



WORDSTAT 9

Text Analytics Software

User's Guide

Provalis Research

Table of Contents

Introduction to WordStat 9.0.....	1
Program Capabilities.....	2
The Content Analysis & Categorization Process.....	7
A Quick Tour.....	9
Performing Content Analysis	9
Starting WordStat from QDA Miner or SimStat.....	12
Creating a Project in WordStat.....	14
Creating a Project from a List of Documents	14
Creating a WordStat Project from Windows Explorer	17
Notes on Importing PDF Files	19
Creating a Project from an Existing Data File or Web Service	19
Importing from a Data File	20
Importing from Web Surveys	24
SurveyMonkey.....	24
Qualtrics.....	26
SurveyGizmo.....	29
Voxco.....	31
QuestionPro.....	33
TripleS v.1.2.....	36
Importing from Social Media	36
Importing from RSS.....	37
Importing from Twitter.....	38
Importing from Facebook	40
Importing from Reddit.....	42
Importing Bibliographic References	43
EndNote	44
Mendeley.....	45
Zotero.....	47
RIS File.....	48
Importing News Transcripts	49
Nexis UNI.....	50
Factiva	53
Importing from Email Servers	55
Outlook.....	56
Hotmail.....	59
Gmail.....	61
MBox	62
Using the Document Conversion Wizard	63
Starting WordStat Document Explorer from Windows Explorer.....	65
The User Interface.....	68
The Data Tab	70
Project Properties	72

Appending Documents	74
Appending from a Data File	76
Monitoring a File or Folder	77
Deleting Cases	80
Identifying Duplicate Cases	81
Setting the Case Descriptor	84
Filtering Cases	85
Editing Project Data	86
Adding New Variables	86
Deleting Existing Variables	87
Reordering Variables	88
Transformation	88
Changing Variable Types.....	89
Recoding Values of a Variable.....	91
Transform Document.....	92
Binning Numerical Variables.....	93
Performing Complex Numeric Transformation.....	95
xBase functions.....	97
Editing Variable Properties	101
Variable Statistics	105
Analyze	108
The Text Processing Tab	111
Language	113
Preprocessing	113
Notes on Preprocessing.....	116
Writing a Preprocessing Routine with Python.....	117
Calling an Executable Program.....	119
Calling a Function in a DLL.....	119
Exclusion	120
Substitution	124
Categorization	127
Creating and Maintaining Dictionaries.....	134
Working with Rules.....	138
Working with Regular Expressions	140
Importing Dictionaries.....	142
Exporting Dictionaries	144
Printing Dictionaries.....	144
Using Lexical Tools for Dictionary Building.....	145
Basic Dictionary-Building Tools.....	146
Advanced Dictionary-Building Tools.....	147
Postprocessing	154
The Frequencies Tab	156
Using the Dictionary Panel	159
Working with the Suggestions Tab	161
Barcharts, Pie Charts, and Word Clouds	162
Keyword Retrieval	165
Running and Creating Post-Processing Scripts	170
The Extraction Tab	177
Topics	177
Phrases	182
Named Entities	187
Misspellings & Unknowns	188
The Cooccurences Tab	195
Hierarchical Clustering and Multidimensional Scaling Options	195
Dendrogram	198
2D and 3D Concept Maps	200
Multidimensional Scaling Plot Options Dialog (COPY).....	202
Proximity Plot	203

Link Analysis	206
Statistics	210
Cooccurrence Matrix	211
The Crosstab Tab	215
Bubble Charts	220
Heatmap Plot	222
Correspondence Analysis	226
Keyword-In-Context Page	229
The Classification Tab	234
Settings	235
Select Features	236
Learn & Test	239
History	241
Classification Experiment Dialog Box.....	244
Apply	245
Miscellaneous.....	250
Preparing and Importing Data	250
Preliminary Text Preparation	250
Importing Spreadsheet or Database Files	251
Importing Plain Text or Word Processor Files from Simstat	254
Creating Comparison Charts	257
Barchart and Line Chart Options	259
Edit Case Descriptors	261
Filtering Cases	262
Geocoding	264
Mapping	265
Customizing Markers	267
Text Editor	269
Publishing Categorization Models	272
Creating and Using Norm Files	273
Program Settings	274
Mode	275
Exporting Frequency Data	276
Performing Multivariate Analysis	278
Performing Analysis on Manually Entered Codes	280
Web Collector	281
Word Frequency Analysis	285
WordStat Software Development Kit (SDK)	288
Report Manager.....	289
Working with the Table of Contents	290
Creating a New Item	290
Importing Items from Files	290
Renaming an Item	291
Deleting an Item	291
Moving an Item	291
Adding or Editing Item Comments	292
Editing Documents	292
Editing Tables	292

Editing Charts	293
Searching and Replacing Text	293
Exporting items to HTML or Word	294
References on Content Analysis.....	297

Introduction to WordStat 9.0

WordStat is a text analysis module specifically designed to study textual information such as responses to open-ended questions, interviews, titles, journal articles, public speeches, electronic communications, etc. WordStat may be used for automated categorization of text using a dictionary approach or various text mining methods. WordStat can apply existing categorization dictionaries to a new text corpus. It also may be used in the development and validation of new categorization dictionaries. WordStat can now be used as a stand-alone software. When used as in conjunction with QDA Miner, this module can provide assistance for a more systematic application of coding rules, help uncover differences in word usage between subgroups of individuals and assist in the revision of existing coding using KWIC (Keyword-In-Context) tables.

WordStat includes numerous exploratory data analysis and graphical tools that may be used to explore the relationship between the content of documents and information stored in categorical or numeric variables such as the gender or the age of the respondent, satisfaction scores, publication date, etc. Relationships among words or categories as well as document similarity may be identified using hierarchical clustering, topic modeling and multidimensional scaling analysis. Correspondence analysis and heatmap plots may be used to explore relationship between keywords and different groups of individuals.

WordStat can be used as a stand-alone software or it can be run as a module of either of the following base products:

QDA Miner - The text management and qualitative analysis program allows you to create and edit data files, import documents, and perform manual coding and annotation of the documents. Several analysis tools are also available to look at the frequency of manually assigned codes and the relationship between these codes and other categorical or numeric variables. When used with QDA Miner, WordStat can perform content analysis on whole documents or selected segments of the documents tagged with specific user defined codes.

Simstat - This statistical software provides a wide range of statistical procedures for the analysis of *quantitative* data. It offers advanced data file management tools such as the ability to merge data files, aggregate cases, perform complex computation of new variables and transformation of existing ones. When used with Simstat, WordStat can analyze textual information stored in any alphanumeric, plain text and rich text memo variable (or field). It includes various tools to explore the relationship between any numeric variable of a data file and the content of alphanumeric ones.

Stata - Stata is a complete, integrated software package that provides all your data science needs—data manipulation, visualization, statistics, and automated reporting.

The WordStat module may be accessed the first two applications from the **CONTENT ANALYSIS** command in the **ANALYSIS** menu. When installed in Stata, WordStat is accessible from the **WORDAT | CONTENT ANALYSIS** command in the **USER** menu.

A few additional utility programs are also included with WordStat that may be run as standalone applications or be accessed directly through WordStat:

- **Report Manager** - This application has been designed to store, edit and organize documents, notes, quotes, tables of results, graphics and images created by QDA Miner and WordStat or imported from other applications.
- **Document Conversion Wizard** - This utility program provides an easy way to import numerous documents and create a project file. It can also be used to split large files into smaller units and to extract various numeric and alphanumeric data from structured documents.
- **Document Classifier** - This utility program is a stand-alone application that may be used to perform content analysis and automatic text classification on a text pasted from the clipboard or stored in a file. For more information on this utility program, see [WordStat Document Explorer](#).
- **Dictionary Builder** - This tools allows the development of comprehensive categorization dictionary for automatic content analysis. The program may be run as standalone application but also from **Text Processing** tab of WordStat by pressing the **Suggest** button. [Click here to obtain more information on the Dictionary Builder](#).

For a detailed listing of WordStat features, see [Program's capabilities](#).

Program Capabilities

Content Analysis Capabilities

- Text analysis of textual documents.
- Full Unicode support allows one to analyze almost any language.
- Performs analyses on alphanumeric variables containing short textual information such as responses to open ended questions, titles, descriptions, etc. as well as on longer documents stored as plain ASCII or as rich text (RTF) document.
- Stemming in 24 human languages (Arabic, Armenian, Basque, Catalan, Czech, Danish, Dutch (Kraaij-Pohlmann), English (Lovins, Porter, Snowball), Finnish, French, German, Hungarian, Irish, Italian, Latin, Norwegian, Portuguese, Romanian, Russian, Slovene, Spanish, Swedish, Tamil and Turkish)
- Automatic lemmatization (available in English, French, Spanish, Swedish, German, Norwegian, Polish and Italian), contact us if you need support of other languages.
- Substitution process for customized lemmatization of words or automatic spell correction of common misspellings.
- Optional exclusion of pronouns, conjunctions, expressions, etc, by the use of existing or user-defined stop word lists (a.k.a. exclusion list).
- Automated categorization of words, word patterns, or phrases using existing or user-defined categorization dictionaries.
- Advanced proximity rules with proximity operators (NEAR, AFTER, BEFORE, etc.) and word distance within sentence, paragraphs or documents may be used for word disambiguation.
- Frequency analysis of words, content categories or concepts.
- Phrase finder allows identification of the most recurring phrases.
- Pattern-based named-entity extraction allows identification of people, product names, places and acronyms.
- Choice between automatic spelling correction and semi-automatic misspelling identification and processing.
- Interactive development and validation of multi-level categorization dictionaries or taxonomies
- Easy drag and drop assignments of words, phrases, or topics to categorization dictionaries.
- 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 cases.
- Integrated spell-checking with support for 20 languages.
- Integrated thesaurus and lexical database (English, French, Spanish, German, Norwegian, Polish, Italian) to assist the creation of comprehensive categorization dictionaries.
- Case filtering on any numeric or alphanumeric variable and on keyword occurrence (with AND, OR, and NOT Boolean operators).
- Printing of presentation quality tables.
- Export any output table to SPSS, Stata, Excel, HTML, JSON, XML, ASCII, Tab separated or comma separated value files.

- All graphics may be saved to disk in BMP, JPEG or PNG file format.

Univariate Keyword Frequency Analysis

- Univariate keyword frequency analysis (keyword count and case occurrence).
- Keyword cooccurrence matrix (within documents, paragraphs, sentences, etc.).
- Keyword by case data matrix.

Relationship to Numerical, Categorical Data, or Dates

- Comparison between any textual variable and values of one or two nominal or ordinal variables (such as sex of the respondent, specific subgroups, years of publication, etc.).
- Automatic recoding of date variables into days, weekdays, weeks, months, quarters, years or decades.
- Choice between 12 different association measures to assess the relationship between keyword 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.
- Deviation table for easy identification of characteristic words or content categories for values of categorical or numerical variables.
- Heatmap plot and dual hierarchical clustering may be used to identify functional relationship between keywords and values of categorical or numerical variables.
- Ability to sort keyword matrix in alphabetical order by keyword frequency or keyword occurrence, on the obtained statistics or on its probability.

Keyword cooccurrence and Analysis

- Integrated clustering and dendrogram display of keyword cooccurrence
- Powerful topic modeling module with unique topic enrichment features for more accurate assessment of topic (including suggestions for topic names, associated phrases, exceptions, and spelling corrections)
- 2-D and 3-D multidimensional scaling on cooccurrence of words or content categories.
- Link analysis using force-based graphs, multidimensional scaling or circular graph displays.
- Proximity plot to easily identify all keywords that cooccurs with one or several target keywords.
- Flexible keyword cooccurrence criteria (within a case, a sentence, a paragraph, a window of n words, a user-defined segment) as well as clustering methods (first- and second-order proximity, choice of similarity measures).
- Easy text retrieval directly from dendrogram or proximity plots.

Advanced Topic Modeling

- Automatic topic extraction with non-negative matrix factorization (NNMF) and factor analysis (additional algorithms such as LDA may also be performed using R or Python scripts).

- Automatic generation of topic names.
- Topics may be merged, deleted and renamed. One may also remove topic items.
- Unique topic enrichment feature for automatic inclusion of associated phrases, and generation of suggested phrases, exceptions and spelling corrections.
- Push topic solutions to the crosstabulation feature of WordStat for computation of comparison statistics and graphics, correspondence analysis, bubble plots, deviation table, etc..
- Perform co-occurrence analysis on topic solutions (hierarchical clustering, multidimensional scaling, proximity plot, etc.)
- Save topic solutions as a categorization model for further editing or to apply topic model to new datasets.

Analysis of Case or Document Similarity

- Hierarchical clustering, multidimensional scaling and proximity plot may be used to explore the similarity between documents or cases.

Norm Creation and Comparison

- Ability to create norm files based on frequency analysis of words or content categories.
- Comparison of obtained frequencies to previously saved norm files.

Multiple Responses and Comparisons Between Variables

- Can perform a single frequency analysis on information stored in several text variables (documents or short alphanumeric variables)
- Comparison of keyword occurrence between variables.

Automated Text Classification

- Machine learning algorithms (Naive Bayes and K-Nearest Neighbors) for document classification.
- Flexible feature selection for automatic selection of best subsets of attributes.
- Numerous validation methods (leave-but-one, n-fold crossvalidation, split sample).
- Experimentation module allows easy comparison of predictive models and fine-tuning of classification models.
- Classification models may be saved to disk and applied later using either a standalone document classification utility program, a command line program or a programming library (the command line and the programming library are part of a new software developer's kit (SDK) sold separately).

Keyword-In-Context

- Ability to display a Keyword-In-Context (KWIC) table of any included keywords, leftover or user defined word, word pattern or phrase.
- KWIC tables may be sorted in ascending order of case number, context, or on values of categorical or numerical variables.

- Ability to jump from a specific occurrence in the KWIC table to the original text variable in order to view or edit the selected document.
- KWIC tables may be saved as a new data file for further processing.
- Customizable KWIC and report function to display all hits as lists of paragraphs, sentences or user defined segments.
- Perform quick word frequency analysis on context words appearing before, after, or both before and after.

Keyword Retrieval Function

- A powerful keyword retrieval function allows identification of text units (documents, paragraph or sentences) containing one keyword or a combination of keywords with optional filtering of cases.
- Ability to attach QDA Miner codes to retrieved text segments.
- Retrieved segments may be exported to disk in tabular format (Excel or delimited text files) or as text reports (Rich Text Format).
- Perform quick word frequency analysis on retrieve results.

Supports R and Python pre- and post-processing scripts

- Real time text transformation may be performed using Python or R scripts, allowing one to perform various tasks such as cleaning text data, tokenizing Asian languages, performing part-of-speech tagging,
- Post-processing scripts in R and Python may be created to analyze either original sources or document-term matrices, allowing one to extend the analytical capabilities of WordStat with additional machine learning algorithms, predictive models, etc.
- Ability to create dialog boxes for customizing the execution of scripts. allowing the execution of flexible scripts by non-programmers.
- Automatically import text files, tables and images outputs for easy review.

Full Integration with QDA Miner and SimStat

- Document variables are stored in the same file as all other numeric and categorical variables.
- Variable selection and analysis are performed within the main statistical program using a simple 3-step operation:
 1. Open the existing data file.
 2. Select one or several alphanumeric fields as dependent variables and, optionally, other nominal or ordinal variables to be treated as independent.
 3. Execute the **CONTENT ANALYSIS** command from the **ANALYZE** drop-down menu.
- 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 Excel, SPSS, Stata, comma or tab separated text files, etc.
- Ability to perform numeric and alphanumeric transformation or to apply filters on cases of the data file to restrict the analysis to specific subgroups.

Integration with Stata

- WordStat may be configured to run from the Stata USER menu to analyze text data stored in the currently open dataset.
- The Document Conversion Wizard may be used to create Stata .dta project files out of MS Word, RTF, HTML, TXT and PDF documents.
- When running WordStat as a standalone, it is possible to import Stata files from version 8 to version 17 as well as export any project to corresponding Stata file formats.

The Content Analysis & Categorization Process

The most basic form of content analysis WordStat can perform is a simple frequency analysis of all words contained in one or several text variables of a data file. However, WordStat offers several features that permit the user to accomplish more advanced forms of content analysis that may involve automated categorization, different weighting of words and phrases, inclusion or exclusion of words based on frequency criteria, etc. To fully understand the possibilities offered by the program, you first need 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 seven consecutive processes:

1- Text Preprocessing (including stemming, n-grams, etc.)

The preprocessing option allows users to access external text preprocessing routines that are not part of the WordStat program. This option is useful to perform custom transformation on the text to be analyzed. WordStat includes a few sample text processing routines, such as a Porter stemmer which remove common English suffixes and prefixes as well as a character n-grams routine which decomposes every word into sequences of 3, 4 or 5 characters. Please note that the Porter stemmer routine available in the preprocessing process is for demonstration purpose only and will greatly reduce the processing speed of WordStat. We recommend using instead the integrated stemming routine (see #2 below)

2- Stemming

The stemming process is a natural language processing routine that reduces inflected words to a common stem or root form. The English stemmer, for example, returns "write" for "write", "writes", "writing", and "writings". Stemming routines are available for these languages: Arabic, Armenian, Basque, Catalan, Czech, Danish, Dutch (Kraaij-Pohlmann), English (Lovins, Porter, Snowball), Finnish, French, German, Hungarian, Irish, Italian, Latin, Norwegian, Portuguese, Romanian, Russian, Slovene, Spanish, Swedish, Tamil and Turkish

3- Substitution Process (including lemmatization and automatic spell correction)

The substitution process takes individual words and replace them with another word form or with a sequence of words. Such a process is typically used for lemmatization, a procedure by which all plurals are transformed into singular forms and past-tense verbs are replaced with present-tense versions. It may also be used for derivational stemming in which different nouns, verbs, adjectives and adverbs derived from the same root word are transformed into this single word. Custom substitution process may also be created to perform automatic spelling corrections of common misspelling.

4- 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 or phrases used too frequently or with little discriminative value.

5- Categorization Process

The categorization process allows you to change specific words, word patterns, or phrases to other words, keywords or content categories and/or to extract a list of specific words or codes. This process requires the specification of a categorization 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".

6- Addition of Frequent Words

The fifth process is the application of a frequency criterion that is used to add to the included words and categories words that are used more than a specific number time or that are found in more than a specific number of cases. When a categorization 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 categorization dictionary is used, all words that meet this minimum

requirement and that have not been excluded (see process #3) 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 categorization dictionary. It cannot be used to remove any of these items (see process #6).

7- Removal or Words or Categories

When this process is applied, all words or categories that do not meet a minimum frequency or case occurrence criterion will be removed from the final list. It can also remove items occurring in too many cases. This process may be combined with the categorization process (#4) to remove infrequent or too common categories. It may also be used in conjunction with the addition criterion (see process #5) to provide a composite criterion of inclusion that involves both a minimum word frequency and a minimum case 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:

Types or Analysis	Preprocessing & Substitution	Exclusion List	Categorization Dictionary	Add Words	Remove Words	Comments
Simple word frequency analysis (most frequent words)				✓*		
Simple frequency analysis of semantically significant words		✓		✓*		
Word count with lemmatization	✓	✓		✓*		
Word count of specific words			✓			
Automatic categorization of texts			✓			
Frequency analysis on the most frequent categories.			✓		✓	
Frequency analysis of manually entered codes or keywords			✓			Codes may optionally be inserted between brackets
Rating of texts on specific attributes			✓			Weights may be assigned to different words

* The use of minimum frequency criteria is recommended when you want to perform analysis on only the most frequent words.

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.PPJ.

Performing Content Analysis

The required data has been stored in a file named SEEKING.PPJ.

Step #1 - Open or create a project.

- Open the data file in either QDA Miner or SimStat, select your variables and run the content analysis module. If you have WordStat as a stand-alone software, create a project in WordStat.

Step #2 - Choose the proper dictionaries

WordStat consists of a single dialog box with seven or eight tabs. The second tab, **Text Processing**, allows you to select, view, and edit the dictionaries used in this specific content analysis. Set the dictionaries to the following values:

Exclusion: ENGLISH

Categorization: SEEKING

and make sure both of them are enabled. (see the checkboxes on the tabs)

Step #3 - Setting the proper options

Still on the **Text Processing** tab **Preprocessing** tab allows you to specify various options such as whether numeric values should be included etc. Move to the **Postprocessing** tab, here you can decide which frequent words should be added etc. Disable all options by removing any check mark in the various check boxes.

Step #4 - Perform a univariate frequency analysis on categories

- Click the third tab (Frequency). The program will perform a categorization of words found in the ads and compute a frequency analysis on the categories.
- By default, the words displayed in the matrix are those specified in the Categorization dictionary. To display words that have been left out, click the **Leftover words** tab.
- To move a word to the categorization dictionary or the exclusion list, right-click the mouse. (Click here for more information on the [creation and maintenance of dictionaries](#)).

Step #5 - Examining the relationship between included categories and the gender of the author.

- Press on the sixth tab (Crosstab).
- Click the WITH drop-down list and select GENDER to display a contingency table of category frequency by gender.

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

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

The DISPLAY option allows you to specify the information displayed:

- Count
- Row percent
- Column percent
- Total percent
- Category percent (for case occurrences)
- Percent of total words (for keyword frequency)

Step #6 - Estimating the strength of the relationship

- Use the STATISTIC drop-down list 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.

Step #7 - Visualizing the relationship between some categories and the gender of the author.

- Use the mouse to highlight cells of the categories you would like to compare.
- Click the  button or right-click the mouse and select the Chart Selected Rows menu item.

Step #8 - Performing correspondence analysis on age groups

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

Step #9 - Displaying a keyword by keyword matrix or a keyword by case matrix

- Click the WITH drop-down list and select <other keywords> to display a keyword by keyword matrix or on <case no> to view a keyword by case matrix.

Step #10 - Viewing a Keyword-In-Context (KWIC) list of specific words or categories

- Click the KEYWORD-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 the  button to display the KWIC table for this word or category.

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

Step #11 - Editing a text from the KWIC list

- To modify the word or keyword or the text surrounding it, select it from the KWIC list, right-click and select the **EDIT** command. (You may also double-click the specific line you wish to edit).

Step #12 - Creating a concordance report

- Make sure the KEYWORD-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 keyword by setting the CONTEXT DELIMITER option.
- Click the  button. Note: If this button is inactive, click the  button to refresh the content of the KWIC table and then click the  button.

Step #13 - Examining relationships between words or content categories using hierarchical clustering and multidimensional scaling

- Go to the Cooccurrence page.
- Click the DENDROGRAM tab to perform a hierarchical cluster analysis on included categories. You may change the number of partitions displayed using the No clusters option.
- Click the MAPPING tab to perform a multidimensional scaling, and display a plot in two or three dimensions.

For more information see [Hierarchical Clustering and Multidimensional Scaling](#)

Step #14 - Saving or exporting frequency statistics to disk

- Move to the **Frequencies** tab.
- Press the  button.
- Set the options and export to the existing data file or a new one.

Step #15 - Quitting the module and returning to QDA Miner or SimStat

- Click the  button in the upper left-hand corner of WordStat and select the EXIT command or click the X mark in the upper right-hand corner.

Related Topics:

[Creating, developing, and maintaining dictionaries](#)

[Performing analysis on manually entered codes](#)

[Automated Text Classification](#)

[Performing multivariate analysis \(PCA, factor analysis, correspondence analysis, etc.\) on words or categories.](#)

[Saving numeric and text results into a data file](#)

Starting WordStat from QDA Miner or SimStat

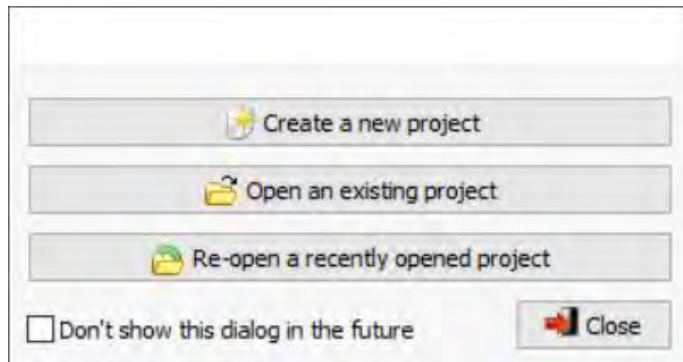
WordStat can run as a stand-alone software or as a module run from either one of the following two base products:

QDA Miner: The text management and qualitative analysis program allows you to create and edit data files, import documents, and perform manual coding of the documents. Several analysis tools are also available to look at the frequency of manually assigned codes and the relationship between these codes and other categorical or numeric variables. When used with QDA Miner, WordStat can perform content analysis on whole documents or selected segments of the documents tagged with specific user defined codes.

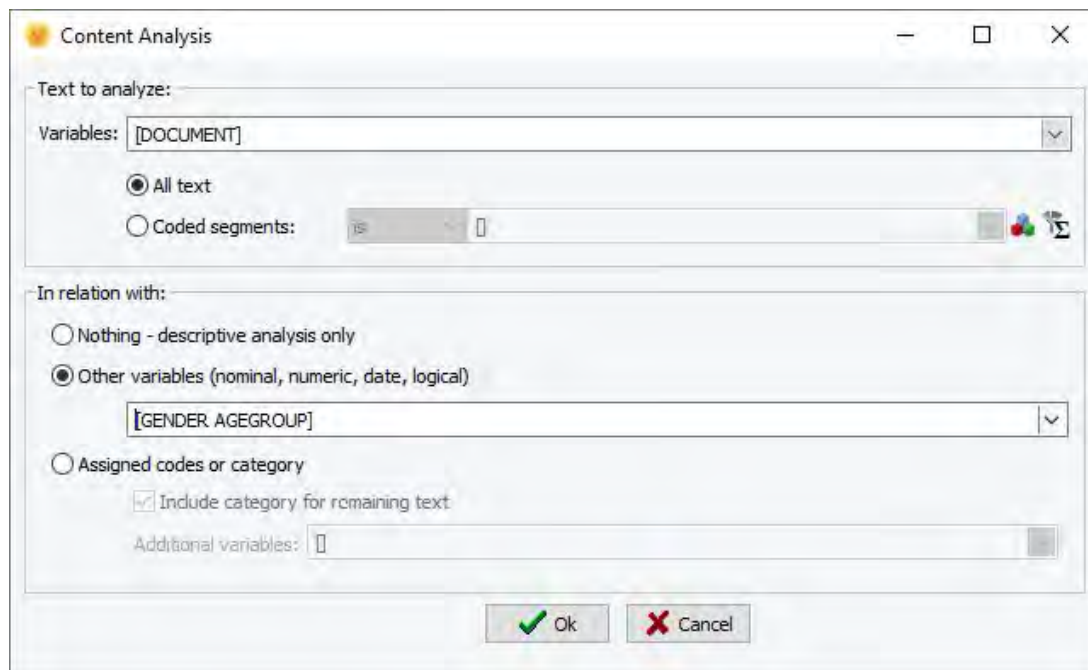
SimStat: This statistical software provides a wide range of statistical procedures for the analysis of quantitative data. It offers advanced data file management tools such as the ability to merge data files, aggregate cases, perform complex computation of new variables and transformation of existing ones. When used with Simstat, WordStat can analyze textual information stored in any alphanumeric, plain text and rich text memo variable (or field). It includes various tools to explore the relationship between any numeric variable of a data file and the content of alphanumeric ones.

To run WordStat from QDA Miner:

- Start QDA Miner. You will be presented with a dialog box like this one:



- Click the **Open an existing project** and select the data file containing the document you want to analyze.
- If you closed or disabled this introductory dialog box, from the main screen select the **OPEN** command from the **PROJECT** menu and select this data file.
- Execute the **CONTENT ANALYSIS** command from the **ANALYZE** menu.
- Select in the **Variables** list box the variable that contains the documents you want to analyze and set the text to analyze to **All text**.
- To examine the relationship between the content of those documents and any categorical or numerical variable, in the **In relation with** group box, select the **Other variables** radio button and click the drop-down list to select those categorical or numerical variables.



- Press the **Ok** button. You will be taken to WordStat's **Text Processing** tab, where you can begin your analysis.

To run WordStat from SimStat:

- From within SimStat, select the **FILE | DATA | OPEN** command sequence and select the Seeking.ppj file
- Execute the **STATISTICS | CHOOSE X-Y** command
- Set the **Variable list box** to **ALL** to view all variable 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
- Execute the **STATISTICS | CONTENT ANALYSIS** command. You will be taken to WordStat's **Text Processing** tab, where you can begin your analysis.

Creating a Project in WordStat

WordStat can be run as a stand-alone software. You can now create your projects directly from within WordStat.

Please see the following sections for more information on:

[Creating a Project from a List of Documents](#)

[Importing from a Data File](#)

[Importing from Web Surveys](#)

[Importing from Social Media](#)

[Importing from Email Servers](#)

[Importing from News Aggregators](#)

[Importing Bibliographic References](#)

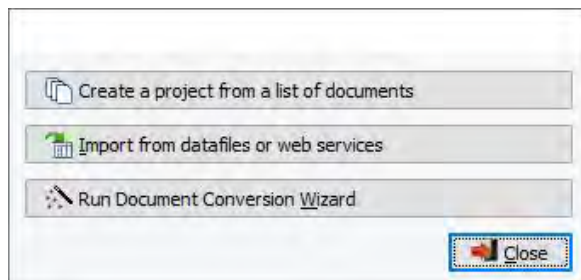
[Using the Document Conversion Wizard](#)

Creating a Project from a List of Documents

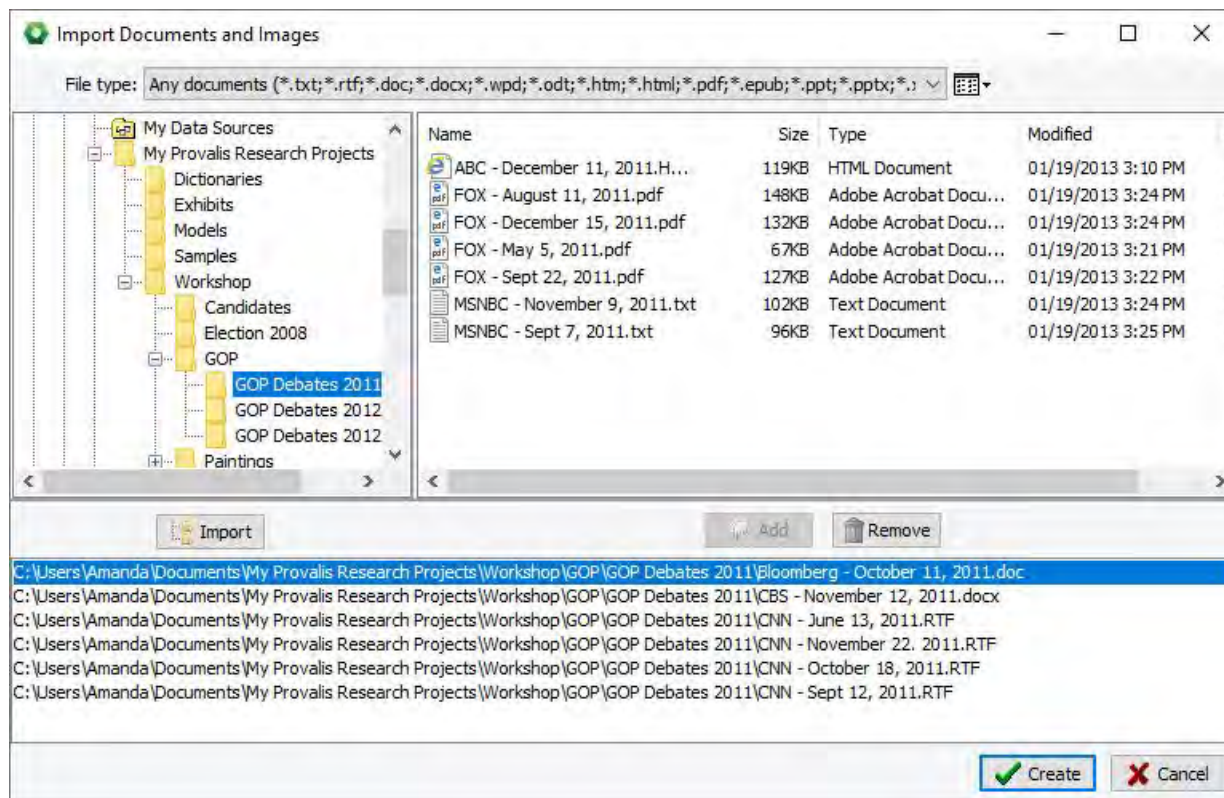
The easiest method to create a new project and start performing analysis in WordStat is by specifying a list of existing documents and importing them into a new project. This method creates a simple project with two-to-four variables: a categorical variable containing the original name of the files from which the data originated, a categorical variable containing the location of the files from which the data originated (optional), a variable containing the date and time that the files were created (optional), and a **DOCUMENT** variable containing imported documents. All text files are stored in different cases, therefore, if 10 files have been imported, the project will have 10 cases with two-to-four variables each. To split long documents into several documents or to extract numerical, categorical, or textual information from documents and store it in additional variables, use the [Document Conversion Wizard](#).

To create a new project from a list of documents:

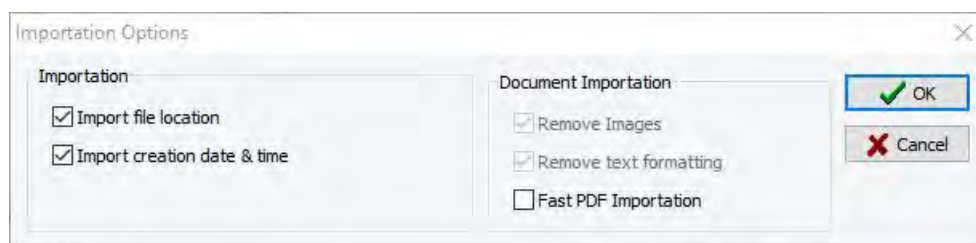
- Select the **NEW** command. A dialog similar to the one below will appear.



- Click the **Create a project from a list of documents** button. A dialog box similar to the one below will appear.



- In the upper right section of the dialog box, WordStat displays all supported document file formats that may be imported, such as MS Word, WordPerfect, RTF, PDF documents, plain text files or HTML etc. To display only files of a specific type, set the **File type** list box to the desired file format.
- Click the file you would like to import. To select multiple files, hold down the **Ctrl** key while clicking on the other files.
- Click the  **Add** button to add the selected files to the list of files to import, located at the bottom of the dialog box. You may also drag the files from the top right section to this list.
- Click the  **Import** button to add numerous files contained within a folder. A dialog box will appear giving you the option of including **subfolders** and choosing **file types** to include in the import.
- To remove a file from the list of files to import, select that file name and click the  **Remove** button.
- Once all files have been selected, click the **Create** button. A dialog box similar to the one below will appear.



You have the options to **Import file location** and **Import creation date & time**. If you select these options, this information will become variables entitled **LOCATION** and **CREATED**.

At this point your importation options differ depending on the documents you are importing. WordStat will only make available the importation options of the file type(s) that you have chosen. Other options will be grayed out.

Document Importation Options:

The **Remove Images** option has been set by default. When importing formatted documents like Word, HTML or PDF files, images stored in the document may significantly increase the resulting document size and slow down the browsing and text-processing speed of WordStat. To keep the images in the imported documents, disable this option.

Check the **Remove Text Formatting** checkbox to further reduce the size of the imported documents if the text formatting of the existing documents such as the font styles and colors or the paragraph formatting are not relevant. Enable this option to convert all documents into plain-text documents without any formatting or images.

The fast **PDF Importation option** removes both images and text formatting from PDF files and is considerably faster than choosing both **Remove Images** and **Remove text formatting** options at the same time.

- Set the importation options.
- Click the **OK** button.
- Choose a name for your project and save it in the appropriate location. The **Data tab** will appear, containing a table with your imported data. From here you can further tailor your data set if necessary, or start analyzing immediately.

Related Topics:

For more information on **importing PDFs**, see [Notes on Importing PDF Files](#).

