basic form of content analysis that WordStat can perform is a simple frequency analysis of all words contained in one or several text fields of a data file. However, WordStat offers several features that permit the user to accomplish more advanced forms of content analysis that may involve automatic categorization, different weighting of words, inclusion or exclusion of words based on frequency criteria, etc. To fully understand the possibilities offered by the program, one first needs to understand the various underlying processes involved in a typical WordStat frequency analysis and how these processes may be combined to achieve various kinds of content analysis tasks.

WordStat's categorization involves up to four successive processes:

1- EXCLUSION PROCESS

An exclusion process may be applied to remove words that you do not want to be included in the content analysis. This process requires the specification of an exclusion list. Such a process is used mainly to remove words with little semantic value such as pronouns, conjunctions, etc., but may also be used to remove some words used too frequently or with little discriminative value.

2- INCLUSION / CATEGORIZATION PROCESS

The inclusion / categorization process allows one to change specific words or word patterns to other words, keywords or categories and/or to extract a list or specific words or codes. This process requires the specification of an inclusion dictionary. This dictionary may be used to remove variant forms of a word in order to treat all of them as a single word. It may also be used as a thesaurus to perform automatic coding of words into categories or concepts. For example, words such as "good", "excellent" or "satisfied" may all be coded as instances of a single category named "positive evaluation", while words like "bad", "unsatisfied" or expressions like "not satisfied" may be categorized as "negative evaluation".

3- ADDITION OF FREQUENT WORDS

The third process is the application of a frequency criterion that is used to add to the included words or categories words that are used more than a specific number time or that are found in more than a specific number of records. When an inclusion dictionary is used, this option will append to this list of included words or categories, all words that meet the minimum frequency criterion. If no inclusion or categorization dictionary is used, all words that meet this minimum requirement and that have not been excluded (see process #1) will be added to the final word/category list. Note that this process can only be used to add new words to the actual list of words and categories found in this inclusion dictionary. It cannot be used to remove any of those items (see process #4).

4- REMOVAL OF WORDS OR CATEGORIES

When this process is active, all words or categories that do not meet a minimum frequency or record occurrence criterion will be removed from the final word/category list. This process may be combined with the inclusion/categorization process to remove infrequent categories. It may also be used in conjunction with the addition criterion (see process #3) to provide a composite criterion of inclusion that involves both a minimum word frequency and a minimum record occurrence.
Since the application of each process is optional, numerous combinations are possible, each combination allowing the researcher to perform different types of content analysis. For example, here are the minimal requirements for different forms of content analysis:

<table>
<thead>
<tr>
<th>TYPE OF ANALYSIS</th>
<th>EXCLUSION DICTIONARY</th>
<th>INCLUSION DICTIONARY</th>
<th>ADD WORDS</th>
<th>REMOVE WORDS</th>
<th>COMMENT</th>
</tr>
</thead>
<tbody>
<tr>
<td>Simple word frequency analysis (most frequent words)</td>
<td></td>
<td></td>
<td>√</td>
<td></td>
<td>√ *</td>
</tr>
<tr>
<td>Simple frequency analysis of semantically significant words</td>
<td>√</td>
<td></td>
<td></td>
<td>√ *</td>
<td></td>
</tr>
<tr>
<td>Word count withlemmatization</td>
<td>√</td>
<td>√</td>
<td>√ *</td>
<td>√</td>
<td></td>
</tr>
<tr>
<td>Word count of specific words</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Automatic categorization of texts</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Frequency analysis on the most frequent categories.</td>
<td>√</td>
<td></td>
<td></td>
<td>√</td>
<td></td>
</tr>
<tr>
<td>Frequency analysis of manually entered codes or keywords</td>
<td>√</td>
<td></td>
<td></td>
<td></td>
<td>Codes may optionally be inserted between brackets.</td>
</tr>
<tr>
<td>Rating of texts on specific attributes</td>
<td>√</td>
<td></td>
<td></td>
<td></td>
<td>Different weights may be assigned to different words</td>
</tr>
</tbody>
</table>

*recommended to restrict the analysis to the most frequent words or categories.
A Quick Tour

A CONTENT ANALYSIS ON PERSONAL ADS

For this example, we will produce a content analysis on personal ads published in a Montreal cultural newspaper on January 22 and January 29, 1998, and we will examine whether there is a relationship between words used and the gender and age of the person who wrote the ad. The required data has been stored in a file named SEEKING.DBF.

Step #1 - Open the data file

- From within SIMSTAT, select the FILE | DATA | OPEN command sequence and select the SEEKING.DBF file

Step #2 - Select the variables

- Execute the STATISTICS | CHOOSE X-Y command
- Set the Variable List box to ALL to view all field types.
- Move the GENDER and AGEGROUP variables to the INDEPENDENT list box
- Move the AD_TEXT variable to the DEPENDENT list box
- Press the OK button

Step #3 - Run the content analysis module

- Execute the STATISTICS | CONTENT ANALYSIS command.

Step #4 - Choose the proper dictionaries

WordStat consists of a single dialog with five pages. The first page allows one to select, view, and edit the dictionaries used in this specific content analysis. Set the dictionaries to the following values:

   Exclusion: DICT\DEFAULT.EXC
   Inclusion: DICT\SEEKING.CAT

and make sure all are active. (see check boxes on the left side of the dictionaries edit boxes)

Step #5 - Setting the proper options

The second page allows you to specify various options such as whether numeric values should be included, whether frequent words should be added, etc.. Disable all options by removing any check mark in the various check boxes.
Step #6 - Perform an univariate frequency analysis on categories

- Click on the third tab (Word count). The program will perform a categorization of words found in the ads and compute a frequency analysis on those categories.

- To sort the word count matrix in alphabetical order, set the SORT BY option to Words. You can also display those words in descending order of Word Frequency or Record Occurrence.

- By default, the words displayed in the matrix are those specified in the Inclusion list. To display words that have been left out, set the WORDS TO DISPLAY option to Not Included.

- To move a word to the inclusion or the exclusion list, or to remove a word from the inclusion list, click on the "+- List" button or press on the right button of the mouse.

Step #7 - Examining the relationship between included categories and the gender of the author.

- Press on the fourth tab (Crosstab).

- Click on the WITH drop-down list box and select GENDER to display a contingency table of categories frequency by gender.

The TABULATE option allows one to choose whether the table should be based on the total frequency of included words or on the total number of records containing those words.

The SORT BY option allows one to sort the table on the word or category name (alphabetic order) or by descending order of word frequency. You may also click on any column header to sort the grid in ascending or descending order of the values found in this column.

The DISPLAY option allows one to specify the information displayed:

- Count
- Row percent
- Column percent
- Total percent

Step #8 - Estimating the strength of the relationship

- Use the STATISTIC drop-down list box to select an association measure, such as a Chi-square or a Pearson’s R statistic.

To sort the table on the chosen statistic or on its probability, use the SORT BY drop-down list box.

Step #9 - Visualizing the relationship between categories and the age of the author.

- Use the mouse to highlight cells of the categories you would like to compare.

- Click on the button or press on the right button of the mouse and select the Chart Selected Rows menu item. Click on Barchart to view compare the groups using bar chart.
Step #10 - Performing correspondence analysis on age groups

- Click on the WITH drop-down list box and select AGEGROUP to display word counts by age group.
- Click on the button to access the correspondence analysis dialog.
- Press on the 2D Map or 3D Map tabs to examine a 2 axis or a 3 axis solution, or on the STATISTICS tab to browse through the correspondence analysis statistics.
- Press on the button to close this dialog and return to WordStat main dialog.

Step #11 - Displaying a word by word matrix or a word by record matrix

- Click on the WITH drop-down list box and select <other words> to display a word by word matrix or on <record no> to view a word by record matrix.

Step #12 - Viewing a Key-Word-In-Context (KWIC) list of specific words or categories

- Press on the Key-Word-In-Context tab to access the KWIC table.
- Set the LIST option to Included and select the word or category for which you would like to obtain a KWIC table.
- Click on the GO button to display the KWIC table for this word or category.

To sort the table on the record number, on the keyword along with the prior or subsequent words, or on the sex of the respondent, use the SORT BY drop-down list box.

- To display KWIC tables of any user-defined word or word pattern, set the LIST option to "User-defined", enter your word pattern (with or without wildcards) in the WORD edit box and click the GO button.

Step #13 - Editing a text from the KWIC list

- To modify the word or keyword or the text surrounding it, select it from the KWIC list and click the EDIT button. (You may also double click on the specific line you wish to edit).
- To save the modified text, click the OK button. Clicking the CANCEL button restores the original text.

Step #14 - Creating a concordance report

- Make sure the Key-Word-In-Context page is active and that the KWIC table displays the proper information.
- Set the amount of context that should be displayed around each word by setting the CONTEXT DELIMITER option.
• Press on the REPORT button. Note: If this button is inactive, click the GO button to refresh the content of the KWIC table and then click on the REPORT button.

Step #15 - Examining relationship between words or categories

• Click on the third tab to activate the Word Count page.

• Press on the Tree/Map button to display the Dendrogram & Concept Maps dialog.

• Press on the Dendrogram tab to perform a hierarchical cluster analysis on included categories. You may change the number of partitions displayed using the Nb Clusters option.

• Press on the 2D Map or 3D Map tabs to perform a multidimensional scaling and display a plot in 2 or 3 dimensions.

• Press on the button to close this dialog and return to WordStat main dialog.

For more information see Hierarchical Clustering and Multidimensional Scaling (page 53).

Step #15 - Saving the word frequencies on disk

• Press on the SAVE button and choose the data from the Save Information box to export to the existing data file or a new one.

Step #16 - Quitting the module and returning to SIMSTAT

• Press on the OK button.

• Select the various outputs you would like to append to the statistical program notebook and click the OK button.
Preparing and Importing Data

SIMSTAT stores its numeric and alphanumeric data into a DBF file format. Memo fields (up to 64k) are stored in a separate file with a FPT extension. The file format used by SIMSTAT is partly compatible with FoxPro 2.x applications. This section provides general information on how to prepare textual data or specific instruction on how to import data into SIMSTAT.

Preliminary Text Preparation

In order to perform content analysis, WordStat required that all information be transformed into raw text (or plain ASCII text). While interview transcripts, responses to open-ended questions, or any other kind of textual information may be typed directly within SIMSTAT, there are many situations where electronic versions already exists either in the form of text files or as data files accessible only through specific applications such as spreadsheet or database programs. All this information must be transferred into a SIMSTAT data file for further processing. However, prior to this transfer, some modification or adjustments may need to be made.

Uppercase and lowercase letters

WordStat is case-insensitive and therefore accepts files in either upper- or lower-case.

Check spelling of documents

The automatic content analysis feature of WordStat involved numerous operations of word recognition and generally requires each word to be spelled correctly. Any misspelled word may be left uncoded and leads to imprecise or invalid conclusions. For this reason, it is strongly suggested to run your text file through a spelling checker to make sure all words are spelled correctly. WordStat provides spell checking for English, French, Spanish, Dutch, German, and Italian texts.

Remove hyphenation

While WordStat can be configured to accept compound words with dashed, it cannot differentiate dashes and hyphens. As a consequence, an hyphenated word will often be treated as two separate words. It is thus recommended to revise the text to insure not hyphenation is present.

Add or remove square brackets ([ ]) and braces ({ })

Square brackets and braces have special meanings for WordStat. For example, braces are used to remove a section of the text that you don't want to process while square brackets may be used to restrict the analysis to specific portions of texts. If these symbols are used in a text for other purposes, they should be substitute by other symbols.

If there are specific parts of your text that you do not what to be processed, such as some explanation notes, interviewer questions and probes, comments, etc.), enclose them in braces (ex. {comment} ). Also, if you want to perform a content analysis on only a small portion of the entire text, such as on manually entered codes, enclose this portion of text in square brackets.
Importing Spreadsheet Files

Most spreadsheet programs allow for entry of both numeric and alphanumeric data into cells of a data grid. SIMSTAT can import spreadsheet files produced by LOTUS 1-2-3 (v1.1 to v5.0), SYMPHONY (v1.0 and v1.1), EXCEL (v2.2 to v5.0), and QUATTRO PRO (v1.0 to v6.0). To import data from any of these applications:

- Choose the DATA | IMPORT command from the FILE menu.
- Select the file format using the List File of Type drop down list.
- Select the file you want to import and click on the OK button.