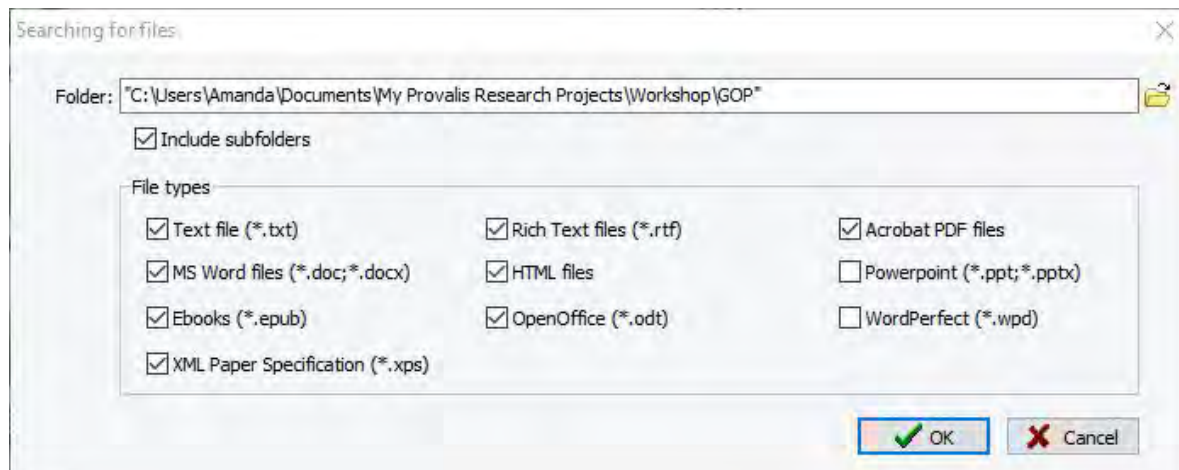
To configure the data set and start analyzing please see [The Data Tab](#).

Creating a WordStat Project from Windows Explorer

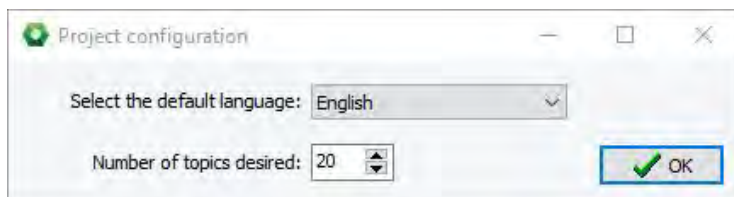
The easiest way to create a project from multiple documents or a folder containing documents is to run WordStat through Windows Explorer. This method allows you to create a project without going through the importation process in QDA Miner or WordStat.

To access WordStat from Windows Explorer:

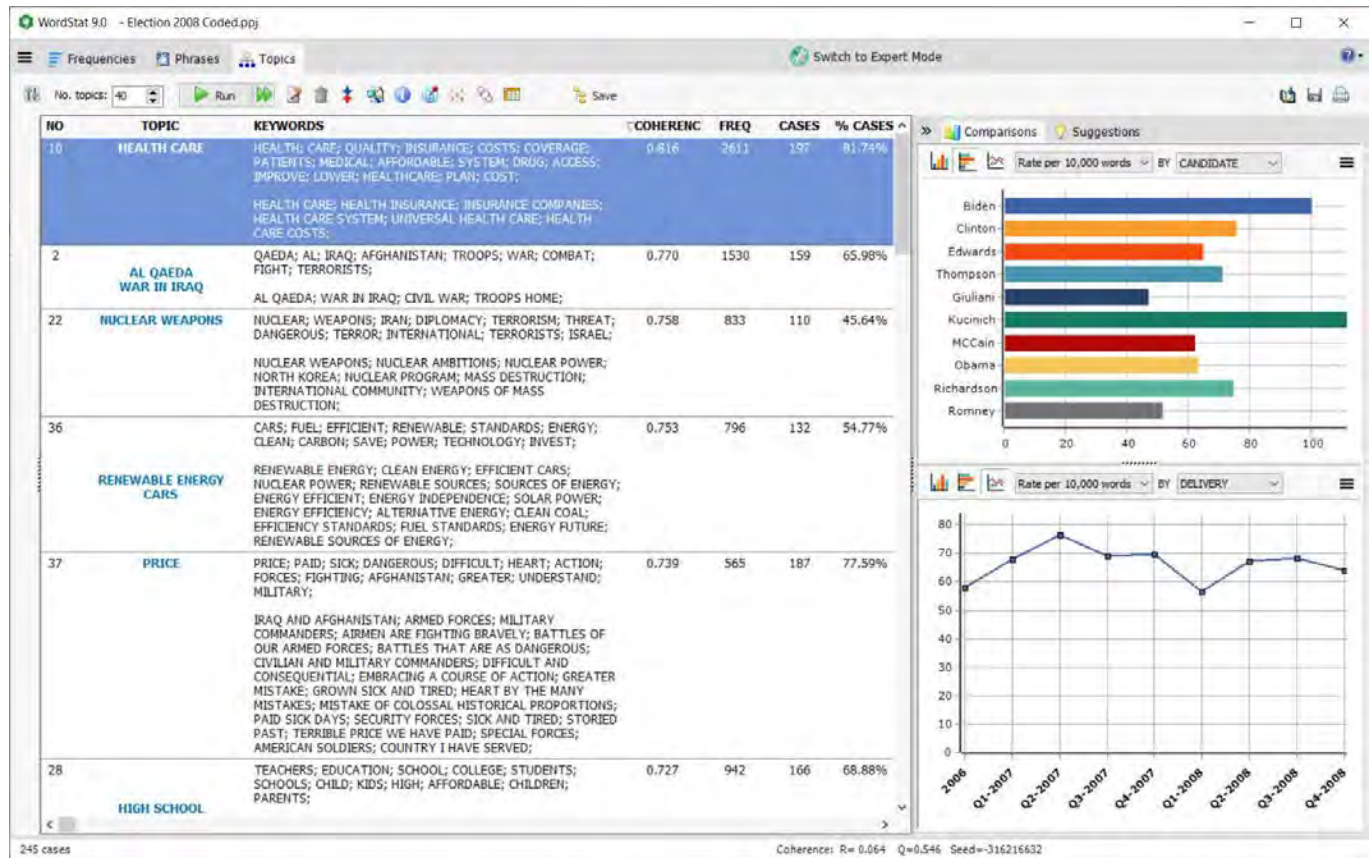
- Select the folder in Windows Explorer containing the subfolders and documents you wish to analyze.
- Right click your mouse, a menu opens, scroll down to **WORDSTAT CONTENT ANALYSIS** and choose **RUN WORDSTAT** from the adjacent menu. A dialog box will appear giving you the option of including **subfolders** and choosing **file types** to include in the import.



- Once all file types have been selected, click the **OK** button. A Project configuration dialog box opens. Rather than selecting a folder you can select the individual documents you would like to explore, using the **Shift** or **Ctrl** key to select numerous documents at the same time. Right click your mouse, a menu opens, scroll down to **WORDSTAT CONTENT ANALYSIS** and choose **RUN WORDSTAT** from the adjacent menu.



- **Select the default language** from the drop down menu.
- Enter the **Number of topics desired** for topic extraction by typing the number in the field or clicking on the up or down arrow beside the field until the desired number is reached. A temporary project opens containing three tabs: **Frequencies**, **Phrases** and **Topics**.



- Click the **≡** button and scroll down the **MODE**. Select **EXPERT** from the adjacent menu or click the **Switch To Export Mode** button. You will now have access to seven tabs and WordStat's full capabilities.
- Select the **X** at the top right of the screen to close the window. A dialog opens.
- You are asked if you would like to save the categorization model. If you would, select **Yes**, if not, select **No**.
- You are asked if you you want to save the data into a new project file. Select **Yes**. A Save As dialog opens.
- Choose a name for your project file and a place to save it. Select **Save**.

Notes on Importing PDF Files

The PDF file format is designed to display text in the same way on various platforms. Different strategies have been considered in QDA Miner and WordStat to support PDF files. They could simply be imported as is and stored in the QDA Miner project, allowing you to view the documents exactly as they appear in Acrobat. You could then code the PDF documents directly. Such an approach has, however, several inconveniences. First, the document may not be edited, preventing you from removing unnecessary information, such as headers and footers, appearing on each page or unrelated information like advertising. Also, PDF files are often created by outputting each page and each line of text separately, often resulting in the loss of the original document structure. Carriage returns will often be inserted at the end of each line, paragraphs spread over two pages remain physically split up, as do hyphenated words. In multi-column documents, lines from different columns may even be mixed up. All these minor imperfections may prevent you from retrieving full sentences and paragraphs in QDA Miner and may reduce recall results when searching for specific words or phrases separated by hyphens, carriage returns or page limitations. Such issues may also undermine the performance of numerous features of WordStat relying on phrase identification, word cooccurrence analysis or proximity rules.

Therefore, we chose to import PDF documents by converting them into editable documents, storing them in rich-text format, just like any other document file type supported by QDA Miner and WordStat. The conversion engine has been carefully designed to remove hyphens and unnecessary carriage returns, adjust the text flow in multi-column documents, and import tables and images correctly. While headers and footers will still be imported, which will break the flow of text across pages, you may now easily identify those and remove them from the text since the imported document will be fully editable.

Despite all the advanced importation features, some PDF documents may still not be imported properly. Several factors may explain such difficulties. Some PDF files consist of only scanned images of the original document and contain no text at all. These can be easily recognized by the fact that it is not possible to select any text segment or that text searches never return any hits. Other documents may have quite complex layout designs, making their proper importation quite difficult. For those documents, we recommend pre-processing them with full-fledged OCR (Optical Character Recognition) software like Abbyy's [FineReader](#) or Nuance's [OmniPage](#) or by some PDF to Word conversion utility tools like [PDF Transformer](#) by Abbyy. Although the latest version of Acrobat Professional does include some OCR features, its performance was deemed not good enough to preserve the document structure.

We also recommend removing images or scaling down the graphic resolution of images to the lowest setting because images will significantly increase the resulting document size and will slow down the browsing and text-processing speed of QDA Miner and WordStat.

Creating a Project from an Existing Data File or Web Service

The program can read data stored in the following file formats:

- MS Access
- MS Excel
- Comma Separated Values
- Tab Separated Values
- SPSS
- Stata v8 - 15
- Triple-S 1.2 and 2.0 XML file (an XML standard to exchange survey data)
- RIS files
- MBox and EML email file formats
- MS Outlook accounts and PST data files
- Reference Information System (RIS) data files
- ODBC
- EndNote

It can also import data from various Internet database and social media services such as:

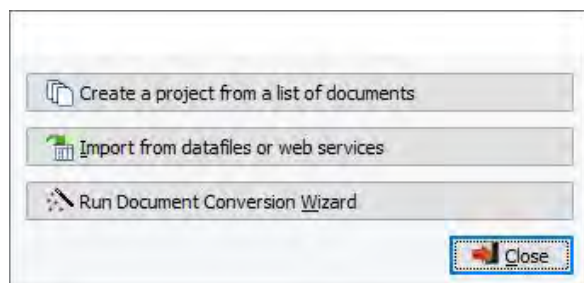
- SurveyMonkey
- Qualtrics
- QuestionPro
- SurveyGizmo
- Voxco
- EndNote
- Zotero
- Mendeley
- Reference Information System (RIS)
- Twitter
- RSS Feeds
- Lexis UNI (from LexisNexis)
- Factiva
- Facebook
- Reddit
- Gmail
- Hotmail
- Outlook

Importing from a Data File

WordStat allows you to directly import data files from spreadsheets and database applications, as well as from plain ASCII data files (comma or tab delimited text).

To import data from a data file:

- Select the **New** button. A dialog box similar to the one below will appear.



- Click the **Import from an existing data files or web service** button.
- Select the desired file format from the menu. An import dialog box will appear.
- Select the corresponding file format from the **Files of type** drop-down list and select the file you want to import.
- Click **Open**.
- Name your project and save it in the appropriate location.

We will now discuss the specific file formats in greater detail, including formatting requirements (if applicable), supported features and limitations.

Excel Data Files

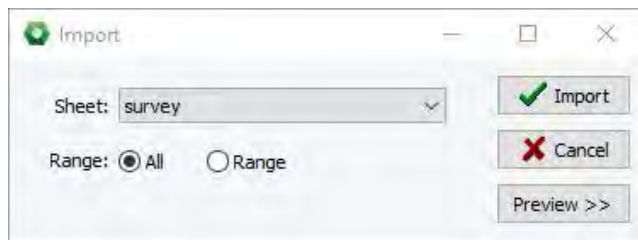
In an Excel spreadsheet you can enter both numeric and alphanumeric data in the cells of a data grid. WordStat can import spreadsheet files created by MS Excel *.xls and *.xlsx files.

Formatting Spreadsheet Data

The selected range must be formatted so that the columns of the spreadsheet represent variables (or fields), while the rows represent cases. As well, the first row should preferably contain the variable names, while the remaining rows should hold the data, one case per row. WordStat will automatically determine the most appropriate format based on the data that 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, WordStat will automatically provide one for each column in the defined range.

To create a new project from an Excel data file:

- Once you have named your project and saved it, a dialog similar to the one below appears.



When you select an Excel file format for importation, the program displays a dialog box in which you can specify the spreadsheet tab 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** radio button to **All**, the program attempts to read the whole tab. You can preview the data before you import by selecting the **Preview** button. Your Excel spreadsheet must be closed before you select **Import**. You can only import one tab at a time. If you have more than one tab to import, containing the same column headers, simply append the other tabs.

- In the **Sheet** drop-down choose the tab you would like to import.
- Select either a **Range** or **All** and set your range, if necessary.
- Select **Import**. An Import Options dialog appears similar to the one below.

Import Options

ROW IDENTIFICATION - STEP 1 OF 2

Does your worksheet contain:

☒ **Variable Names:** Start on row:

☐ **Variable Descriptions:** Start on row:

Data: Starts on row:

Preview

	ID	WHERE	GENDER	AGEGROUP	ETHNICITY	JOB	SKILLS
2	6	Los Angeles, CA	Male	25-39	caucasian	entertainment industry	None that I am aware of. LOL
3	8	Scotland	Male	19-24	Scottish	Student	Leading huge PvP groups and
4	10		Male	25-39	black	auto worker	n/a
5	11	USA	Male	14-18	American	Student	It helps me to think things
6	12	Moose Jaw, Canada	Male	25-39	Irish	Accountant	Jokes and COH humor cross
7	13	Birmingham, Alabama- USA	Male	19-24	Caucasian	College Student/ Restaurant	I've learned a little more
8	15	lakenheath, england	Male	25-39	caucasian	retail	typing skills if any
9	17	Bremerton, Washington USA	Male	40-54	American	Programmer	none, games do not effect
10	18	Amherst, Massachusetts,	Male	25-39	Caucasian	Molecular biologist	At most, the game has taken
11	19	Arkansas	Male	25-39	Black	Govt	Humor. Im a real serious
12	24	Sacramento, CA.	Male	25-39	Caucasian	State Worker	I really don't think much I've

This dialog is a two-step process that helps you properly configure your data for import. On the first page you are asked to identify the location of the variable names and variable descriptions, if present. You are also asked to identify the row on which the data starts.

- If your spreadsheet contains variable names, check the checkbox called **Variable Names** and enter the row number in which they are located.
- If your spreadsheet contains variable descriptions, check the checkbox called **Variable Descriptions** and enter the row number in which they are located.
- In the **Data** field enter the row number on which the data starts.
- Select **Next**. The second page appears.

Import Options

SELECTION OF VARIABLES & SETTINGS - STEP 2 OF 2

Select the variables that you would like to import. You can customize the name and description if necessary by typing in the associated cell. Change the variable type by selecting from the dropdown menu.

Variables

Selected	Name	Description	Type
☒	ID	ID	Integer
☒	WHERE	WHERE	Short String
☒	GENDER	GENDER	Nominal
☒	AGEGROUP	AGEGROUP	Nominal
☒	ETHNICITY	ETHNICITY	Document
☒	JOB	JOB	Short String
☒	SKILLS	SKILLS	Document

The second page allows you to choose the variables you would like to import. You can change the names and descriptions by typing directly in the cells. To change the data type select from the drop-down list.

- Select the variables you want to import by checking their checkboxes.
- Modify the variable names, descriptions and data types as necessary.
- Select **Import**. The data will be imported and WordStat's **Data** tab will appear containing a table with your imported Excel data. From here you can further tailor your data set if necessary, or start analyzing immediately.

MS Access Database Files

When you select a MS Access data file, the program displays a dialog box in which you can specify the table where the data are located. Once the table has been selected, click the **Import** button to import the data file.

ASCII Data Files

WordStat will read up to 2000 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 spaces, tabs and/or commas.
- Variable names may not be longer than ten characters. Longer strings are truncated at ten characters.
- The remaining lines must include alphanumeric or numeric scores separated by spaces, tabs and/or commas.
- Each line must contain data for one case; variables must be in the same order for all cases.
- All invalid scores and all blanks encountered between commas or tabs are treated as missing values. A single dot can also be used to represent a missing value.

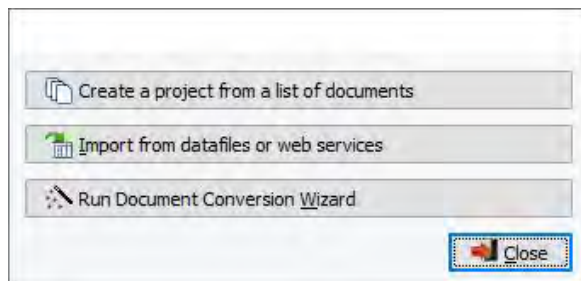
- Comments can be inserted anywhere in the file by putting an asterisk ('*') at the beginning of the line.
- Blank lines can also be inserted anywhere in the file.

Importing from Web Surveys

WordStat allows you to import survey data directly from various web survey platforms, enabling you to analyze open-ended responses, using the text mining and content analysis features of WordStat. Importing open-ended and closed-ended questions permits you to quickly compare results by demographics and find out how they are related to ratings, or any other numeric, categorical or date variables. WordStat supports SurveyMonkey, Qualtrics, SurveyGizmo, Voxco, and QuestionPro survey platforms. It also supports TripleS v1.2 survey files. Importing data from the various survey platforms is generally done in a similar fashion. You will be asked to log in to your account or provide a token before importing survey data. QuestionPro, Qualtrics and Voxco survey platforms require tokens to import data into WordStat.

Creating a New Project Using Survey Data

- Select the **New** button on the **Data** tab. This command calls up a dialog box similar to the one below.



- Select the **Import from data files or web services** button. A menu will appear on the right-hand side of the dialog box.
- Scroll down to the **SURVEY** menu item and choose the survey platform from which you wish to import the data.

For instruction on how to import survey data from the various survey platforms, please see:

Importing from [SurveyMonkey](#)

Importing from [Qualtrics](#)

Importing from [SurveyGizmo](#)

Importing from [Voxco](#)

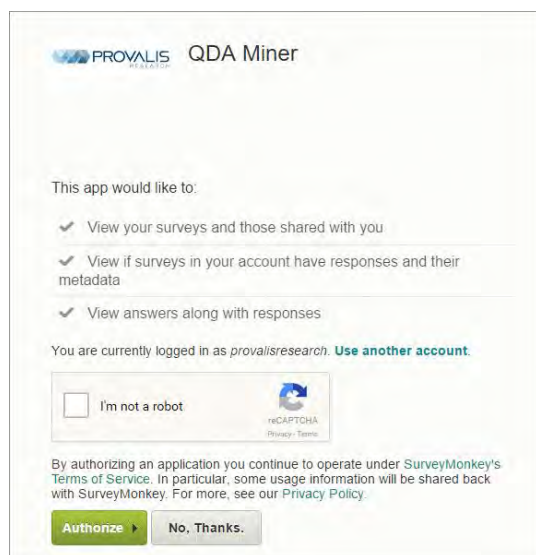
Importing from [QuestionPro](#)

Importing from a [TripleS v1.2](#) file

SurveyMonkey

Connecting to SurveyMonkey

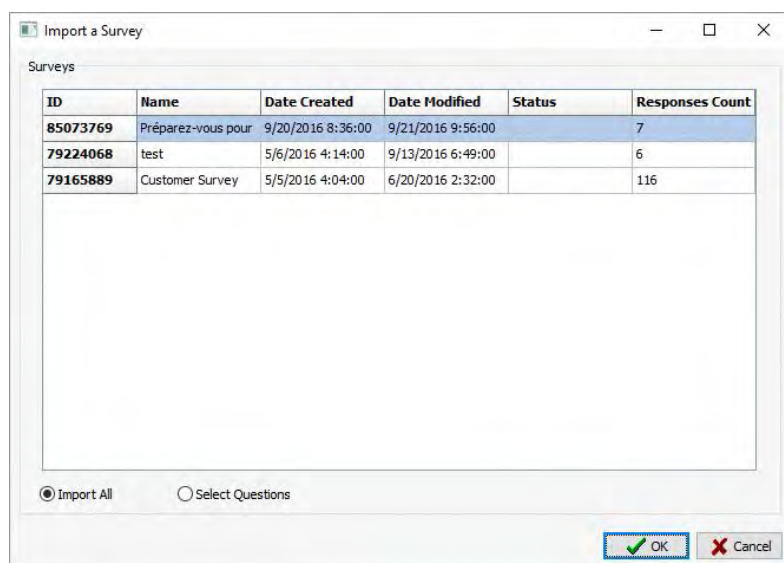
- Select **SURVEY | SURVEYMONKEY** from the menu. You will be prompted to log in to your SurveyMonkey account.
- Log in to your account. A dialog box requesting authorization for WordStat to access your account will appear.



- Select the **I'm not a robot** checkbox and perform the requested task.
- Click **Authorize**.

Importing SurveyMonkey Survey Data

Once a connection has been established an **Import a Survey** dialog box will appear.



The dialog box lists the **ID**, **Name**, **Date Created**, **Date Modified**, **Status** and the **Response Count** of each survey on the platform.

- Choose the survey that you wish to analyze. You have two options. You can import all questions from the survey or you can select specific survey questions that are relevant for your analysis.

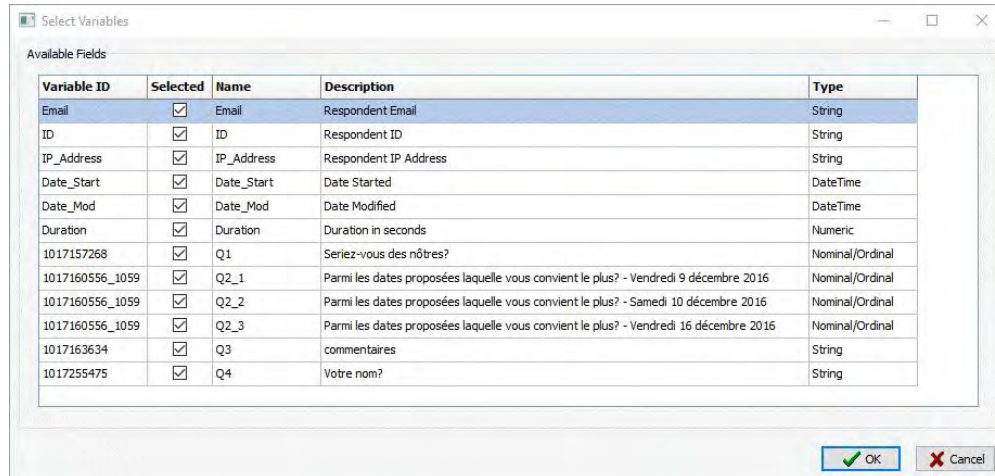
To import all survey questions:

- Select the **Import All** radio button at the bottom of the dialog box.
- Click **OK**.

- Choose a name for your project and save it in the appropriate location. The data will be imported and WordStat's **Data** tab will appear, containing a table with your imported survey data. From here you can further tailor your data set if necessary, or start analyzing immediately.

To import only specific questions:

- Select the **Select Questions** radio button at the bottom of the dialog box.
- Click **OK**. A dialog box like the one below will appear.



- The **Variable IDs** will become your project variables. Select the checkboxes of the variables you wish to import.
- Click **OK**.
- Choose a name for your project and save it in the appropriate location. The data will be imported and WordStat's **Data** tab will appear, containing a table with your imported survey data. From here you can further tailor your data set if necessary, or start analyzing immediately.

Related Topics:

To configure the data set and start analyzing please see [The Data Tab](#)

To append new survey responses to your project please [Monitoring Online Resources](#).

Qualtrics

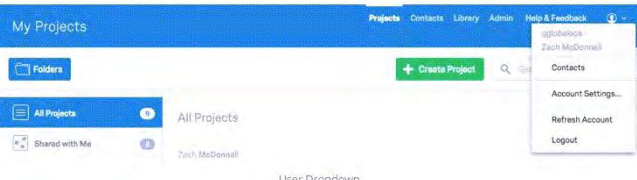
Connecting to Qualtrics

You must enter a token to link with your Qualtrics survey. To get your Qualtrics token:

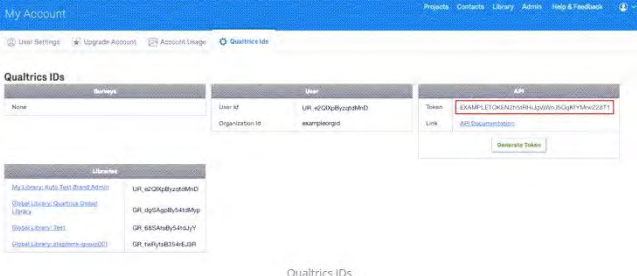
- Log in to your Qualtrics account.
- Go to the **Account Settings** menu item in the **User** drop-down menu.
- Go to the **Qualtrics IDs**.
- Click **Generate** to receive a token.

How to Find Your Token

1. Login to Qualtrics
2. Go to Account Settings in the user dropdown



3. Go to Qualtrics IDs

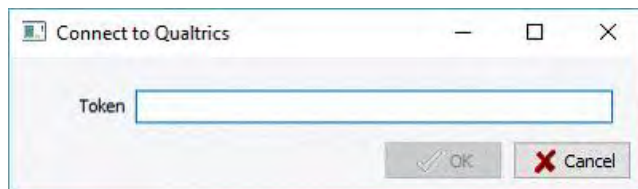


4. Click Generate if you haven't generated your token yet

Generate Token

If you already have an API token, "Generate Token" will replace it with a new one. Any existing API calls will not work until they are updated to use the new token.

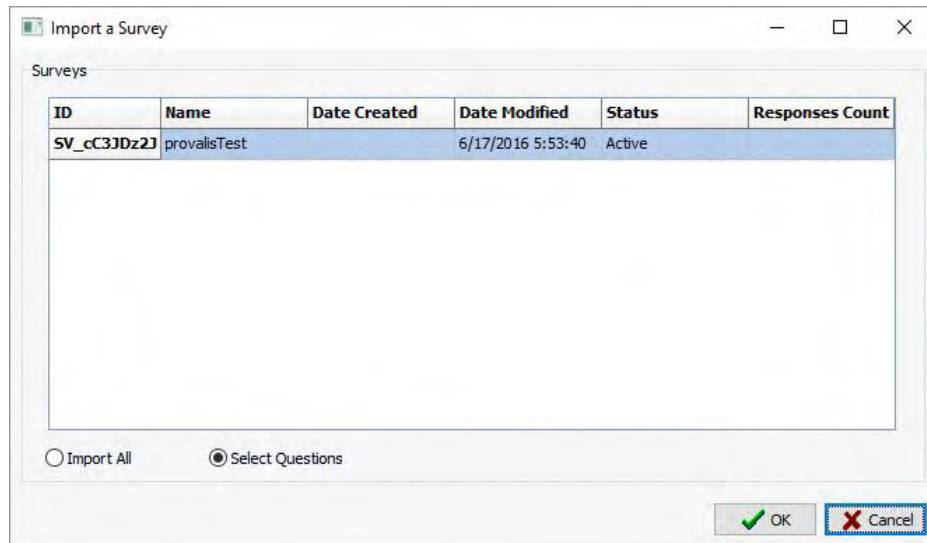
- In WordStat, select **SURVEY | QUALTRICS** from the menu.
- Enter your token in the **Connect to Qualtrics** dialog box.



- Click **OK**.

Importing Qualtrics Survey Data

Once a connection has been established an **Import a Survey** dialog box will appear.



The dialog box lists the **ID**, **Name**, **Date Created**, **Date Modified**, **Status** and the **Response Count** of each survey on the platform.

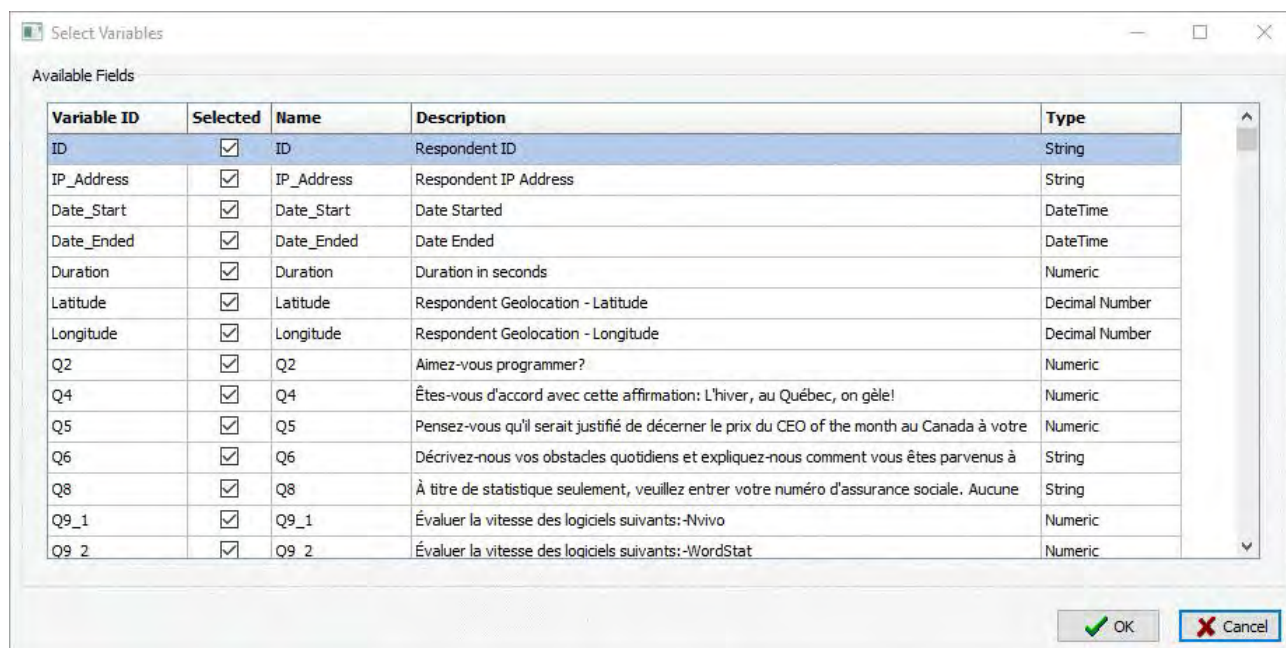
- Choose the survey that you wish to analyze. You have two options. You can import all questions from the survey or you can select specific survey questions that are relevant for your analysis.

To import all survey questions:

- Select the **Import All** radio button at the bottom of the dialog box.
- Click **OK**.
- Choose a name for your project and save it in the appropriate location. The data will be imported and WordStat's **Data** tab will appear, containing a table with your imported survey data. From here you can further tailor your data set if necessary, or start analyzing immediately.

To import only specific questions:

- Select the **Select Questions** radio button at the bottom of the dialog box.
- Click **OK**. A dialog box like the one below will appear.



- The **Variable IDs** will become your project variables. Select the checkboxes of the variables you wish to import.
- Click **OK**.
- Choose a name for your project and save it in the appropriate location. The data will be imported and WordStat's **Data** tab will appear, containing a table with your imported survey data. From here you can further tailor your data set if necessary, or start analyzing immediately.

Related Topics:

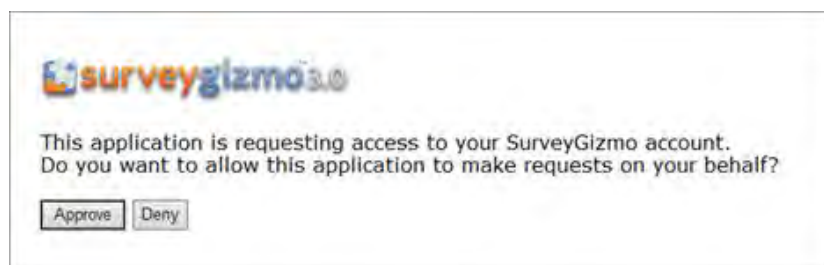
To configure the data set and start analyzing please see [The Data Tab](#)

To append new survey responses to your project please [Monitoring Online Resources](#).

SurveyGizmo

Connecting to SurveyGizmo

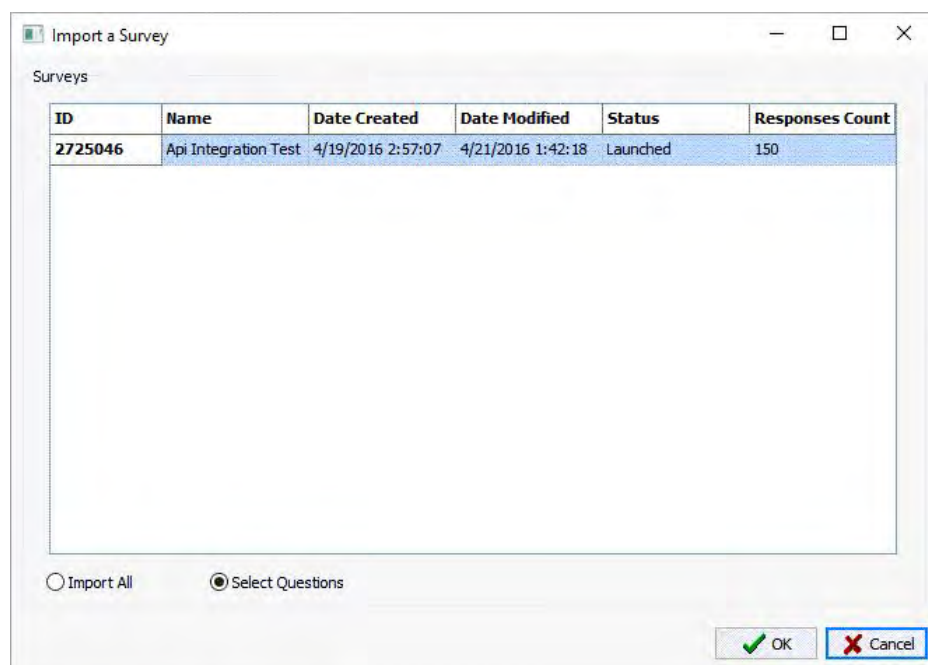
- Select **SURVEY | SURVEYGIZMO** from the menu. You will be prompted to log in to your SurveyGizmo 3.0 account.
- Log into your account. A dialog box requesting authorization for WordStat to access your account will appear.



- Click **Approve**.

Importing SurveyGizmo Survey Data

Once a connection has been established **Import a Survey** dialog box will appear.



The dialog box lists the **ID**, **Name**, **Date Created**, **Date Modified**, **Status** and the **Response Count** of each survey on the platform.

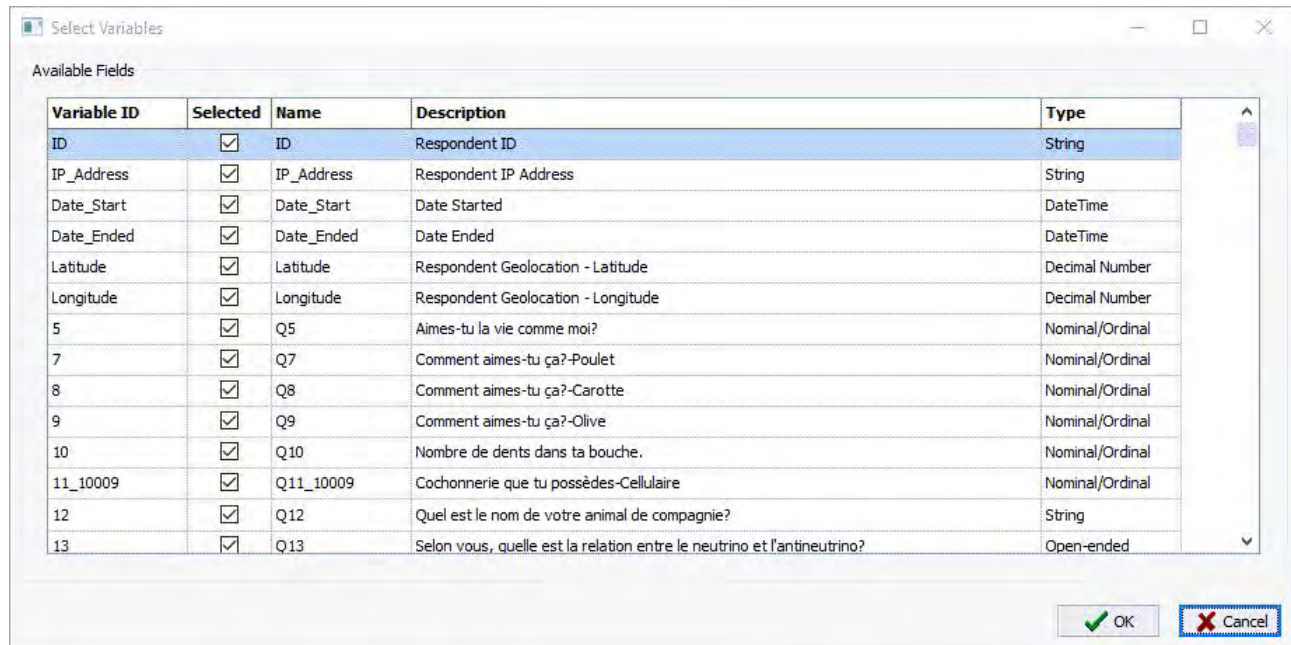
- Choose the survey that you wish to analyze. You have two options. You can import all questions from the survey or you can select specific survey questions that are relevant for your analysis.