The program displays a dialog box where one can specify the spreadsheet page and the range of cells where the data are located. You must specify a valid range name or provide upper left and lower right cells, separated by two periods (such as A1..H20). If you set the Range Name list box to ALL, the program attempts to read the whole page.

Formatting spreadsheet data

The selected range must be formatted such that the columns of the spreadsheet represent variables (or fields) while the rows represent cases or records. Also, the first row should preferably contain the variable names while the remaining rows hold the data, one case per row. SIMSTAT will automatically determine the most appropriate format based on the data it finds in the worksheet columns. Cells in the first row of the selected range are treated as field names. If no variable name is encountered, SIMSTAT will automatically provide one for each column in the defined range.

When reading the data for analysis, all blank cells and all cells that do not correspond to the variable type (e.g., alphanumeric entries under a numeric variable, or a numeric value under a string variable) are treated as missing values.

Limitations

If the maximum length of the text contained in any of the cells does not exceed 256 characters, you may import spreadsheet files directly from SIMSTAT. However, spreadsheet files with longer cell contents should preferably be saved as TAB delimited text files (with a .TAB file extension) and then imported from within SIMSTAT.
Importing Database Files

dBase and Paradox files
SIMSTAT can directly import dBase and Paradox data files. However, the length of alphanumeric fields should not exceed 256 characters and memo fields are not supported. If you do have such kind of data, you may use the exporting capabilities of your database program to create a data file more compatible with Simstat (such as a FoxPro 2.x data file or a tab delimited text file).

To import data from one of these applications:

- Choose the DATA | IMPORT command from the FILE menu.
- Select the file format using the List File of Type drop down list.
- Select the file you want to import and click on the OK button.

MS Access data files
The data file format used by SIMSTAT is compatible with FoxPro 2.x data files. While SIMSTAT cannot read or import MS Access data files, this program offers the capability to export its data file to a FoxPro 2.x file format. For instruction on how to export MS Access files to FoxPro 2.x files, please refer to the MS ACCESS manual or help file.

Other database files
Most database applications provide exporting capabilities that allow the user to save a copy of the data file in several file formats. The recommended file formats are, in descending order of preference, FoxPro 2.x data file and Tab separated text files. However, if your data file contains no memo fields and no alphanumeric fields larger than 256 character you may also export your file to a dBase, a Paradox, or any supported spreadsheet data file.

Importing memo fields
Memo fields that have not been successfully imported, may be transferred to the data file either by using cut and paste operations or by retrieving text files from disk. For more information on this topic, see Importing Plain Text and Word Processor Files (page 16)
Importing Plain Text or Word Processor Files

Probably one of the easier ways to transfer data from a word processor document into SIMSTAT is to open simultaneously both applications and use cut and paste operations to transfer data through the clipboard. However, this may not be the most efficient way, especially when one needs to import a large amount of information. The following section presents four additional methods to transfer text information into memo fields:

- Using the Document Conversion Wizard program
- Retrieving a text file into a memo field
- Importing comma or tab delimited text file
- Importing page delimited memo files

While the first method can read textual data stored in word processor documents, the last three methods require the data to be stored on disk in plain ASCII files without any formatting or typesetting code. Most word processors offer an option to save a document as a plain text file. If you don't know how to create such a text file, please refer to your word processor manual.

Using the Document Conversion Wizard program

WordStat includes a conversion utility program that can assist you in the importation of text files stored in either word processor documents such as MS Word, MS Write, WordPerfect or RTF files, but also of text stored in ASCII (plain text), HTML or even Excel spreadsheet files. To run this program:

- Point to the Programs folder in the Windows' Start menu, then select Provalis Research and then click on Document Conversion Wizard.

This utility program will guide you through the process of importing one or numerous text files.

Retrieving a text file into a memo field

This method should be used to retrieve a single unit of text into a memo field for a specific record. If textual data for several records need to be retrieved, they should be stored in different text files. To retrieve the text file from SIMSTAT:

- Open the data file where the information should be stored.
- Position the cursor on cell in which you would like to store the text. A memo editor should appear at the bottom of the data sheet.
- Click inside the memo editor or press F2.

- Click on the Import Text Into Memo button , select the text file you wish to retrieved and click OK.
Importing comma or tab delimited text files

If you wish to retrieve a text file containing several numeric and alphanumeric variables, you may have to transform this file into a comma or tab delimited text file. There are, however, several limitations to this transfer method. If commas are used as delimiters, then all existing commas within text fields should ideally be removed. If a tab delimited format is chosen, all tab characters already present in a text field should be removed. Another important limitation is that all the information of a single record must be stored in a single line. For this reason, hard returns in long texts should be removed so that the entire text is stored on a single line. (There is no limitation on the total number of columns per line, so it is possible to store very long texts on a single line).

SIMSTAT can read up to 500 numeric and alphanumeric variables from a plain ASCII file (text file). The file must have the following format:

- Every line must end with a carriage-return.
- The first line must include the variable names, separated by tabs or commas.
- Variable names may have a length of not more than 10 characters. Longer strings are truncated to 10 characters.
- The remaining lines must include the numeric or alphanumeric values, separated by tabs or commas.
- Each line must contain data for one case and variables must be in the same order for all cases. All invalid data and all blanks encountered between commas or tabs are treated as missing values. A single dot can also be used to represent a missing numeric value.
- Comments can be inserted anywhere in the file by putting a * at the beginning of the line.
- Blank lines can also be inserted anywhere in the file.
- Comma delimited text files require a .CSV extension while tab delimited files require a .TAB extension.

Importing page delimited memo files

SIMSTAT provides a simple method to import numerous records of texts by the use of page delimited memo files. This file format consists of a plain text file which contain the textual data of numerous individuals for a single memo field. The text for each record must be separated by page break characters (ASCII 12). The file name extension of this text file should be .MMO. To import such a file:

- Choose the DATA | IMPORT command from the FILE menu.
- Set the file format to Page Delimited Memo using the List File of Type drop down list.
- Select the file you want to import and click on the OK button.

The resulting file consists of a SIMSTAT data file with two fields: RECNO, a numeric field containing a sequential number going from 1 up to the total number of records encountered in the input file, and TEXT, a memo field containing the textual data for this record.

Note: Importation of numerous text fields may be achieved by performing successive importations of page delimited memo files and then using the APPEND VARIABLES command to merge the resulting files into a single one. In order to achieve this, great care should be taken to give unique names to the various TEXT fields and to assure that the record sequence of the various text files is identical.
Without further information, WordStat can perform a word count analysis on each of the words encountered in the chosen alphanumeric fields. However, it is also possible to apply various transformations on the words before performing the frequency analysis. The first two pages of the main dialog box (see below) allow one to specify how the textual information should be processed. For example, you can tell the program to preprocess words using two dictionaries: an exclusion and an inclusion/categorization dictionary. The Dictionaries page also allows you to use or create those dictionaries, add, remove or edit existing entries in those dictionaries.

(For more information of other analysis options available see Options Page)

The following section provides a description of those two dictionaries. Additional information about dictionary creation and maintenance can be found on page 38.

**EXCLUSION DICTIONARY**

The exclusion dictionary (also known as a stop list) is used to remove all words that are not to be included in the word count operation. It is used mainly to remove words with little semantic value such as pronouns, conjunctions, etc. Wildcards such as * and ? are supported.
For example, the following expression:

```
REPORT*
```

will exclude all words beginning with REPORT (such as, REPORT, REPORTS, REPORTER).

The next example:

```
EXP?RT
```

will remove both EXPORT and EXPERT.

An expression that includes several words may also be excluded by joining the various words with underline characters. For example:

```
NOT_*
```

Will exclude all words preceded by the word "not".

The currently opened exclusion dictionary may be deactivated by removing the check mark in the check box at the left of the exclusion dictionary name.

**INCLUSION DICTIONARY**

The inclusion dictionary allows one to change specific words, word patterns (such as REPORT* or EXP?RT), or expressions, to another word, category or concept. This feature may be used to remove variant forms of a word in order to treat them as a single instance or to group related words under meaningful categories. Inclusion dictionaries may also be used to perform a frequency analysis on manually entered codes. By manually entering specific keywords (such as "EVAL_POS", or "EVAL_NEG") in a text field and by entering those same keywords in the inclusion list, it becomes possible to extract those codes and perform frequency and contingency analysis on them.

The inclusion dictionary is structured as a hierarchical tree where words, word patterns, and expressions are grouped in a folder that represents a category name. Categories and individual words may also be included in a higher order category, allowing one to create multi-level dictionaries like the following one:

```
COUNTRY
    NORTH-AMERICA
        • CANADA (1)
        • UNITED-STATES (1)
        • USA (1)
        • MEXICO (1)
    SOUTH-AMERICA
        • BRAZIL (1)
        • CHILI (1)
```

In the above example, words like CANADA, USA, or MEXICO may be coded as either NORTH-AMERICA or COUNTRY, depending on whether the categorization is performed up to the first or second level of the dictionary (see Level of Analysis, page 21).
Wildcards such as * and ? are supported. For example, the following item under the support category:

```
 superhero
   * SUPERHERO*
```

will change SUPPORT, SUPPORTS, SUPPORTING, SUPPORTIVE, SUPPORTER, etc. into a single word SUPPORT, while the following word pattern:

```
 superhero
   *SUPERHERO*
```

will also substitute all words with the substring "SUPERHERO" in it, such as UNSUPERHORABLY, UNSUPERHORED, etc.

An expression that includes several words may also be substituted by joining the various words with underline characters. For example, you may change the expression "going out" with the category "NIGHTLIFE" by specifying the following item:

```
 nightlife
   * GO_OUT
```

You may also use wildcard in expressions such as:

```
 nightlife
   * GO_ OUT
```

to substitute several forms of an expression at once.

Integer weights can also be assigned to specific items so that a specific word or word pattern may count for more than one instance of the category. For example, in order to compute an aggressiveness score on specific texts, you may choose to assign a weight of 5 points to word patterns such as KILL* or MURDER* but only a single point to word patterns like INSULT*.
INCLUSION SETTINGS

LEVEL OF ANALYSIS - This option allows one to specify up to which level the coding should be performed. For example, in the following dictionary:

- COUNTRY
  - NORTH-AMERICA
    • CANADA (1)
    • UNITED STATES (1)
    • USA (1)
    • MEXICO (1)
  - SOUTH-AMERICA
    • BRAZIL (1)
    • CHILI (1)

if a level of 1 is specified, all words that are stored at a higher level that the root level will be coded as the parent category at this first level. For example, words like CANADA and MEXICO will be coded as COUNTRY along with other country names like BRAZIL. Setting the level of analysis to a numeric value of 2 will results in the coding of those two words as NORTH-AMERICA, while BRAZIL will be coded as SOUTH-AMERICA. Items stored at the same or at a lower level than this option will remain unchanged.

CATEGORIES ONLY - When the LEVEL OF ANALYSIS option is set to a value higher than one, this option instructs WordStat to limit the level increase to the coding of the last category at or below the specified level. This option is especially useful when working with unbalanced hierarchical categorization systems where individual words are stored at different levels. For example, in the following dictionary:

- SENSATION
  - ODOR
    • AROMA (1)
    • BREATH (1)
    • FRAGRANCE (1)
    • NOSE (1)
  - ANXIETY
    • AFRAID (1)
    • TREMOR (1)

setting the level of analysis to 2 without enabling this option would code words like AROMA or BREATH as ODOR, but would include in the final results individual words like TREMOR or
AFRAID. Enabling the CATEGORIES ONLY option insures that individual words won't be included but will be coded as their parent category.

**USE FULL PATH AS CATEGORY NAME** - When the LEVEL OF ANALYSIS option is set to a value higher than one, this option instructs WordStat to substitute the full path of an item as the category name. The slash (/) characters is used to separate the various levels. For example, in the above example, setting this option to true and the level analysis to 2 will code the word AROMA as SENSATION/ODOR. Increasing the level of analysis up to 3 will return SENSATION/ODOR/AROMA.

**ALLOW OVERLAP** - By default, categories are mutually exclusive such that a word can only be entered in a single category. Enabling this option allows one to create overlapping categories where words can be classified simultaneously into two or more categories. However, please take note that current multivariate techniques available in WordStat such as clustering, correspondence analysis and multidimensional scaling as well as other multivariate statistical procedures make the assumption that categories are statistically independent. Using overlapping categories creates data that clearly violate this assumption and may yield dubious results.

For more information on how to open, activate or deactivate a dictionary or how to add, edit or remove an entry in a dictionary, see Creating and maintaining dictionaries, page 38)
Options Page

This page offers different options that control how the textual information should be processed. The options are grouped under three different pages:

1) Analysis
2) Speller/Thesaurus
3) Others

ANALYSIS OPTIONS

ADD WORDS - When the inclusion dictionary is disabled, all words that are not found in the exclusion list will be included in the final word frequency analysis. This option allows one to restrict the number of words included to the most frequent ones by setting a minimum Frequency or Record Occurrence criterion for inclusion. This option may also be used while the inclusion list is active to add to this list, other words that are used at a high frequency. However, this option can only be used to add new words to the list of words and categories found in this inclusion dictionary and cannot be used to remove any of those items. To remove items in this inclusion dictionary based on a frequency or record occurrence criterion see the REMOVE WORDS option below.
REMOVE WORDS - This option allows one to restrict the number of included words or categories to the most frequent ones by setting a minimum Frequency or Record Occurrence criterion for inclusion. This criterion is applied both to items in the inclusion dictionary and words that meet the criterion specified with the ADD WORDS option.

Examples:

- If no inclusion dictionary is used and you want to include any word that appears at least 10 times, but in no less than 5 different records, you need to activate the ADD WORDS option and set its criterion to a minimum FREQUENCY of 10. You then have to set the REMOVE WORDS criterion to a minimum Record Occurrence of 5. Only words that meet both criteria will be included.

- When an inclusion list is used to lemmatize words, but you only want to obtain frequency information on those words that appear a specific number of times, you have to activate the inclusion dictionary and set the minimum frequency criterion of both the ADD WORDS and REMOVE WORDS options to the required frequency.

- When an inclusion list is used to categorize words, but you only want to analyse the most frequent categories, you have to activate the inclusion dictionary and set the REMOVE WORDS option to the required frequency. In this situation, the ADD WORDS option should be deactivated.

DON'T PROCESS TEXT WITHIN BRACES - This option can be used to instruct the program to skip all text found between braces (i.e. { and }). This option is especially useful when you want to insert comments or annotations in the text field without affecting the word count analysis. It can also be used to ignore in an interview transcript all questions, prompts, and other verbal interventions made by the interviewer.

PROCESS ONLY TEXT WITHIN BRACKETS - This option can be used to instruct the program to process only the text found between brackets (i.e. [ and ]). This option may be used to perform an analysis on keywords entered manually in the text by one or several coders.

RANDOM SAMPLE - When this option is activated, the program will randomly select a fraction of all records and performs the word counting analysis on this subsample. The proportion of records can be specified using the spin button located at the right of the checkbox. This option reduces the processing time for large files and is especially useful during the initial phase of an analysis where dictionaries are constructed and categorization schema are developed and revised. It also allows one to preview the kind of results that would be obtained on very large data files.

ACCEPT NUMERIC CHARACTERS - By default, every word consisting of numeric values or of a mix of letters and numbers is excluded from the analysis. This option can be used to include those words.
VALID CHARACTERS - This option allows one to specify which characters, beside letters of the alphabet, should be considered as an integral part of a word. For example, the word "ex-wife" can be treated as a single word or as two separate words (i.e. "ex" and "wife") if the hyphen is included in the list of valid characters.

CASE SENSITIVE - By default, WordStat internally convert all text to uppercase letters so that processing of words is cases insensitive. This may be inappropriate if one wants to identify proper nouns or analyze text written in some European languages like German where difference in letter cases may denote different meaning. Enabling this option prevent the internal conversion to uppercase letters and will treat two instances of the same word different in their case (lower or upper case) as two distinct words.

DISABLE CROSSTAB AND CLUSTER - If you are only interested in obtaining a frequency analysis of words or categories or in exploring the usage of specific words using KWIC lists, then you may use this option to disable the collection of information used for crosstabulation and clustering. When this option is set, the memory requirement for an analysis is drastically reduced and the performance is increased.

SPELLER / THESAURUS PAGE

ACTIVE DICTIONARIES - WordStat make use of language dictionaries in order to spell-check existing textual data and to suggest inflected forms of words found in the user dictionary. This group of options let you specify the dictionary to use with the current data file. If the chosen dictionary is not installed, WordStat will display a message informing you of this as well as instructions on how to obtain it.

SPELLER OPTIONS

CONFIRM ADDITION TO USER DICTIONARY - When this option is activated, you will be prompted after clicking the Add button of the spell-checking dialog to confirm word additions to the custom dictionary.

IGNORE WORDS CONTAINING NUMBERS - Enabling this option instructs WordStat to ignore any word that contains one or more numeric characters.

IGNORE WORDS IN UPPER CASES - Enabling this option instructs WordStat to ignore all words fully in uppercase.

THESAURUS

DISABLE ENGLISH THESAURUS - WordStat's Suggest feature uses an English thesaurus to suggest synonyms of existing words in the user directory. This option allows one to disable this feature. This may be especially useful when developing a dictionary in another language than English or when one only want the program to suggest inflected forms. (for more information on this feature, see Finding new words using the integrated thesaurus and dictionary)
NOTE TAKING FIELD - This option allows to select a memo or an alphanumeric field where one can type comments or codes related to the actual record. When a note taking field is selected, the text editor is split horizontally in two parts, the lower part displaying the content of this field. Any change made to the field is automatically saved with the data file. Its content may later be browsed, revised, and even submitted to a new content analysis.

WORKING SPACE - By default, WordStat uses available computer memory to stores all temporary lists and data. This option speeds up computation, but you might run out of memory on very large projects. Selecting the DISK option frees up memory and allows you to analyze larger projects, at the cost of slower computation.
The Word Count page is used to display a frequency table of words or category names. This can be used to perform an univariate frequency analysis on words or categories and also to modify any of the three dictionaries.

By default, the table shows the included words in descending order of frequency. The table includes the following statistics:

- **NB WORDS**: Number of occurrence of the word or category names.
- **% SHOWED**: Percentage based on the total number of words displayed in the table.
- **% TOTAL**: Percent based on the total number of words (except those excluded).
- **NB RECORDS**: Number of records where this word appears.
- **% RECORDS**: Percentage of records where this word appears.

**WORD TO DISPLAY** - While by default, the table shows the frequency of all included word, it allows one to also display all words not explicitly excluded but left out. This option greatly facilitates the development and maintenance of the various dictionaries. Setting the WORD TO DISPLAY option to NOT INCLUDED allows for identification of important words that should have been put in the inclusion or substituted but were ignored. This display may also be used to identify common words that might be added to the exclusion list.
SORT BY - This option allows a display of words in the frequency table in alphabetic order, on word endings, or by descending order of word frequency (NB WORDS column) or record occurrence (NB RECORDS column). Sorting the table by word endings facilitates the identification of plural form of words that should be substituted by their singular form or the substitution of verbs by their infinitive form.

+/- LIST - This button can be used to move one or several words to the exclusion or substitution list or to add or remove a word from the inclusion list. The permitted moves depend on the words currently displayed. If you want to remove a word from the inclusion list, the WORD TO DISPLAY option should be set to INCLUDED or ALL. To add a word to the inclusion list, the WORD TO DISPLAY should be set to NOT INCLUDED or ALL. This button is also used to display a Key-Word-In-Context table of the selected word.

It is also possible to quickly access the popup menu invoked by this button by pressing the right button of the mouse anywhere on the grid (see below).

TREE/MAP - This button allows one to perform cluster analysis and multidimensional scaling on all included words or categories and display a dendrogram or concept map of those items based on their proximity. For further information see Hierarchical Clustering and Multidimensional Scaling page 53).
The Crosstab page is used to display a contingency table of words or categories. This contingency table is computed only on items that have been included. If an inclusion dictionary has been specified, this grid will display only the words or keywords in this list. If no inclusion list has been specified, the grid will display all words that have not been explicitly excluded. Along with absolute and relative frequency of word occurrence or word frequency, several statistics may be displayed to assess the relationship between independent variables and word usage or to assess the reliability of coding made by several human coders or a single coder at different times.

Options

TABULATE - The TABULATE option allows choosing whether the values in the table should be based on the total frequency of keywords or the number of records containing the keywords.

WITH - The WITH drop down list allows choices on how the word count should be broken down. The following options are available:

- <other words> - display a square table showing the number of co-occurrence of words in the same record.
- <record number> - display the word occurrence or frequency for each individual records.
• **ANY INDEPENDENT VARIABLE** - If numeric independent variables were selected as independent variables, their names will appear at the bottom of this list box. Selecting any of those variable names will display a contingency table allowing for the assessment of the relationship between this independent variable and the words or categories.

• **<variables>** - When the word count is performed on several alphanumeric variables, this option will allow for comparison of the occurrence of words according to variable name.

**SORT BY** - The SORT BY option presents the opportunity to sort the table by word or category names (alphabetic order) or by descending order of frequency or record occurrence. When a statistic is displayed (see option STATISTICS), the table can also be sorted based on the value of this statistic or on its statistical probability. It is also possible to sort on the values of any specific column by clicking on this column heading. Clicking several times of the same column heading toggles between ascending and descending sort orders.

**DISPLAY** - The DISPLAY list box allows one to specify the information displayed in the table. The following options are available:

- Count
- Row percent
- Column percent
- Total percent

When the TABULATE option is set to Record Occurrence, two additional statistics are also available:

- Percent of cases (percentage of all records or individuals)
- Category percent (percentage of records or individuals in this subgroup)

**STATISTIC** - When word frequency or occurrence is broken down by an independent variable (see WITH option), a drop down list box will appear. This list box allows one to choose among 12 association measures to assess the relationship between this independent variable and the utilization of each word or category.

**Nominal level statistics**

- Chi-square
- Likelihood ratio
- Student's F

**Ordinal or internal level statistics**

- Tau-a
- Tau-b
- Tau-c
- Somers' D (symmetric)
- Somers' Dxy (asymmetric)
- Somers' Dyx (asymmetric)
- Gamma
- Spearman's Rho
- Pearson's R
PROBABILITY - The probability option allows one to select whether the probability value should be computed using a 1-tailed or 2-tailed test. Probabilities of Chi-square, Likelihood ratio, and Student’s F are always computed using a 2-tailed test.

AGREEMENT - When comparing word or category usage between different alphanumeric variables, a drop-down list box will appear. This list box allows one to choose among 8 different inter-raters agreement measures to assess the reliability of coding.

Nominal level agreement statistics
- Percentage of agreement
- Cohen's Kappa
- Scott's pi
- Free marginal (nominal)

Ordinal or internal level agreement statistics
- Krippendorf's R
- Krippendorf's r-bar
- Free marginal (nominal)
- Intraclass correlation

For more information on how to assess reliability with those statistics see Computing Inter-Raters Agreement Statistics (page 73).

Creating bar charts or line charts
The Crosstab page also allows one to produce bar charts or line charts to visually compare the distribution of specific words or categories among values of an independent variable such as subgroup of individuals (male vs female) or time periods. To produce such charts:

- Set the TABULATE and DISPLAY options so that the information you want to visualize is displayed in the table.
- Using the mouse, select the rows you would like to display. Multiple disjunct rows can be selected through clicking while holding down the CTRL key.
- Click on the button or press the right button of the mouse and select the Chart Selected Rows command.
A dialog like this one will appear:

This window allows one to graphically examine the relationship between codes and values of an independent variable. The bar chart should preferably be used to display the distribution of various categories within subgroups as defined by a nominal independent variable, while the line chart should preferably be used to examine the relationship between those categories and an ordinal or quantitative variable.

The following table provides a short description of available buttons and controls:

<table>
<thead>
<tr>
<th>Controls</th>
<th>Description</th>
</tr>
</thead>
<tbody>
<tr>
<td>![Button]</td>
<td>Press this button to retrieve a chart previously saved on disk.</td>
</tr>
<tr>
<td>![Button]</td>
<td>Press this button to save a chart on disk. Charts are saved in a proprietary format and may be edited and customized using the Chart Editor.</td>
</tr>
<tr>
<td>![Button]</td>
<td>Pressing this button allows you to print a copy of the displayed chart.</td>
</tr>
<tr>
<td>![Button]</td>
<td>This button allows you to edit various features of the chart such as the left and bottom axis, the chart and axis titles, the location of the legend, etc.</td>
</tr>
</tbody>
</table>
This button is used to create a copy of the chart to the clipboard. When this button is clicked, a popup menu appears allowing you to select whether the chart should be copied as a bitmap or as a metafile.