To import all survey questions:

- Select the **Import All** radio button at the bottom of the dialog box.
- Click **OK**.
- Choose a name for your project and save it in the appropriate location. The data will be imported and WordStat's **Data** tab will appear, containing a table with your imported survey data. From here you can further tailor your data set if necessary, or start analyzing immediately.

To import only specific questions:

- Select the **Select Questions** radio button at the bottom of the dialog box.
- Click **OK**. A dialog box like the one below will appear.



- The **Variable IDs** will become your project variables. Select the checkboxes of the variables you wish to import.
- Click **OK**.
- Choose a name for your project and save it in the appropriate location. The data will be imported and WordStat's **Data** tab will appear, containing a table with your imported survey data. From here you can further tailor your data set if necessary, or start analyzing immediately.

Related Topics:

To configure the data set and start analyzing please see [The Data Tab](#)

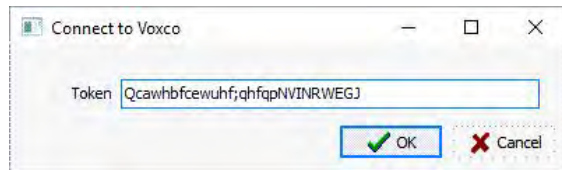
To append new survey responses to your project please [Monitoring Online Resources](#).

Voxco

Connecting to Voxco

You must enter a token to link to your Voxco survey. To get your Voxco token:

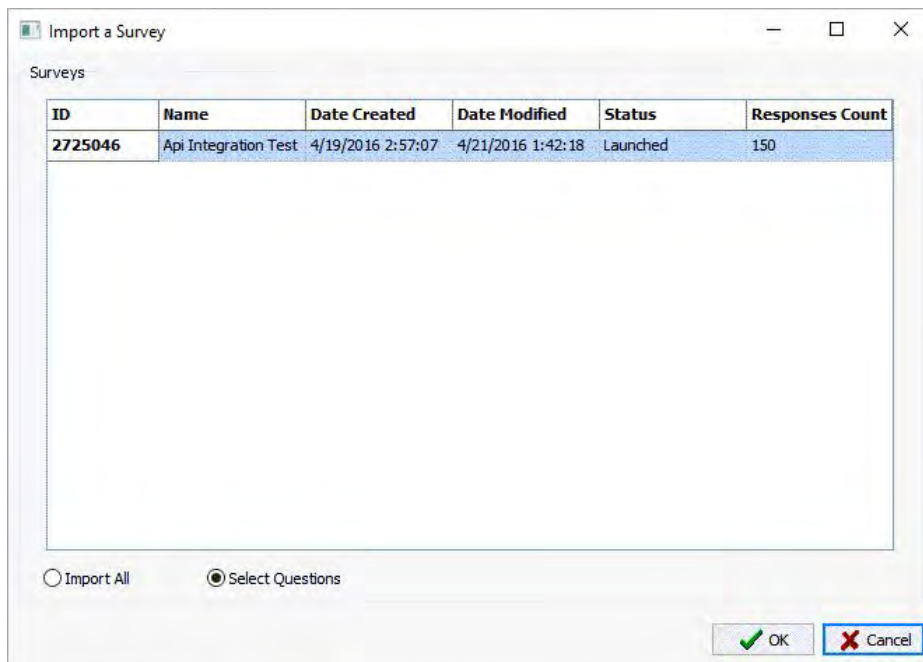
- Log in to your account on Voxco's A4s platform.
- Select **User Settings** from the **User** drop-down menu.
- On the **User Settings** page you will find the **API Access Key** in the list of **Security** options. Copy this token.
- In WordStat, select **SURVEY | VOXCO** from the menu.
- Enter your token into the **Connect to Voxco** dialog box.



- Click **OK**.

Importing Voxco Survey Data

Once a connection has been established **Import a Survey** dialog box will appear.



The dialog box lists the **ID**, **Name**, **Date Created**, **Date Modified**, **Status** and the **Response Count** of each survey on the platform.

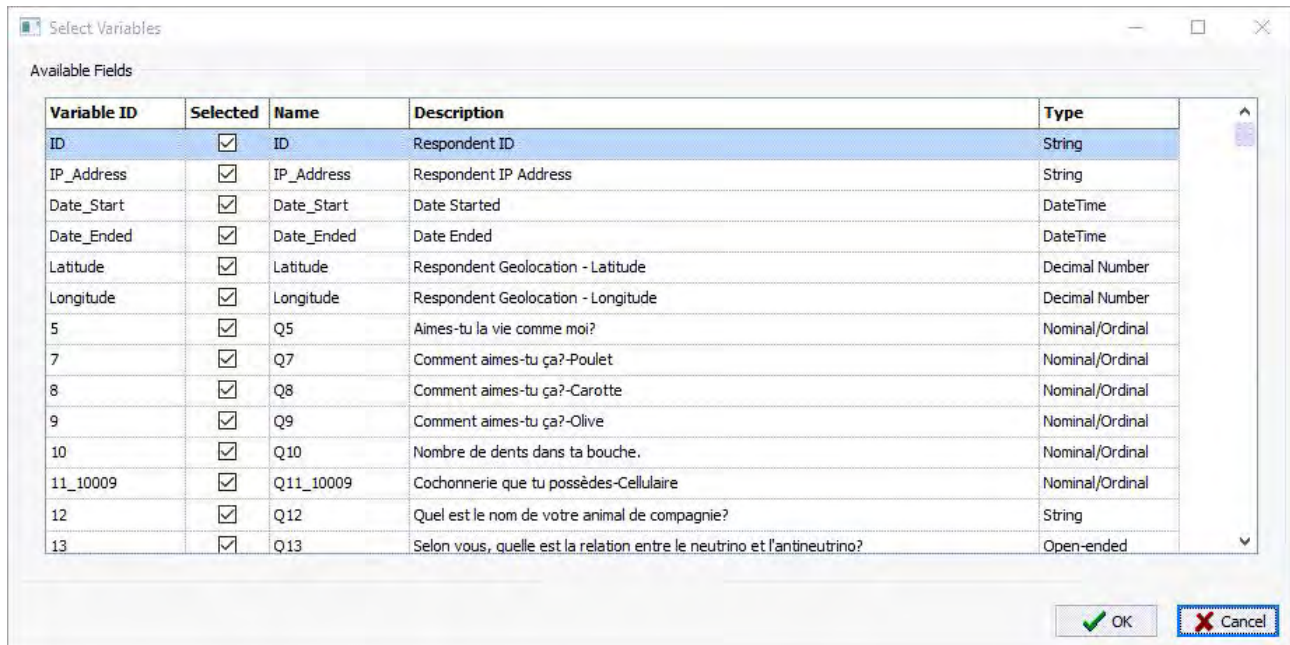
- Choose the survey that you wish to analyze. You have two options. You can import all questions from the survey or you can select specific survey questions that are relevant for your analysis.

To import all survey questions:

- Select the **Import All** radio button at the bottom of the dialog box.
- Click **OK**.
- Choose a name for your project and save it in the appropriate location. The data will be imported and WordStat's **Data** tab will appear, containing a table with your imported survey data. From here you can further tailor your data set if necessary, or start analyzing immediately.

To import only specific questions:

- Select the **Select Questions** radio button at the bottom of the dialog box.
- Click **OK**. A dialog box like the one below will appear.



- The **Variable IDs** will become your project variables. Select the checkboxes of the variables you wish to import.
- Click **OK**.
- Choose a name for your project and save it in the appropriate location. The data will be imported and WordStat's **Data** tab will appear, containing a table with your imported survey data. From here you can further tailor your data set if necessary, or start analyzing immediately.

Related Topics:

To configure the data set and start analyzing please see [The Data Tab](#)

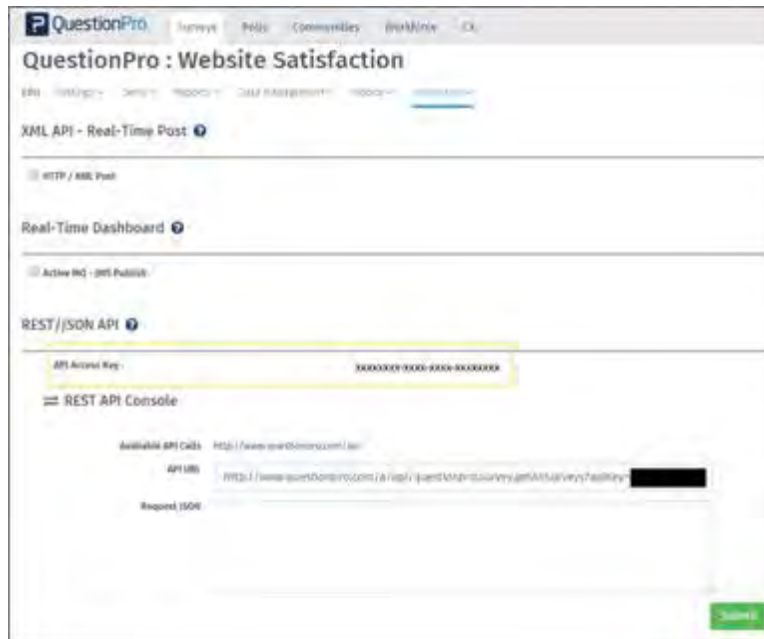
To append new survey responses to your project please [Monitoring Online Resources](#).

QuestionPro

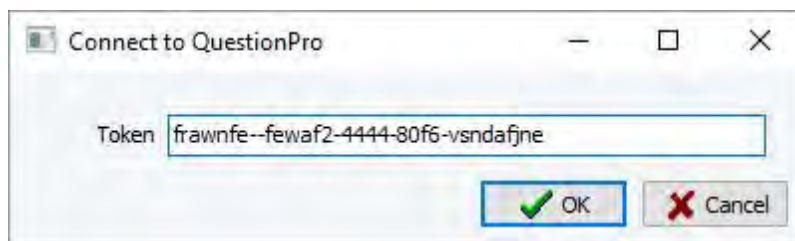
Connecting to QuestionPro

You must enter a token to link to your QuestionPro survey. To get your QuestionPro token:

- Log in to your QuestionPro account.
- Select **Surveys** from the navigation bar at the top of the screen.
- Select the survey you wish to import.
- Click on **Developer API** in the **Integration** drop-down menu. The **API Access Key** is your token.



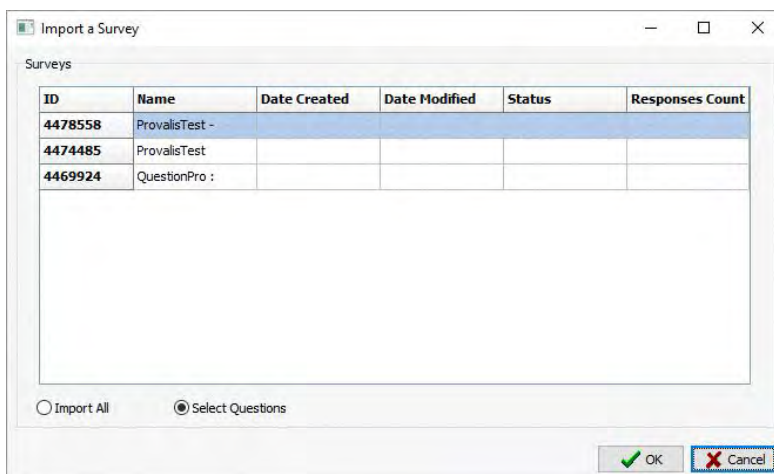
- In WordStat, select **SURVEY | QUESTIONPRO** from the menu.
- Copy and paste this **key** into the **Connect to QuestionPro** dialog box.



- Click **OK**.

Importing QuestionPro Survey Data

Once a connection has been established **Import a Survey** dialog box will appear.



The dialog box lists the **ID**, **Name**, **Date Created**, **Date Modified**, **Status** and the **Response Count** of each survey on the platform.

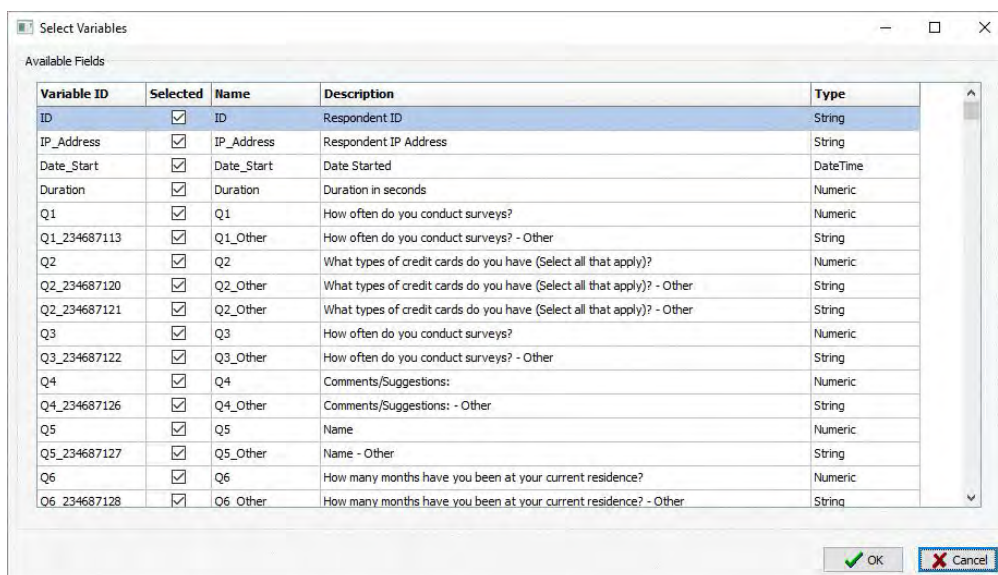
- Choose the survey that you wish to analyze. You have two options. You can import all questions from the survey or you can select specific survey questions that are relevant for your analysis.

To import all survey questions:

- Select the **Import All** radio button at the bottom of the dialog box.
- Click **OK**.
- Choose a name for your project and save it in the appropriate location. The data will be imported and WordStat's **Data** tab will appear, containing a table with your imported survey data. From here you can further tailor your data set if necessary, or start analyzing immediately.

To import only specific questions:

- Select the **Select Questions** radio button at the bottom of the dialog box.
- Click **OK**. A dialog box like the one below will appear.



- The **Variable IDs** will become your project variables. Select the checkboxes of the variables you wish to import.
- Click **OK** and the selected questions will be imported into your project.
- Choose a name for your project and save it in the appropriate location. The data will be imported and WordStat's **Data** tab will appear, containing a table with your imported survey data. From here you can further tailor your data set if necessary, or start analyzing immediately.

Related Topics:

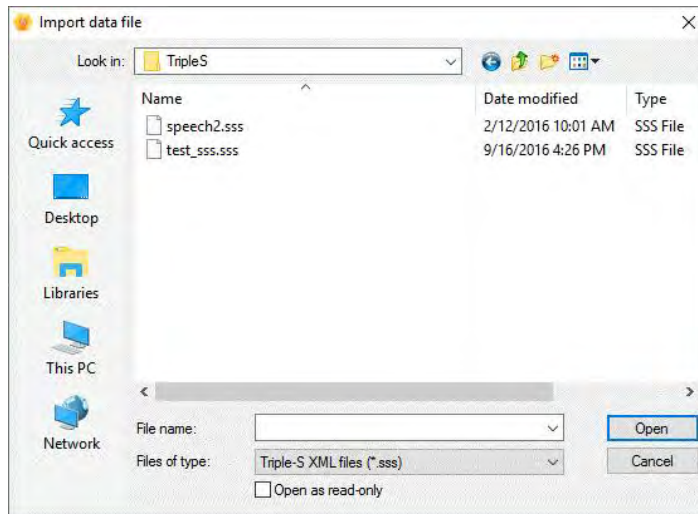
To configure the data set and start analyzing please see [The Data Tab](#)

To append new survey responses to your project please [Monitoring Online Resources](#).

TripleS v.1.2

Importing TripleS v1.2 files

- Select **SURVEY | TRIPLES V1.2** from the menu. An **Import data file** dialog box will appear.



- Choose the triple-s file that you would like to import.
- Click **Open**.
- Choose a name for your project and save it in the appropriate location. The data will be imported and WordStat's **Data** tab will appear, containing a table with your imported survey data. From here you can further tailor your data set if necessary, or start analyzing immediately.

Related Topics:

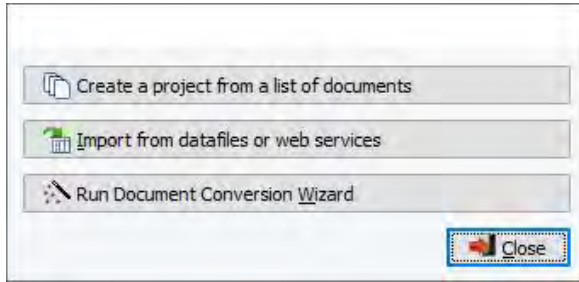
To configure the data set and start analyzing please see [The Data Tab](#)

Importing from Social Media

Social media is a popular networking and marketing medium. Many companies, organizations, public figures etc. have a social media presence. WordStat allows you to import social media data directly from various social media platforms, allowing you to use the text mining and content analysis features of WordStat to analyze and monitor social media data over time. WordStat supports RSS, Twitter, Facebook pages and Reddit. You must have access to a Twitter account to use the Twitter importation feature.

Creating a New Project Using Social Media Data

- Select the **New** button on the **Data** tab. This command calls up a dialog box similar to the one below.



- Select the **Import from data files or web services** button. A menu will appear on the right-hand side of the dialog box.
- Scroll down to the **SOCIAL MEDIA** menu item and choose the social media platform from which you wish to import the data.

For instruction on how to import survey data from the various survey platforms, please see:

[Importing from RSS](#)

[Importing from Twitter](#)

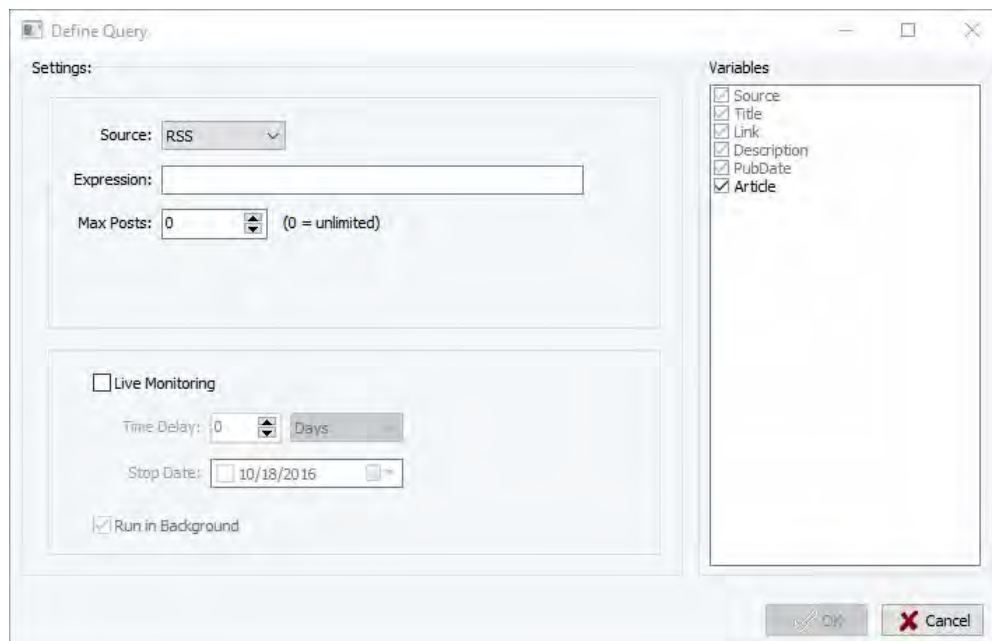
[Importing from Facebook pages](#)

Importing from RSS

WordStat allows you to create projects by importing RSS feeds and use the text mining and content analysis features of WordStat to analyze and monitor your RSS sources over time.

Creating a New Project Using a RSS Feed

- Select **SOCIAL MEDIA | RSS FEED** from the menu. A **Define Query** dialog box will appear.



- Select **RSS** as the **Source** of your data import.
- Copy the URL of the RSS feed of your choice into the **Expression** field.

- The **Max Posts** field will be set to 0 by default, which will give you an unlimited number of results. You can limit the amount of RSS data collected by changing the number in the **Max Posts** field.
- The **Source**, **Title**, **Link Description** and **PubDate** variables in the **Variables** list box are grayed out as they are imported into the project by default. Check the **Article** checkbox from the **Variables** list box if you would like to import the full text of the article in a document variable.
- The **Live Monitoring** option instructs WordStat to monitor the RSS feed over time and collect for import RSS posts published after the initial data capture. Enabling this option calls the Web Collector, an external tool, which collects your RSS data. You can access the Web Collector through the system tray and the Windows start menu. When you restart your computer the Web Collector will start automatically and continue collecting data if the query status is active.
- Set the **Time Delay** frequency and the **Stop Date**. The Web Collector will collect new RSS posts published according to the selected time interval until the stop date is reached. Data will be collected as long as the Web Collector is running. As the Web Collector is a completely separate application WordStat does not have to be running for it to work.
- Once all the options have been set, click **OK**.
- Enter the project name and save it in the location of your choosing. The data will be imported and WordStat's **Data** tab will appear, containing a table with your imported RSS data. From here you can further tailor your data set if necessary, or start analyzing immediately.

Related Topics:

To configure the data set and start analyzing please see [The Data Tab](#).

To append RSS data to your project please [Monitoring Online Resources](#).

For more information on the Web Collector please see the section entitled [Web Collector](#).

Importing from Twitter

WordStat allows you to create a project by importing Twitter data and use the text mining and content analysis features of WordStat to analyze and monitor your Twitter sources over time. You must have access to a Twitter account to use this feature.

Creating a New Project Using Twitter Data

- Select **SOCIAL MEDIA | TWITTER** from the menu. A **Define Query** dialog box will appear.

- Select **Twitter** as the **Source** of your data import.
- Enter your Twitter query in the **Expression** field. You can enter a simple Boolean query directly into this field or you can click the  button to the right of the field and use the advanced search option to help you create your query. The **Build your expression** dialog box is an advanced search option. It mirrors the advanced search in Twitter and works on the same principles.
- The **Max Posts** field will be set to 0 by default, which will give you an unlimited number of results. You can limit the amount of Twitter data collected by changing the number in the **Max Posts** field.
- Check the corresponding boxes to **Include Retweets** and **Geocode locations** in the import if you desire. **Including Retweets** may result duplicate cases. The number of duplicates indicates the virality of the Tweet. Importing **Geocode Locations** will result in the addition of three variables: Longitude, Latitude and User Location. User Location is the user defined. Including these variables in the import allows you to use the **GEOCODING** function to obtain the geographic locations of the users and map them using WordStat's mapping feature.
- The **Date Created**, **Tweet ID** and **Tweet Text** variables in the **Variables** list box are grayed out as they are imported into the project by default. Check the variables that you would like to import from the **Variables** list box.
- The **Live Monitoring** option instructs WordStat to monitor the Twitter search over time and collect for import Tweets published after the initial data capture. Enabling this option calls the Web Collector, an external tool, which collects your Twitter data. You can access the Web Collector through the system tray and the Windows start menu. When you restart your computer the Web Collector will start automatically and continue collecting data if the query status is active.
- Set the **Time Delay** frequency and the **Stop Date**. The Web Collector will collect new Tweets published according to the selected time interval until the stop date is reached. Data will be collected as long as the Web Collector is running. As the Web Collector is a completely separate application WordStat does not have to be running for it to work.
- Click **OK**. If it is the first time you are creating a WordStat project from Twitter data, you will need to authorize the Provalis **Web Collector** to connect to your Twitter account. A dialog box like below will appear.



- Enter your log in details and click **Authorize app**.
- Enter the project name and save it in the location of your choosing. The data will be imported and WordStat's **Data** tab will appear, containing a table with your imported Twitter data. From here you can further tailor your data set if necessary, or start analyzing immediately.

Note: Twitter will automatically go back seven days in its database to retrieve information. The maximum number of Tweets accessible each 15 minutes by a single Twitter account is 18,000. If you have reached your limit, the Web Collector will stop and display a dialog box to inform you that you have reached your limit.

Related Topics:

To configure the data set and start analyzing please see [The Data Tab](#).

To append Twitter data to your project please [Monitoring Online Resources](#).

For more information on the Web Collector please see the section entitled [Web Collector](#).

Importing from Facebook

WordStat allows you to create a project by importing Facebook page data and use the text mining and content analysis features of WordStat to analyze and monitor your Facebook page sources over time. WordStat cannot access data from password protected profiles, only from public pages.

Creating a New Project Using Facebook Page Data

- Select **SOCIAL MEDIA | FACEBOOK** from the menu. A **Define Query** dialog box will appear.

- Select **Facebook** as the **Source** of your data import.
- Enter the URL of the Facebook page you would like to analyze.
- Set the **From** and **Until** date options by clicking the calendar to the right of the fields and selecting start and end date. WordStat will retrieve posts that were posted during the chosen time span. This time span must not exceed 6 months. The last possible day of the date range is the present day. **It is not possible to set this option to capture future posts.** Comments associated with the posts will also be imported. Imported comments, however, may have been published after the chosen date range. It is possible to capture future comments by setting the **Live Monitoring** option.
- Check the variables that you would like to import from the **Variables** list box.
- The **Live Monitoring** option instructs WordStat to monitor the Facebook posts published during the chosen time period and collect for import any comments associated with the posts that have been published after the initial data capture. Enabling this option calls the Web Collector, an external tool, which collects your Facebook data. You can access the Web Collector through the system tray and the Windows start menu. When you restart your computer the Web Collector will start automatically and continue collecting data if the query status is active. It is important to note that the Web Collector will not collect any new posts, only new comments associated with the posts from the chosen date range.
- Set the **Time Delay** frequency and the **Stop Date**. The Web Collector will collect comments published according to the selected time interval until the stop date is reached. Data will be collected as long as the Web Collector is running. As the Web Collector is a completely separate application WordStat does not have to be running for it to work.
- Once all the options have been set, click **OK**.
- Enter the project name and save it in the location of your choosing. The data will be imported and WordStat's **Data** tab will appear, containing a table with your imported Facebook data. From here you can further tailor your data set if necessary, or start analyzing immediately.

Related Topics:

To configure the data set and start analyzing please see [The Data Tab](#).

To append Facebook data to your project please [Monitoring Online Resources](#).

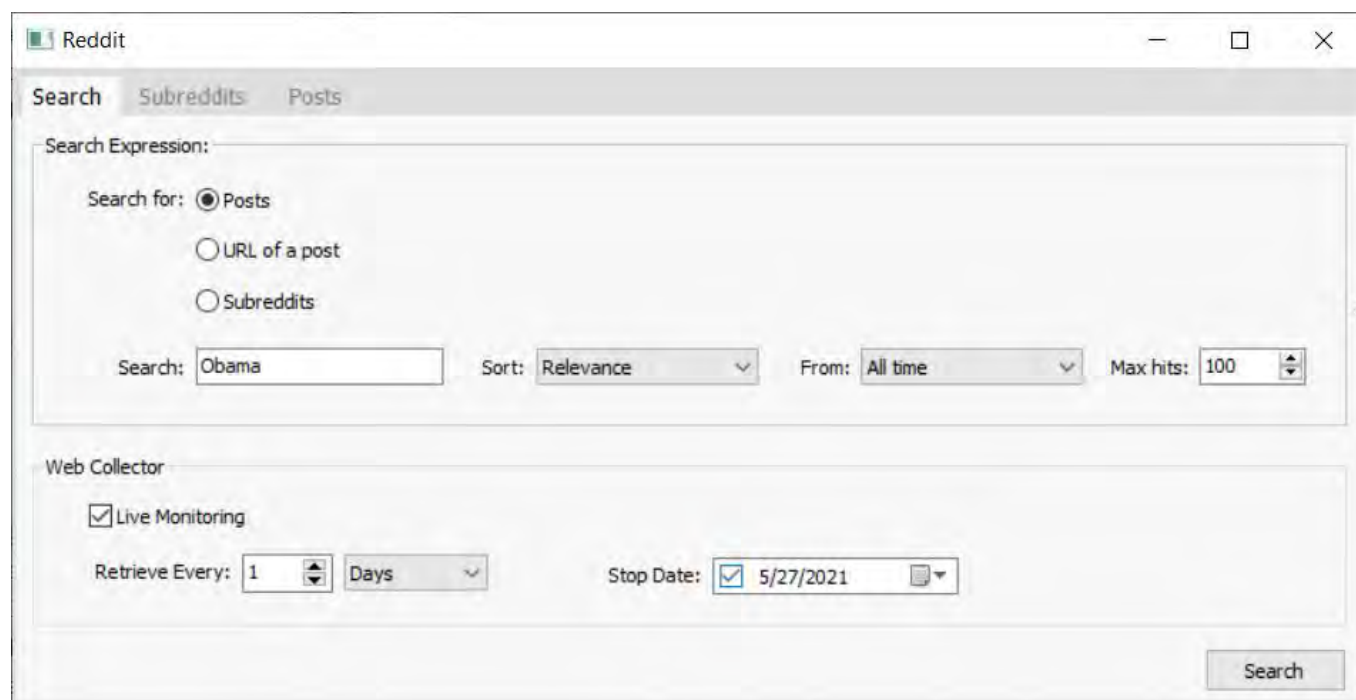
For more information on the Web Collector please see the section entitled [Web Collector](#).

Importing from Reddit

WordStat allows you to create projects by importing Reddit data and use the text mining and content analysis features of WordStat to analyze and monitor your Reddit sources over time.

Creating a New Project Using a Reddit

- Select **SOCIAL MEDIA | REDDIT** from the menu. A **Define Query** dialog box will appear.



- Select what you are searching for **Posts**, **URL of a post** or **Subreddits**.
- If you are looking for **Posts** or **Subreddit** type your query into the **Search for** field. If you have the **URL** of a Reddit that you would like to import type that in the **Search for** field.
- Choose the the sorting order of your imported data by selecting from the drop down in the **Sort** field. You can sort by **Relevance**, **Hot**, **Top**, **New** and **Comments**.
- Choose the start date of your import by selecting from the drop down in the **From** field.
- Set the **Max hits** field to determine the maximum number of posts for the project.
- The **Live Monitoring** option instructs WordStat to monitor the Reddit search over time and collect for import Reddit posts published after the initial data capture. Enabling this option calls the Web Collector, an external tool, which collects your Reddit data. You can access the Web Collector through the system tray and the Windows start menu. When you restart your computer the Web Collector will start automatically and continue collecting data if the query status is active.
- Set the time delay frequency in the **Retrieve Every** field and the **Stop Date**. The Web Collector will collect new Reddit posts published according to the selected time interval until the stop date is reached. Data will be collected as

long as the Web Collector is running. As the Web Collector is a completely separate application WordStat does not have to be running for it to work.

- Once all the options have been set, click **Search**. If it is the first time you are creating a WordStat project from Reddit data, you will need to authorize the Provalis **Web Collector** to connect to your Reddit account. A dialog box like below will appear.

- Enter your log in details and click **Authorize app**.
- Enter the project name and save it in the location of your choosing. The data will be imported and WordStat's **Data** tab will appear, containing a table with your imported Reddit data. From here you can further tailor your data set if necessary, or start analyzing immediately.

Related Topics:

To configure the data set and start analyzing please see [The Data Tab](#).

To append Reddit data to your project please [Monitoring Online Resources](#).

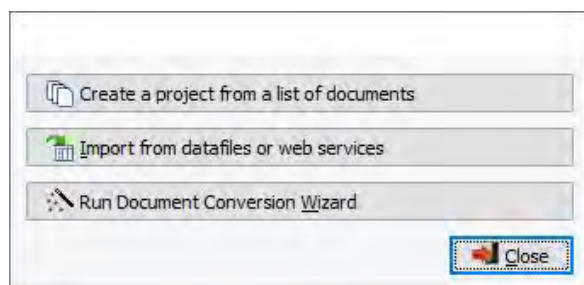
For more information on the Web Collector please see the section entitled [Web Collector](#).

Importing Bibliographic References

Reference Management tools are used to store, organize and search bibliographic data, abstracts and complete papers. WordStat allows you to create projects by importing bibliographic data and use the text mining and content analysis features of WordStat to effectively analyze your bibliographic data. WordStat supports EndNote, Mendeley and Zotero. It also supports the RIS (*.ris) file format.

Creating a New Project Using Bibliographic Data

- Select the **New** button on the **Data** tab. This command calls up a dialog box similar to the one below.



- Select the **Import from data files or web services** button.
- Select the **REFERENCE MANAGERS** menu item and choose the desired reference management software from its submenu.

For instruction on how to import bibliographic reference data from various sources, please see:

Importing from [EndNote](#)

Importing from [Mendeley](#)

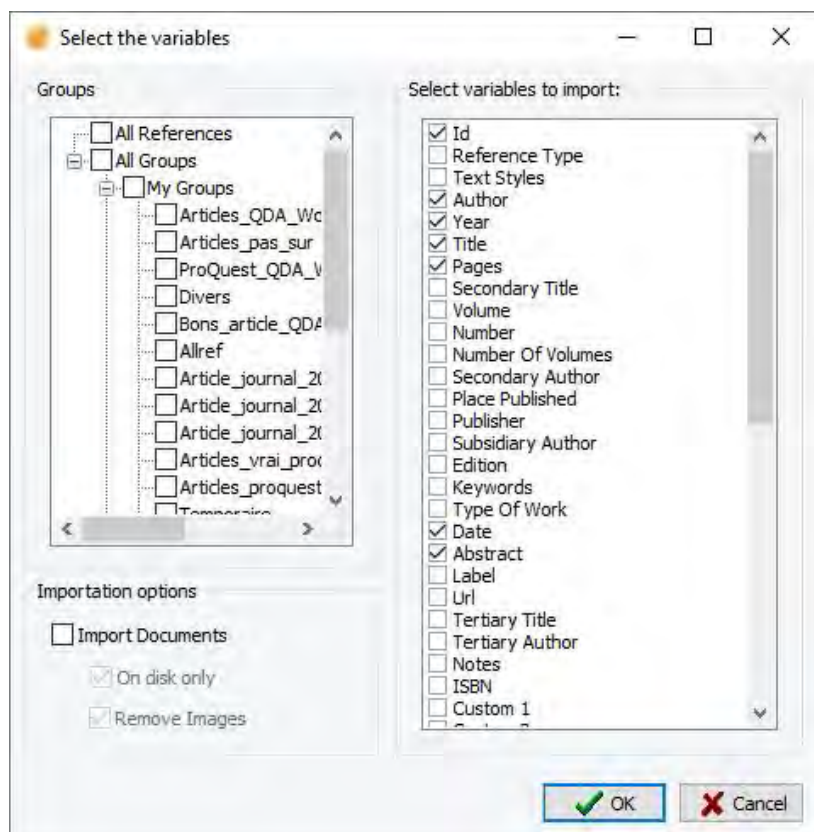
Importing from [Zotero](#)

Importing a [RIS \(*.ris\)](#) file

EndNote

Creating a New Project Using EndNote

- Select **REFERENCE MANAGERS | ENDNOTE** from the menu and an **Import data file** dialog box will appear.
- Select the EndNote file that you wish to import.
- Click **Open**.
- Enter the project name, save it in the location of your choosing. A **Select the variables** dialog box will appear.



- On the right-hand side of the dialog box is the **Groups** list box. A group in EndNote basically separates your library into categories. Select the check-boxes of the groups you want to import from.
- You are given the option to **Import Documents**. Check the **Import Documents** check box if you would like to import documents.
- Documents in EndNote are saved on disk, URLs that access online documents are also saved within the files. You have the option of importing documents that are saved **On disk only** and ignoring the URLs that point to online documents.
- Check the **Remove Images** checkbox if you would like to remove the images contained in the documents you are importing.
- On the right side of the dialog box is a list box of variables you can select for importation. Check the boxes of the variables you wish to import.
- Click **OK**. The data will be imported and WordStat's **Data** tab will appear, containing a table with your imported bibliographic data. From here you can further tailor your data set if necessary, or start analyzing immediately.

Related Topics:

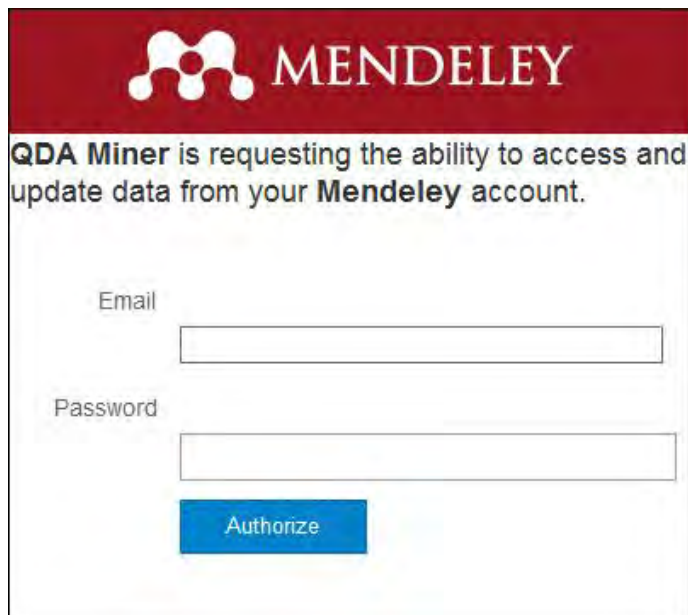
To configure the data set and start analyzing please see [The Data Tab](#).

To append new bibliographic references to your project please [Monitoring Online Resources](#).

Mendeley

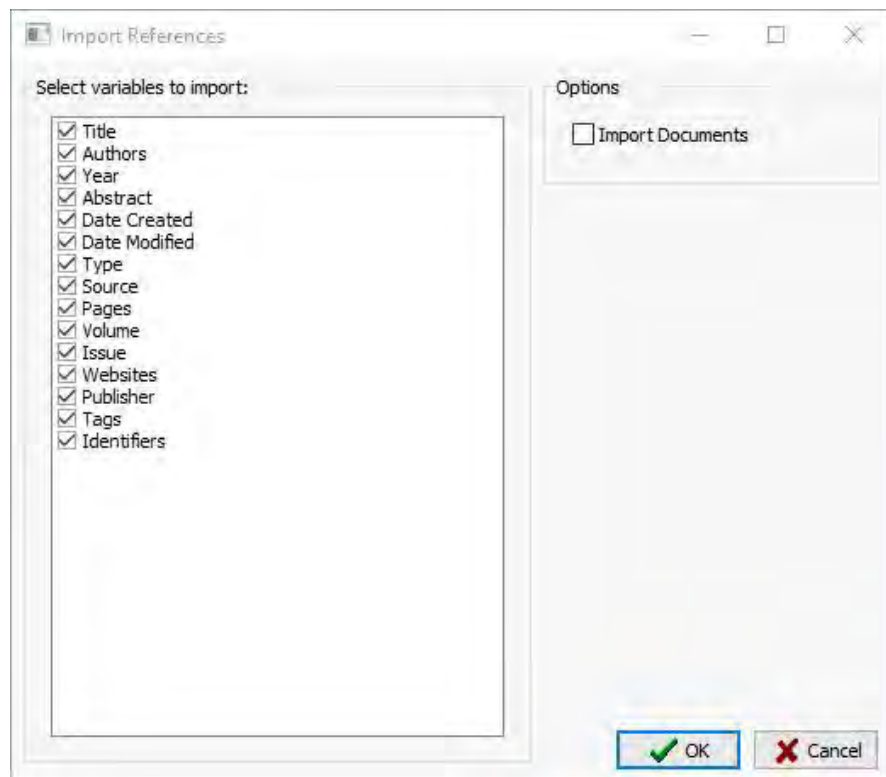
Creating a New Project Using Mendeley

- Select **REFERENCE MANAGERS | MENDELEY** from the menu and a dialog box requesting access to your Mendeley account appears.



The image shows a Mendeley authorization dialog box. At the top is the Mendeley logo and name. Below it, a message states: "QDA Miner is requesting the ability to access and update data from your Mendeley account." There are two input fields: "Email" and "Password". Below these fields is a blue button labeled "Authorize".

- Enter your **Email** and your **Password**.
- Click **Authorize** if you give permission for WordStat to access your account. An **Import References** dialog box will appear.



The image shows the "Import References" dialog box. It has a title bar with "Import References" and standard window controls. The main area is divided into two sections. On the left, under "Select variables to import:", there is a list of variables with checkboxes next to them: Title, Authors, Year, Abstract, Date Created, Date Modified, Type, Source, Pages, Volume, Issue, Websites, Publisher, Tags, and Identifiers. All checkboxes are checked. On the right, under "Options", there is a checkbox labeled "Import Documents" which is currently unchecked. At the bottom right, there are two buttons: "OK" (with a green checkmark icon) and "Cancel" (with a red X icon).

- On the left side of the dialog box is a list box of variables you can select for importation. Check the boxes of the variables you wish to import.
- On the right-hand side of the dialog box you are given the option to **Import Documents**. Select the **Import Documents** checkbox if you would like to import documents.

- Click **OK**.
- Enter the project name and save it in the location of your choosing. The data will be imported and WordStat's **Data** tab will appear, containing a table with your imported bibliographic data. From here you can further tailor your data set if necessary, or start analyzing immediately.

Related Topics:

To configure the data set and start analyzing please see [The Data Tab](#).

To append new bibliographic references to your project please [Monitoring Online Resources](#).

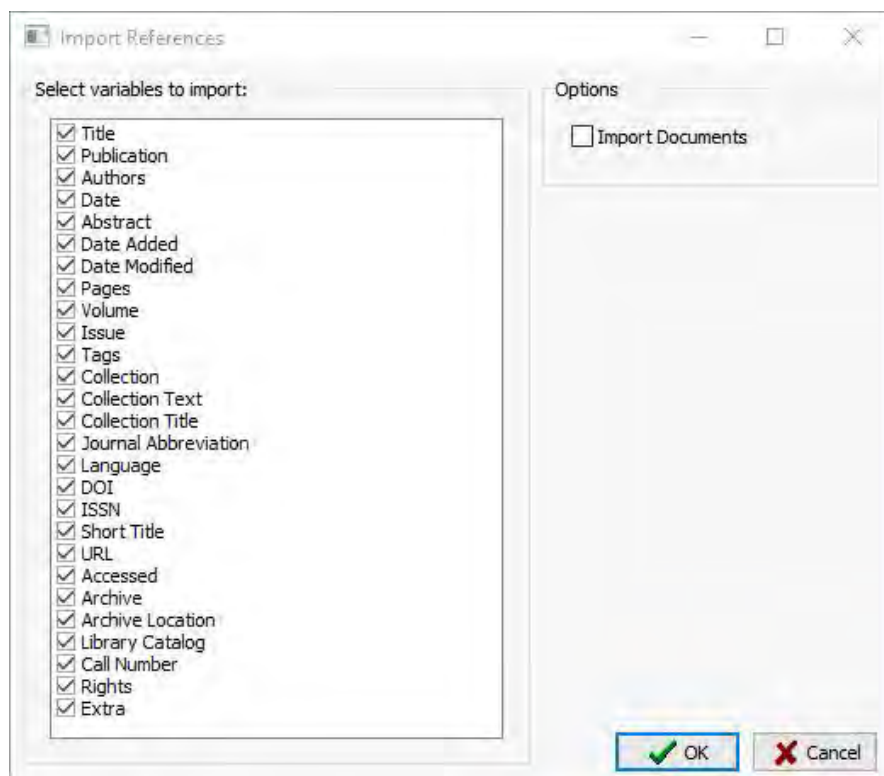
Zotero

Creating a New Project Using Zotero

- Select **REFERENCE MANAGERS | ZOTERO** from the menu and a dialog box asking you to sign in to your Zotero account appears.
- Log in to your Zotero account. A dialog box notifying you that WordStat would like to access your account appears.



- Click **Accept Defaults** if you give permission for WordStat to access your account. An **Import References** dialog box will appear.



- On the left side of the dialog box is a list box of variables you can select for importation. Check the boxes of the variables you wish to import.
- On the right-hand side of the dialog box you are given the option to **Import Documents**. Check the **Import Documents** checkbox if you would like to import documents.
- Click **OK**.
- Enter the project name and save it in the location of your choosing. The data will be imported and WordStat's **Data** tab will appear, containing a table with your imported bibliographic data. From here you can further tailor your data set if necessary, or start analyzing immediately.

Related Topics:

To configure the data set and start analyzing please see [The Data Tab](#).

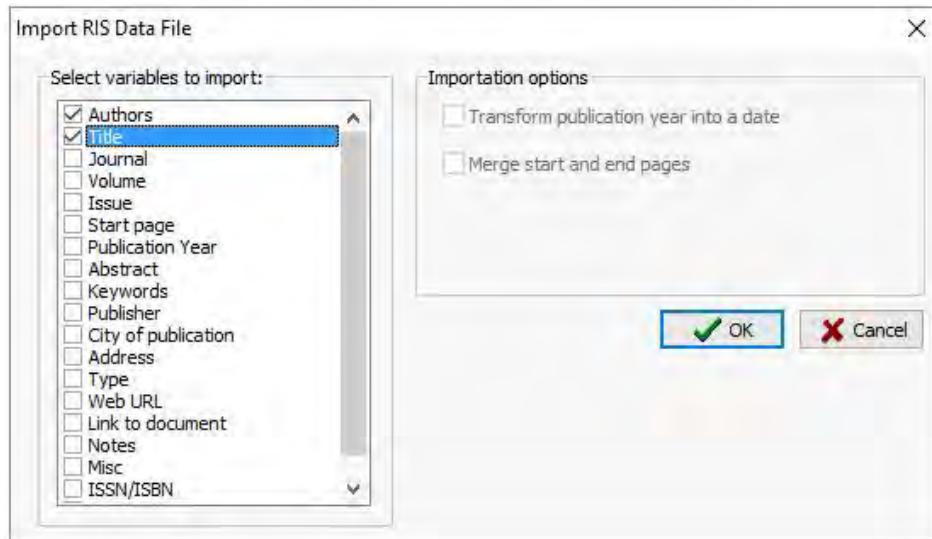
To append new bibliographic references to your project please [Monitoring Online Resources](#).

RIS File

Creating a New Project from a .RIS File

Bibliographic data from Reference Management tools and digital libraries is often stored in the RIS (*.ris) file format. RIS files are plain-text files and have by default a TXT file extension. To import such a file, set the **Files of type** list box to Reference Information Management (RIS) and then select the file you wish to import. Another approach is to change the file extension to RIS, and WordStat will automatically recognize this file extension and import the references in this file without the need to specify the proper file type.

- Once a file has been selected for importation, WordStat will display an **Import RIS Data File** dialog box.



The following importation options are available:

Select variables to import: This list box allows you to put check marks beside the variables you want to import. All unchecked items will be ignored.

Transform publication year into a date: The publication dates in RIS files can consist of a full date, with days and months, or only the year, or sometimes the publication month and year. By default, WordStat will import just the publication year and store it in an integer variable. Choosing this option will store all dates into a **Date** variable. If only the year is specified, WordStat will set the day and month to January 1. If the publication date consists of a month and a year only, then WordStat will set the day to the first day of the month.

Merge start and end pages: By default, start-page and end-page numbers are stored in separate variables. Selecting this option will join both numbers with a hyphen character and store in a single string variable.

- Once your options are set, select **OK**. The data will be imported and WordStat's **Data** tab will appear, containing a table with your imported bibliographic data. From here you can further tailor your data set if necessary, or start analyzing immediately.

Related Topics:

To configure the data set and start analyzing please see [The Data Tab](#).

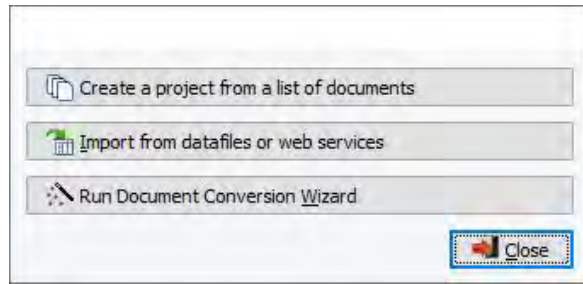
To append new bibliographic references to your project please [Monitoring Online Resources](#).

Importing News Transcripts

Factiva and **Nexis UNI** are online database giving their users access to previously published news from newspapers and magazines around the world, TV and radio news transcripts as well as additional information specific to each service such as online blogs, financial and market news, company information, case law, etc. WordStat can import files obtained from those two service, allowing one to analyze media coverage of specific topics, compare it by media outlets, countries, asses changes over time, etc.

Creating a New Project with News Transcripts

- Select the **New** button on the **Data** tab. This command calls up a dialog box similar to the one below.



- Select the **Import from data files or web services** button.
- Select the **NEWS** menu item and choose the desired news aggregation service from its submenu.

For instruction on how to import news transcripts from those two sources, please see:

Importing from [Nexis Uni](#)

Importing from [Factiva](#)

Nexis UNI

Nexis UNI from LexisNexis contains information from over 10,000 news, legal and business sources, including international and domestic newspaper, magazines, broadcast news transcripts, company financial information, industry and market news, law reviews, medical news, etc. It also includes business information on over 80 million public & private international companies and 75 million executives. Non-English language news sources are available in Spanish, French, German, Italian, and Dutch. Campus news from some 400 college/university papers and over 50 wire services are also available.

For the sake of this presentation, we will make references below to news transcripts, yet similar procedures may be followed to import other types of text data from Nexis UNI.

To import news transcripts from Nexis Uni, you first need to access the online service, perform your searches, select the relevant news transcripts and then save those on disk using the download command. A dialog box like this one will appear:

What do you want to download?

- ☒ Full documents (10)
 - ☐ Include document attachments, where available
- ☐ Full document attachments only (where available)
- ☐ Results list for 'News': (10)
- ☐ Include Bibliography

File type

- ☐ PDF
100 document limit.
- ☒ MS Word (docx)
100 document limit.
- ☐ Rich Text Format (rtf)
100 document limit.

When downloading multiple documents

- ☒ Group and save documents as a single file (Note: Attachments will be delivered as separate documents)
- ☐ Save as individual files
- ☐ Compress files in .ZIP format [What's this?](#)

Filename

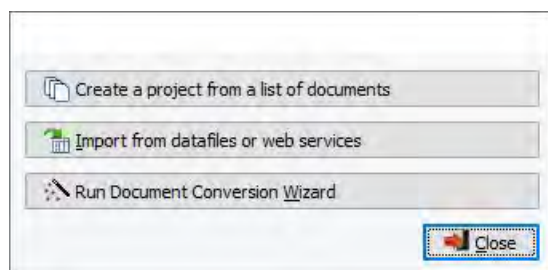
100-character limit

Files(10)

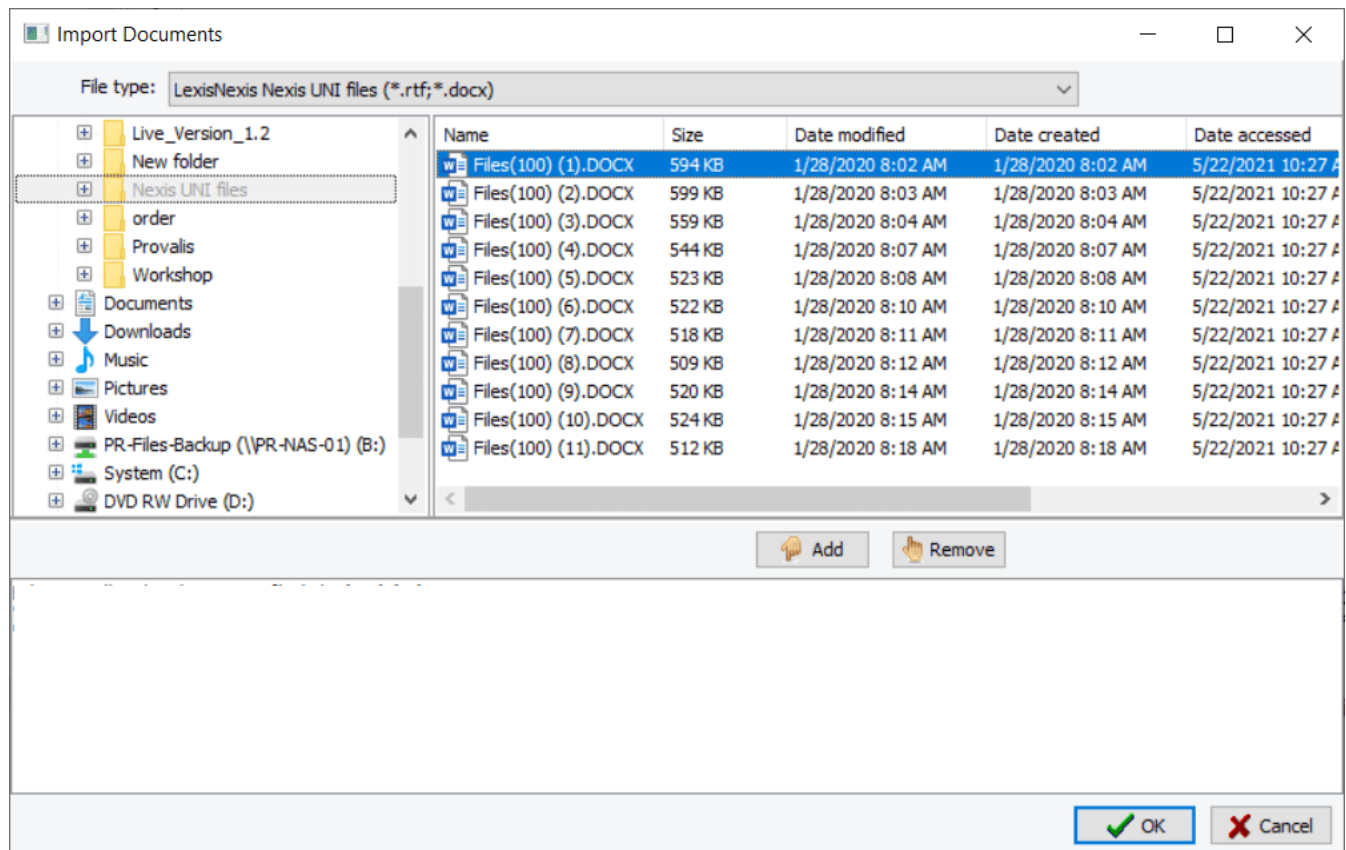
Make sure the **Full Documents** option is selected, and choose either the **MS Word (docx)** or the **Rich Text Format (rtf)** file format is selected. It is also recommended to keep documents grouped in a single file rather than individual files. You may need to create several of those files if you need to import more than the maximum number of documents allowed per download.

Once all the news transcripts have been downloaded, you can extract the news transcripts from those files with WordStat by following these steps:

- Select the **New** button on the **Data** tab. This command calls up a dialog box similar to the one below.



- Select the **Import from data files or web services** button.
- Select the **NEXIS UNI** menu item from the **NEWS** menu.
- A dialog box similar to the one below will appear allowing you to select the files containing the news transcripts.



- Using the upper left panel, navigate to the folder containing the downloaded files.
- In the upper right panel, files in the selected folder that match the file type will be listed. Select the files you want to import and click the  **Add** button to move them to the bottom file list.
- Once the files to import have been moved to the bottom section of the dialog box, click the **OK** button. WordStat will import all files by extracting each news transcript separately. It will also store various information in as many variables such as the title, the body of the article, the publication type, the source, the author, etc. WordStat will also save all indexing tags that have been assigned to the specific news respective categories (subject, geographic, organization, person, industry, etc.). If needed, variables containing superfluous information may then be deleted (see [Deleting Existing Variables](#)).

WordStat 9.0.7 - Test.pptj

Data

Open New Save Save as Export Properties Analyze

Data Editor Structure

	FILENAME	PUB_TYPE	SOURCE	DATEPUB	AUTHOR	TITLE	BODY	LANGUAGE	LOAD_DATE	NBWORDS	SUBJECT	GEOGRAPHIC	ORGANIZAT	INDUSTRY	PERSON
File(100) [9]	Journal	Cameron Tribune (Yaounde)	1/27/2010	Brice Mbete	[DOCUM...]	[DOCUM...]	FRENCH	2/12/2010	893	[DOCUM...]	[DOCUMENT]	[document]	[document]	[document]	[document]
File(100) [9]	Journal	Cameron Tribune (Yaounde)	5/3/2013	Alliance Nyobia	[DOCUM...]	[DOCUM...]	FRENCH	5/3/2013	442	[DOCUM...]	[DOCUMENT]	[document]	[DOCUME...]	[document]	[document]
File(100) [9]	Journal	Cameron Tribune (Yaounde)	11/14/2013	Badjang Ba Nken	[DOCUM...]	[DOCUM...]	FRENCH	11/15/2013	600	[DOCUM...]	[DOCUMENT]	[document]	[DOCUME...]	[document]	[document]
File(100) [9]	Journal	Cameron Tribune (Yaounde)	11/18/2013	Rita Diba	[DOCUM...]	[DOCUM...]	FRENCH	11/19/2013	431	[DOCUM...]	[DOCUMENT]	[document]	[DOCUME...]	[document]	[document]
File(100) [9]	Journal	Cameron Tribune (Yaounde)	7/28/2010	Josiane Tchakounte	[DOCUM...]	[DOCUM...]	FRENCH	7/28/2010	404	[DOCUM...]	[DOCUMENT]	[document]	[DOCUME...]	[document]	[document]
File(100) [9]	Journal	Cameron Tribune (Yaounde)	9/21/2010	Essama Essomba	[DOCUM...]	[DOCUM...]	FRENCH	9/21/2010	304	[DOCUM...]	[DOCUMENT]	[DOCUMENT]	[DOCUME...]	[document]	[document]
File(100) [9]	Journal	Cameron Tribune (Yaounde)	3/7/2014	Alain Tchakounte	[DOCUM...]	[DOCUM...]	FRENCH	3/10/2014	556	[DOCUM...]	[DOCUMENT]	[document]	[DOCUME...]	[document]	[document]
File(100) [9]	Journal	Cameron Tribune (Yaounde)	2/7/2011		[DOCUM...]	[DOCUM...]	FRENCH	2/8/2011	372	[DOCUM...]	[DOCUMENT]	[document]	[DOCUME...]	[document]	[document]
File(100) [9]	Journal	Cameron Tribune (Yaounde)	5/21/2012	Alliance Nyobia	[DOCUM...]	[DOCUM...]	FRENCH	5/22/2012	435	[DOCUM...]	[DOCUMENT]	[document]	[DOCUME...]	[document]	[document]
File(100) [9]	Journal	Cameron Tribune (Yaounde)	11/16/2011	Makou Ma Pondi	[DOCUM...]	[DOCUM...]	FRENCH	11/16/2011	682	[DOCUM...]	[DOCUMENT]	[document]	[DOCUME...]	[document]	[document]
File(100) [9]	Journal	Cameron Tribune (Yaounde)	4/6/2010	Alain Tchakounte	[DOCUM...]	[DOCUM...]	FRENCH	4/6/2010	438	[DOCUM...]	[DOCUMENT]	[DOCUMENT]	[document]	[document]	[document]
File(100) [9]	Journal	Cameron Tribune (Yaounde)	7/16/2010	Elise Ziemine	[DOCUM...]	[DOCUM...]	FRENCH	7/16/2010	368	[DOCUM...]	[DOCUMENT]	[document]	[DOCUME...]	[document]	[document]
File(100) [9]	Journal	Cameron Tribune (Yaounde)	12/18/2014	Monda Bakoa	[DOCUM...]	[DOCUM...]	FRENCH	12/18/2014	536	[DOCUM...]	[document]	[document]	[DOCUME...]	[document]	[document]
File(100) [9]	Journal	Cameron Tribune (Yaounde)	4/4/2011	Sancilair Mezing	[DOCUM...]	[DOCUM...]	FRENCH	4/5/2011	336	[DOCUM...]	[DOCUMENT]	[document]	[DOCUME...]	[document]	[document]
File(100) [9]	Journal	Cameron Tribune (Yaounde)	3/1/2010	Jeanine Fankam	[DOCUM...]	[DOCUM...]	FRENCH	3/1/2010	392	[DOCUM...]	[DOCUMENT]	[document]	[document]	[document]	[document]
File(100) [9]	Journal	Cameron Tribune (Yaounde)	10/27/2010		[DOCUM...]	[DOCUM...]	FRENCH	10/28/2010	328	[DOCUM...]	[DOCUMENT]	[document]	[document]	[document]	[document]
File(100) [9]	Journal	Cameron Tribune (Yaounde)	2/3/2010	Simon Pierre Etoundi	[DOCUM...]	[DOCUM...]	FRENCH	2/12/2010	679	[DOCUM...]	[DOCUMENT]	[document]	[DOCUME...]	[document]	[document]
File(100) [9]	Journal	Cameron Tribune (Yaounde)	5/6/2010	Eric Vincent Fono	[DOCUM...]	[DOCUM...]	FRENCH	5/6/2010	600	[DOCUM...]	[DOCUMENT]	[document]	[DOCUME...]	[document]	[document]
File(100) [9]	Journal	Cameron Tribune (Yaounde)	5/25/2010	Alfred Mvogo Biyeck	[DOCUM...]	[DOCUM...]	FRENCH	6/4/2010	449	[DOCUM...]	[DOCUMENT]	[document]	[document]	[document]	[document]
File(100) [9]	Journal	Cameron Tribune (Yaounde)	3/25/2010	Steve Libam	[DOCUM...]	[DOCUM...]	FRENCH	3/25/2010	490	[DOCUM...]	[DOCUMENT]	[document]	[DOCUME...]	[document]	[document]
File(100) [9]	Journal	Cameron Tribune (Yaounde)	5/21/2012	Wallo Mongo	[DOCUM...]	[DOCUM...]	FRENCH	5/22/2012	562	[DOCUM...]	[DOCUMENT]	[document]	[DOCUME...]	[document]	[document]
File(100) [9]	Journal	Cameron Tribune (Yaounde)	7/31/2013	Messi Bala	[DOCUM...]	[DOCUM...]	FRENCH	7/31/2013	302	[DOCUM...]	[DOCUMENT]	[document]	[DOCUME...]	[document]	[document]
File(100) [9]	Journal	Cameron Tribune (Yaounde)	5/3/2013	Monda Bakoa	[DOCUM...]	[DOCUM...]	FRENCH	5/3/2013	588	[DOCUM...]	[DOCUMENT]	[DOCUMENT]	[DOCUME...]	[document]	[document]
File(100) [9]	Journal	Cameron Tribune (Yaounde)	2/22/2010	Josiane Tchakounte	[DOCUM...]	[DOCUM...]	FRENCH	2/22/2010	478	[DOCUM...]	[DOCUMENT]	[document]	[document]	[document]	[document]
File(100) [9]	Journal	Cameron Tribune (Yaounde)	10/3/2013	Jean Francis Belibi	[DOCUM...]	[DOCUM...]	FRENCH	10/3/2013	317	[DOCUM...]	[DOCUMENT]	[document]	[DOCUME...]	[document]	[document]
File(100) [9]	Journal	Cameron Tribune (Yaounde)	2/24/2014	Josiane Tchakounte	[DOCUM...]	[DOCUM...]	FRENCH	2/25/2014	624	[DOCUM...]	[DOCUMENT]	[document]	[document]	[document]	[document]
File(100) [9]	Journal	Cameron Tribune (Yaounde)	5/10/2010	Steve Libam	[DOCUM...]	[DOCUM...]	FRENCH	5/10/2010	477	[DOCUM...]	[DOCUMENT]	[document]	[DOCUME...]	[document]	[document]
File(100) [9]	Journal	Cameron Tribune (Yaounde)	6/18/2010	Augustin Fogang	[DOCUM...]	[DOCUM...]	FRENCH	6/18/2010	638	[DOCUM...]	[DOCUMENT]	[document]	[DOCUME...]	[document]	[document]
File(100) [9]	Journal	Cameron Tribune (Yaounde)	5/7/2010	Josiane Tchakounte	[DOCUM...]	[DOCUM...]	FRENCH	5/7/2010	255	[DOCUM...]	[DOCUMENT]	[DOCUMENT]	[document]	[document]	[document]
File(100) [9]	Journal	Cameron Tribune (Yaounde)	10/28/2014	Eric Elouga	[DOCUM...]	[DOCUM...]	FRENCH	10/28/2014	438	[DOCUM...]	[DOCUMENT]	[document]	[DOCUME...]	[document]	[document]
File(100) [9]	Journal	Cameron Tribune (Yaounde)	5/29/2012	Jean Baptiste Ketchateng	[DOCUM...]	[DOCUM...]	FRENCH	5/29/2012	358	[DOCUM...]	[DOCUMENT]	[document]	[DOCUME...]	[document]	[document]
File(100) [9]	Journal	Cameron Tribune (Yaounde)	6/21/2010	Patrice Mbossa	[DOCUM...]	[DOCUM...]	FRENCH	6/22/2010	425	[DOCUM...]	[DOCUMENT]	[document]	[DOCUME...]	[document]	[document]
File(100) [9]	Journal	Cameron Tribune (Yaounde)	3/7/2012	Steve Libam	[DOCUM...]	[DOCUM...]	FRENCH	3/7/2012	435	[DOCUM...]	[DOCUMENT]	[document]	[DOCUME...]	[document]	[document]
File(100) [9]	Journal	Cameron Tribune (Yaounde)	5/7/2010	Eric Vincent Fono	[DOCUM...]	[DOCUM...]	FRENCH	5/7/2010	462	[DOCUM...]	[DOCUMENT]	[DOCUMENT]	[DOCUME...]	[document]	[document]
File(100) [9]	Journal	Cameron Tribune (Yaounde)	7/16/2014	Angile Bepede	[DOCUM...]	[DOCUM...]	FRENCH	7/16/2014	651	[DOCUM...]	[DOCUMENT]	[document]	[document]	[document]	[document]
File(100) [9]	Journal	Cameron Tribune (Yaounde)	6/29/2011	Sancilair Mezing	[DOCUM...]	[DOCUM...]	FRENCH	6/29/2011	466	[DOCUM...]	[DOCUMENT]	[document]	[DOCUME...]	[document]	[document]

1200 / 1200

Factiva

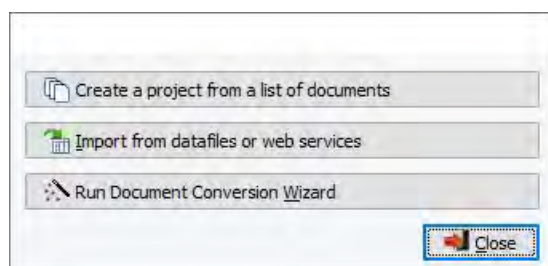
Factiva is a business information and research tool owned by Dow Jones & Company that aggregates content from both licensed and free sources. Factiva provides access to more than 32,000 sources (such as newspapers, journals, magazines, television and radio transcripts, photos, etc.) from nearly every country worldwide in 28 languages, including more than 600 continuously updated newswires. It can also retrieve press releases and financial information about public and private companies as well as historical and current market indices.

For the sake of this presentation, we will make references below to news transcripts, yet similar procedures may be followed to import other types data retrieved from Factiva.

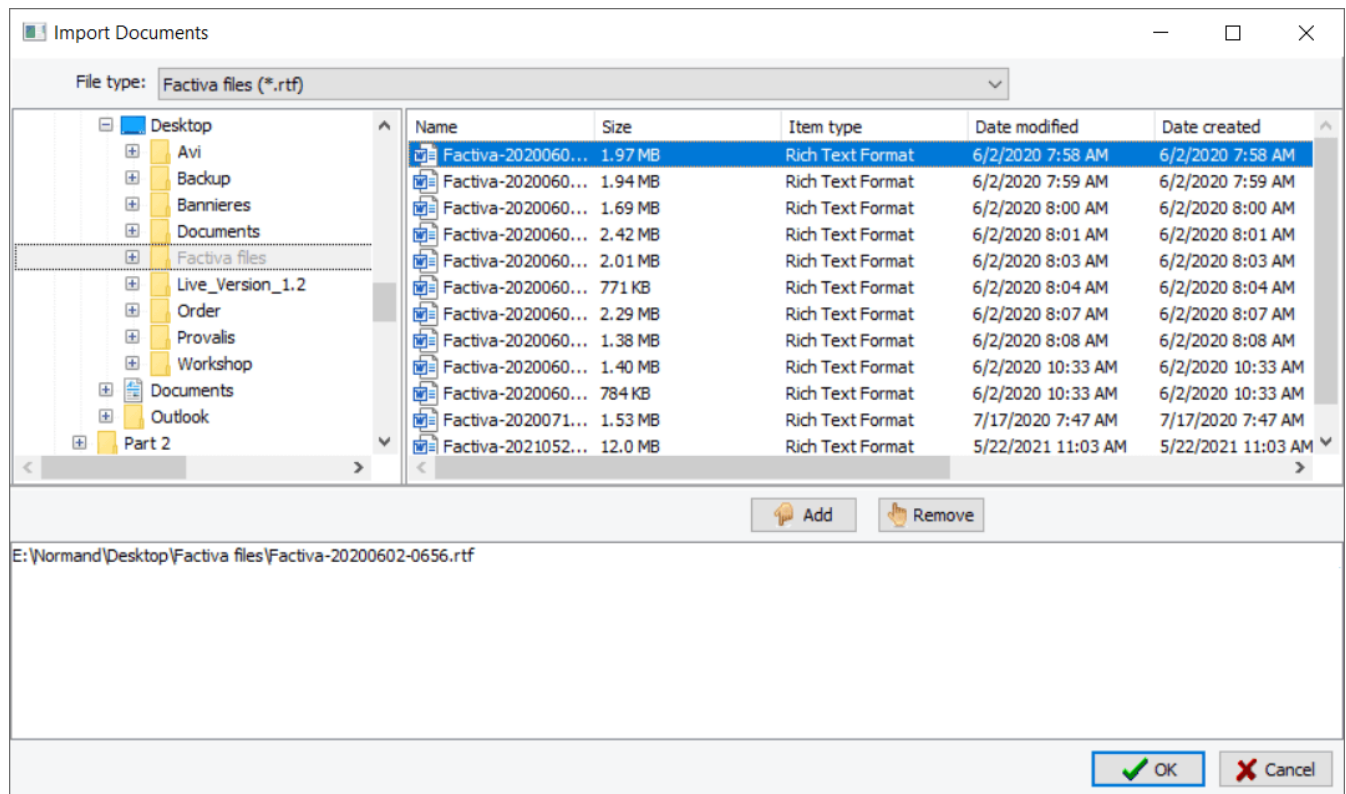
To import news transcripts from Factiva, you first need to access the online service, perform your searches, select the relevant news transcripts and then save those on disk in RTF format by click the  button and selecting **Article Format**.

Once all the news transcripts have been downloaded, you can extract the news transcripts from those files with WordStat by following these steps:

- Select the **New** button on the **Data** tab. This command calls up a dialog box similar to the one below.



- Select the **Import from data files or web services** button.
- Select the **FACTIVA** menu item from the **NEWS** menu.
- A dialog box similar to the one below will appear allowing you to select the files containing the news transcripts.



- Using the upper left panel, navigate to the folder containing the downloaded files.
- In the upper right panel, files in the selected folder that match the file type will be listed. Select the files you want to import and click the  **Add** button to move them to the bottom file list.
- Once the files to import have been moved to the bottom section of the dialog box, click the **OK** button. WordStat will import all files by extracting each news transcript separately. It will also store various information in as many variables such as the title, the body of the article, the publication type, the source, the author, etc. If needed, variables containing superfluous information may then be deleted (see [Deleting Existing Variables](#)).