Pressing this button closes the chart dialog and returns to WordStat's main screen.

**Performing correspondence analysis**

Correspondence analysis is an exploratory technique that provides a graphic overview of relationships in large crosstabulation tables of frequency. To perform a correspondence analysis:

- Set the WITH option to an independent variable.
- Click on the button to access the correspondence analysis dialog.

For more information on correspondence analysis, see Correspondence Analysis (page 58).
Key-Word-In-Context Page

The Key-Word-In-Context (KWIC) technique allows one to display in a table the occurrences of either a specific word, or of all words related to a category, with the textual environment in which they occur. The text is aligned so that all keywords appear aligned in the middle of the table. This technique is useful to assess the consistency (or lack of consistency) of meanings associated with a word, word pattern or category. In the example below, we can see that the word pattern KILL*, which may have been assigned to a category like "aggressiveness", refers to words with different meanings, some of them quite distant from the concept of "aggressiveness":

I have decided to KILL a few hours before...
He said that he would KILL me if I call the police.
Too much garlic KILL the taste of the meat.
The Black Death was a disease that KILLED millions.
My shoes are KILLING me
The French skier Jean Claude KILLY won 3 gold medals.

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 substitution or inclusion dictionary. The word KILLY may also be added to the dictionary of excluded words.

The KWIC technique is also useful to highlight syntactical or semantic differences in word usage between individuals or subgroup of individuals. For example, candidates from two different political parties may use the word "rights" in their discourses at the same relative frequency, but we may find that these two groups use this word with quite different meanings. We may also find that the meaning of a word like "moral" evolves with the age of a child.

The Key-Word-In-Context page has been designed to facilitate the various tasks involved in content analysis. The page looks like this (see next page):
LIST - This option allows for specifying whether the words for display in the KWIC table either should be selected from the list of included words or from the list of all remaining words that have not been explicitly excluded. The option User Specified allows one to enter a word or word pattern at the keyboard and search for all instances of this expression.

WORD - This option allows for choosing among all words belonging to the list of Included or Not Included words (see above). When the LIST option is set to User Specified, this option become an edit box where one can type a word or word pattern. (Wildcards such as * and ? are supported).

SORT BY - This option allows for sorting the keyword-in-context table in either ascending order on any of the following options:

- Record number - The KWIC table is sorted in ascending order of record position.
- Keyword & Before - The KWIC table is sorted on the keyword as well as the words appearing immediately before it
- Keyword & After - The KWIC table is sorted on the keyword and the words appearing immediately after it.

• **Keyword & Variable** - When several text fields have been selected, the KWIC table includes a column indicating in which field (or variable) the word was found. When this option is selected, the KWIC table is sorted so that all words associated with a category or matching a word pattern are displayed in alphabetic order. Lines with identical words are sorted on the field name from which they come. This display is useful to examine whether specific words are used with the same meaning in different fields.

• **Variable & Keyword** - When several textual fields have been selected, the KWIC table includes a column indicating in which field (or variable) the keyword was found. This option displays a KWIC table where all lines are sorted on the field name from which they originate. Lines extracted from a single field are sorted by keywords. This display is useful to establish whether different fields contain different information. For a more detailed analysis of differences in usage of specific words, use the Keyword & Variable sort order.

• **Keyword & VARNAME** - This option displays a KWIC table where lines are sorted by words. Lines with identical words are sorted on the value of the selected independent variable. This display is useful to highlight differences between subgroups in the meanings associated with a specific word.

• **VARNAME & Keyword** - This option displays a KWIC table where lines are sorted by the values of the selected independent variable. Lines with identical values on this variable are sorted by keywords. This display is useful to establish whether subgroups differ on the use of words associated with a category. For a more detailed comparison of usage of specific word, use the Keyword & VARNAME sort order.

**CONTEXT DELIMITER** - This option allows to select the amount of context displayed in the KWIC table as well as in the concordance report. In the KWIC table, context strings, either before or after the keyword, are limited to 255 characters.

• None - This option instructs WordStat to display as much context as possible, up to a limit of 255 characters.

• Paragraph - When this option is selected, the program will limit the context displayed to the paragraph in which a specific keyword appears.

• Sentence - When this option is selected, the program will limit the context displayed to the sentence in which a specific keyword appears. A sentence must end with a period followed by a space or a hard return, or by an exclamation or a question mark.

• User defined - When this option is selected, the program will retrieve text found before and after the keyword until a slash character is encountered.
REPORT - Clicking this button produces a concordance report on the keywords currently displayed in the KWIC table. The sort order and context delimiter of the current KWIC table are used to determine the display order and the amount of context displayed in this concordance report. This report is displayed in a text editor dialog (see below) and may be modified, stored on disk in RTF, HTML or plain text format, printed, or cut and pasted to another application. Graphics may also be pasted anywhere in this report.

[Record #9  SEX = Women]
This is Lynn. I'm 30 yrs. old, 5'2" tall, with shoulder-length blonde hair, green eyes, & I'm considered pretty. I'm not slim, nor am I overweight. I'm financially secure, & attend night classes. I'm a simple girl, with a good personality & a good sense of humor. I enjoy movies, music, dancing, going for long walks, travel & more. I'd like to meet new people & make new friends. If you're interested, leave a message at Box 3317.

[Record #29  SEX = Men]
My name's Garth. I'm 20 yrs. old, 6' tall & well built with one-length hair. I'm good-looking of Greek background. I've just finished college & I'm heading for university. I'm looking for a good-looking girl, preferably of European background, who likes winning, dining, going to the clubs on St. Lawrence, & who likes to live the good life. Leave a message at Box 1625.

[Record #52  SEX = Women]
My name's Michelle. I'm a Montreal-born, French-speaking black woman of exotic background. My father's black, my mother's European. I consider myself to be a citizen of the world. I've traveled extensively & lived in different places, including Los Angeles, Brazil & Paris. I'm 5'7'' tall, 120 lbs., & I'm said I'm attractive, but beauty is in the eye of the beholder. I'm feminine, youthful, intelligent, cultured & alluring. I'm seeking a mature man, over 40 yrs. old, over 5'10'' tall, who has a great sense of humor, gallantry & romance. I have a preference for cultured men who are financially secure. Box 1917.

[Record #52  SEX = Women]
My name's Michelle. I'm a Montreal-born, French-speaking black woman of exotic background. My
Creating and Maintaining Dictionaries

To open an existing dictionary

- Enter the dictionary file name and location in the proper edit box, or click on the button to display a dialog box that will let you browse through directories and select a file name with the proper extension.

To create a new dictionary

- Enter the new dictionary file name and location in the proper edit box. If the file name does not exist, you will be asked to confirm the creation of a new directory file. If a dictionary file is already active, you will be asked whether existing entries should be copies to the new dictionary file. If you answer Yes, all entries in the previously opened dictionary will be retrieved and stored in the new one. Answering No will result in an empty dictionary.

To add new words to an existing dictionary

From the Dictionaries page

- In the Dictionary Viewer group box, select the dictionary to which you would like to add new words.
- Press on the ADD button.

From the Word Count page

- Select the rows containing the words you would like to add.
- Press on the +/- List button or click on the right button of the mouse.
- Select the dictionary to with you would like to add the selected words.

From the Crosstab page

- Select the rows containing the words you would like to add.
- Press on the right button of the mouse to display a popup menu.
- Select the dictionary to with you would like to add the selected words.

From the text editor

- Select the word you would like to add.
- Press on the right button of the mouse to display a popup menu.
- Select the dictionary to with you would like to add the selected words.
If you choose to add a word to the exclusion list, the word will automatically be stored in this file without any dialog. If the Inclusion dictionary is selected, the program will display the following dialog:

![Add Words / Categories Dialog](image)

To add a new word to an existing category:

- Make sure the Words radio box is active
- Select the proper category from the Location outline.
- Type the word you would like to add in the edit box and click the Add button.

Wildcards such as * and ? are supported.

You can specify several words at a time by separating each word with a space (see example above).

Integer weights can also be assigned to specific categorization, so that a specific word, word pattern, or expression may count for more than one instance of the concept. The default value for this option is 1. To use a higher value, edit the Weight option either by entering a new numeric value in the edit box or by clicking on the spin buttons to increase or decrease this value.

If you want to add a word to an unexisting category, you first need to create such a category (see below) and then follow the above steps to add the word to this new category.
To add a category to the inclusion dictionary

From the Dictionaries page

- In the Dictionary Viewer group box, select the Inclusion radio button.
- Press on the ADD button. The Add Words / Categories dialog box will appear.
- Select the Main Category or the Sub-Category radio button depending on whether you want this new category to appear at the main level or whether you want it to be created under an existing category. If you choose to create a sub-category, you then need to select from the Location outline the category under which you would like to store it.
- Type the category name you would like to add in the edit box and click the Add button.

You can specify several categories at a time by separating each category name by a space (see example above).

To remove a word or a category from a dictionary

- Select the Dictionaries page by clicking on the first tab at the top of the dialog.
- In the Dictionary Viewer group box, select the dictionary from which you would like to remove words or categories.
- Select the words or categories you would like to delete and click on the Delete button. If a non-empty category is selected, you will be asked to confirm its deletion. If you answer Yes, all words and subcategories belonging to this category will also be erased.

To edit an entry in a dictionary

- Select the Dictionaries page by clicking on the first tab at the top of the dialog.
- In the Dictionary Viewer group box, select the dictionary containing the words or categories you would like to edit.
- Select the item you want to modify and click on the EDIT button.

Moving words or categories using drag and drop

The easiest way to change the structure of an inclusion dictionary is by using drag and drop operations. Using the mouse, you can move a word to a different category, move an existing category or sub-category to another location on the main level or under an existing category. To perform such operations, you first need to enable the drag & drop editing feature:

- Select the Dictionaries page by clicking on the first tab at the top of the dialog.
- In the Dictionary Viewer group box, select the Inclusion dictionary.
- Check the Drag & Drop Editing checkbox.
Once activated, you just have to click on the item you want to move, and hold the mouse button down. Then, simply drag the item over its new location and release the mouse button. By default, the dragged item is stored under the category at the cursor position. To move a word or a category to the main level or to the same level as the category under the cursor, simply hold the ALT key while dropping the dragged item.
Using lexical tools to for dictionary-building

Creating a comprehensive categorization dictionary is quite often a difficult, time-consuming and subjective task. WordStat can assist you in finding words that may be related to existing words in your categories by the use of three lexical tools:

- A spelling dictionary is used to propose inflected forms of existing words already in your dictionary. Several dictionaries are currently available for different human languages such as English, French, Italian, Dutch, etc.

- An English thesaurus is also used to propose synonyms of words already in your dictionary. The thesaurus data file contains over 8,700 context topics (indexed words) and offers a database of over 75,000 synonyms.

- A WordNet based lexical database is used to find synonyms, antonyms as well as hypernyms, hyponyms, coordinate terms, holonyms, meronyms, etc.. This database contains over 120,000 root words (including many proper nouns) and offers over 100,000 synonym sets. The availability of word sense definitions allows for manual as well as automatic filtering of proper word senses.

These three tools are available through two different dictionary-building commands.

- The Basic command uses the selected spelling dictionaries and the basic English thesaurus to identify related synonyms and inflected forms.

- The Advanced command gives you access to a more powerful dictionary-building tool that uses a WordNet based lexical database to find related words and the selected spell-checking dictionaries to find inflected forms of those words.

You will find below a description of these two dictionary-building tools:

Basic Dictionary-building Tools

To access the basic dictionary-building tool:

- Select the Dictionaries page by clicking on the first tab at the top of the main WordStat screen.

- Press on the Suggest button.

- Select the Basic command.

WordStat will immediately start looking for synonyms and inflected forms of all words in your inclusion dictionary and will report them in a dialog like this one:
This dialog displays on the first page a list of synonyms that were found to be related to existing words in the various categories. Synonyms for a specific category are sorted so that those that were related to several existing words in this category are located at the top of the list while synonyms related to only a single word are located at the bottom. The numeric value in parenthesis indicates the number of existing words to which it was related.