WordStat 9.0 - Factiva.ppj

Open New Save Save as Export Properties Analyze

Data Editor Structure

FILENAME	SOURCE	DATEPUB	AUTHOR	TITLE	BODY	LANGUAGE	NBWORDS
Factiva-20200602-0656	The New York Times	6/28/2020	By Javier C. Hernández and Benjamin Mueller	[DOCUM...	[DOCUM...	Anglais	1519
Factiva-20200602-0656	NYTimes.com Feed	6/2/2020		[DOCUM...	[DOCUM...	Anglais	1754
Factiva-20200602-0656	NYTimes.com Feed	6/2/2020		[DOCUM...	[DOCUM...	Anglais	3526
Factiva-20200602-0656	NYTimes.com Feed	6/2/2020	By David Leonhardt	[DOCUM...	[DOCUM...	Anglais	1867
Factiva-20200602-0656	NYTimes.com Feed	6/2/2020	By Giovanni Russonello	[DOCUM...	[DOCUM...	Anglais	1313
Factiva-20200602-0656	NYTimes.com Feed	6/2/2020	By Ross Douthat	[DOCUM...	[DOCUM...	Anglais	1031
Factiva-20200602-0656	NYTimes.com Feed	6/2/2020		[DOCUM...	[DOCUM...	Anglais	295
Factiva-20200602-0656	NYTimes.com Feed	6/2/2020		[DOCUM...	[DOCUM...	Anglais	1369
Factiva-20200602-0656	NYTimes.com Feed	6/2/2020	By John Branch	[DOCUM...	[DOCUM...	Anglais	2981
Factiva-20200602-0656	NYTimes.com Feed	6/2/2020	By A.O. Scott	[DOCUM...	[DOCUM...	Anglais	1517
Factiva-20200602-0656	NYTimes.com Feed	6/2/2020	By Kevin Draper and Julie Creswell	[DOCUM...	[DOCUM...	Anglais	848
Factiva-20200602-0656	NYTimes.com Feed	6/2/2020	By Tonya Russell	[DOCUM...	[DOCUM...	Anglais	1052
Factiva-20200602-0656	NYTimes.com Feed	6/2/2020	By Nick Corasaniti	[DOCUM...	[DOCUM...	Anglais	1555
Factiva-20200602-0656	NYTimes.com Feed	6/2/2020	By Gail Collins and Bret Stephens	[DOCUM...	[DOCUM...	Anglais	1668
Factiva-20200602-0656	NYTimes.com Feed	6/2/2020	By Janelle Bouie	[DOCUM...	[DOCUM...	Anglais	1211
Factiva-20200602-0656	NYTimes.com Feed	6/2/2020		[DOCUM...	[DOCUM...	Anglais	535
Factiva-20200602-0656	NYTimes.com Feed	6/2/2020	By Trish Bendix	[DOCUM...	[DOCUM...	Anglais	912
Factiva-20200602-0656	The New York Times	6/2/2020	By Declan Walsh	[DOCUM...	[DOCUM...	Anglais	1007
Factiva-20200602-0656	The New York Times	6/2/2020		[DOCUM...	[DOCUM...	Anglais	72
Factiva-20200602-0656	The New York Times	6/2/2020		[DOCUM...	[DOCUM...	Anglais	838
Factiva-20200602-0656	The New York Times	6/2/2020		[DOCUM...	[DOCUM...	Anglais	239
Factiva-20200602-0656	The New York Times	6/2/2020		[DOCUM...	[DOCUM...	Anglais	49
Factiva-20200602-0656	The New York Times	6/2/2020	By Evan Hill, Ainara Tiefenthaler, Christian Triebert, Drew Jordan...	[DOCUM...	[DOCUM...	Anglais	1318
Factiva-20200602-0656	The New York Times	6/2/2020	By Sanna Maheshwari and Michael Corkran	[DOCUM...	[DOCUM...	Anglais	1372

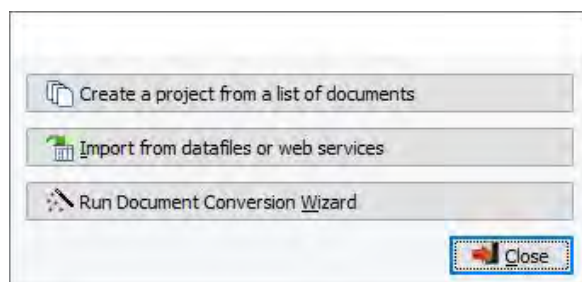
1 / 185

Importing from Email Servers

Email has become a major source of communication and may also be used as a cost-efficient data collection tool. As a result, electronic mailboxes often contain a wealth of useful information which may be the basis of a research project. WordStat allows you to create a project by importing email data and use the text mining and content analysis features of WordStat to analyze your email data. You can create projects from Outlook, Gmail and Hotmail cloud-based email servers. You can also create projects from the Outlook account on your PC and from PST and MBOX files.

Creating a New Project Using Email Data

- Select the **New** button on the **Data** tab. This command calls up a dialog box similar to the one below.



- Select the **Import from data files or web services** button.
- Select the **E-MAILS** menu item and choose the desired email service from its submenu.

For instruction on how to import email from various sources, please see:

Importing from [Outlook](#)

Importing from [Gmail](#)

Importing from [Hotmail](#)

Importing from an [Outlook Account](#)

Importing from a [PST](#) file

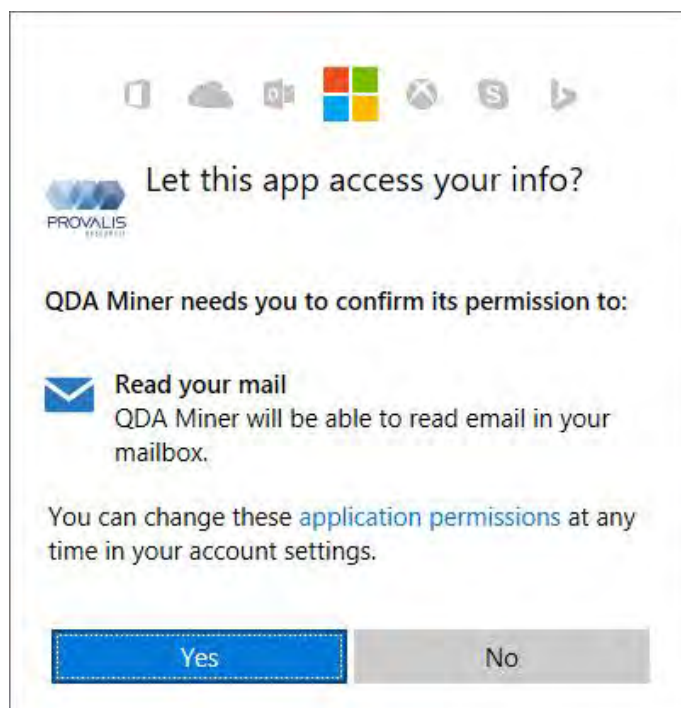
Importing from a [MBOX](#) file

Outlook

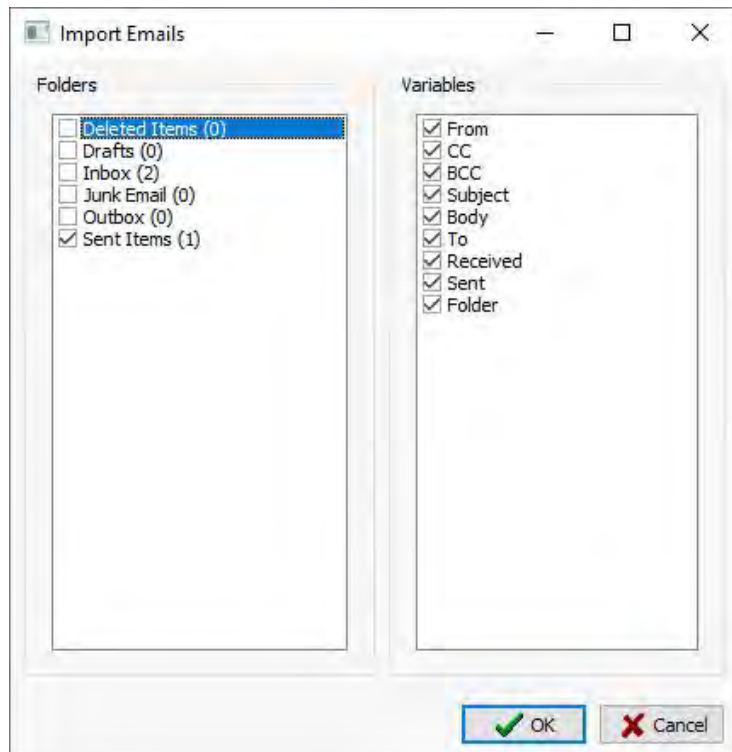
There are three ways to access emails stored in Outlook. You can connect through the Outlook.com email server, connect directly with the outlook application on your computer or import Outlook data contained in PST files. PST files can be imported regardless of whether or not you have access to the account from which they originated. Two examples of when a PST importation could be used are to import backups of old emails or emails that are not yours and for which you don't have access to the email account.

Connecting to Outlook.com

- Select **E-MAILS | OUTLOOK | OUTLOOK.COM** from the menu to log in to Microsoft Outlook's cloud-based email service. A log in dialog box will appear.
- Log in to your Outlook.com account. A permissions dialog box will appear.



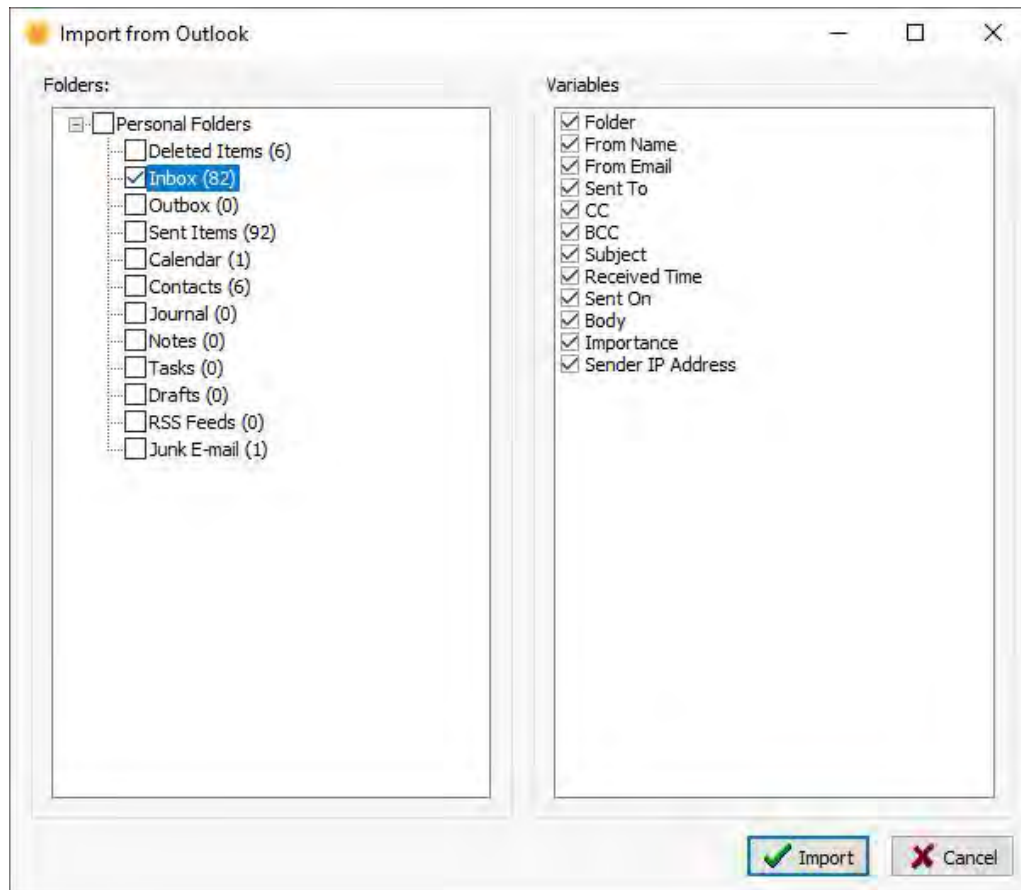
- Click **Yes** if you give permission for WordStat to access your account. An **Import Emails** dialog box will appear.



- The items in the **Folders** list box on the left side of the dialog box correspond to the folders in your email account. Check the **Folders** you wish to import.
- The items in the **Variables** list box on the right side of the dialog box will become project variables if imported. Check the **Variables** you want to include in your project.
- Click **OK**.
- Enter the project name and save it in the location of your choosing. The data will be imported and WordStat's **Data** tab will appear, containing a table with your imported email data. From here you can further tailor your data set if necessary, or start analyzing immediately.

Importing Data from an Outlook Email Application on Your Computer

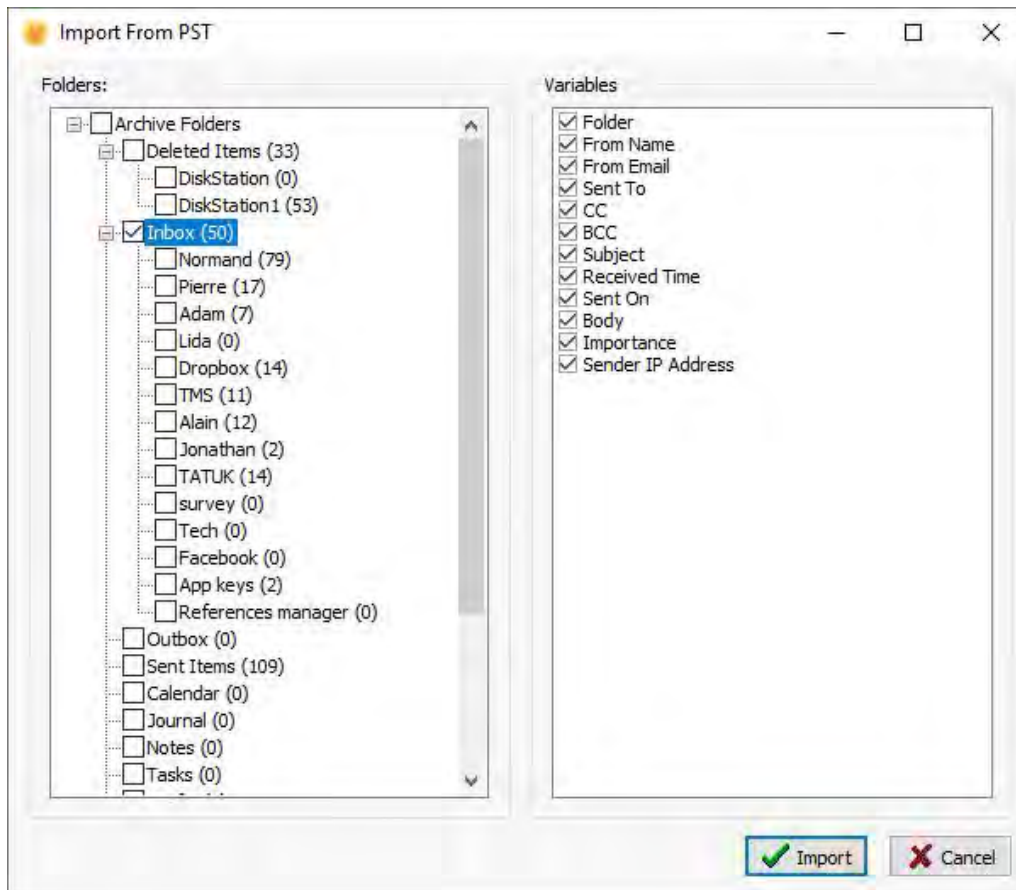
- Select **E-MAILS | OUTLOOK | ACCOUNT** from the menu to import email from the Outlook application that is installed on your computer. An **Import from Outlook** dialog box will appear.



- The items in the **Folders** list box on the left side of the dialog box correspond to the folders in your email account. Check the **Folders** you wish to import.
- The items in the **Variables** list box on the right side of the dialog box will become project variables if imported. Check the **Variables** you wish to import.
- Click **Import**.
- Enter the project name and save it in the location of your choosing. The data will be imported and WordStat's **Data** tab will appear, containing a table with your imported email data. From here you can further tailor your data set if necessary, or start analyzing immediately.

Importing Data From a PST File

- Select **E-MAILS | OUTLOOK | PST FILE** from the menu to import a file saved on your PC in the Personal Storage Table (*.pst) format. An **Import data file** dialog box will appear.
- Select the PST file you want to import.
- Click **Open**. An **Import from PST** dialog box will appear.



- The items in the **Folders** list box on the left side of the dialog box correspond to the folders in your email account. Check the **Folders** you wish to import.
- The items in the **Variables** list box on the right side of the dialog box will become project variables if imported. Check the **Variables** you wish to import.
- Click **Import**.
- Enter the project name and save it in the location of your choosing. The data will be imported and WordStat's **Data** tab will appear, containing a table with your imported email data. From here you can further tailor your data set if necessary, or start analyzing immediately.

Related Topics:

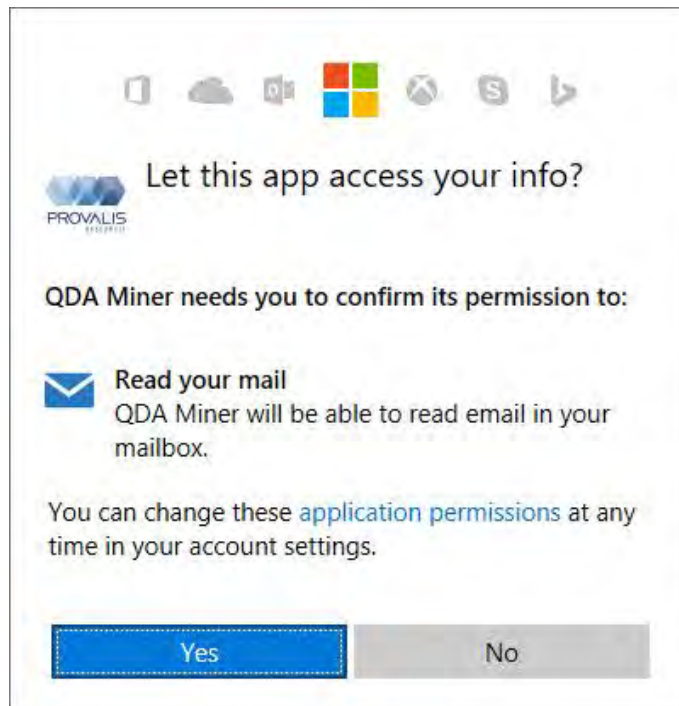
To configure the data set and start analyzing please see [The Data Tab](#).

To append new email data to your project please [Monitoring Online Resources](#).

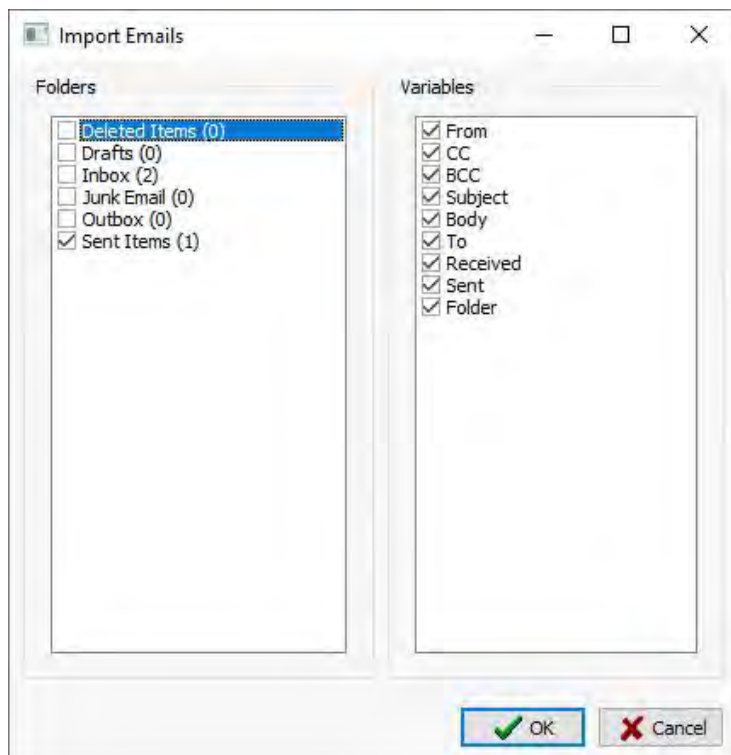
Hotmail

Connecting to Hotmail

- Select **E-MAILS | HOTMAIL** from the menu. A log in dialog box will appear.
- Log in to your Outlook.com account. A permissions dialog box will appear.



- Click **Yes** if you give permission for WordStat to access your account and an **Import Emails** dialog box will appear.



- The items in the **Folders** list box on the left side of the dialog box correspond to the folders in your email account. Check the **Folders** you wish to import.
- The items in the **Variables** list box on the right side of the dialog box will become project variables if imported. Check the **Variables** you want to include in your project.
- Click **OK**.

- Enter the project name and save it in the location of your choosing. The data will be imported and WordStat's **Data** tab will appear, containing a table with your imported email data. From here you can further tailor your data set if necessary, or start analyzing immediately.

Related Topics:

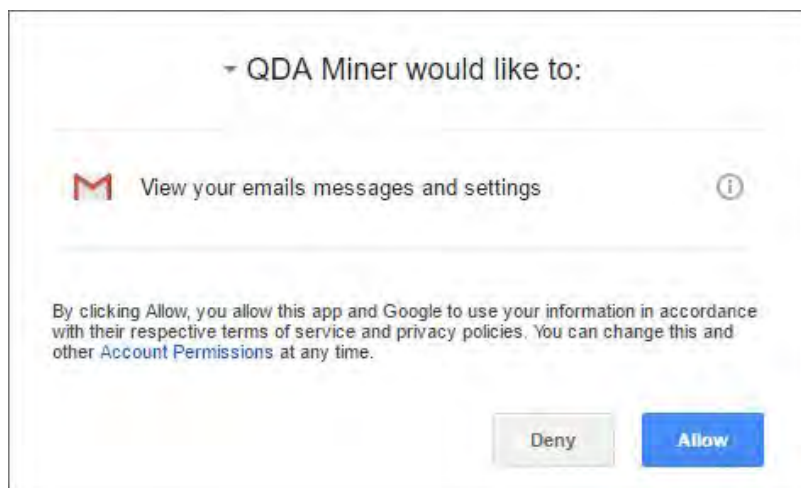
To configure the data set and start analyzing please see [The Data Tab](#).

To append new email data to your project please [Monitoring Online Resources](#).

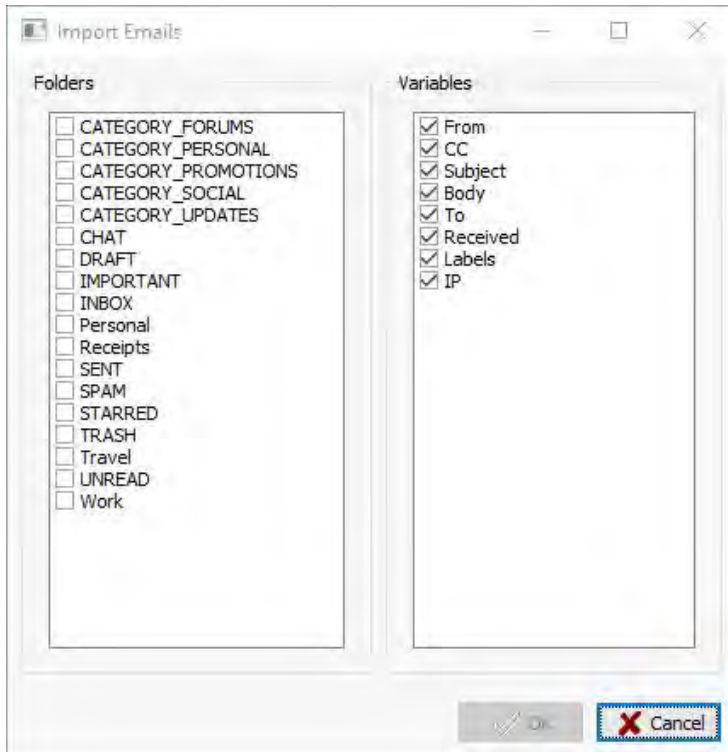
Gmail

Connecting to Gmail

- Select **E-MAILS | GMAIL** from the menu and you will be prompted to log in to your Gmail account.
- Log in to your account. A permissions dialog box will appear similar to the one below.



- Click **Allow** if you give permission for WordStat to access your Gmail account. An **Import Emails** dialog box will appear.



- The items in the **Folders** list box on the left side of the dialog box correspond to the folders in your Gmail account. Check the **Folders** you wish to import.
- The items in the **Variables** list box on the right side of the dialog box will become project variables if imported. Check the **Variables** you want to include in your project.
- Click **OK**.
- Enter the project name and save it in the location of your choosing. The data will be imported and WordStat's **Data** tab will appear, containing a table with your imported email data. From here you can further tailor your data set if necessary, or start analyzing immediately.

Related Topics:

To configure the data set and start analyzing please see [The Data Tab](#).

To append new email data to your project please [Monitoring Online Resources](#).

MBox

MailBox (*.mbox) is a file extension for a text file used to store email. It was first implemented for Fifth Edition Unix, and though not formally defined by Internet Engineering Task Force (IETF) and the Internet Society (ISOC), it is the most common format for storing email messages on a hard drive. All messages are stored in a single plain text file. Each message starts with a "From" line and ends with a blank line.

Importing Data From a MBOX File

- Choose the **E-MAILS | MBOX** from the menu and **Import data file** dialog box will appear.
- Select the MBOX file that you would like to import.
- Click **Open**.

- Enter the project name and save it in the location of your choosing. An **Import Mbox** box will appear.



- The items in the **Select Fields** list box will become project variables when imported. Select the fields you wish to import.
- Click **Import** and WordStat's **Data** tab will appear, containing a table with your imported email data. From here you can further tailor your data set if necessary, or start analyzing immediately.

Related Topics:

To configure the data set and start analyzing please see [The Data Tab](#).

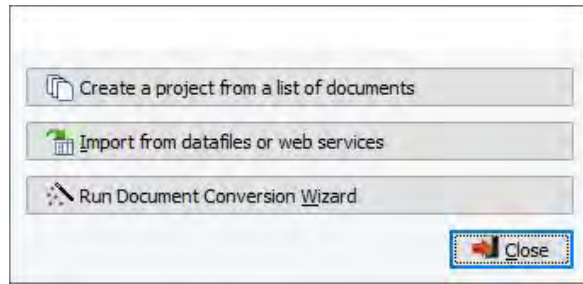
To append new email data to your project please [Monitoring Online Resources](#).

Using the Document Conversion Wizard

The Document Conversion Wizard is a utility program used to import one or more documents into a new project file. This tool supports the importation of numerous file formats including plain ASCII, Rich Text Format, MS Word, HTML, Acrobat PDF files, and WordPerfect documents. It may be instructed to split large files into several cases and to extract numeric and alphanumeric data from these files. The Document Conversion Wizard may be run either as a stand-alone application or from within WordStat. When no splitting, transformation or extraction of information from documents are necessary, you can also use the easier to use [Creating a Project from a List of Documents](#) importation method.

To run the Document Conversation Wizard from WordStat:

- Select the **New** button on the **Data** tab. A dialog box similar to the one below will appear.



- Click the **Run Document Conversion Wizard** button. The program then guides you through the necessary steps to import the documents and extract relevant data.

For more detailed information on the document importing process using this utility program, consult the Document Conversion Wizard help file.

Starting WordStat Document Explorer from Windows Explorer

WordStat Document Explorer can be started directly from Windows Explorer. WordStat Document Explorer, a pared-down version of WordStat, allows you to explore your documents, listing word and phrase frequencies and associated text to quickly reveal themes present in the selected documents. You can also apply a previously saved categorization or classification model when exploring documents.

The screenshot shows the WordStat Document Explorer window with the following components:

- Top Bar:** 19 documents, Documents, Model, Help.
- Left Panel (Frequencies):** Data, Frequencies, Documents. Sub-tabs: Words, Phrases. Table showing word frequencies and TF-I values.
- Middle Panel (Word Cloud):** Visual representation of word frequencies with 'AMERICA' and 'AMERICAN' being the most prominent.
- Right Panel (Documents):** DOCUMENT, FREQ. Table showing document frequencies.
- Bottom Panel (Paragraphs):** Paragraphs, Document. Text view of selected paragraphs.

	FREQ	% SHOWN	% WORDS	CASES	% CASES	TF-I
AMERICA	152	7.0%	0.6%	11	84.62	11.0
AMERICAN	115	5.3%	0.4%	11	84.62	8.3
PEOPLE	98	4.5%	0.4%	11	84.62	7.1
CHINA	53	2.4%	0.2%	6	46.15	17.4
FREEDOM	53	2.4%	0.2%	8	61.54	11.0
GREAT	53	2.4%	0.2%	10	76.92	6.0
CENTURY	51	2.3%	0.2%	9	69.23	8.1
COUNTRY	50	2.3%	0.2%	10	76.92	5.7
WAR	46	2.1%	0.2%	6	46.15	15.0
PEACE	44	2.0%	0.2%	6	46.15	14.0
GOVERNMENT	42	1.9%	0.2%	8	61.54	8.5
INTERESTS	41	1.9%	0.1%	8	61.54	8.6
AMERICANS	40	1.8%	0.1%	8	61.54	8.4

DOCUMENT	FREQ
McCain - Foreign Policy	17
Bush - Foreign Policy	10
Gore - Foreign Policy	5
McCain - Announcement	4
Bush - Announcement	2
Buchanan - Announcement	1
Forbes - Announcement	1
Bradley - Foreign Policy	1

WordStat Document Explorer

WordStat Document Explorer contains three or four tabs. The **Data** tab lists the files you have imported. On the top left of the **Frequencies** tab, WordStat Document Explorer allows you to view the most frequent words and phrases or categories (if a categorization module has been applied) in your selected documents. Below the frequencies list in the left panel is an interactive word cloud. A list of documents containing the selected words, phrases or categories appears in the middle panel. On the right you can toggle between viewing the paragraphs or full documents containing the selected word, phrase or category. The **Document** tab contains the word or category count per document. The **Classification** tab, only visible if you have run a classification model, lists the selected documents classified according to the chosen model.

WordStat Document Explorer with a Categorization Model

WordStat allows you to save and publish categorization models so you can analyze new projects with a previously created model. WordStat Document Explorer will apply the categorization dictionary, exclusion list and substitution list etc. contained in the chosen model, allowing you to see the most frequent categories, the documents containing the words and phrases

and rules in those categories, and the associated text. A categorization model can be chosen when opening WordStat Document Explorer from the Windows Explorer menu or can be applied while WordStat Document Explorer is already open. For more information please see [Categorization](#).

WordStat Document Explorer with a Classification Model (WordStat Document Classifier)

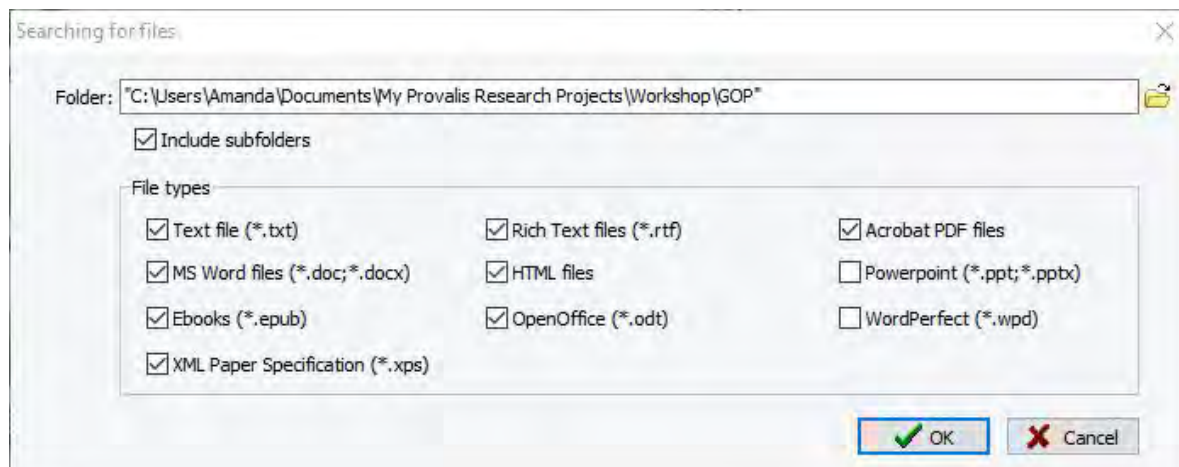
WordStat also allows you to save classification models so that you can classify documents with a previously tested model. You can see the most frequent words and phrases or categories, the documents containing the words and phrases, and the associated text on the **Frequencies** tab. A **Classification** tab has been added that lists the documents classified according to the chosen model. For more information please see [Classification](#). A classification model can be chosen when opening WordStat Document Explorer from the Windows Explorer menu or can be applied while WordStat Document Explorer is already open.

To open WordStat Document Explorer from individual documents in Windows Explorer:

- In Windows Explorer, select the individual documents you would like to explore, using the **Shift** or **Ctrl** key to select numerous documents at the same time.
- Right click your mouse, a menu opens, scroll down to **WORDSTAT CONTENT ANALYSIS** and choose **EXPLORE**, **CATEGORIZE** or **CLASSIFY** from the adjacent menu.
- If you have chosen **CATEGORIZE** or **CLASSIFY**, select from the adjacent menu one of your previously saved categorization or classification models. WordStat Document Explorer opens, pausing briefly on the **Data** tab, moving automatically to the **Frequencies** tab once the data has been processed.

To open WordStat Document Explorer from a folder in Windows Explorer:

- If you wish to analyze all documents in a folder, select the folder.
- Right-click your mouse to display the context menu.
- Scroll down to **WORDSTAT CONTENT ANALYSIS** and choose either **EXPLORE**, **CATEGORIZE** or **CLASSIFY** from the adjacent menu.
- If you have chosen **CATEGORIZE** or **CLASSIFY**, select from the adjacent menu one of your previously saved categorization or classification models.
- A dialog box will appear allowing you to choose which **file types** to import and analyze and giving you the option of including documents in **subfolders** as well.



- Once all files types have been selected, click the **OK** button. WordStat Document Explorer opens, pausing briefly on the **Data** tab, moving automatically to the **Frequencies** tab once the data has been processed.

To apply a categorization or classification model while in WordStat Explorer:

- Select the  button. A dialog box will appear containing your previously saved categorization and categorization models.
- Select the model you would like to apply and click the **Open** button.
- Select the  button.

The User Interface

The WordStat user interface is built around a multi-tab workspace that provides an integrated environment for project creation, data management, developing, transformation, editing, testing, validating, applying a content analysis dictionary and performing various text mining tasks.

[Data](#) - This tab appears only when running WordStat as a standalone application and will not be visible if you run WordStat from QDA Miner, Simstat or Stata. This is where all the data you have imported into your project is displayed. The rows in the data tab represent cases and the columns are variables. You can append and delete cases, as well as edit the data within. The structure tab allows you to view and modify your project variables.

[Text Processing](#) - This tab contains various options controlling how the text data will be processed. It also includes options affecting linguistic tools. It allows you to create and modify dictionaries, exclusion and substitution lists, as well as add, remove and edit existing dictionaries.

[Frequencies](#) – This tab displays a table of the frequency of keywords or content categories. You may also view a list of leftover words, allowing you to modify the current categorization dictionary, the exclusion list or the substitution list. There is also a suggestions tab displaying leftover words potentially related to the currently selected item and comparison graphs which visualize the relationship between the currently selected item and the chosen variables.

[Extraction](#) - This tab allows you to perform topic modeling, find the most common phrases, extract named entities, as well as misspellings and uncommon words, and assign them to the current categorization dictionary, the exclusion list or the substitution list. You may also use the misspelling tab to batch replaced misspellings in the original documents.

[Cooccurrences](#) - This tab allows you to explore connections between words, keywords, phrases or content categories using hierarchical clustering, multidimensional scaling, link analysis and proximity plots.

[Crosstab](#) – This tab allows you to compare keyword frequencies across values of numerical, categorical or date variables. Along with a table of frequencies, several statistics and graphical techniques may be applied including correspondence analysis, heatmaps, bubble charts, bar charts and line charts.

[Keyword-in-Context](#) – This tab allows you to display a concordance table of specific words, word patterns or phrases, or of all items related to a content category. This is very useful to validate a dictionary by allowing you to examine, in context, how words are being used.

[Classification](#) - This tab gives access to the automated text classification module that allows you to apply a machine-learning approach to the existing textual database. Options allow you to develop a classification model that can later be used to accurately classify uncategorized documents into predefined classes.

Note: When you are using the stand-alone version of WordStat in the Explore mode only the Data, Frequencies, Phrases, and Topics tabs are visible.

Two additional drop-down menus can be accessed to perform various tasks:

- Clicking the  button in the upper left corner of the main window displays a menu that allows you to leave WordStat or return to the calling application as well as perform various tasks such as:

[Editing documents](#)

[Editing case descriptors](#)

[Filtering cases](#)

[Geocoding](#)

[Accessing the Report Manager](#)

Program Setting

Mode

- The  button in the upper right corner provides access to this help file, which can also be accessed at any time by pressing the **F1** key. In addition, this menu allows you to check whether you are using the latest version of WordStat and also gives access to specific important information and some useful links to the Provalis Research website.

The Data Tab

The **Data** tab is visible only when you open WordStat directly. It does not appear when you call WordStat from QDA Miner, Simstat, or Stata. This tab allows you to open existing projects, modify those, or create new ones. Once your project had been created or opened, the data will be displayed on this tab. The **Data** tab allows you to append and delete cases, as well as edit the data within. You can also view and modify your project variables. Once you are satisfied with the content and structure of your data you can start your analysis.