The second page lists all words whose spelling begins with the same letters as existing words and that were not already included in the actual dictionary. For example, if the word "understanding" is found in the dictionary, the program will suggest words like "understandings", "understandingly", "understands", "understanded", "understandable", and "understandably".

To add suggested words to the dictionary, place a check mark beside the words you would like to add and click on the Add button.

Click on the button to return to WordStat.

**Advanced Dictionary-Building Tools**

The advanced dictionary-building tool can be accessed either as a stand-alone application of from within WordStat. To run the stand-alone version:
Point to the Programs folder in the Windows' Start menu, then select Provalis Research and then click on Dictionary Builder.

To access the advanced dictionary-building tool from within WordStat:

- Select the **Dictionaries** page by clicking on the first tab at the top of the main WordStat screen.
- Press on the **Suggest** button.
- Select the **Advanced** command.

A dialog like this one will appear:

![WordStat Dictionary Builder](image)

The first page is used to set various dictionary and search options. The second and third pages are used to find words and idioms semantically related to existing entries in the dictionary, while the last page is used to find derived form of those entries.

**DICTIONARY PAGE**

The first page of the dictionary builder program allows you to select or change the WordStat dictionary, specify the words and categories you want to work with, along with the type of relationship to look for. It also allows you to specify how the program will search for inflected forms of existing words in your dictionary.
To select a dictionary

- Click on the button. A standard Open dialog will appear.
- Select the WordStat dictionary file you want to work with.

Selecting words and/or categories

By default, the dictionary-building program will search for related words and idioms for all existing words and categories in your WordStat dictionary. To restrict the search to specific categories or words within a category, simply deselect the words and categories you want to exclude by removing the check marks beside them. Clicking on a category check box to change its state also changes the check box state of all words and subcategories within this category.

Specifying the type of relationship to look for

The Search for group box allows you to specify what type of relationship the program will look for. For example, you may choose to search only for synonyms and similar terms or decide to also search for hypernyms, hyponyms, coordinate terms, etc.

Setting how Inflected forms will be retrieved

The Match Partial Word option affects how inflected forms are found. When this option is deactivated, the program only retrieves words that start with the whole word. For example, if the dictionary includes the word INTELLIGENT, the program will suggest words like INTELLIGENTLY and INTELLIGENTSIA. If the Match Partial Word option is activated, the program will also suggest words like INTELLIGENCE, INTELLIGENCES, INTELLIGIBLE, and INTELLIGIBLY.

DEFINITIONS PAGE

Using a comprehensive lexical database such as WordNet to find related words and phrases has one major drawback. Searching for numerous types of relationship for even small WordStat dictionaries can yield a huge number of suggested words. For example, when searching for suggested words for a dictionary containing 129 words grouped under 13 categories, more than 12,000 new words and phrases were obtained, many of them unrelated to the existing categories. Browsing through such a huge number of suggestions to find the most relevant ones can be an overwhelming task. The Definitions page was created to somewhat reduce this burden by providing an intermediary step where the user can select, for each of the words, the word senses that are the most relevant to the containing category. The program offers both manual and automatic selections of word senses and also allows one to combine both methods.
Automatic selection of word senses

WordStat dictionary builder uses a basic disambiguation algorithm to try to identify, among all word senses, those that are the most likely to be related to the containing category. This algorithm involves the computation for each word sense of a relevance score. The higher is this score, the more likely the word sense will be related to the category, while a score equal to zero suggests that this word sense is unrelated to the category. Once those relevance scores have been computed, the program can use one of three different rules to select proper word senses.

- **Best** - This rule instructs the program to select for each word, the sense that has obtained the highest relevance score. When selecting the highest score, a 20% tolerance is used so that, in some occasion, more than one word sense will be selected. This selection rule is the most conservative one and insures that relevant word senses are the most likely to be selected. However, we have also found that this selection method may lack some sensitivity and may fail to select other relevant word senses (false negatives).

- **Relevance > 0** - This rule instructs the program to select all word senses that have been found to be related, even slightly, to the category. This selection rule is very liberal in that it is the most likely to select most relevant word senses at the cost of a lack of specificity (too much false positives).

- **Relevance > 0.1** - This rule is slightly more conservative that the previous one, in that it also rejects all word senses that have obtained a score of 0.1. Beside a score of zero, 0.1 is the lowest score that may be obtained. Experiences have shown that, very often, word senses with such a low score are unrelated to the category. Removing those word senses thus results in an increase in specificity along with only a marginal decrease in sensitivity.

The application of any of these three rules is performed by selecting the proper rule from the Select drop down list. This list box may also be used to select or unselect all definitions.
Manual selection of word senses

Manual selection of word senses can be carried out either alone or after an automatic selection has been made by the program. Manual selection is performed simply by browsing through the list of all definitions and selecting those that are related to the current category while making sure unrelated definitions are unselected. The decision to include or exclude a specific word sense may rely on the displayed definition, on the relevance score, and also on the examination of all words that have been found to be related to this specific word sense. Those suggested words are automatically displayed in the right panel of the Definition page when the word definition is highlighted.

Selected word senses may be saved on disk by clicking on the \[\text{button}\] and later retrieve by clicking on the \[\text{button}\].

Once the word senses have been chosen, activating the Words page will start the search, extract all words and phrases related to the selected word senses, and will display them by categories and by the type of relationship (synonyms, antonyms, etc.)

**WORDS PAGE**

The Words page displays a list of suggested words and idioms that were found to be related to existing words in the various categories and allows you to select suggestions and add them to the existing dictionary. The "All words" page includes a list of all words and idioms that were suggested, irrespective of their relationship with the existing entries. The remaining pages allow one to examine those same words by the nature of their relationship with existing entries.
Relevance ranking and sorting

For each suggestion, a Word relevance score is computed that takes into account the number of times an item has been suggested as well as the relevance score obtained by the word senses from which it was derived. Those suggestions are presented in descending order of relevance so that the suggestions that are the most likely related to the containing category are located at the top of the list while suggestions that are less likely to be relevant are found at the bottom of this list.

Specificity index

Very often, a word is suggested in more than one category. This is especially true when the dictionary includes categories that are semantically close each other. One good example of such a categorization system is the Lasswell dictionary that tries to differentiate ten different forms of power relations (power gain, power loss, cooperation, authoritative, conflict, doctrine, etc.). When making a decision on whether a word should be added to a given category, it is important to consider whether this word is specific to this category or whether it has also been suggested in other categories. The Compute Specificity button allows one to obtain a specificity index as well as a list of all the other categories in which this item also appears. This specificity index is computed by making the sum of all relevance scores obtained by this word in the various categories and computing the proportion of this total score that is related to the current category. A specificity of 1.0 indicates that this item has only been suggested for this category. When the item has been found to be related to more than one category, a list of all other categories in which it also appears is displayed in the Other Categories column along with the relevance score obtained in each of those categories. You can use this information to decide to which category this word should be added.

To add words or idioms to categories

- Place check marks beside the item you would like to add.
- Click on the Add button.

INFLECTED FORMS PAGE

The Inflected Form page lists all words whose spelling begins with the same letters as existing words and that were not already included in the actual dictionary. For example, if the word "understanding" is found in the dictionary, the program will suggest words like "understandings", "understandingly". If the Math Partial Word option is enabled (see Dictionary page), this same word will also yield words like "understands", "understandable", and "understandably". The From column displays the original word from which the inflected form has been derived.
To add suggested words to the dictionary, place a check mark beside the words you would like to add and click on the Add button.

Click on the button to return to quit the dictionary builder program and return to WordStat.
Viewer and editing text

WordStat's text editor allows browsing and editing text fields submitted to content analysis, as well as spell checking of text found in a specific record or in the entire data file. The keyword highlighting feature also allows to identify all words that have been coded as well as those belonging to specific coding categories. The text editor can also be used for dictionary maintenance tasks by allowing the addition of words or expressions to active dictionaries as well as easy display of KWIC tables for selected words.

The following table provides a short description of available buttons and controls:

<table>
<thead>
<tr>
<th>CONTROL</th>
<th>DESCRIPTION</th>
</tr>
</thead>
<tbody>
<tr>
<td>![Control]</td>
<td>Cut the selected text to the clipboard.</td>
</tr>
<tr>
<td>![Control]</td>
<td>Copy the selected text to the clipboard.</td>
</tr>
<tr>
<td>![Control]</td>
<td>Paste text from the clipboard at the current cursor position.</td>
</tr>
<tr>
<td>![Control]</td>
<td>Search for and replace a specified word or phrase.</td>
</tr>
<tr>
<td>![Control]</td>
<td>Spell-check all records.</td>
</tr>
<tr>
<td>![Control]</td>
<td>Spell check only the current text.</td>
</tr>
<tr>
<td>![Control]</td>
<td>Pressing this button allows you to add the selected word to an active dictionary. It may also be used to produce a KWIC table of the currently selected word or expression.</td>
</tr>
</tbody>
</table>
Field: This drop-down list box allows to select the alphanumeric or memo field displayed in the edit box.

Highlight: WordStat’s text editor displays all words that have been coded using bold characters while words belonging to the active category are showed in blue. To change the active category and highlight all words that belong to a selected category, simply choose the proper category from this list box. To highlight all categories using different colors, set this option to <all categories>.

This button allows you to access the color coding dialog that lets you to assign to each specific category in the dictionary specific font and background colors (see below).

Move to the first record of the data file.

Move to the last record of the data file.

Move to the previous record.

Move to the next record.

Assigning color codes to categories

The color code dialog allows you to assign specific font and background colors to each category of the current inclusion dictionary.

![Color codes dialog](image)
To access the color code dialog

- Set the Highlight drop down list to <all categories>.
- Click on the button.

To change the colors of a category

- Select the category in the categories list box
- Use the Font and Background color selectors to choose from a list of predefined colors or click on the Other button to define a custom color.

Information on color codes are automatically saved with the project options
Hierarchical Clustering and Multidimensional Scaling

WordStat allows one to further develop categorization by providing various graphic tools to assist the identification of related words or categories. Those tools are obtained by the application of hierarchical cluster analysis and multidimensional scaling on all included words or categories and are displayed in the form of dendrograms and concept maps. The first page of the dialog is used to set various analysis and display options for both types of analysis.

**CO-OCCURRENCE** - This option allows you to specify how a co-occurrence will be defined. By default, a co-occurrence is said to happen every time two words or two categories appear in the same record (by record option). You may also restrict the definition of co-occurrence to entries that appear in the same paragraph or the same sentence, or to words or categories that are located in the same user defined section (delimited by a / character). Finally, you may restrict even further the definition of co-occurrences by limiting the co-occurrence to a small window of words of specified length. Such a small window is especially useful when doing an analysis directly on words (rather than categories) since it allows to identify idioms or phrases that may need to be added to the categorization dictionary. Co-occurrence on larger text segments such as records or paragraphs may be more appropriate to identify the co-occurrence of themes in individual subjects.

**INDEX** - The Index option lets you select the similarity measure used in clustering and in multidimensional scaling. Four measures are available. The first three measures are based on the mere occurrences of specific words or categories in a record and do not take into account their frequency. In all those indexes, joint absences are excluded from consideration.

- **Jaccard's coefficient** - This coefficient is computed from a fourfold table as \( \frac{a}{(a+b+c)} \) where \( a \) represents cases where both items occur, and \( b \) and \( c \) represent cases where one item is found but not the other. In this coefficient equal weight is given to matches and non matches.

- **Sorensen's coefficient** - This coefficient is similar to Jaccard's but matches are weighted double. Its computing formula is \( \frac{2a}{(2a+b+c)} \) where \( a \) represents cases where both items occur, and \( b \) and \( c \) represent cases where one item is present but the other one is absent.

- **Ochiai's coefficient** - This index is the binary form of the cosine measure. Its computing formula is \( \sqrt{a^2/(a+b)(a+c))} \) where \( a \) represents cases where both items occur, and \( b \) and \( c \) represent cases where one item is present but not the other one.

The last coefficient takes into account not only the presence of a word or category in a record, but also how often it appears in this record.

- **Cosine theta** - This coefficient measures the cosine of the angle between two vectors of values. It ranges from -1 to +1.

**REAL TIME ANIMATION** - When this option is enabled, the multidimensional plots are updated after every iteration allowing the user to monitor the progress made during the analysis at the cost of higher computing time.