The Data Editor Tab

Project data will be displayed in rows and columns. The rows on the **Data Editor** tab represent cases and the columns are project variables. You can append, delete and filter cases on this tab as well as identifying duplicates and changing case descriptors. You can also edit the project data directly on this tab

The screenshot shows the WordStat 9.0 interface with the 'Data Editor' tab selected. The window title is 'WordStat 9.0 - Election 2008b.pprj'. The menu bar includes 'Data', 'Text Processing', 'Frequencies', 'Extraction', 'Cooccurrences', 'Crosstab', 'Keyword-In-Context', and 'Classification'. The toolbar has buttons for 'Open', 'New', 'Save', 'Save as', 'Export', 'Properties', and 'Analyze'. On the left, the 'Data Editor' panel shows 'Structure' as the active view, with buttons for 'Append', 'Delete', 'Duplicates', 'Descriptor', 'Filter', and an 'Allow editing' checkbox. The main area displays a table with the following data:

CANDIDATE	FILE	DOCUMENT	PARTY	DELIVERY
Biden	1-Biden20060316	[DOCUMENT]	Democratic	2006
Biden	1-Biden20060719	[DOCUMENT]	Democratic	2006
Biden	1-Biden20060907	[DOCUMENT]	Democratic	2006
Biden	1-Biden20060920	[DOCUMENT]	Democratic	2006
Biden	1-Biden20061205	[DOCUMENT]	Democratic	2006
Biden	1-Biden20070110	[DOCUMENT]	Democratic	Q1-2007
Biden	1-Biden20070111	[DOCUMENT]	Democratic	Q1-2007
Biden	1-Biden20070203	[DOCUMENT]	Democratic	Q1-2007
Biden	1-Biden20070205	[DOCUMENT]	Democratic	Q1-2007
Biden	1-Biden20070215	[DOCUMENT]	Democratic	Q1-2007

Below the table, a text preview is shown for the selected document, titled 'Five Years After 9/11: Rethinking America's Future Security'. The text reads: 'Five years ago, on September 10th, 2001, standing at this podium, I argued against this administration's fixation on national missile defense. I said: "We will have diverted all that money to address the least likely threat while the real threats come into this country in the hold of a ship, or the belly of a plane, or are smuggled into a city in the middle of the night in a vial in a backpack." I wasn't clairvoyant. I was making a point that was valid then and remains valid today: when it comes to America's national security, this administration has the wrong premises and the wrong priorities. The President is right, as he put it this week: we're "a nation at war." That makes it all the more incomprehensible that, five years after 9/11, he has

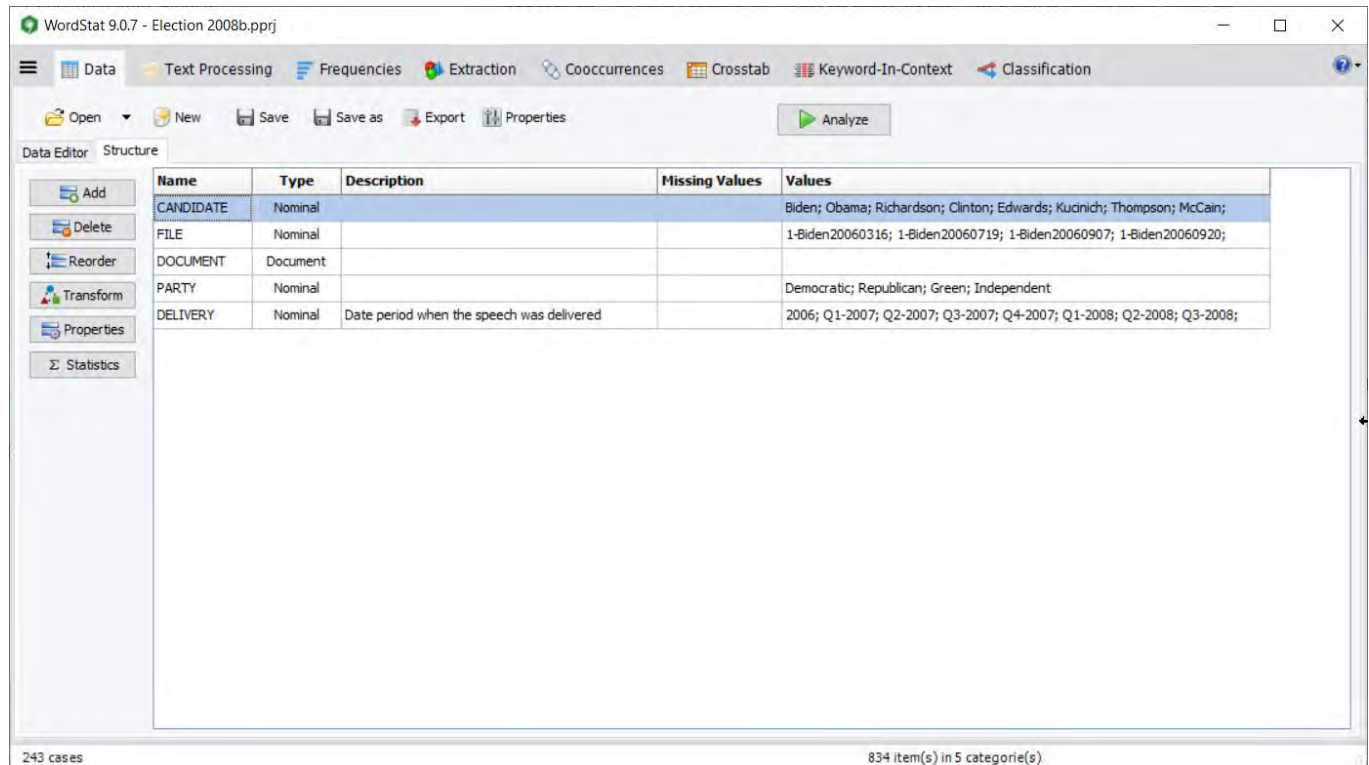
The status bar at the bottom indicates '3 / 243' and '834 item(s) in 5 categorie(s)'.

Related Topics:

- [Appending Documents](#)
- [Appending from a Data File](#)
- [Monitoring a File or Folder](#)
- [Deleting Cases](#)
- [Identifying Duplicate Cases](#)
- [Filtering Cases](#)
- [Editing Project Data](#)

The Structure Tab

The **Structure** tab displays project variables. Each row presents information about a variable. From this tab you can add, delete and transform variables as well as view variable properties and obtain the frequency and cross-frequency distribution of numerical, categorical, date and short-string variables.



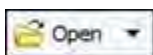
Related Topics:

[Adding New Variables](#)
[Deleting Existing Variables](#)
[Transforming Variables](#)
[Recoding Values of a Variable](#)
[Editing Variable Properties](#)
[Variable Statistics](#)

The Toolbar

Control

Description



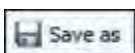
This button opens a dialog that allow you to select and open previously created projects. Click the arrow to access a list of recently opened project.



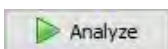
This button opens a dialog that allows you to create a new project. See [Creating a Project in WordStat](#).



This button allows you to save the currently opened project to disk.



This button opens a dialog which allows you to name your file and save the project to disk.



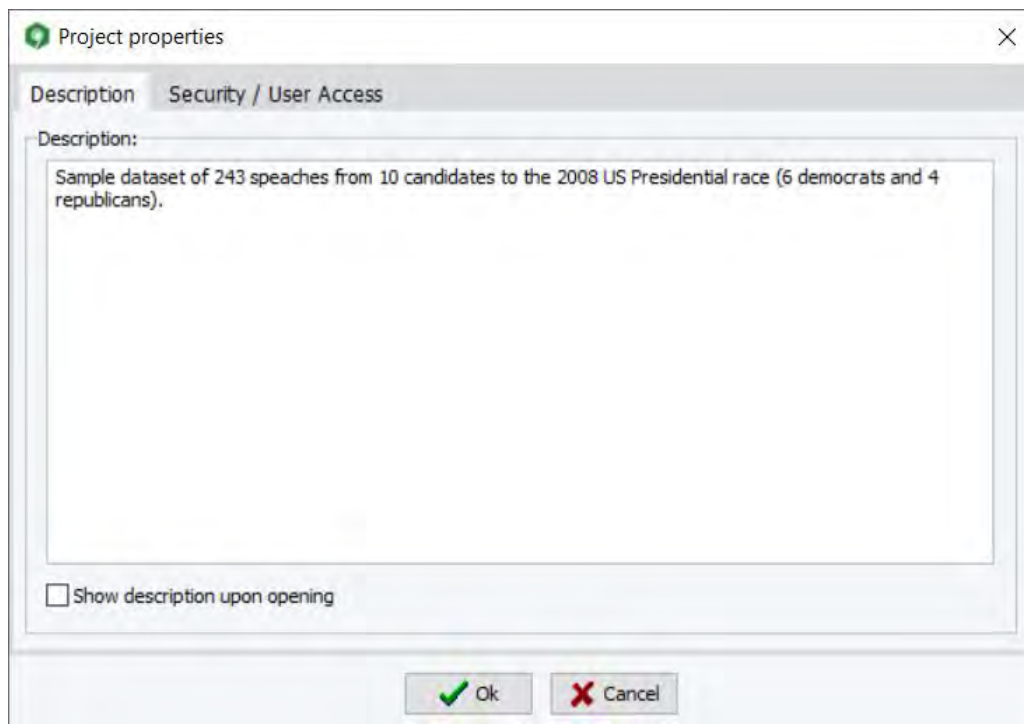
This button opens a dialog that allows you to [analyze](#) your data in both Explore mode and Expert mode.

Project Properties

The project properties function allows you to create a description of your project that may be displayed automatically when opening the project file as well as set access privileges to the project to limit the number of users accessing the data and performing specific operations.

To access the project properties function:

- Selecting the  **Properties** button on the tool bar of the data page will display a dialog box similar to the one below.



The **Project properties** dialog box contains two tabs: Description and Security/User Access.

To save or edit a project description:

- Select the **Description** tab.
- Enter or edit a project description.
- If you wish to display this description automatically when the project file is opened, select the **Show description upon opening** checkbox.
- Click the **OK** button to save the changes you made. If you choose to quit this dialog box without saving the changes, click the **Cancel** button.

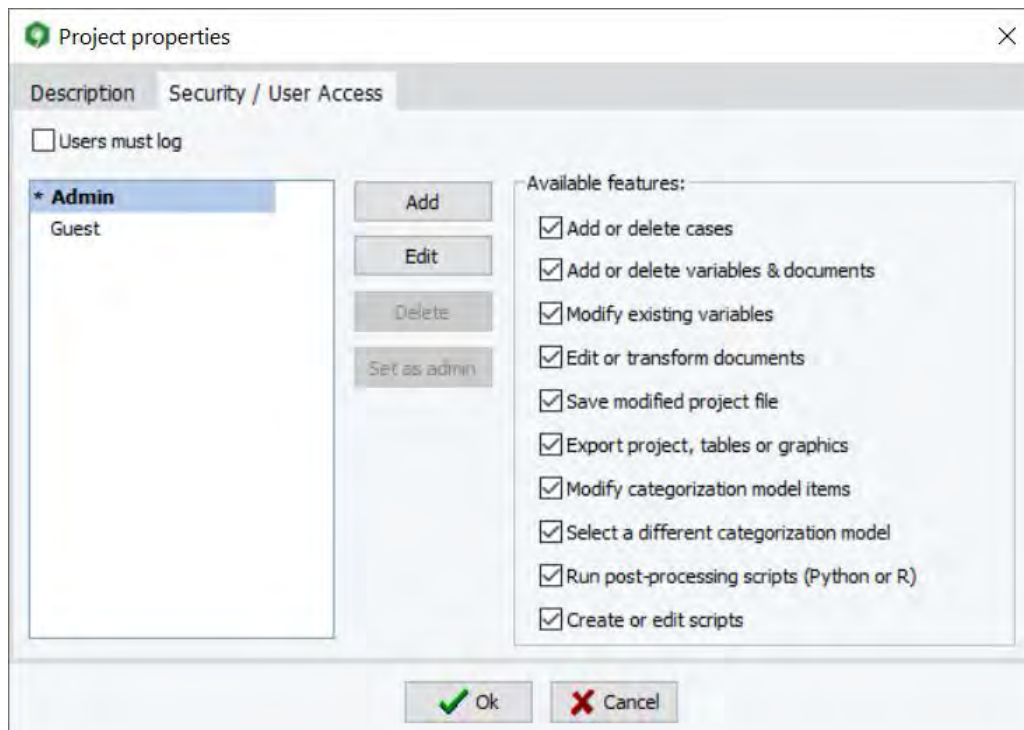
Adjusting security and user access settings

It is possible to limit who can access a specific project as well as limit the type of operations that can be performed by certain individuals by creating user accounts and requiring people to provide a username and password to access the project. This feature is useful for security or confidentiality issues, but it is also useful to prevent the deletion or editing of existing cases or variables.

When setting up a project to support multiple users, one of these users must be able to control access to the project, create and delete user accounts and define passwords. This user is commonly known as the **administrator**. When creating a new project, an administrator account is automatically created. Both the default username and password for this account is **ADMIN**. If you choose to restrict access to some users, it is highly recommended that you change this username and password to prevent unauthorized changes to user access rights.

To change the User Access settings of a project:

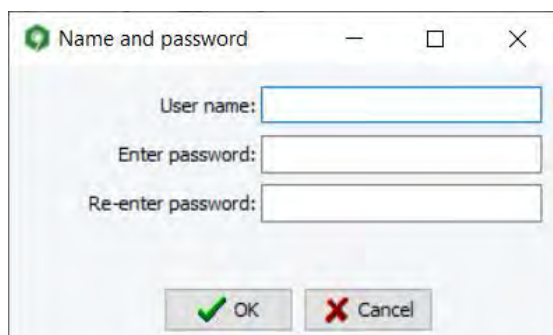
- Click the  **Properties** button on the tool bar of the data page to display the project property dialog box and then move to the **Security / User Access** page. The dialog box should look like this:



By default, opening a project without using the user log screen gives the user all administrator access rights. Enabling the **Users must log** option will display a **Name and password** dialog box prompting the user to enter a valid username and password to access the project file.

To add a new user account:

- Click the **Add** button. The following dialog box will appear:



- Enter the **Username**.

- In the **Enter password** edit box, enter this user's password.
- Enter the password a second time in the next edit box to make sure it was entered correctly.
- Click **OK** to save the new user account. Once a user's account has been created, you may specify which specific features this user will be able to access.

To define the user access rights:

- Select the user for which you want to define or edit access rights.
- In the list of available features to the right of the dialog box, select the features you want this user to have access to and clear the features that you do not want them to access.
- To prevent users from modifying existing documents or values stored in variables, clear the check box beside the **Modify variables** and **Edit or transform documents** options. It may also be a good idea to disable the **Add or delete cases** and **Add or delete variables & documents** options.
- Alternatively, you may allow the user to perform any modification they want on the data set but prevent them from saving and exporting the project.
- You may prevent the user from modifying any categorization models accessed from this project, as well force this user to use a specific categorization model by preventing him from selecting a different one.
- Despite the presence of other restrictions such as the inability to export data, create variables, etc, Python and R scripts could circumvent some of those restrictions and allow a user to export sensitive data. Removing the **Create or edit scripts** will prevent the user from editing existing pre- or post-processing scripts or creating new ones. Please note that if you disable only this option, the user will still be able to execute existing post-processing scripts. Disabling the **Run post-processing scripts** will prevent this user from running those scripts. It will also prevent the creation or editing of existing post-processing scripts, even if the **Create or edit scripts** option is enabled. However, the creation and editing of pre-processing scripts will remain possible unless one disables this feature.
- Once the restrictions have been set, Click **OK**.

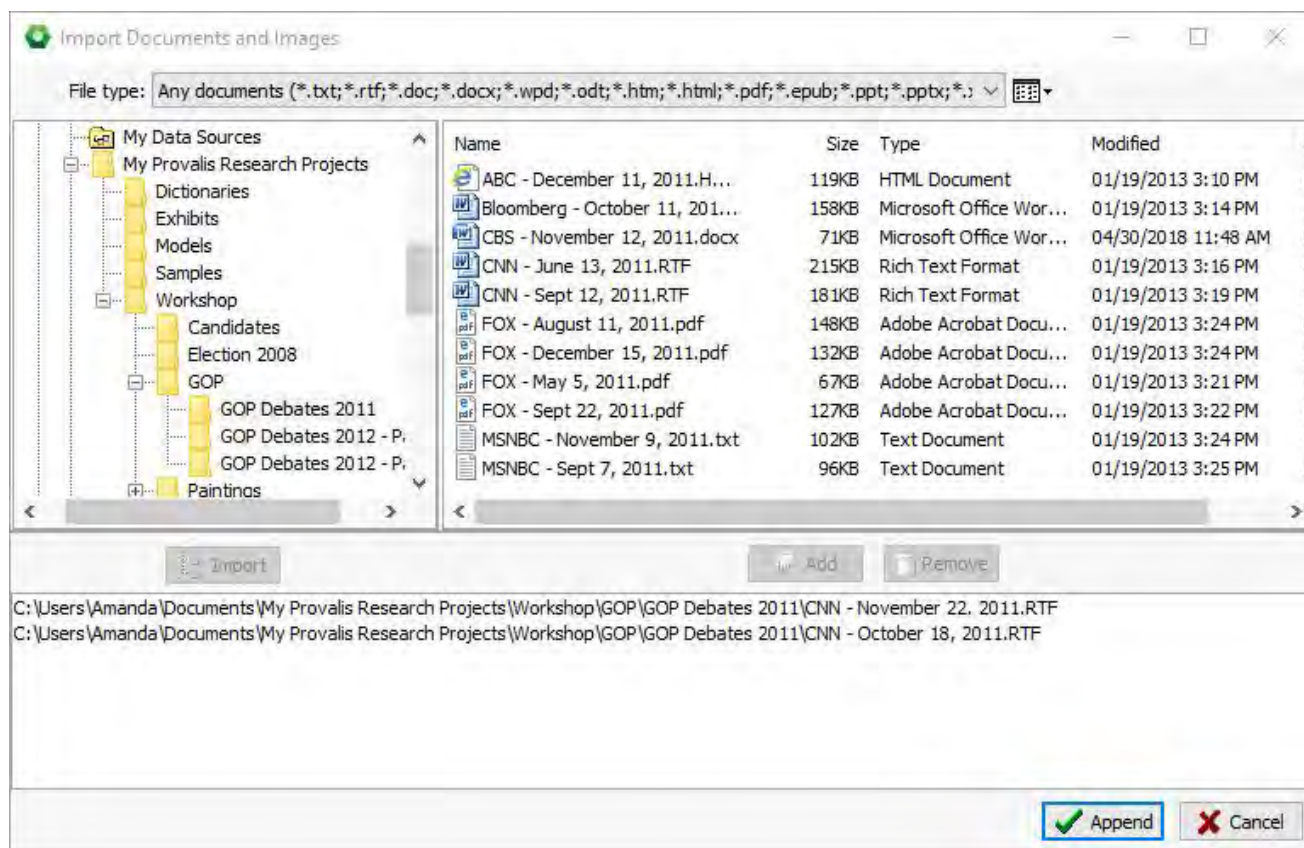
To delete a user account:

- In the list of existing accounts to the left of the dialog box, select the account you want to delete.
- Click the **Delete** button.
- To save the changes you made to the accounts and close this dialog box, click the **OK** button. To close this dialog box without saving any changes, click the **Cancel** button.

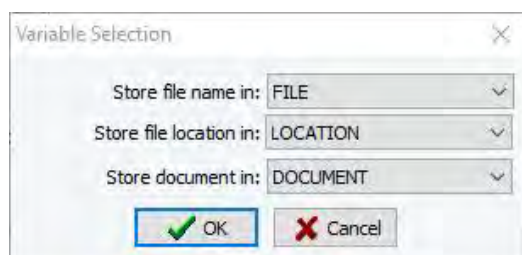
Appending Documents

To append documents and store them as new cases:

- Select the **Append** button and choose **DOCUMENTS** from the adjacent menu. A dialog box similar to the one below will appear:



- Click on a folder in the folders list on the upper left section of the dialog box to display its contents. If you want to see the contents of a drive, go to the folders list, click **My Computer**, and then double-click on a drive.
- In the upper right section of the dialog box, WordStat displays all supported document file formats that may be imported. To display only documents of a specific type, set the **File Type** list box to the desired file format.
- Click the file you would like to import. To select multiple files, hold down the **CTRL** key while clicking the other files.
- Click the **Add** button to add the files to the list of files to import, located at the bottom of this dialog box. You may also drag the files from the top right section to this list.
- Click the **Import** button to add numerous files contained within a folder. A dialog box will appear giving you the option of including **subfolders** and choosing **file types** to include in the import.
- To remove a file from the list of files to import, select that file name and click the **Remove** button.
- Once all files have been selected, click the **Append** button. If the project contains more than one categorical or document variable, a dialog box similar to this one will appear:



- Select the categorical variable in which the file name will be stored. Set this list box to **<none>** to prevent the program from storing this information in the project.

- If your project contains a LOCATION variable the **Store file location** option will be available. Select the variable in which you want the file location to be stored. This variable is created upon initially importing your documents by selecting the **Import file location** option when creating your project.
- Select the document variable where the imported documents should be stored.
- Click the **OK** button.

Related Topics:

[Creating a project from a list of documents](#)
[Importing an Existing Data File or Web Service](#)
[Appending from a Data File](#)
[Monitoring a File or Folder](#)

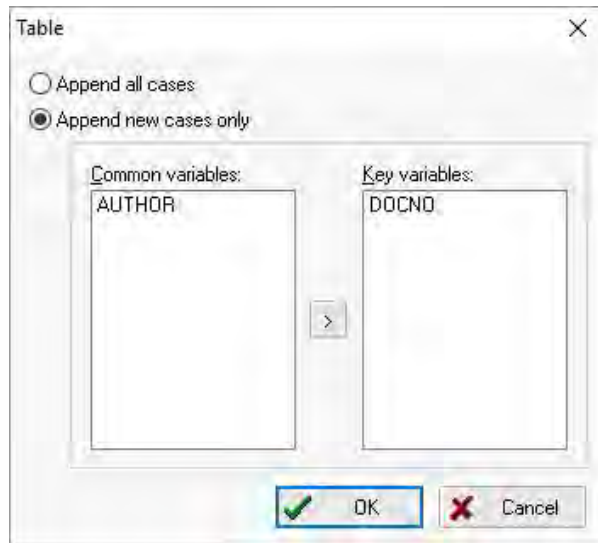
Appending from a Data File

The **Append from a data file** function allows you to append cases stored in an external data file to the current project. In order for data to be properly imported, both data files need to share variables with identical names and compatible data types (numerous type conversions are supported). If variables do not exist in the external file or if their types do not match or cannot be converted, the value of these variables will be set to missing. WordStat can import cases stored in the following file formats:

- QDA Miner projects (*.ppj)
- QDA Miner 1.0 to 4.0 project file (*.wpj)
- MS Excel spreadsheets (*.xls or *.xlsx)
- MS Access data files (*.mdb)
- Tab delimited files (*.tab)
- Comma separated value files (*.csv)
- Stata 8 to 15 (*.dta)
- SPSS (*.sav)

To append from a data file:

- Select the **Append** button and choose **FROM A DATA FILE** from the adjacent menu.
- QDA Miner will first ask you to select the data file containing the cases to be imported. It will then display a dialog box similar to this one:



If the **Append all cases** option is chosen, QDA Miner will import all cases found in this other file and store them as new cases in the current project.

If the **Append new cases only** option is chosen, QDA Miner will require the identification of one or several key variables that will be used to differentiate already existing cases that will be omitted from the append. Only new cases that will be appended to the current project.

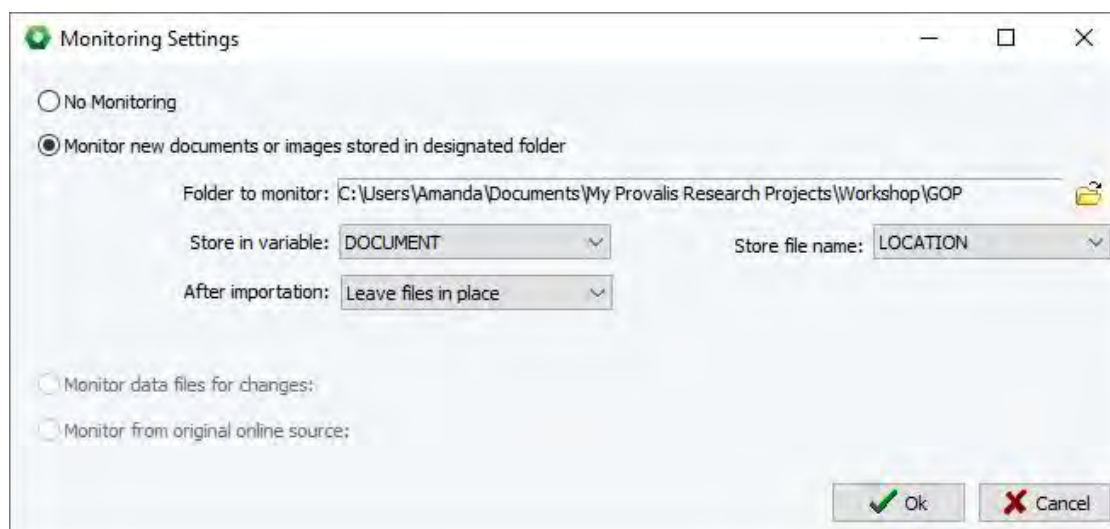
To identify a key variable:

- Select the key variable in the **Common variables** list box
- Click the  button to move it to the **Key variables** list box. If a single key variable is identified, then all cases matching existing values in the current project will be ignored while all the other ones containing new values will be imported as new cases. If several key variables are chosen, a case will be considered as existing and be ignored only if it matches values on ALL the key variables.
- Click **OK**.

Monitoring a File or Folder

QDA Miner allows you to monitor a file, folder or online service for recent additions. You can configure your project to monitor a specific folder for newly added documents and import the documents. You can also monitor changes to data files or online services such as web surveys, email accounts or reference manager tools, allowing you to import any new cases, responses, emails, Tweets etc.

- To access this function select **Append** and choose **MONITORING** from the adjacent menu. A **Monitoring Settings** dialog box similar to the one below will appear.



There are three types of monitoring that can be performed: monitoring a folder for new documents that have been added, monitoring individual data files for changes and monitoring from an online source. The options available to you are dependent on the data source of your project. Options that are not available to you will be grayed out.

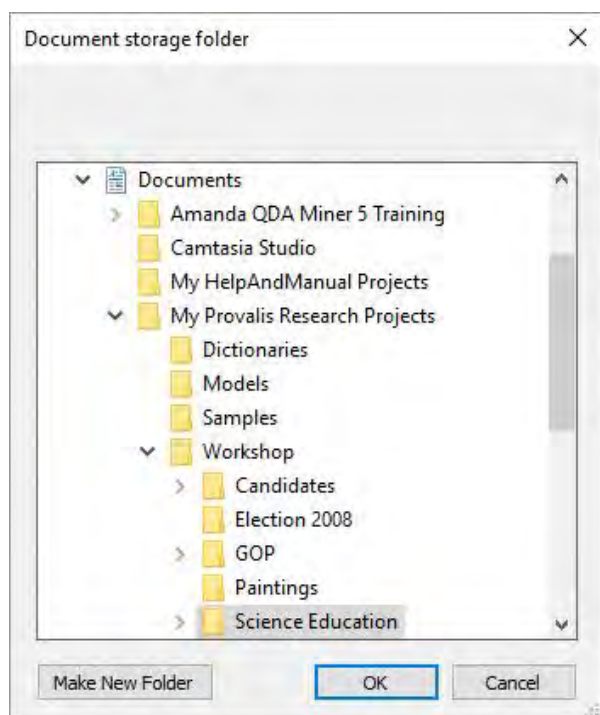
- Select the radio button that applies your project.

Monitoring a Folder

Adding relevant documents to an existing project normally requires you to open the project and perform some operations to append the newly found documents in the project. An easier way to import documents is to instruct QDA Miner to monitor a specific folder for any newly stored documents. This allows you to quickly add new documents to existing projects, by simply saving or copying the desired documents to the folder being monitored, without the need to open them or even run QDA Miner. By specifying a folder on a cloud storage service such as Dropbox, Google Drive, or OneDrive, you can even add documents to a QDA Miner project located on a remote computer.

To monitor a folder for new documents or images:

- Click on the  button to the right of the **Folder to monitor** field. A **Document storage folder** dialog box appears.



- Select the folder that you would like to monitor or click the **Make New Folder** button to create a new one, and then click **OK**
- Set the **Store In variable** list box to the variable in which you would like to store the document.
- Set the **Store file name** list box to the variable in which you would like to store the name of the file.
- Using the **After Importation** list box, choose what you would like to happen once the new files are appended. You can delete files, move files to another folder or leave files where they are once they have been appended. If you choose to move a file once it is appended another field will appear. Click on the  button and select folder to which you would like the newly appended file to be moved.
- Click **OK**. Upon reopening the project, you will be notified if any new documents have been added to the folder you are monitoring.

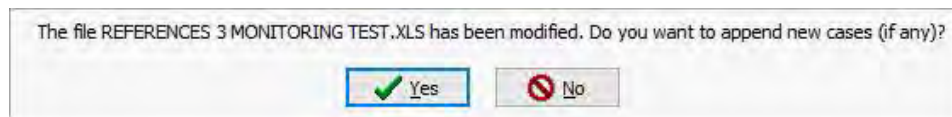
Monitoring a Data File

When a project has been created from a data file such as Excel, MS Access, Endnote, etc., QDA Miner stores the name and location of this file as well as its size and the date it was last modified. It can be set to monitor changes to the file and, if needed, import new cases from the files. When such an operation is available, the **Monitoring data file for changes** option becomes enabled and may be chosen.

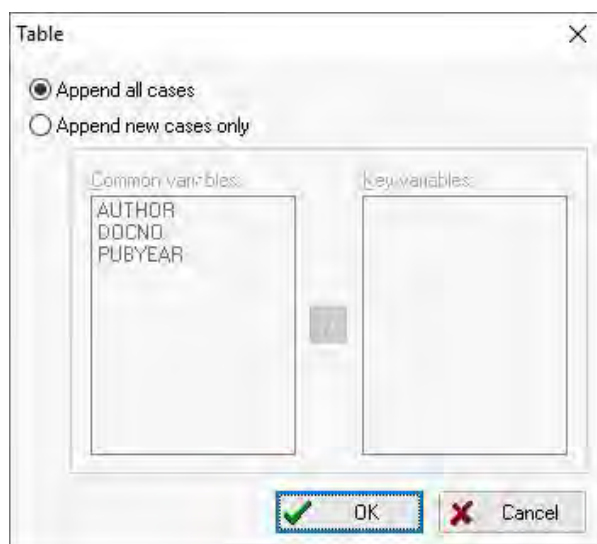
To monitor changes to individual data files:

The original source file path should appear next to this option. If the source file is not on your PC this option will not be available to you.

- Click the **Check Now** button to see if the file has been modified since you created your project or last appended from the data file. If the file has been modified a dialog box asking you if you want to append the new cases will appear.



- Select **Yes** and a **Table** dialog box will appear.



- Select whether you want to **Append all cases** or **Append new cases only**. If you choose to **append all cases**, this may result in projects containing duplicates, as some of the cases may have been imported previously. If you choose

to **append new cases only**, only new rows that were added to the data file since the last importation will be appended. If you choose to append only new cases you must choose the common variable for the cases to be matched on.

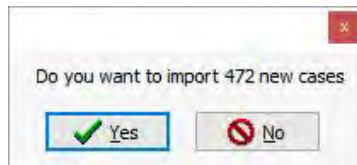
- Click **OK** and the cases will be added to your project.

Monitoring an Online Source

When a project has been created from an online source such as a web survey platform, an online reference management tool or from Email services, QDA Miner stores with the project, the connection information needed to retrieve additional cases. When this information is available, the **Monitoring original online source** option becomes enabled and may be selected.

To monitor from an online source:

- The online source will appear next to this option.
- Click the **Check Now** button to see if new data has been collected by the online source. If so, a dialog box, similar to the one below, asking if you want to append the new cases will appear.



- Select **Yes** and the cases will be appended to the project.

Note: If your online source is **Twitter**, **RSS**, **Facebook** or **Reddit** once the data has been imported the **Posts** column in the **Web Collector** for the query in question will be reset to zero. Please see the [Web Collector](#) section for more information.

For more information on projects created from online sources please see:

[Importing from Web Surveys](#)

[Importing from RSS](#)

[Importing from Twitter](#)

[Importing from email servers](#)

[Importing bibliographic references](#)

Deleting Cases

To delete the current case:

- On the **Data Editor** tab, select the case you would like to delete by clicking it.
- Select the **Delete** button on the left of the screen and choose **CURRENT CASE** from the adjacent menu.

To delete multiple cases:

- Select the **Delete** button on the left of the **Data Editor** tab and choose **MULTIPLE CASES** from the adjacent menu. A dialog box appears showing all cases in the project.



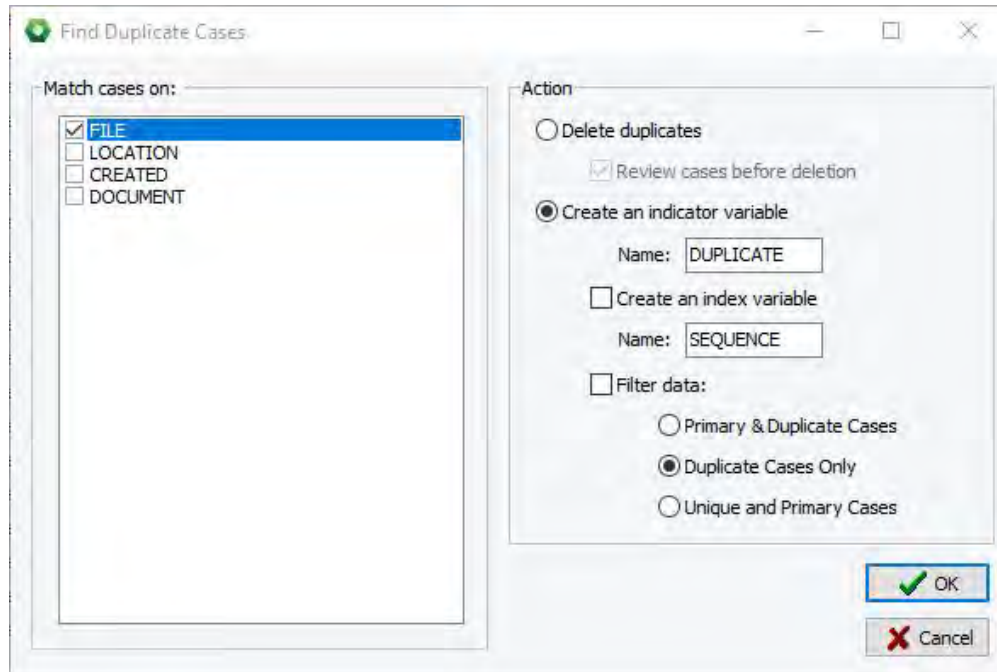
- Select the checkbox beside the cases you want to delete. By default, the current case is automatically selected.
- Once you have selected all cases you want to delete, click the **OK** button.

Identifying Duplicate Cases

Duplicate cases may occur for many reasons, including data importation, data entry or data management errors, or may naturally occur in Twitter and other online sources (ads, news lines, press releases etc.). Duplication may affect WordStat's ability to extract relevant features (topics, phrases, etc.). Use the **Duplicates** button to identify, tag, select, filter out or delete duplicate cases.

To identify duplicate cases:

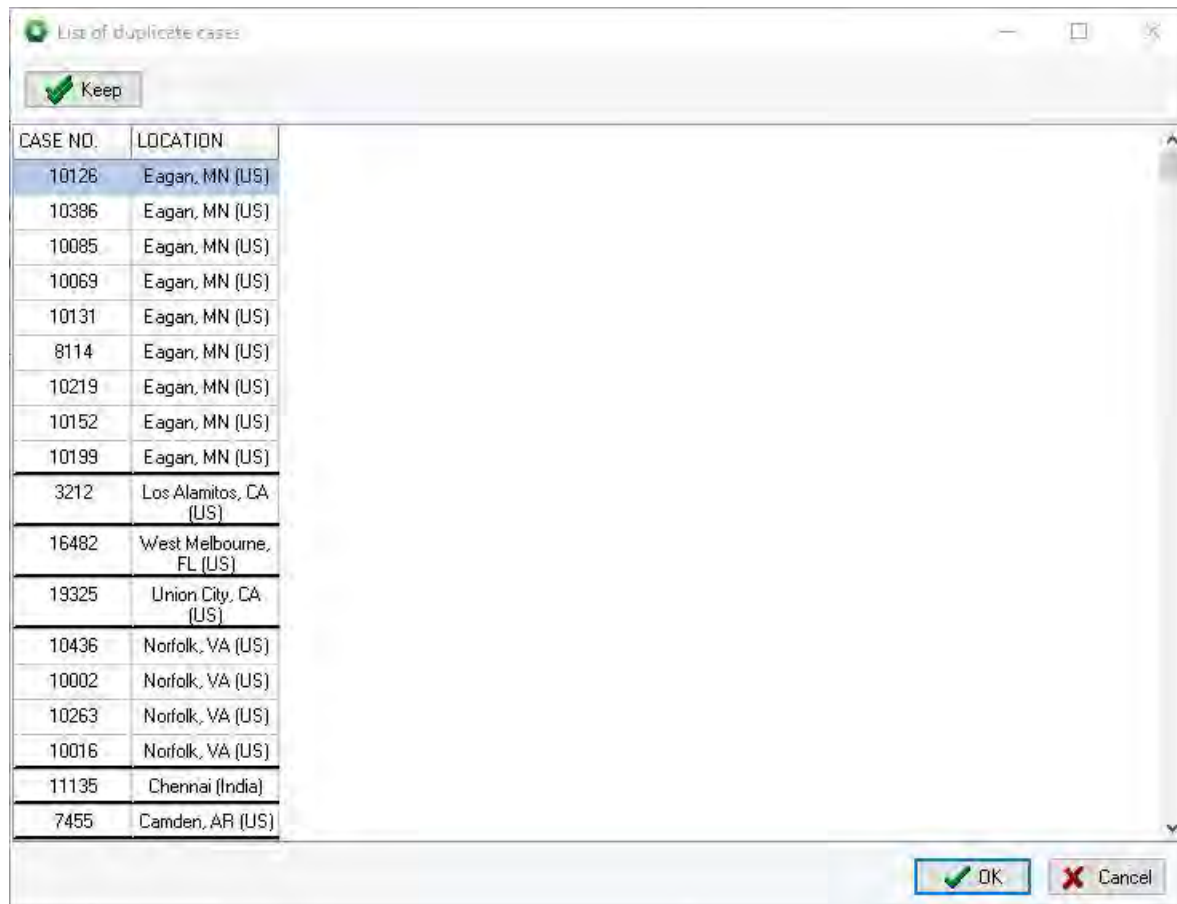
- Select the **Duplicate** button on the left of the Data Editor tab. A dialog box similar to the one below will appear.



There are two actions that you can perform. You can delete duplicate cases, with or without reviewing them first, or you can create an indicator variable that identifies duplicate, primary or unique cases. Selecting the second option also offers you possibility of creating an index variable for easy identification of primary and duplicated cases, and to filter cases based on the new indicator variable.

To delete duplicate cases:

- Select the items from the **Match cases on** list box on the left side of the dialog box. These items are the variables that are compared to determine if a case is a duplicate.
- Choose the **Delete duplicates** option from the **Action** options.
- If you wish to immediately delete duplicate cases without reviewing them click **OK**. If you wish to **review the cases before deletion** check the corresponding box and click **OK**.
- A **List of duplicate cases** dialog box, like the one below will appear.



- Select cases you wish to retain, if any, and click **Keep**.
- When you are finished reviewing the list click **OK**.
- You will be asked to confirm that you want to delete duplicate case. If so, click **Yes**.

To create an indicator variable that identifies duplicate cases:

- Select the items from the **Match cases on** checkbox list on the left side of the dialog box. These items are the variables that are compared to determine if a case is a duplicate.
- Choose the **Create an indicator variable** option in the **Action** options and enter a **Name** for your new variable. The name of this variable should be descriptive so it is easily identifiable. We have chosen **DUPLICATE** as it clearly describes the information in the variable. For each case, this variable will have a value of unique, primary or duplicate. This allows you to easily identify duplicates.

When you create an indicator variable you have the options of **Creating an index variable** and **Filtering the data**.

To create an index variable:

- Select the **Create an index variable** checkbox if you would like to assign a number to each unique case. Assigning an index variable can prove useful when using the [Case Descriptor](#) function. This will allow you to sort by sequence or group items on this variable to easily identify which cases have duplicates.
- Enter a **name** for the index variable in the corresponding field. The name of this variable should be descriptive so it is easily identifiable. We have chosen **SEQUENCE** in the example above as it clearly describes the information in the variable.

To filter cases on the indicator variable:

- Select the **filter data** checkbox to filter the data to upon the creation of an indicator variable. Choose the filtering parameters from the three options:

Primary & Duplicate Cases: Shows which cases have duplicates and how many.

Duplicate Cases Only: Allows you to see only the duplicates, without the primary cases.

Unique and Primary Cases: Filters the duplicates from the dataset.

- Click **OK**.

For more information on filtering data see [Filtering Cases](#).

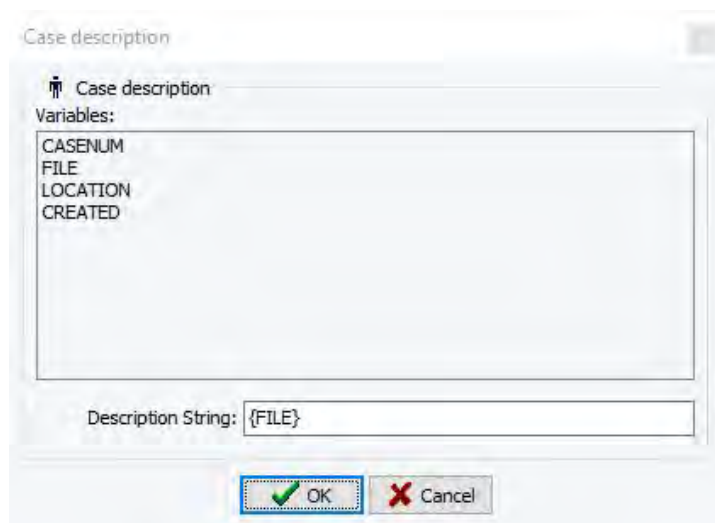
For more information on grouping and descriptors see [Setting the Case Descriptor and Grouping](#).

Setting the Case Descriptor

When retrieving results from a project using the keyword retrieval or report features of WordStat, each hit will be associated with a text description of the case it come from. For this reason, it is recommended to adjust these descriptors so that cases are easily identifiable. To build such a descriptor, you can use existing variable values as well as text strings.

To adjust the descriptors :

- Select the **Descriptor** button at the left of the **Data Editor** tab. The following dialog box will appear:



This dialog box allows you to specify a label that will be used to describe each case. The label may be changed by editing the text in the **Description String** edit box. To insert the value stored in a specific variable into the description, simply enter the variable name in uppercase letters and enclose this name between braces. Alternatively, you can insert a variable name at the current cursor location by clicking the corresponding item in the **Variables** list located just above the edit box.

If you enter the following string:

`{GENDER} subject - {AGE} years old`

The {GENDER} and {AGE} strings will be replaced with their corresponding value for this specific case. If the current case contains information about a seventeen-year-old male, the above string will be displayed as:

Male subject - 17 years old

It is also possible to insert the following string:

{CASENUM}

This string will display a unique case number, representing the physical order of this case in the project file.

Filtering Cases

The **Filter** button allows you to temporarily select cases according to logical conditions. You can use this command to restrict analysis to a subsample of cases or to temporarily exclude some subjects. The filtering condition may consist of a simple expression, or include up to four expressions joined by logical operators (AND, OR).

To filter cases:

- Select the **Filter** button on the left of the Data Editor tab. A dialog like the one below appears.

	Variable:	Operator:	Criteria:
	CANDIDATE	equals	[Bradley;Bush;Buchanan]
AND	TOPIC	equals	[Announcement]
AND			
AND			

Apply Ignore

The following table shows the various operators available for each data type

DATA TYPE	AVAILABLE OPERATORS
NOMINAL / ORDINAL	Equals Does not equal Is empty Is not empty
NUMERIC, DATE & DATETIME	Equals Does not equal Is greater than Is lesser than Is greater than or equal to Is lesser than or equal to Is empty Is not empty
BOOLEAN	Is true Is false
STRING	Contains Does not contain

	Is empty Is not empty
DOCUMENT	Is empty Is not empty Is coded Is uncoded
IMAGE	Is empty Is not empty

- Once a filtering expression has been entered, you can apply the filter and leave this dialog box by clicking the **Apply** button. If the filter expression is invalid, a message will appear and exiting from the dialog box will not occur.

To temporarily deactivate the current filter expression, click the **Ignore** button. The filter expression will be kept in memory and may be reactivated by selecting the **Filter** button again and clicking **Apply**.

To store the filtering expression, click the  button and specify the name under which the filtering options will be saved.

To retrieve a previously saved filter, click the  button and select from the displayed list the name of the filter you would like to retrieve.

To exit from the dialog box and restore the previously active filtering expression, click the **Close** button.

Editing Project Data

To edit project data:

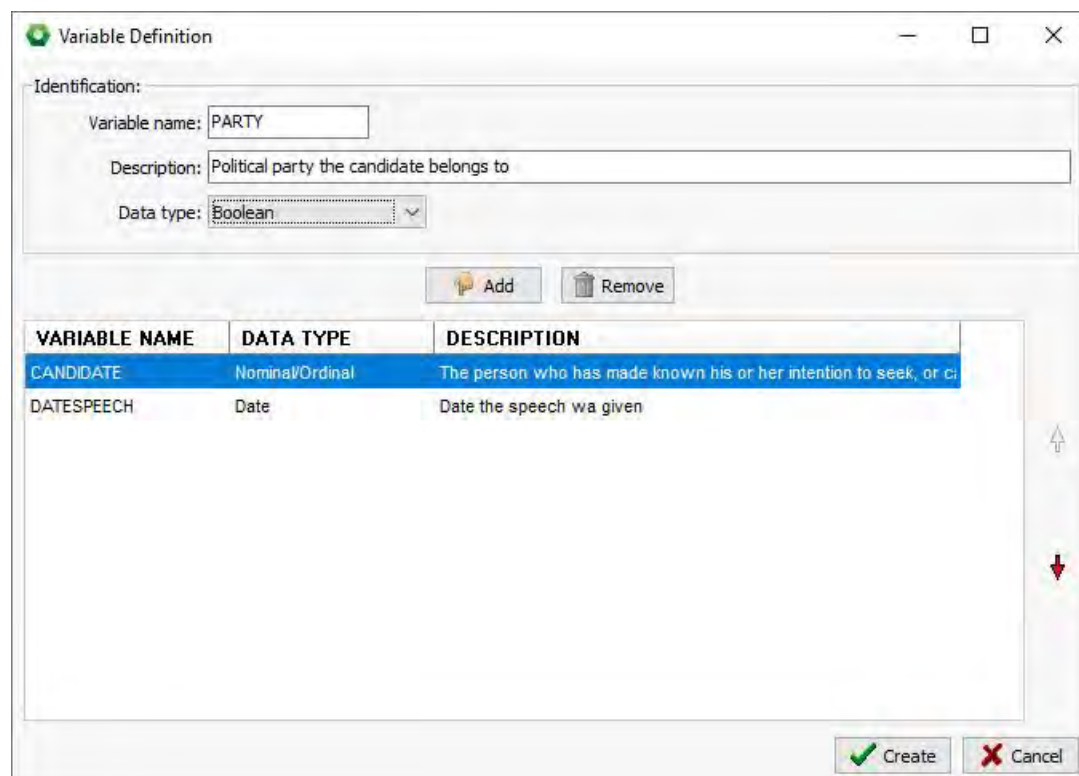
- Check the **Read only** checkbox on the left side of the screen
- Select the cell you would like to edit.
- Populate the cell with new data.
- Select the **Save** button at the top of the screen.

Adding New Variables

The **Add** button on the Structure tab is used to add new variables to the project file.

To add new variables:

- Select the **Add** button. This command displays a dialog box similar to the one below.



The Variable Definition dialog box is used to create new variables. It includes fields for variable name, description, and data type, as well as a table to manage existing variables.

Identification:

Variable name:

Description:

Data type:

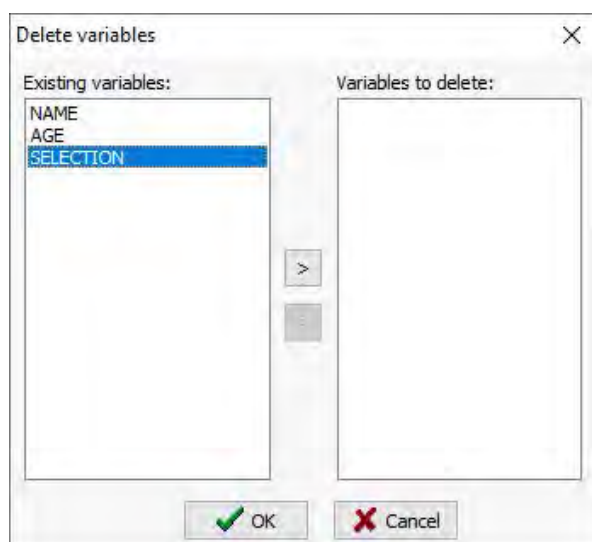
VARIABLE NAME	DATA TYPE	DESCRIPTION
CANDIDATE	Nominal/Ordinal	The person who has made known his or her intention to seek, or c
DATESPEECH	Date	Date the speech wa given

- Specify the name of the new variables and set the variable types, sizes and descriptions.
- Click the **OK** button to create these new variables and add them to the end of the current data set.

Deleting Existing Variables

To delete one or more variables:

- Select the **Delete** button on the **Structure** tab. The following dialog box will appear.



The Delete variables dialog box allows users to select variables to be removed from the dataset.

Existing variables:

- NAME
- AGE
- SELECTION

Variables to delete:

- Highlight the names of the variables that you want to delete and click the  button to move them to the **Variables to delete** list box.
- To delete successive variables, click the first variable, drag the mouse cursor down the list to highlight multiple variables, and then click the  button.
- To remove a variable from the list of variables to delete, select the name of this variable in the **Variables to delete** list box and click  the button.
- Click the **OK** button to delete all selected variables.

Reordering Variables

To reorder variables:

- Select the **Reorder** button on the **Structure** tab. The following dialog box will appear.



- Select the variable you would like to move.
- Click on the up or down arrow buttons until the variable is in the desired position.
- Click the **OK** button.

Transformation

Certain operations require specific variable types or transformations. The **Transformation** function allows you to alter variables in a number of ways, and either override the current variable or save the transformation as a new variable.

Wordstat provides numerous functions for [changing variable types](#) allowing you to change the data type of the selected variable.

The [Recode](#) function allows you to apply multiple changes to the values of numeric, categorical or string variables or to create new variables based on a new grouping of the values of an existing variable.

The [Transform Document](#) function allows you to transform document variables stored in your project using any one of the preprocessing routines available in WordStat.

The [Binning](#) function allows you group numeric or floating point variable variables into different categories of nominal variables that consists of a range of numbers.

The [Compute Number](#) command allows transformations using various functions on one or more variables. WordStat offers more than 50 operations and functions including numerical operators, trigonometric transformations (cos, sin, log, etc.), statistical functions (mean, minimum, maximum across variables or cases, etc.), date and random number operations. Conditional transformation can also be performed using an IF-THEN-ELSE logical structure.

Changing Variable Types

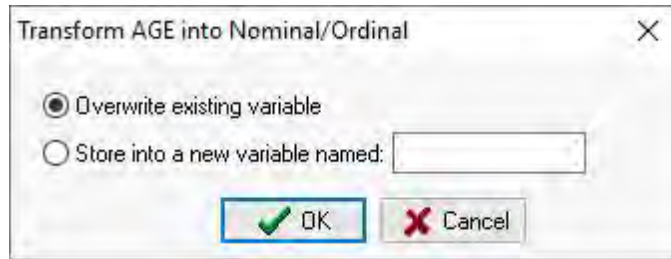
Some operations require specific variable types. For example, a change in data type may be required when you need to add decimal values to a numeric variable created as an integer. A date or a numeric variable may have been imported as a string and may need to be transformed into the correct data type for processing.

WordStat offers the ability to change the type of an existing variable or to create a new variable containing the values of the existing variable stored in a different data type. The following transformations are currently supported:

- Float -> Integer
- Float -> String
- Integer -> Float
- Integer -> String
- Integer -> Nominal
- Nominal -> String
- Nominal -> Date
- Nominal -> DateTime
- String -> Nominal
- String -> Document (codable)
- String -> Date
- String -> DateTime
- String -> Integer
- String -> Float
- Document -> String
- Date -> String
- DateTime -> String

To change the type of a specific variable:

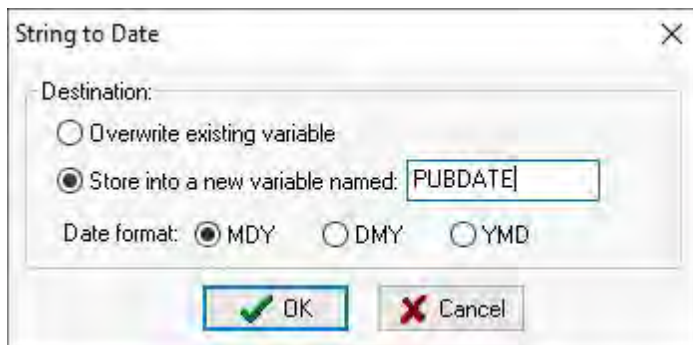
- Select the variable you want to transform by clicking on it on the **Structure** tab.
- Select the **Transform** button. The allowed transformations will be listed in a submenu.
- Choose the desired data-type transformation. For most transformations, a dialog box similar to the one shown below will appear.



- To change the type of the selected variable, choose **Overwrite existing variable** and click the **OK** button.
- To copy the values of the selected variable into a variable of the new data type, choose **Store into a new variable named**, type the name of the new variable in the edit box and click the **OK** button. If the new variable name already exists, you will be prompted to confirm the overwriting of this variable.

Transforming Strings Into Dates

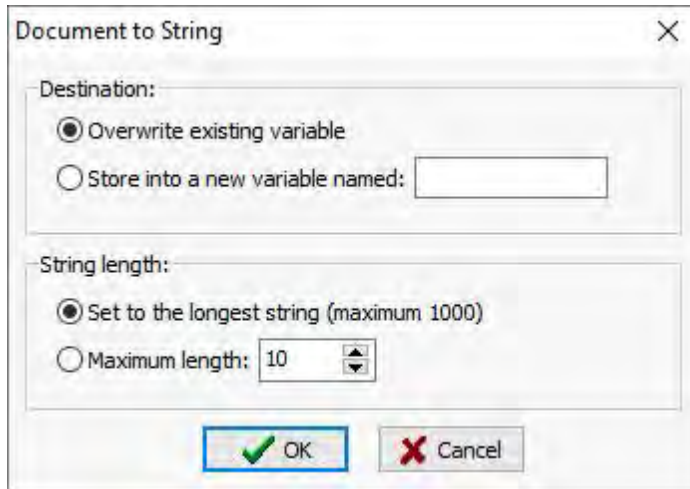
WordStat can extract date information from a string and store the date into a new variable. It can recognize in a string various date formats such as 01/05/2011, January 5th, 2011, 5 JAN 2011, or 2011-01-05 and will ignore surrounding text. When this transformation is performed, a dialog box similar to the one shown below will appear:



- To change the selected string variable into a date, choose **Overwrite existing variable**. To store the date values that are extracted from the selected variable into a new date variable, choose **Store into a new variable named** and type the name of the new variable in the edit box and then click the **OK** button. If the new variable name already exists, you will be prompted to confirm the overwriting of this variable.
- The **Date format** option allows you to choose the specific sequence that is used for displaying dates. QDA Miner will recognize only one date sequence format at a time, either month/day/ year (MDY), day/month/year (DMY) or year/month/day (YMD). Select the sequence used in the most common date format. If no date with a specific format is found for a case, the date variable will remain empty.

Transforming Documents Into Strings

WordStat can transform a document variable into a string with a maximum length of 1000 characters. Several situations may justify such a transformation. For example, you cannot filter cases based on the content of document variables, but you may perform such a filter on string variables. Also, case descriptors cannot include text stored in documents, while string variables may be used as part of the case descriptors. When this transformation is requested, a dialog box similar to the one shown below will appear:



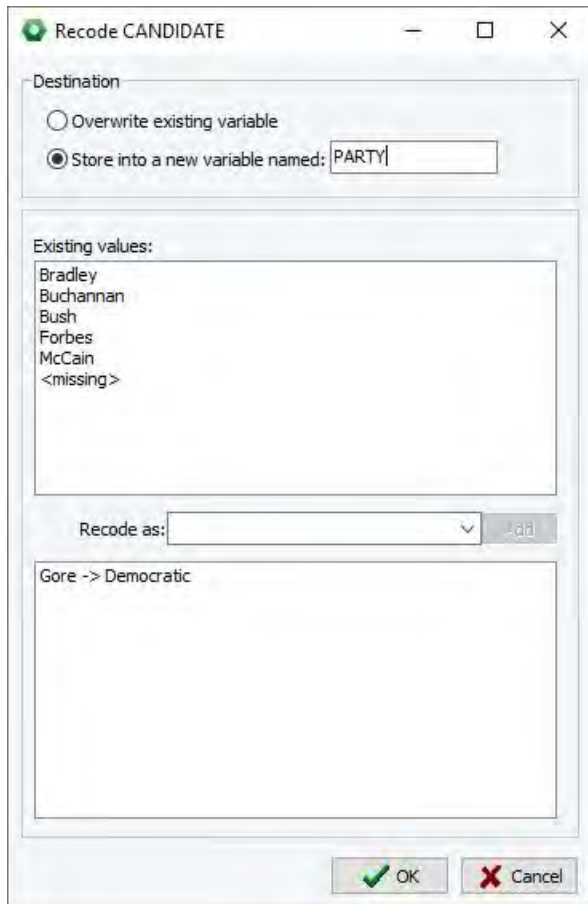
- To change the selected document variable into a string type, choose **Overwrite existing variable**. To store the date values that are extracted from the selected variable into a new date variable, choose **Store into a new variable named** and type the name of the new variable in the edit box. If the new variable name already exists, you will be prompted to confirm the overwriting of this variable.
- WordStat sets the size of the new string variable automatically by selecting the **Set to the longest string** option. WordStat will go through all cases and set the size to the longest string encountered. If a document is longer than 255 characters, the variable size will be set to 255 and the document will be truncated to this length. You can also set the size to a fixed length by setting the **Maximum length** option to the desired length. All documents longer than the set length will be truncated to this length.

Recoding Values of a Variable

The **RECODE** command, available when you select the **Transform** button, provides an easy way to apply multiple changes to the values of numeric, categorical or string variables or to create new variables based on a new grouping of the values of an existing variable.

To recode a variable:

- Select the **Transform** button and choose the **RECODE** command from the adjacent menu. A recoding dialog box will appear with the following regions.



- At the top of the dialog box the **Destination** section allows you to specify where the computation results will be stored. To transform the values of the selected variable, choose the **Overwrite existing variable** option. To keep the selected variable intact and store the result in another variable, select the **Store in a new variable named** option and type in the name of the new target variable in the edit box. If you enter the name of an existing variable, the program will ask you if you want to replace its values with those produced by the change. If the variable does not exist, the program will ask you to confirm the creation of the new variable.
- The **Existing values** list box displays the list of all values found in the selected variable. Select one or several items from this list and recode them by either typing a new value in the **Recode as** edit box or selecting another existing value from its drop-down list. To confirm the recoding, click the **Add** button. The value transformations to be performed will be listed in the recoding list box. Repeat this operation until all desired transformations have been specified. Untransformed values remaining in the **Existing values** list box will remain unchanged or will be copied to the destination variable if you selected to store values in another variable. QDA Miner can automatically detect when a nominal variable is needed and perform the variable type transformation and recoding in a single operation. An example of when this may be useful is when you recode ages to age ranges.

The special keyword **<missing>** is used to replace specific values with an empty cell and further treat those values as missing. Cases with missing values are ignored when performing some analysis involving these variables.

To remove a transformation from the list of recoding, select it and press the **<Delete>** keyboard key.

- Once a valid destination and a recoding list have been entered, perform the recoding by clicking the **OK** button.

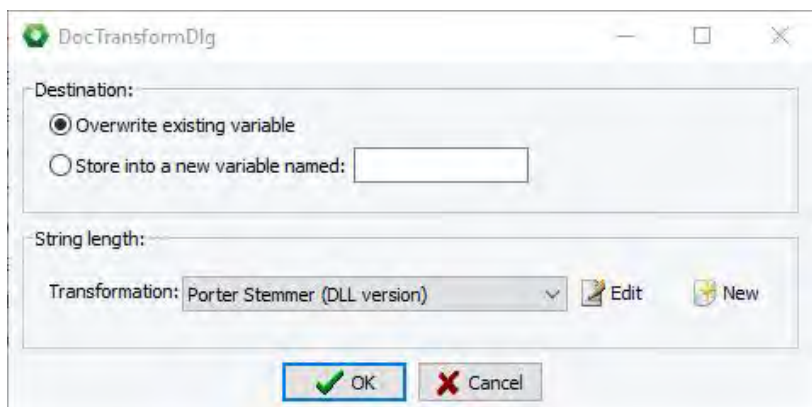
Transform Document

Preprocessing routines allow for the live transformation of text prior to the standard text analysis processes provided by WordStat (for more information on this feature see [Preprocessing](#)). Live transformation typically results in an increase in

the processing time which may be significant. For example, performing part-of-speech tagging followed by a lemmatization using the NLTK Python library could add several minutes to the analysis of a small text collection. Rather than applying this transformation repeatedly on the original documents, you can apply a transformation only once and either replace the original document with the transformed one, or store the result of the transformation into a new document variable.

The **Transform document** function allows you to transform documents stored in your project using any one of the preprocessing routines available in WordStat. To apply a transformation on a document variable:

- Select the document variable on which to apply the transformation.
- Click the **Transform** button and choose the **TRANSFORM DOCUMENT** command from the adjacent menu. A dialog box similar to this one will appear:



- The **Destination** section of this dialog box allows you to choose if you would like to overwrite the selected variable with the transformed document or to store the transformed text in a new document variable. If you choose this second option, you will need to type the name of the new variable in the edit box.
- The **Transformation** option allows to you choose an existing preprocessing routine from a drop down list. You can also **Edit** an existing routine or add a **New** one by selecting the corresponding button. For more information on creating a preprocessing routine please see [Preprocessing](#).
- Once your options have been chosen select he **OK** button.

Binning Numerical Variables

The **Binning** function allows you to "bin" or group variables into different categories. This function transforms a numeric or floating point variable to a nominal variable that consists of a range of numbers. For example, if you have a numeric variable with values ranging from 1 to 100, and you want to transform this variable into only 5 different classes, you can use this function to create 5 bins: 1 to 20, 21 to 40, 41 to 60, 61 to 80 and 81 to 100

To perform binning on a numerical variable:

- On the **Data** tab move to the **Structure** tab
- Select a numeric or floating point variable
- Select the **Transform** button.
- Scroll down to **BINNING** in the adjacent menu. A dialog box like the one below opens.

Destination

☒ Overwrite existing variable
☐ Store into a new variable named:

Binning:

Number of bins:

Method: ☒ Equal intervals Start: Increment:
☐ Equal frequencies

Boundaries: ☒ Include upper boundary (<=) ☐ Exclude upper boundary (<)

Label	Upper boundary	Frequency
0 - 40	40	22323
>40 - 80	80	1
>80 - 120	120	3
>120 - 160	160	0
>160 - 200	200	1

At the top of the dialog box the **Destination** section allows you to specify where the computation results will be stored. To transform the values of the selected variable, choose the **Overwrite existing variable** option. To keep the selected variable intact and store the result in another variable, select the **Store in a new variable named** option and type in the name of the new target variable in the edit box. If you enter the name of an existing variable, the program will ask you if you want to replace its values with those produced by the change. If the variable does not exist, the program will ask you to confirm the creation of the new variable.

The **Binning** section allows you to choose the **Number of bins** and the binning **Method**. You have two **Method** available:

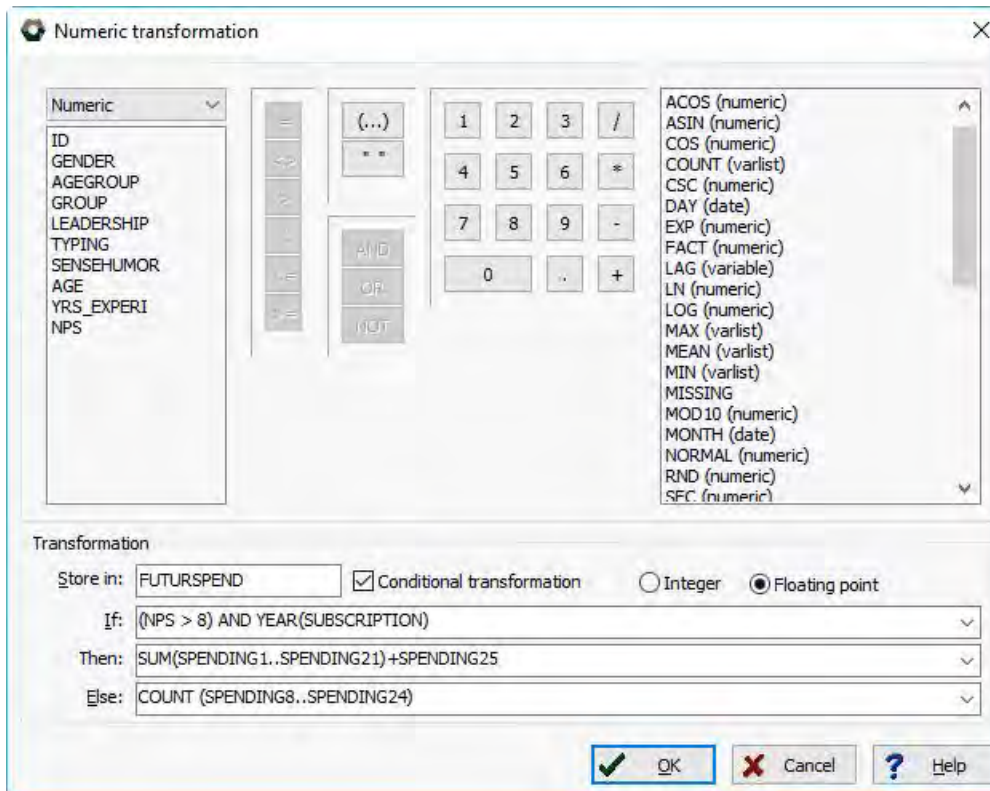
1. **Equal Intervals** means the length of the interval for each bin will be the same. In the example above, we have an interval of 20 for each bin. With this option you can set the **Start** and **Increment**. These values are automatically generated but you have the choice to change them by typing the chosen value in the field.
2. The **Equal Frequencies** option automatically calculates the intervals of each bin and attempt to produce bin with a similar number of cases. The **Boundaries** option determines if the upper boundary is included in the bin or not.

- Decide if you want to **Overwrite the existing variable** or **Store in a new variable**. If you chose the latter option, type the name of your new variable.
- The **table** displays the currently set bins.
- Choose the number of bins.
- Set your **Method** and your **Boundaries**.

- Select the **Generate** button.
- Click **OK** to perform the transformation.

Performing Complex Numeric Transformation

The **Compute Number** command allows transformations using various functions on one or more variables. WordStat offers more than 50 operations and functions including numerical operators, trigonometric transformations (cos, sin, log, etc.), statistical functions (mean, minimum, maximum across variables or cases, etc.), date and random number operations. Conditional transformation can also be performed using an IF-THEN-ELSE logical structure. When this command is selected, a dialog box similar to this one is displayed.



Store In- This option allows you to specify the target variable where the computation results will be stored. If the target variable already exists, WordStat asks you whether you want to replace its values with those produced by the transformation. If the variable does not exist, it is appended at the right end of the data file. The new numerical variable can be either an **Integer** or a **floating point** numerical variable. If the computed numerical values contains decimal and is stored into a integer data type, the value is rounded to the nearest integer. The TRUNC function (see below) may however be used to store only the integer part of the computed numerical value.

Conditional transformation - This option allows you to select whether the same transformation will be performed on all active records or whether different transformations should be performed on specific records according to a logical condition.

Formula - When the **Conditional Transformation** option is deactivated, this edit box is used to store the transformation expression that will be applied to all active records.

If - When a conditional transformation is requested, this option contains a logical expression up to 250 characters long specifying criteria for cases on which to apply a specific transformation. This expression must be evaluated as true or false. It can be a simple logical expression including the following 3 components: a variable name, a relational operator, and a variable name or a numeric constant. It can also be a complex expression composed of many simple expressions related with logical operators (AND, OR, XOR). Parentheses can be used to specify the order in which expression are evaluated. In the following example, the last two conditions are assessed before the first one.

(GROUP = 1) OR ((GROUP = 2) AND (AGE > 30))

For a list of all available functions and operations available for logical expressions see xBase functions.

Then / Else - If the expression is true the transformation expression in the **THEN** field is computed, if not the expression in the **ELSE** field is used.

Expression Operators and Rules

ARITHMETIC OPERATOR

+	Addition
-	Subtraction
*	Multiplication
/	Division
^	Exponentiation

CONSTANT

PI	3.1415926535897932385
----	-----------------------

MISSING VALUE FUNCTIONS

SYSMIS	Blank cells
MISSING	Variable missing values

NUMERIC AND TRIGONOMETRIC FUNCTIONS

Syntax FUNCTION (value, variable or expression)

ABS	Absolute
ACOS	ArcCosine
ASIN	ArcSine
ATAN	Arctangent
CSC	Cosecant
COS	Cosine
EXP	Exponential
FACT	Factorial
LN	Natural logarithm
LOG	Base-10 logarithm
MOD10	Modulus
RND	Round
SEC	Secant
SQRT	Square root
SQR	Square
SIN	Sine
TAN	Tangent
TRUNC	Truncate

STATISTICAL FUNCTIONS (ACROSS VARIABLES)

Syntax FUNCTION (Var [Var Var..Var])

MEAN	Mean
COUNT	Count (missing value excluded)
SUM	Sum

STDEV	Standard deviation
VAR	Variance
MIN	Minimum
MAX	Maximum

STATISTICAL FUNCTIONS (ACROSS CASES)

Syntax FUNCTION (Variable)

VMEAN	Mean
VCOUNT	Count
VSUM	Sum
VSTDEV	Standard deviation
VVAR	Variance
VMIN	Minimum
VMAX	Maximum
ZSCORE	Normalized score
LAG	Lag

RANDOM NUMBER FUNCTIONS

Syntax FUNCTION (value, variable or expression)

NORMAL	Normal pseudo-random number with mean of 0 and standard deviation of X
UNIFORM	Uniform pseudo-random number between 0 and X

DATE FUNCTIONS

YRMODA	(yy,mm,dd) convert 3 values, variables or expressions into a julian date
YEAR	(julian date) return the year of a julian date
MONTH	(julian date) return the month of a julian date
DAY	(julian date) return the day of a julian date
TODAY	return current julian date

xBase functions

ALIAS()

Returns the Alias name of the current workarea as a string.

ALLTRIM(String)

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

ALLTRIM(" Provalis ") 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).

AT("gh", "defghij") returns 4.

CHR(Val)

Converts a decimal value to its ASCII equivalent.

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.

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(Date)

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

DESCEND(DTOS(billdate)) + CUSTNO

DTOC(Date)

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(Date)

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.

IF(DATE() - CTOD("12/31/93") > 0, "This Year", "Last Year")

IIF(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(Date)

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("Scott", 9, "-") + ']' returns '[-Scott-]'.

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("Scott", 8, "***") + ']' returns '['***Scott]'.

'[' + PADL("Loren Scott", 8, " ") + ']' returns '[Loren 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("Scott", 8, " ") + ']' returns '[Scott]'.

'[' + PADR("Loren Scott", 8, " ") + ']' returns '[Loren 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 workarea 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 decimals 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 decimals. 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.

`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..

`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:

- B** left justify;
- C** display CR after positive numbers;
- X** display DR after negative numbers;
- Z** blank a zero value;
- (** encloses negative numbers in parentheses.

Function string characters allowed for strings are:

- R** inserts unassigned template characters;
- !** converts all alpha characters to upper case.