**TOLERANCE** - This option specifies the tolerance factor that is used to determine when the algorithm has converged to a solution. Reducing the tolerance value may produce a slightly more accurate result but will increase the number of iterations and the running time.
**MAXIMUM ITERATIONS** - This option allows one to specify the maximum number of iterations that are to be performed during the fitting procedure. If the solution does not converge to the limit specified by the TOLERANCE option before the maximum number of iterations is reached, the process is stopped and the results are displayed.

**INITIAL CONFIGURATION** - This option allows one to specify whether the multidimensional scaling will be applied on a random configuration of points or on the result of a classical scaling.

Selecting the **Classical Scaling** option instructs WordStat to perform a classical scaling first on the similarity matrix, and then use the derived configuration as initial values for the ordinal multidimensional scaling analysis.

Selecting the **Randomized Location** option instructs WordStat to perform the multidimensional scaling analysis on a random configuration of points. By default, WordStat initializes the random routine before each analysis with a new random value. The seed value used for the creation of this initial configuration is stored along with the final stress value in the history list box, located at the bottom of the dialog. The **Seed** option may be used to specify a starting number that will be used to initialize the randomization process and produce a fixed random sequence. To recall a specific seed value used previously, double-click on the proper line in the history list box.

**Dendrogram**

WordStat use an average-linkage hierarchical clustering method to create clusters from a similarity matrix. The result is presented in the form of a dendrogram (see below), also know as a tree graph. In such a graph, the vertical axis is made up of the items and the horizontal axis represents the clusters formed at each step of the clustering procedure. Words or categories that tend to appear together are combined at an early stage while those that are independent from one another or those that don't appear together tend to be combined at the end of the agglomeration process.
NB CLUSTERS - This option allows you to set how many clusters 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.

VERTICAL LINES - This option lets one choose whether the vertical lines of the dendrogram represents the agglomeration schedule or the similarity indices.

FONT SIZE - This option allows one to alter the font size used to display the dendrogram. Reducing the font size allows to visualize a larger portion of the tree.

SAVE - This button allows you to store the cluster solution currently displayed into a new inclusion dictionary where folders at the first level correspond to different clusters, and where each of those folder contains the associated words or expressions.

Note: Clustering using other similarity or distance measures or agglomeration methods may be achieved using the MVSP cluster analysis procedure (see Performing Multivariate Analysis, page 79).

2D and 3D concept maps

The concept maps are graphic representations of the proximity values computed on all included words using multidimensional scaling (MDS). In those maps, a point represents an item (word or category) and the distances between pairs of items indicate how likely those items tend to appear together. In other words, items that appear close together on the plot usually tend to occur together, while words or categories that are independent from one other or that don't appear together are located on the chart far from each other. Colors are used to represent membership of specific items to different partitions created.
using hierarchical clustering. The resulting maps are useful to detect meaningful underlying dimensions that may explain observed similarities between items.

Please note that since multidimensional scaling attempts to represent the various points into a two- or three-dimensional space, some distortion may result, especially when this analysis is performed on a large number of items. As a consequence, some items that tend to appear together or that are parts of the same cluster may still be plotted far from each other. Also, performing a multidimensional scaling on a large number of items usually produces a cluttered map that is hard to interpret. For these reasons, interpretation of concept maps may be feasible only when applied to a relatively limited number of items.

2D and 3D Map buttons and controls

**NB CLUSTERS** - This option allows you to set how many clusters 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 actual orientation of axes in the final solution is arbitrary. The map may be rotated in any way we want, the distances between items remain the same. The rotating knob can be used to adjust the final orientation of axes in the plane or space in order to obtain an orientation than can be most easily interpreted.

Clicking this button enables to zoom in a plot. To zoom a area of the plot, hold the left mouse button and drag the mouse down/right. You’ll see a rectangle around the selected area. Release the left mouse button to zoom.

Clicking this button restores the original viewing area of the plot.

Clicking this button performs another multidimensional scaling on a new random configuration of points. This bottom is visible only when the initial configuration is set to Random Location.

This button is used to create a copy of the chart to the clipboard. When this button is clicked, a popup menu appears, allowing you to select whether the chart should be copied as a bitmap or as a metafile.

Clicking this button allows you to print a copy of the displayed chart.

Clicking this button closes the Dendrogram & Concept Map dialog and returns to WordStat's main dialog.
3D Map buttons

This button can be used to show or hide left, bottom and back walls.

Clicking this button exchanges the data of the X, Y and Z axis.

Clicking this button allows you to change the viewing angle of the chart. To rotate the chart, make sure this button is selected, click on any area of the chart, hold the mouse button and drag the mouse to apply the desired rotation.
Correspondence analysis

Correspondence analysis is a descriptive and exploratory technique designed to analyze relationships among entries in large frequency crosstabulation tables. Its objective is to represent the relationship among all entries in the table using a low-dimensional Euclidean space such that the locations of the row and column points are consistent with their associations in the table. The correspondence analysis procedure implemented in WordStat allows one to graphically examine the relationship between words or content categories and subgroups of an independent variable. The results are presented using a 2 or 3-dimensional map. Correspondence analysis statistics are also provided to assess the quality of the solution. WordStat currently restricts the extraction to the first three axes. This insures that the results remain easily interpretable. This limitation may however be removed in future versions upon users request.

The first two pages of the dialog, provides graphical displays of correspondence maps. When the number of words or categories or the number of subgroups is less than four, only the 2D Map page can be accessed. When the comparison is restricted to two groups of individuals, only one axis can be extracted. In this situation, the axis is plotted diagonally in the two-dimensional space. We have done this mainly for readability reason: plotting all the words on a single horizontal axis would have produced a cluttered list of words that would have made the graph useless.

You will find immediately below a description of the various controls in the dialog, followed by some guidelines for interpreting correspondence plots.

2D Map control

**PLOT** When more than two axes have been extracted, this control allows you to select all the possible axes combinations that can be graphed on the two axes of the plot.

2D and 3D Map Controls

**Words** This checkbox allows you to display or hide the row points (i.e., words or category names)

**Groups** This checkbox allows you to display or hide the column points (i.e., subgroup labels).

Clicking this button enables to zoom in a plot.. To zoom a area of the plot, hold the left mouse button and drag the mouse down/right. You'll see a rectangle around the selected area. Release the left mouse button to zoom

Clicking this button restores the original viewing area of the plot.

This button is used to create a copy of the chart to the clipboard. When this button is clicked, a popup menu appears, allowing you to select whether the chart should be copied as a bitmap or as a metafile.

Clicking this button allows you to print a copy of the displayed chart.
Clicking this button closes the Dendrogram & Concept Map dialog and returns to WordStat's main dialog.

**3D Plot controls**

This button can be used to show or hide left, bottom and back walls.

Clicking this button allows you to change the viewing angle of the chart. To rotate the chart, make sure this button is selected, click on any area of the chart, hold the mouse button and drag the mouse to apply the desired rotation.

**Interpreting correspondence analysis results**

Interpretation of correspondence analysis maps can be somewhat tricky and should be made with great care, especially when examining the relationship between row points and column points. Here are some basic rules that should help you interpret such maps:

**Relationship among words or categories (row points):**

- The more similar is the distribution of a word or a content category among subgroups to the total distribution of all words within subgroups, the closer it will be to the origin. Words or categories that are plotted far from this point of origin have singular distributions.

- If two words or computed categories have similar distributions (or profiles) among subgroups of the independent variable (columns), their points in the correspondence analysis plot will be close together. For example, if the words consist of artist names and the studied subgroups represent different age groups, then if the form of the distribution of two different artists among those age groups is similar, then they will tend to appear near each other. Words with different profiles will be plotted far from each other. Please note however that two points may appear close to each other on a two or three axes solution, but may in fact be far apart when taking into account an additional dimension.

**Relationship among subgroups (column points)**

- The more singular a profile of words/categories for a subgroup is, compared to the distribution of those words/categories for the entire sample, the farther this subgroup will be from the point of origin.

- If two subgroups of individuals have similar profiles of word usage or content categories, they will be plotted near each other.

- Subgroups with different profiles will be plotted fare from each others.
Relationship between words/categories and subgroups (row and column points):

- Great caution should be made when interpreting the distances between two sets of points (row and column points). The fact that the name of a subgroup is near a specific word or content category should not necessarily be interpreted as an indication that they are closely related.

- While the distance between words or content categories and subgroups has no interpretable meaning, the angle between such a word point and a subgroup point from the origin is meaningful:
  - An acute angle indicates that the two characteristics are correlated.
  - An obtuse angle, near 180 degrees, indicates that the two characteristics are negatively correlated.

In the example below, REM, U2 and RadioHead could be viewed as related to 18 years old listeners. However, while both REM and U2 points are closer to the 18 years old subgroup, Radiohead should be identified as more characteristic of those listeners since it is farther from the origin.

- Words or categories closely associated with two subgroups will be plotted in an angle from the origin that will lie between those two groups. In the above example, the rock groups Korn and Metallica seems to be characteristics of both 15 and 16 years old listeners.
For a more comprehensive description of this method, its computation and applications, see Greenacre (1984). For an application of correspondence analysis to the analysis of textual data see Lebart, Salem, and Berry (1998).
Filtering records

The FILTER RECORDS command temporarily selects cases according to some logical condition. You can use this command to restrict your analysis to a subsample of cases or to temporarily exclude some subjects. The filtering condition may consist of a simple expression, or include many expressions related by logical operators (i.e. AND, OR, NOT). The condition expression should be a valid xBase expression evaluated as true or false and may not exceed 240 characters. (see pages 63 to 70 for more instructions on expression operators and evaluation rules and supported xBase functions)

When selecting the Filter command, a filter builder dialog box is displayed.

You can directly type the filtering expression in the Record Filter edit box using the proper syntax, or use any elements displayed on the upper part of the dialog box to build a valid expression. To restore previously used filtering conditions, click on the down arrow button located to the right of the Filter edit box.

Records can also be filtered based on the presence or absence of specific codes resulting from the current content analysis. The Codes section allows you to select records that contains a single category or a specific combination of two categories. The AND, OR and NOT boolean operators are used to determine how those two categories should be combined.
For example:

**APPEARANCE AND ART** selects only records that contain words in both categories.

**APPEARANCE OR ART** selects records that contain words in either one of these two categories.

**APPEARANCE NOT ART** selects all records that contain a word in category APPEARANCE but that do not contain any word found in the ART category.

Once a filtering condition has been entered you can apply the filter and leave this dialog box by clicking on the APPLY button. If the filter expression is invalid, a message is displayed and the exit is not performed.

To temporarily deactivate the current filter expression, click on the IGNORE button. The filter string will be kept in memory and may be reactivated by choosing the FILTER command again.

To exit from the dialog box and restore the previous active filtering condition, click on the CANCEL button.

The upper part of the Filter dialog box contains various elements to help you build a valid xBase filtering expression:

**VARIABLE NAME LIST BOX** - Double-clicking a variable name from the list box located to the left of the dialog box inserts that name in the edit box at the current caret position.

**FUNCTION LIST BOX** - A list of valid xBase expression is displayed to the right of the dialog box. Double-clicking on a xBase function in the list box, inserts that function at the current caret position. When a function requires one or several arguments, the argument section remains highlighted. To replace the highlighted text with a value, an expression or a variable name, simply type the proper text on the keyboard or select a variable name or function.

**NUMERIC, BOOLEAN AND RELATIONAL OPERATORS BUTTONS** - Clicking on any relational or boolean operation or on any numeric button inserts the corresponding symbol in the edit box at the current caret position.

The following section provides a description of xBase syntax rules used in the FILTER command and a detailed description of each xBase function.

### Expression Operators and Rules

Operators used in xBase expressions are standard in every xBase dialect.

**String Operators**

- **+** Joins two strings. Trailing spaces in the strings are placed at the end of each string.

- **-** Joins two strings and removes trailing spaces from the string preceding the operator and places them at the end of the string following the minus sign operator.
Numeric Operators

+  Addition
-  Subtraction
*  Multiplication
/  Division
^  Exponentiation (or **)  

Relational Operators

=  Equal to
==  Exactly equal to
<>  Not equal to
#  Not equal to
!=  Not equal to
<  Less than
>  Greater than
<=  Less than or equal to
>=  Greater than or equal to
$  Is contained in

Logical Operators (Notice the periods surrounding the operator)

AND  both expressions are true
OR   either expression is true
NOT  either expression is false

Evaluation Order

When more than one type of operator appears in an xBase expression, the order of evaluation is as follows: Expressions containing more than one operator are evaluated from left to right. Parentheses are used to change the evaluation order. If parentheses are nested, the innermost set is evaluated first.

Numeric operators are evaluated according to generally accepted arithmetic principles:

operators contained in parentheses
exponentiation
multiplication and division
addition and subtraction

Order of evaluation may be altered with parentheses:

\[ 3+4*5+6 = 29 \]
\[ (3+4)*5+6 = 41 \]
\[ (3+4)*(5+6) = 77 \]
**Logical operators** are evaluated as .NOT. first, .AND. second, and .OR. last. Logical evaluation order may also be altered with parentheses. In multiple conditional expressions that contain the .NOT. operator, always use parentheses to enclose the .NOT. operator with the expression to which it applies.

**Supported Xbase functions**

The following xBase functions are supported in the SORT RECORDS and FILTER RECORDS command.

**NOTE:** Memo field names are not allowed in SIMSTAT xBase expressions.

**ALIAS()**

Returns the Alias name of the current work area as a string.

**ALLTRIM (String)**

Trims both leading and trailing spaces from a string. The string may be derived from any valid xBase expression.

\[ \text{ALLTRIM(" Provalis ")} \text{ returns 'Provalis'.} \]

**AT (SearchString, TargetString)**

Determine whether a search string is contained within a target. If found, the function returns the position of the search string within the target string (relative to 1). If not found, the function returns 0 (zero).

\[ \text{AT("gh", "defghij") returns 4.} \]

**CHR (Val)**

Converts a decimal value to its ASCII equivalent.

\[ \text{CHR(83) returns 'S'} \]

**CTOD (String)**

Converts a character string into an xBase date. The string must be formatted according to the Windows date format settings.

\[ \text{CTOD("12/31/94")} \]

**DATE ()**

Returns the system date (today). Use DTOC(DATE()) to retrieve today's date formatted according to the Windows settings.
DAY (DateField)

Returns the day portion of an xBase date as an integer.

DELETED()

Returns True if the record is deleted and False if not deleted.

DESCEND (String)

An xBase function that inverts a key value using 2's complement arithmetic. The result of the operation is the arithmetic inverse of the key value. When inverted keys are sorted in ascending sequence, the result is in descending order. A filter expression could be

\[ \text{DESCEND} \left( \text{DTOS(billdate)} \right) + \text{CUSTNO} \]

DTOC (DateField)

Converts an xBase date into a character string formatted according to the Windows settings. For example, if the date format was American and the date field contained March 21, 1995, DTOC(datefield) would return '03/21/1995'.

DTOS (DateField)

Converts an xBase date into a string formatted according to standard xBase storage conventions (CCYYMMDD). For example, December 21, 1993 would be returned as '19931221'. Indexes that contain date elements should use the DTOS () function, which naturally collates into oldest date first.

EMPTY (Field)

Reports the empty status of any xBase field. Character and date fields are empty if they consist entirely of spaces. Numeric fields are empty if they evaluate to zero. Logical fields are empty if they evaluate to False.

Memo fields that contain no reference to a memo block in the associated memo file are empty.

IF (Logical, True Result, False Result)

This is the immediate if function. If the Logical expression is true, return the True result, otherwise return the False result. The types of the True Result and the False Result must be the same (i.e., both numeric, or both strings, etc.) The logical expression must of course evaluate as True or False.

\[ \text{IF(DATE()} - \text{CTOD("12/31/93")} > 0, "This Year", "Last Year") \]

IFF (Logical, True Result, False Result)

Supported exactly like IF() as noted above.

INDEXKEY()

Returns the current index key as a string. (Same as ORDKEY()).
**LEFT (String, Length)**

Returns the leftmost characters of the expression for the defined length.

`LEFT("xyzabc", 3) returns 'xyz'`.  

**LEN (Expression)**

Returns the length of the expression result as an integer.

**LOWER (String)**

Converts the string expression into lower case.

**MONTH (DateField)**

Returns the month portion of an xBase date as an integer.

**ORDER ()**

Returns the current index order as an integer.

**ORDKEY ()**

Returns the current index key as a string. (Same as INDEXKEY())

**PADC (String, Length, Character)**

Centers the passed string between a number of the passed character to make the string the specified length.

`'[' + PADC("Smith", 9,"-") + ']' returns '[--Smith--]'`.  

**PADL (String, Length, Character)**

Pads the passed string to the specified length with the specified characters. If the string is longer than the value specified by Length, the string is truncated to this length.

`'[' + PADL("Smith", 8,"*" ) + ']' returns '[***Smith]'`.  

`'[' + PADL("John Smith", 8, " ") + ']' returns '[John Sc]'`.  

**PADR (String, Length, Character)**

Pads the passed string to the specified length using the specified character. If the string is longer than the value specified by Length, the string is truncated to this length.

`'[' + PADR("Smith", 8, " ") + ']' returns '[Smith  ]'`.  

`'[' + PADR("John Smith", 8, " ") + ']' returns '[John Sc]'`.  

---

RAT (SearchString, TargetString)

Determine whether a search string is contained within a target, starting from the right side of the target string. If found, the function returns the position of the search string within the target string (relative to 1). If not found, the function returns 0 (zero).

RAT ("ab", "abzaba") returns 4.

RECCOUNT ()

Returns the number of records in the table as a long integer.

RECNO ()

Returns the current physical record number as a long integer.

RIGHT (String, Length)

Returns the rightmost characters of the expression for the defined length.

RIGHT ("xyzabc", 3) returns 'abc'.

SELECT ()

Returns the workarea number for the current work area as a long integer.

SPACE (Length)

Returns a string consisting entirely of spaces for the defined length.

STOD (String)

The inverse of DTOS(). STOD() converts a string formatted according to standard xBase storage conventions (CCYYMMDD) to an xBase Date formatted according to the Windows settings.

STR (Number, Length, Decimals)

Converts a number into a right-justified string with decimal digits following the decimal point. The total length of the string is defined by the length parameter. STR(RECNO(), 5, 0) is a common indexing element that ensures creation of unique keys if appended to another field element.

An index key using this expression could be built with NAME + STR(RECNO(), 5, 0)

If the decimals parameter is omitted, the function defaults to zero decimal places. If the length parameter is omitted as well, the length of the result is the length of the field.
STRZERO (Number, Length, Decimals)

Converts a number into a, zero-padded right justified string with decimals digits following the decimal point. The total length of the string is defined by the length parameter.

\texttt{STRZERO( 1234, 10, 2 ) returns '0001234.00'}

If the decimals parameter is omitted, the function defaults to zero decimals. If the length parameter is omitted as well, the length of the result is the length of the field.

SUBSTR (String, Start, Length)

Returns a portion of the string expression starting at the defined start location for the defined length.

\texttt{SUBSTR('xyzabcd', 3, 4) returns 'zabc'}.

TIME()

Returns the system time as a string in the form HH:MM:SS.

TRANSFORM (Expression, Picture)

Transform converts strings and numeric values into formatted character strings. The function transforms the result of the first expression in accordance with the second picture string.

The picture string is made up of two parts. The first part is the Function string and it is optional for both strings and numeric values (as long as the second Template string is present).

A character string transformation picture may consist of only a Function string or only a Template or both.

A numeric picture must contain a Template string; the Function string is optional.

A logical value must contain only a Template string with Template characters L or Y.

The Function string consists of a leading @ character followed by one or more formatting characters. If the Function string is present, the @ character must be the first character in the picture string with its formatting characters immediately following and it may not contain spaces.

If a Template string exists as well, it follows the Function string. A single space separates the Function string and the Template string.

Function string characters allowed for numeric values are:

\begin{itemize}
  \item \texttt{B} left justify;
  \item \texttt{C} display CR after positive numbers;
  \item \texttt{X} display DR after negative numbers;
  \item \texttt{Z} blank a zero value;
  \item \texttt{(} enclose negative numbers in parentheses.
\end{itemize}

Function string characters allowed for strings are:

\begin{itemize}
  \item \texttt{R} inserts unassigned template characters;
  \item \texttt{!} converts all alpha characters to upper case.
\end{itemize}
The @R Function requires a Template; the ! Function does not.
The Template string describes the format on a character by character basis. The Template string is made up of special characters which have specific results and optional unassigned characters which either replace characters or are inserted in the formatted string depending upon the absence or presence of the @R Function string.

Template assigned characters are as follows:

A,N,X,(,# are place holders and are interchangeable;
L displays logical values as T or F;
Y displays logical values as Y or N;
! converts the corresponding character to upper case;
, (comma) or a space (in Europe) in a numeric template separate the elements of a number;
. (period) or , (comma - in Europe) in a numeric template specify the decimal position;
* fills leading spaces with asterisks in a numeric template;
$ as the leading character in a numeric template results in a floating dollar sign being placed in front of the formatted number.

Example: Where "phone" is a character field holding a phone number with no formatting characters.
'transform(phone, "@R (###) ###-####")' returns '909) 699-6776'.

If the formatting characters were actually present in the field, the "@R" function would be omitted.

For numeric fields,
'transform(123456.78, "$9,999,999.99")' returns '123,456.78'.

TRIM (String)

Removes trailing spaces from the string expression.

UPPER (String)

Converts the string expression into upper case. Character fields used in index expressions should always be converted to upper case to insure correct collating sequence.

VAL (String)

Converts a string of numeric characters into its equivalent numeric value. The conversion stops at the first non-numeric character encountered (or the end of the string).

VAL("123ABC") returns a value of 123.

YEAR (DateField)

Returns the year portion of an xBase date as an integer.
Performing Analysis on Manually Entered Codes

While WordStat may be used to perform automatic content analysis of textual data by the development and application of categorization schema, the program may also be used to assist human coders in their efforts to develop coding rules and manually assign codes to text.

The most basic use of the module for manual coding is probably its possibility to extract manually entered codes and perform frequency analysis on those codes as well as the possibility to examine relationships between those codes and other associated characteristics, such as the sex of respondent or date of publication. For example, one can compare codes assigned to different subgroups of subjects (male vs female, immigrant vs non-immigrant, job position, etc.) or establish the existence of a trend relating the presence of those codes and a numeric variable (age, education or income level, level of expertise, publication date, etc.). At least two methods of coding may be used to achieve this:

Unique Keywords

Unique keywords or code name may be inserted anywhere in the text. Those keywords should preferably not be an existing dictionary word. For example, it may consists of an abbreviation of one or several words, includes special symbols (such as #, &^, _ etc.) or numeric digits. The retrieval of those codes can then be achieved by adding all those keywords to the inclusion list. If special symbols have been used, they should also appear in the VALID CHARACTERS edit box.

Codes in square brackets

WordStat provides an option to process only the text found between square brackets (i.e. [ and ]). Codes corresponding to categories may be typed directly in the text and placed within square brackets. By disabling all dictionaries and setting this option, the program simply ignores all text outside those brackets and performs a frequency count of all words found inside square brackets. This method has several advantages over the other method. First, there is no need to enter all existing codes in the inclusion dictionary since they will be processed automatically. Misspelled keywords will always be extracted and may thus be easily identified and changed. Also, the processing time and memory requirement can be much lower than the other approach since WordStat won't keep a list of uncoded words and won't process any text found outside those brackets.

Beside the possibility to extract manually entered codes and perform frequency and comparison on those codes, there are several other ways in which the program may be used to assist the work of human coders. Here are just a few examples of possible uses:

- During the exploratory phase of the analysis, word frequency and crosstabulation may be used to identify differences between subgroups in word usages, differences that may have remained unidentified.

- The KWIC list may be used to locate and visualize text associated with specific codes either to validate the coding or identify associated themes. The KWIC list may also be used to perform a systematic search of words frequently found along with a specific code. The examination of all records containing those words may then help identify instances where the code should have been used.
- A dendrogram of word similarities or a crosstabulation of manually entered codes against all words in a text may allow identification of specific words associated with codes and permit the development of dictionaries. Such dictionaries may then be used either to insure a more systematic application and verification of manually entered codes (by locating individuals who have used those words but didn't receive the associated code) or to develop and validate a dictionary that will later be used for automatic content analysis.

- Reliability of coding made by a single coder at different times or by several coders may also be assessed using inter-raters agreement statistics. See Computing Inter-Raters Agreement Statistics (page 73) for more information on how to compute those statistics.
Computing Inter-Raters Agreement Statistics

The reliability problem in manual coding

When coding of a text is performed automatically using dictionaries, the reliability of the coding is irrelevant since the rules used for coding are explicit and are applied systematically on the entire corpus of texts. However, when the coding is performed manually by human coders, individual differences in interpretation of codes between human coders often occur no matter how explicit, unambiguous and precise the coding rules are. Even a single coder is often unable to apply the same coding rules systematically across time. One way to insure the reliability of the application of coding rules is to ask different raters to code the same content or to ask a single coder to code the same content at different times. The comparison of codes is then used to identify differences in interpretation, clarify ambiguous rules, identify ambiguity in the text, and ultimately quantify the final level of agreement obtained by those raters.

Inter-raters agreement measures

WordStat provides eight different inter-raters agreement measures to assess the reliability in coding. The assessment made by WordStat is based on the presence or absence of a specific code and is performed on each code individually. WordStat cannot assess the reliability of multinomial coding. However, this type of assessment can be achieved within SIMSTAT by storing those codes into numeric variables and using the TABLES | INTER-RATERS command.

The simplest measure of agreement for nominal level variables is the proportion of concordant coding out of the total number of codings made. Unfortunately, this measure often yields spuriously high values because it does not take into account chance agreements that occur from guessing. Several adjustment techniques have been proposed in the literature to correct for the chance factor, three of which are available in the WordStat. The following are the assumptions made by each of these correction techniques:

**Free marginal adjustment** assumes that all categories on a given scale have equal probability of being observed. It assumes that coders decisions were not influenced by information about the distribution of the codes. This coefficient is equivalent to the Bennett, Alpert and Goldstein's (1954) S coefficient, Jason and Vegelius's (1979) C Coefficient, and Brennan and Prediger's (1981) kn Index (Zwick, 1988).

**Scott's pi** adjustment does not assume that all categories have equal probability of being observed, but does assume that the distributions of the categories observed by the coders are equal.

**Cohen's kappa** adjustment also does not assume that all categories have equal probability of being observed. Contrary to the pi measure, it does not assume that the distribution of the various categories is equal for all coders. In the computation of the chance factor, Kappa takes into account the differential tendencies or preferences of coders.

WordStat also offers three additional adjustments for ordinal level variables. These are similar to the previous measures except that they also take into account the ordinal nature of the scales by weighting the various levels
of agreement. They apply the same tree model of chance agreement used in the previous measures for nominal data.

Free marginal adjustment for ordinal level variables also assumes that all categories on a given scale have equal probability of being observed.

Krippendorf's R-bar adjustment is the ordinal extension of Scott's pi and assumes that the distributions of the categories are equal for the two sets of ratings.

Krippendorf's r adjustment is the ordinal extension of Cohen's Kappa in that it adjusts for the differential tendencies of the judges in the computation of the chance factor. The following table illustrates the difference between those six inter-raters measures:

<table>
<thead>
<tr>
<th>DISTRIBUTION ASSUMPTION</th>
<th>LEVEL OF MEASUREMENT</th>
</tr>
</thead>
<tbody>
<tr>
<td>Theoretical uniform distribution</td>
<td>Free marginal</td>
</tr>
<tr>
<td>Observed distribution of each rater</td>
<td>Cohen's Kappa</td>
</tr>
<tr>
<td>Observed distribution of all raters</td>
<td>Scott's pi</td>
</tr>
<tr>
<td></td>
<td>Krippendorf's r</td>
</tr>
</tbody>
</table>

Finally, the intraclass correlation statistic is used to assess the level of agreement of

**WARNING:** When the computation is based on word frequencies (rather than record occurrences) and when codes can be used more than one time per record, it is usually recommended to use ordinal or interval level agreement measures. Otherwise a difference in frequency for a specific code will be treated as a single disagreement. For example, if for a single record a coder assigns a code twice and another coder uses this same code three times, nominal level agreement measures will treat this difference as a single disagreement and will ignore the fact that the both raters may be in agreement in two instances. As a result, the overall agreement level will be underestimated. However, nominal level measures may still be used in those situations if the researcher wishes to treat any difference in frequency as a disagreement.

**How to compute inter-raters agreement statistics**

In order to compute inter-raters agreement measures within WordStat, codes assigned by each coder should be stored in different alphanumeric fields (character or memo fields). For example, in four different raters are assessed, four textual fields should be created, each one containing the codes of a single coder. Two different methods can be used to achieve this. The text field where the content to be analyzed is stored, may be replicated as many times as there are raters. Coders are then asked to insert their codes directly in the text field. The second method requires the creation of empty alphanumeric fields which will contain the codes assigned by coders. Coders then read the content to be analyzed and write down all their codes in this empty field.
To compare the codes of different raters

In the SIMSTAT program, use the CHOOSE X1-Y command from the STATISTICS menu to assign all fields containing codes of different coders to the list of dependent variables.

- Select the CONTENT ANALYSIS command from the STATISTICS menu.
- Set the various options in the DICTIONARIES page so that all codes are extracted from the text fields.
- Click on the CROSSTAB page.
- Set the WITH list box to <variables>. The AGREEMENT list box should appear.
- Set the TABULATE list box to either Word Frequency or Record Occurrence.
- Select the appropriate measure of agreement. (Tables based on word frequency should preferably be assessed using ordinal or interval level inter-raters agreement measures while tables based on record occurrence can be analyzed using either nominal or ordinal level measures).

To identify the source of disagreement

When codes are inserted directly in the original text, it becomes possible to verify the source of disagreement by extracting the content in which those codes appear. To perform such a task:

- Click on the KEY-WORD-IN-CONTEXT page.
- Set the LIST option to Included Words and set the word to the category you would like to examine.
- Set the SORT BY option to Record Number.

You can then use the KWIC table to identify specific contents that were associated with a code by a specific coder but not by other coders.
Saving Numeric and Text Results into a Data File

Record occurrence or word frequency for each included word or category may be appended to the existing data file or stored in a new data file for further processing. WordStat can also store KWIC tables into a new data file allowing the user to examine more closely those KWIC lists either from within SIMSTAT or from other applications. The SAVE button located at the bottom of the WordStat window is used to access the following dialog:

![Save information dialog](image)

**Note:** When the WordStat window first appears, the Save button is disabled. It will become enabled as soon as a word count operation is performed.

**Saving word frequency statistics**

**SAVE WORD COUNT** - This option allows you to specify whether data obtained from a word count operation should be saved into a data file. When selected, WordStat will create as many numeric fields as included words and will store for each record in the original data file, the number of times a word was found (word frequency) or a binary value indicating whether it appeared in this specific record (see DATA TO SAVE option).
DESTINATION - This option allows you to choose whether the new variables should be appended to the current data file or into a new file. If this last option is selected, a dialog box will appear allowing you to specify the name and location of the new file. When data are saved to a new data file, additional fields are created to store the record number and the numerical values of each independent variable.

DATA TO SAVE - This option allows you to specify whether the data to be saved should be record occurrences (i.e. as a dummy variable with 0 for absent, 1 for present) or word frequencies.

VARIABLE NAMES - This option lets you determine what method should be used by WordStat to create new variable names. When set to WORD, the program will attempt to use each word 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 WORD COUNT page.

Saving Key-Word-In-Context tables in a data file

SAVE KWIC LIST - This option allows you to save a KWIC table as a new data file. When selected, WordStat will create a DBF file with the following fields:

- RECORDNO: Numeric field containing the record number where the keyword was found.
- CATEGORY: Alphanumeric field containing the name of the included word or category (if a substitution dictionary was used, a single category might include several word forms or keywords).
- VARIABLE: When more than one dependent variable was selected, this alphanumeric field is created to store the name of the text field where the keyword appeared.
- BEFORE: Alphanumeric field containing the text that appears just before the keyword.
- KEYWORD: Alphanumeric field containing the selected keyword.
- AFTER: Alphanumeric field containing the text appearing just after the keyword.
- <other>: When independent numeric variables have been selected, additional fields with the same name and of the same type will be created. Each of those fields will be used to store the numerical value of the corresponding independent variable.

The KWIC table to be saved can consist of all included words or categories or may be restricted to the currently active KWIC table. For example, if the KWIC table displays all words corresponding to the user-defined pattern SUPPORT*, setting the option to CURRENT KWIC LIST will create a data file which will contain information on all words matching this word pattern.
Saving WordStat tables in RTF, ASCII or HTML files

SAVE TABLES TO DISK - This option allows you to save the displayed WORD FREQUENCY, CROSSTABULATION, or KWIC table to disk in HTML, RTF or plain text file format. For each selected table, a Windows Save As dialog will be displayed allowing you to specify the location and name of the file, as well as its file format.
Performing Multivariate Analysis

One benefit of the integration of a content analysis module within an existing statistical program is the ability to easily perform on numerical results of content analysis, various statistical analyses such as frequency, crosstabulation, multiple regressions, reliability analysis, etc. Among the various multivariate analysis techniques used in content analysis, cluster analysis, factor analysis, and, to a lesser degree, correspondence analysis are often used to establish the relationship existing between different words or categories of words. The table below illustrates some types analysis that may be performed with SIMSTAT and, if needed, the required module to perform those analysis.

<table>
<thead>
<tr>
<th>TYPE OF ANALYSIS</th>
<th>REQUIRED MODULE</th>
</tr>
</thead>
<tbody>
<tr>
<td>Multiple regression analysis</td>
<td>None</td>
</tr>
<tr>
<td>n-way ANOVA/ANCOVA</td>
<td>None</td>
</tr>
<tr>
<td>Reliability analysis</td>
<td>None</td>
</tr>
<tr>
<td>Inter-raters agreement</td>
<td>None</td>
</tr>
<tr>
<td>Factor Analysis (with or without varimax rotation)</td>
<td>EFA</td>
</tr>
<tr>
<td>Principal Component Analysis (without rotation)</td>
<td>MVSP</td>
</tr>
<tr>
<td>Correspondence analysis</td>
<td>MVSP</td>
</tr>
<tr>
<td>Advanced cluster analysis</td>
<td>MVSP</td>
</tr>
</tbody>
</table>

First step - Saving numerical results into a data file

In order to perform a statistical analysis on categories or words, you first need to create new numeric variables that will contain, for each record in the data file, the occurrence or frequency of specific words or categories. To create those variables:

- Set the various options of the DICTIONARIES page.
- Perform the word count analysis by clicking on the WORD COUNT or CROSSTAB page.
- Click on the SAVE button to access the Save Data dialog box.
- Activate the SAVE WORD COUNT checkbox.
- Set the DESTINATION option to Append To Current File.
- Set the DATA TO SAVE list box to Word Frequency.
- Set the VARIABLE NAMES option to Word to instruct WordStat to use the names of the categories (or included words) as the names for the new variables.
- Click on the OK button to proceed to the saving of those data and return to WordStat.
- Click on the OK button of the WordStat dialog to return to SIMSTAT.

For more information on how to store numeric or textual results into the current data file or into a new data file see Saving Numeric and Text Results into a Data File (page 76).
Performing a Cluster Analysis of words or categories

- Select the CHOOSE X-Y command from the STATISTICS menu and assign all the newly created variables to the list of independent or dependent variables. (The distinction between dependent and independent variables is not relevant for this kind of analysis. However, all variables assigned to a single category will be processed together.)

- Choose the OTHER | CLUSTER ANALYSIS command from the statistics menu to display the option dialog.

- Set the various analysis options to one's preferences. (Please take note that in order to perform a cluster analysis on words rather than on records, the Transpose Data option should be left deactivated).

- Click on the OK button to perform the statistical analysis.

Performing a Factor Analysis of words or categories

- Select the CHOOSE X-Y command from the STATISTICS menu and assign all the newly created variables to the list of independent or dependent variables. (The distinction between dependent and independent variables is not relevant for this kind of analysis. However, all variables assigned to a single category will be processed together.)

- Choose the OTHER | FACTOR ANALYSIS command from the statistics menu.

- Set the various options to one's preferences.

- Click on the OK button to perform the statistical analysis.
References

Introduction to Content Analysis


Interrater agreement statistics


Others

If you have any comment or suggestion for further improvement please contact Provalis Research:

By Phone: 514-899-1672
By FAX: 514-899-1750
By Email: support@simstat.com
Web site: http://www.simstat.com
Mail: Provalis Research
      2414 Bennett Street
      Montreal, QC
      H1V 3S4