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(Date)

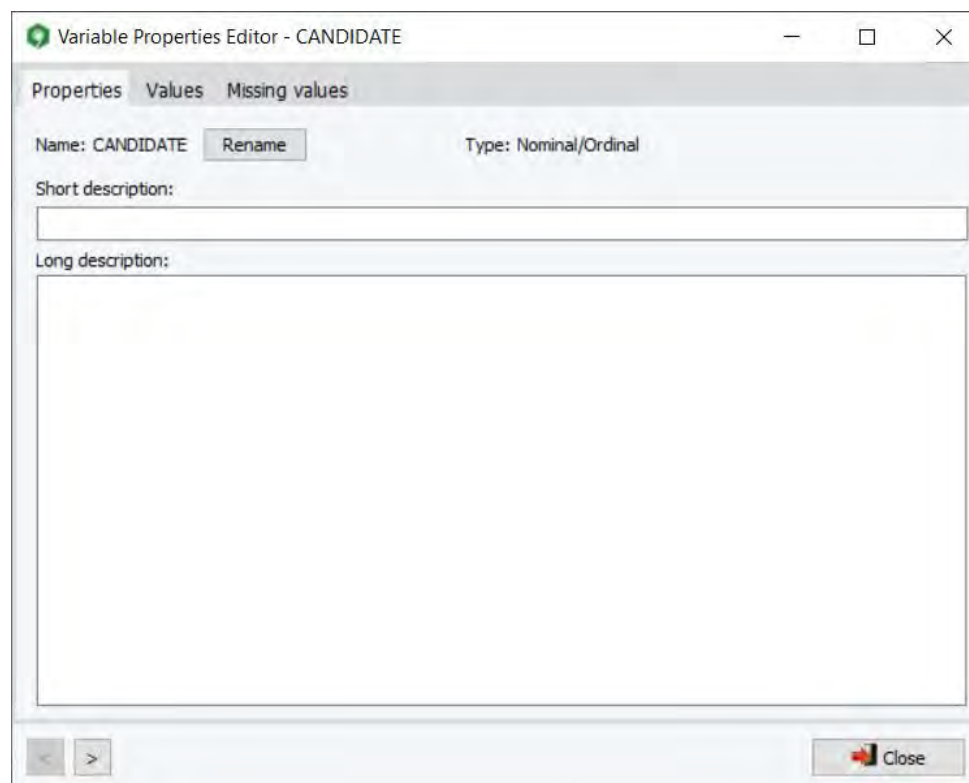
Returns the year portion of an xBase date as an integer.

Editing Variable Properties

WordStat allows you to edit various properties of existing variables in your project. For example, you can attach a short and a long description to a variable, set the number of decimal places for floating-point numerical values, edit values of categorical variables, etc. You can also set an individual variable as "read only" or change its name.

To access the Variable Properties Editor dialog box:

- On the **Structure** tab, position the cursor on the variable that you want to edit.
- Click the **Properties** button. This displays the **Variable Properties** dialog box as shown below.



Once in the dialog box, you can move through variables by clicking either the  or the  button. When viewing or editing the properties of a categorical variable, a second tab appears. You can add new values, as well as edit or delete existing ones on this page (see below).

The Properties Tab

The **Properties** tab of the dialog box offers the following options:

Read only: When selected, this option prevents a variable from being modified. This option is useful to prevent accidental or unauthorized changes to the values of a variable. Setting a variable to "read only" only affects the editing of values in existing records, but still allows users to create new cases and assign values to these new cases.

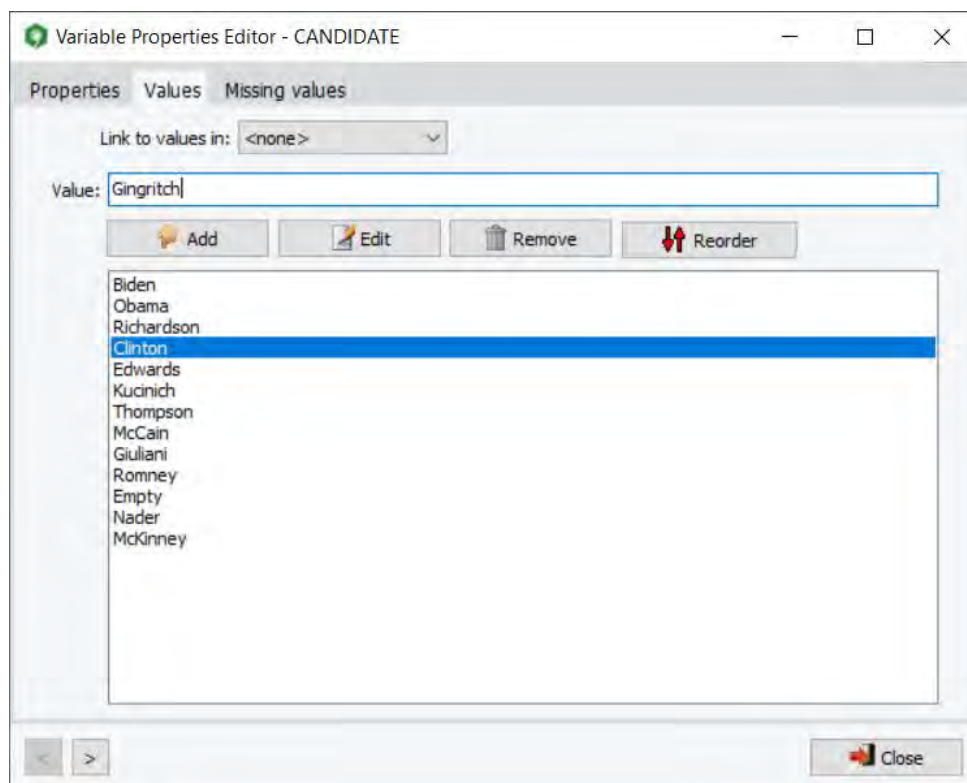
Type: The type of variable is shown here. If it is a string variable, the number of characters permitted appears beside the variable type in brackets.

Description: This option lets you enter both a short single-line alphanumeric description as well as a detailed description of the variable. The short description is displayed in various locations in the program to remind users of the exact content of this variable.

Rename button: You can use this button to change the name of the current variable. When you click this button, you will be prompted for a new variable name. This new name must not exist in the current data file and should follow the basic rules for valid variable names.

The Values Tab

When viewing the properties of a categorical variable, a second tab is shown. The **Values** tab allows you to add new values, as well as edit or delete existing ones. The **Values** tab allows you to add new values, as well as edit or delete existing ones.



To add a new value labels:

- In the **Value** edit box, enter the string that will be used to describe this new value.
- Click the **Add** button.

To remove an existing value label:

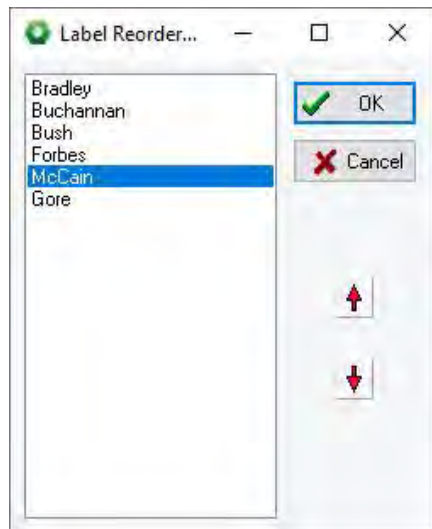
- In the list box located in the lower half of this page, select the label you want to delete.
- Click the **Remove** button

To edit an existing value label:

- In the list box located in the lower half of the page, select the label you want to edit.
- Click the **Edit** button.
- Once you have finished editing the label, click the **OK** button to apply the change.

Reordering Value Labels

Ordinal variables assume a clear ordering of values. A good example of such a variable would be responses to a satisfaction questionnaire with values ranging from “Very Unsatisfied” to “Very Satisfied”, or an age group variable containing several age ranges going from “18 or less” to “60 or more.” To display the values in a proper order in various tables and to be able to apply some of the ordinal statistics available in WordStat, values of these variables should be properly ordered. The **Reorder** button allows you to change the natural order of values of ordinal variables. When this button is clicked, a dialog box like this one will appear:



- To reorder values, simply select the value you would like to move and click the up or down arrow button until the label is in the proper order.
- To confirm the reordering, click **OK**. To return to the original ordering of values, click the **Cancel** button.

Using an Existing Value Labels Definition

In some projects, several variables share the same value labels. For example, a questionnaire may use a common ordinal scale for several questions. Rather than re-entering the same value labels over and over again, WordStat can establish a link between a variable without value labels and an existing one that already contains labels.

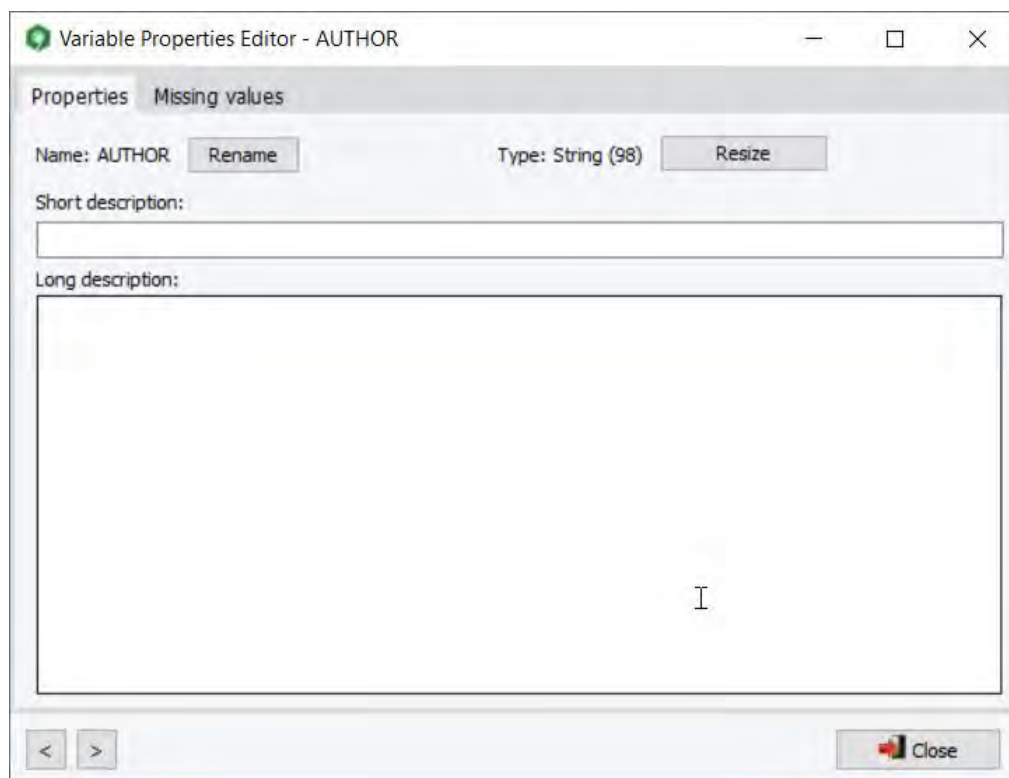
- To establish a link, click the down arrow button of the **Link to values in** list box.
- Then select from the list of variables the one containing the value labels you want to use. These labels will appear in the value list box.
- Once a link has been established, every change made to the value labels list will affect the labels associated with the original variable as well as with all other variables currently linked to this variable.

You may also use this feature to copy value labels from one variable to another. To do this, follow the previous instructions to link the current variable to the one from which you want to copy labels. The labels should appear in the value labels list. Then, remove this link by setting the **Link to values in** option to **none**. You may now edit the newly copied labels without affecting the labels of other variables.

Resizing a String Variable

When you create a string variable you are asked to determine the maximum number of characters the string can contain. If you find it necessary to modify the size of the string, you can do so by using the **Variables Properties Editor**.

- On the Structure tab choose the string variable that you want to edit.
- Click the **Properties** button. This displays the **Variable Properties** dialog box as shown below.



- The size of the string will be indicated in brackets beside the variable **Type**.
- Click the **Resize** button and a dialog box similar to the one below will appear.



- Adjust the number of characters to fit your needs using the **Up** and **Down** arrow buttons.
- Click **OK** when finished.

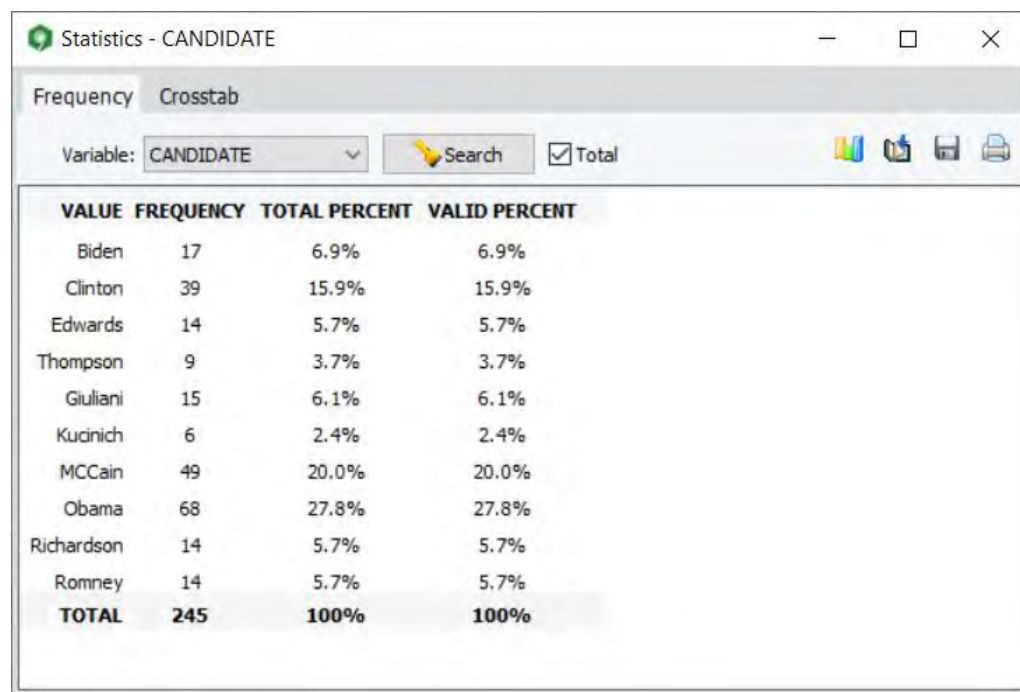
Variable Statistics

Selecting the **Statistics** button allows you to quickly obtain the frequency and cross-frequency distribution of numerical, categorical, date and short-string variables. The univariate frequency table includes the frequency count for each value of the selected variable as well as the percentage of the count over all cases and over valid cases only. The contingency table describes the distribution of two variables simultaneously by displaying either the frequency or the row, the columns or the total percentages. Several types of charts may be created to illustrate the distribution or cross-frequency distribution of variables. You will find below a list of those charts.

Frequency Distribution

To obtain the frequency distribution of a variable:

- Select the **Statistics** button on the **Structure** tab.



You may obtain a frequency table on any other numerical, categorical, alphanumeric, and date variable by selecting its name in the **Variable** list box and then clicking the **Search** button.

To chart the distribution of a variable:

Clicking the  button allows you to obtain up to five types of charts to visually display the distribution of specific codes.

Some of the charts are available only for numerical and date variables (histograms and box-&-whiskers plots), while others like the bar charts and pie charts will be available in all situations when the total number of values is less than 100.



The vertical bar chart is the default chart used to display the frequencies of distinct values of a nominal or ordinal variable. It is especially useful to compare two or more values.



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



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



The donut chart, similar to the pie chart, is useful to display the relative frequency of each value and compare individual values to other values and to the whole. Donut charts are considered easier to read as you can focus on reading the length of the arcs, rather than comparing the proportions between slices.



The histogram graphically displays the distribution of a numeric variable. When selected, the program first separates the values into non-overlapping intervals of equal width, and then plots bars that represent the frequencies of each interval.

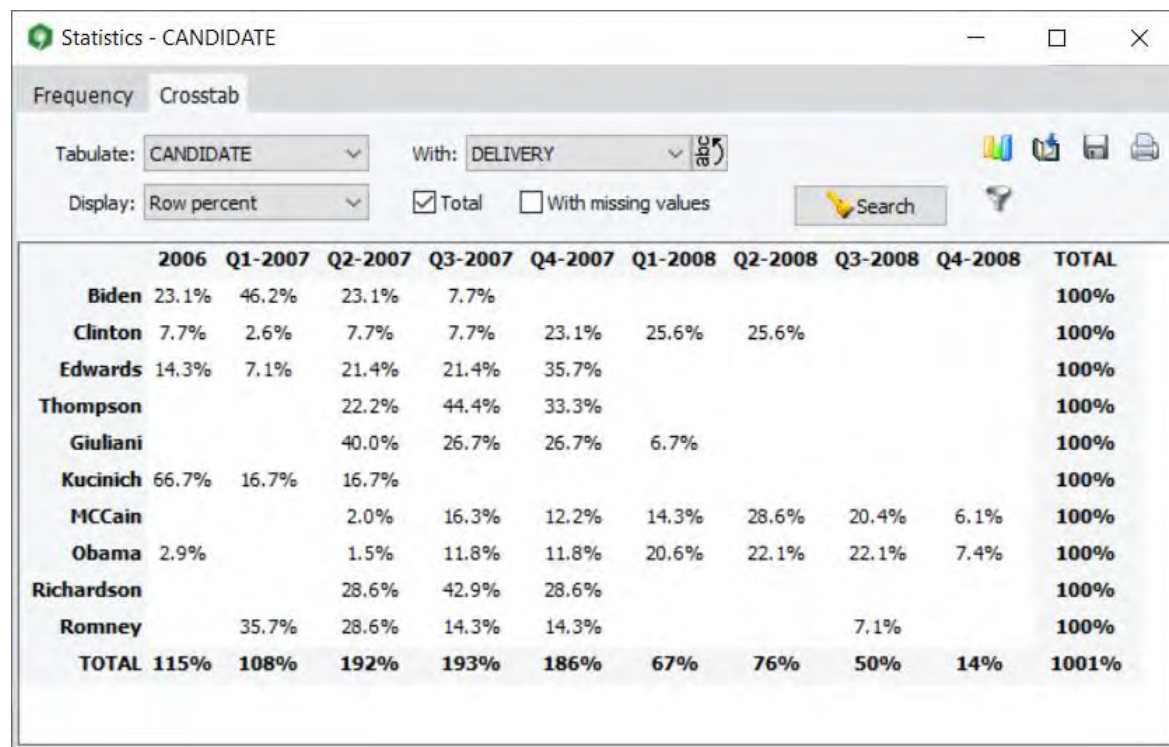


The box-&-whiskers plot can be used to examine the distribution of numerical variables. It is especially useful to detect the presence of outliers and asymmetry in the data distribution. The box includes values that fall between the first and the third quartiles (about 50 percent of the values). The line in the middle of the box represents the median value while the whiskers extend to the farthest observations within 1.5 times the interquartile range measured from the nearest quartiles. Values that are situated farther than 1.5 times the interquartile range but within three times this distance are represented by a dot, while values farther than three times the interquartile range from the nearest quartile are represented by the letter X (for extreme).

Joint Distribution of Two Variables

To obtain the joint distribution of two variables:

- Select the **Statistics** button on the Structure tab, to display the **Statistics** dialog box (see above).
- Move to the **Crosstab** tab. The dialog box will be similar to the one below:



	2006	Q1-2007	Q2-2007	Q3-2007	Q4-2007	Q1-2008	Q2-2008	Q3-2008	Q4-2008	TOTAL
Biden	23.1%	46.2%	23.1%	7.7%						100%
Clinton	7.7%	2.6%	7.7%	7.7%	23.1%	25.6%	25.6%			100%
Edwards	14.3%	7.1%	21.4%	21.4%	35.7%					100%
Thompson			22.2%	44.4%	33.3%					100%
Giuliani			40.0%	26.7%	26.7%	6.7%				100%
Kucinich	66.7%	16.7%	16.7%							100%
MCCain			2.0%	16.3%	12.2%	14.3%	28.6%	20.4%	6.1%	100%
Obama	2.9%		1.5%	11.8%	11.8%	20.6%	22.1%	22.1%	7.4%	100%
Richardson			28.6%	42.9%	28.6%					100%
Romney		35.7%	28.6%	14.3%	14.3%			7.1%		100%
TOTAL	115%	108%	192%	193%	186%	67%	76%	50%	14%	1001%

- Select from the **Tabulate** list box the variable you would like to be displayed on the rows of the table.
- Select the variable you would like to be displayed at the top of the table by selecting its name in the **With** list box.
- In the **Display** list box, select the statistics you would like to be displayed in the table.
- Click the **Search** button.

To chart the joint distribution of two variables:

Bar charts or line charts are useful for visually comparing the joint distribution of two variables. To produce these types of charts:

- Set the **Tabulate**, **With**, and **Display** options so that the information to be viewed is displayed in the table.
- Click the  button.

For more information, see [bar chart and Pie Chart](#).

To append a table to the Report Manager:

- Click the  button. A descriptive title will be provided automatically for the table. To edit this title or to enter a new one, hold down the **Shift** key while clicking this button.

For more information on the Report Manager, see the [Report Manager Feature](#) topic.

To exporting retrieved codes to disk:

- Click the  button. A **Save File** dialog box will appear.
- In the **Save as type** list box select the file format under which you would like to save the table. The following formats are supported: ASCII file (*.TXT); Tab delimited file (*.TAB); Comma delimited file (*.CSV); MS Word (*.DOC); HTML file (*.HTM; *.HTML); XML file (*.XML); Excel spreadsheet file (*.XLS; *.XLSX), and SPSS data file (*.SAV).
- Type a valid filename with the proper file extension.
- Click the **Save** button.
- The table can also be stored in a new project file, where each row becomes a new case and each column is transformed into a variable. This type of project file may be useful for performing a more detailed analysis of coded segments.

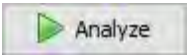
To print the table:

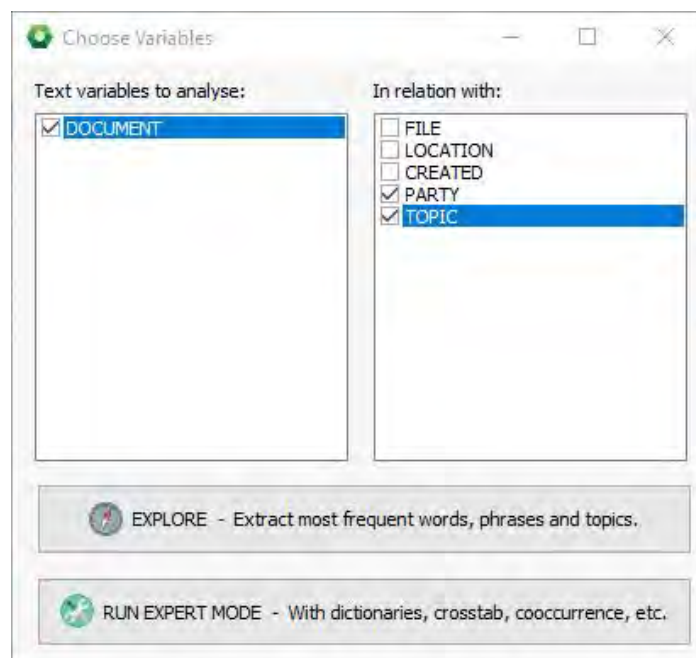
- Click the  button.

Analyze

Once you are satisfied with the content and the structure of the data you have imported into your WordStat project you can begin your analysis. Data can be analyzed in two different modes: **Explore** mode and **Expert** mode. The **Explore** mode lets you see at a glance what is in your data set. It will provide the list of the most frequent words and most frequent phrases, and will extract the most salient topics using topic modeling. Each of those results will allow quick comparisons with up to two numerical, categorical or date variables. For a more comprehensive analysis, run the **Expert** mode. This mode provides, in addition to the basic descriptive tools of the **Explore** mode, additional features such as co-occurrence analysis (clustering, multidimensional scaling, link analysis), crosstabulation (heatmap, correspondence analysis, bubble charts, etc.), machine learning, as well as all the dictionary building features of WordStat (exclusion list, categorization dictionaries, keyword-in-context, etc). It also gives you access to all processing options (stemming, lemmatization, character processing, etc.).

To begin the analysis process :

- Select the  button. A Choose Variables dialog box will appear similar to the one below:

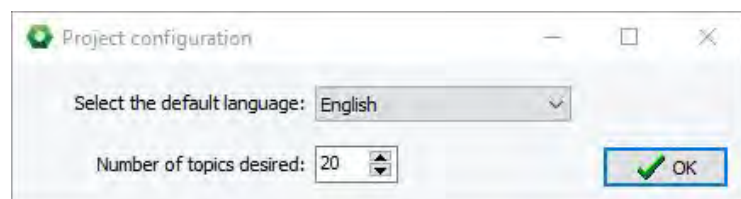


The text variables (document and string) are listed on the right side of the dialog. On the left side of the dialog box are listed the categorical, numeric or date variables that you can analyze in relation to the text variables. Two modes of analysis are available, **Explore** and **Expert**.

- Select the **text variables** by checking the boxes.
- If you want to perform comparisons with numerical, categorical or date variables, select their checkboxes in the **In relation with** panel. Please note that, by default, cases with a missing value on any of the selected variables are excluded from the text analysis. To include all cases, select the **Include cases with missing values** option on the [Preprocessing tab](#).
- Select the mode in which you would like to perform your analysis.

To analyze in Explore mode:

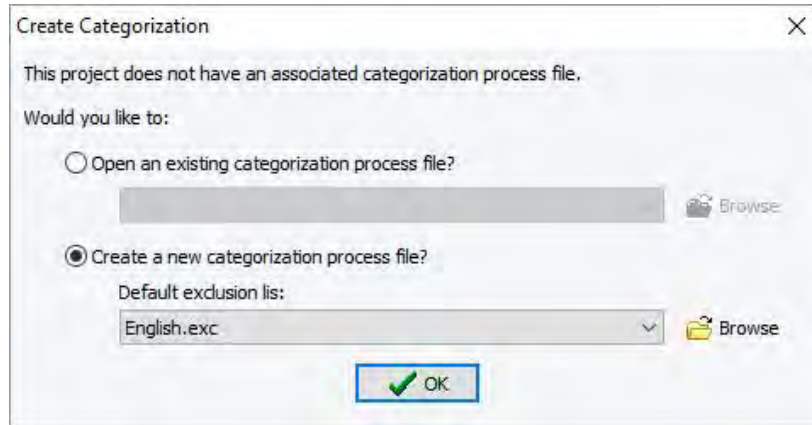
- Select **Explore**. A Project configuration dialog will appear similar to the one below.



- A default language will be chosen for you. Modify the language by choosing one from the drop-down list.
- The number of topics to be extracted will automatically be calculated for you. You can change this number by clicking on the up or down buttons beside the number field or simply by typing the desired number of topics in the field.
- Select **OK**. The word frequencies will be calculated, and phrases and topics will be extracted. Once the software is done processing you will be taken to the **Topics** tab and you can begin to explore your data.

To analyze in Expert mode:

- Select **Run Expert Mode**. If you are analyzing the project for the first time, a Create Configuration dialog box will appear similar to the one below.



Here you have the option of using an existing categorization model file or creating a new one.

To use an exiting categorization model file:

- Choose a categorization file from the drop-down menu or select **Browse** and find the categorization file you wish to use and select it.
- Select **OK**. WordStat will open and you will be taken to the **Text Processing** tab to start your analysis.

If you choose to create a new categorization model file:

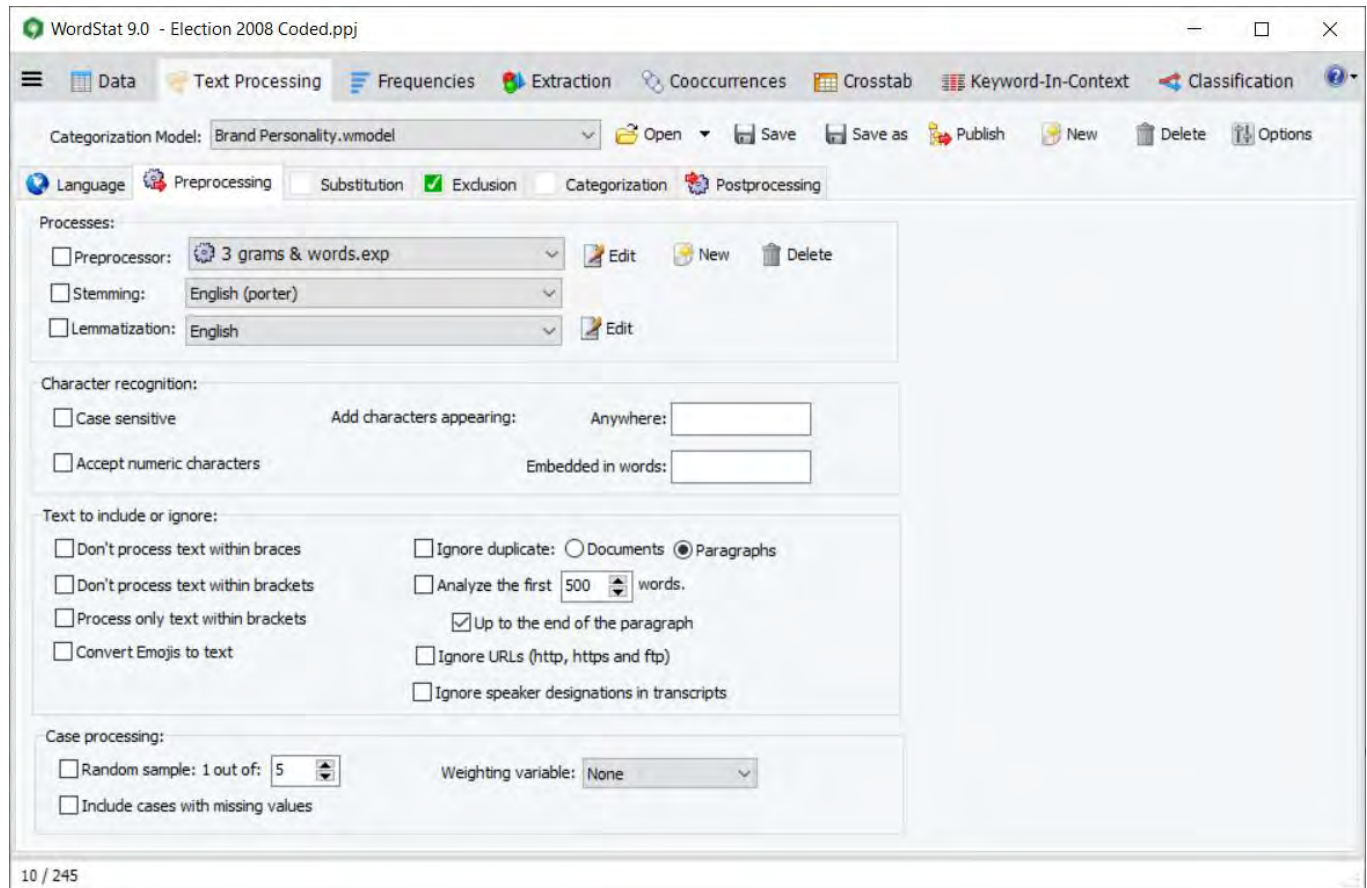
- Choose an exclusion list from the drop-down or select **Browse** and find the exclusion list you wish to use and select it.
- Select **OK**. A dialog will open.
- Type a name for your categorization model and a place to save it. Select **Save**.
- WordStat will open and you will be taken to the **Text Processing** tab to start creating your categorization dictionary, exclusion and substitution lists, and begin analyzing your data.

Related Topics:

For more information on **categorization models** please see [The Text Processing Tab](#).

The Text Processing Tab

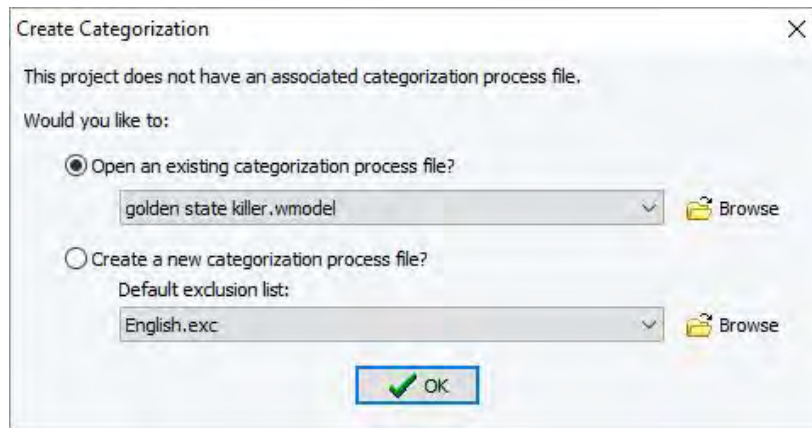
Without further information, WordStat can perform a frequency analysis on each of the words encountered in the chosen document or alphanumeric variables. However, it is also possible to apply various transformations on the words before performing the frequency analysis. The **Text Processing** tab allows you to specify how the textual information will be processed. It contains all the options related to text preprocessing (stemming, lemmatization, etc), text postprocessing (frequency criteria), as well as those involved in the creation and management of exclusion and substitution lists and categorization dictionaries.



A WordStat categorization model consist of various text processing settings and routines such as lemmatization,exclusion, categorization, etc. They are grouped under six tabs: Language, [Preprocessing](#), [Exclusion](#), [Substitution](#), [Categorization](#) and [Postprocessing](#), allowing you control how text will be processed, transformed and reported. All the settings are stored on disk in a single file with a **.WMODEL** file extension. Categorization model files are independent of WordStat projects, allowing you to apply a categorization model to several projects as well as apply, if needed, various categorization models to the same project file.

Initial Setting of the Categorization Model

While models are stored independently of projects files, a project cannot be processed without a categorization model. For this reason, when you attempt to analyze a text analysis project for the first time, you will be asked to either select an existing categorization model or create a new one. A dialog box similar to this one will appear:



To use an existing categorization model:

- Select the **Open an existing categorization process file** radio button.
- Choose a categorization file from the drop-down menu or select **Browse** and find the categorization file you wish to use and select it.
- Select **OK**. WordStat will open and you will be taken to the **Text Processing** tab to containing the categorization dictionary, exclusion list and substitution list present in the chosen model. From here you can tailor the lists to suit your current project or start analyzing your data immediately.

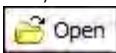
To create a new categorization model:

- Select the **Create a new categorization process file** radio button.
- Choose an exclusion list from the drop-down menu or select **Browse** and find the exclusion list you wish to use and select it. While only one exclusion (stop word) list may be selected, one may later import additional ones.
- Select **OK**. A dialog will open.
- Choose a name for your categorization model and a place to save it. Select **Save**.
- WordStat will open and you will be taken to the **Text Processing** tab to start creating your categorization dictionary, exclusion and substitution lists, and begin analyzing your data.

While categorization models and project files are independent, WordStat stores in the project file the information about the last categorization model used, so that when this project is reopen, this model will be automatically retrieved.

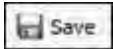
Using the Categorization Model Toolbar

The toolbar across the top of the Text Processing tab pertains to the selection, creation and publication of categorization models etc. The **Categorization Model** drop-down list box allows you to access previously saved categorization models. Select the arrow on the right and choose the desired model from the drop-down list.

By default, the models available from this list are those stored in the **Dictionaries** folder, under **My Provalis Research Projects**. To open a categorization model stored elsewhere on your computer, use the  **Open** button to browse through your system and select the proper **.WMODEL** file.

The following table describes the other functions available on this tool bar.

Control:	Description:
----------	--------------



Press this button to save the currently displayed categorization model to disk.



Pressing this button allows you to save the categorization model **.WMODEL** under a new name. A Save File dialog box will ask you to type the name and select the location where you would like the model to be saved.



Categorization models stored as **.WMODEL** files can only be used in the WordStat desktop software. To use a categorization model within QDA Miner, from Windows Explorer, in the [WordStat Document Explorer](#) utility program, or in any third party application making use of the [WordStat Software Developer's kit \(SDK\)](#), the categorization model has to be stored on disk in a portable file format. with the **.WCAT** file extension. This file is an encrypted and compressed version of the model file along with additional settings, information and resources relevant for its successful application.

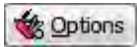
Pressing the Publish button displays a Save File dialog box allowing you to type the name of the file and select the location where it will be saved. However, in order to be used in QDA Miner or from Windows Explorer, the **.WCAT** file should imperatively be stored in the **Models** folder under your **My Provalis Research Project** folder.



This button allows you to create a new categorization model. You will be asked whether to keep the current exclusion list, the substitutions as well as the categorization dictionary. Select those you would like to use in your new categorization model and then type **OK**. A Save File dialog box will appear, asking you to type the name and select the location where you would like this new model to be saved.



This button allows you to delete the currently selected categorization model.



This button allows you to assign a description to a categorization model. It also allows you lock specific features of your model and prevent either accidental or undesired changes that may affect negatively the accuracy of the categorization process.

Language

The **Language** page allows one to choose which language resources will be available for the analysis, such as the spell-checking dictionaries used for automatic spelling corrections, lemmatization routines, as well as the thesaurus that will be used to provide suggestions. It offers the following options:

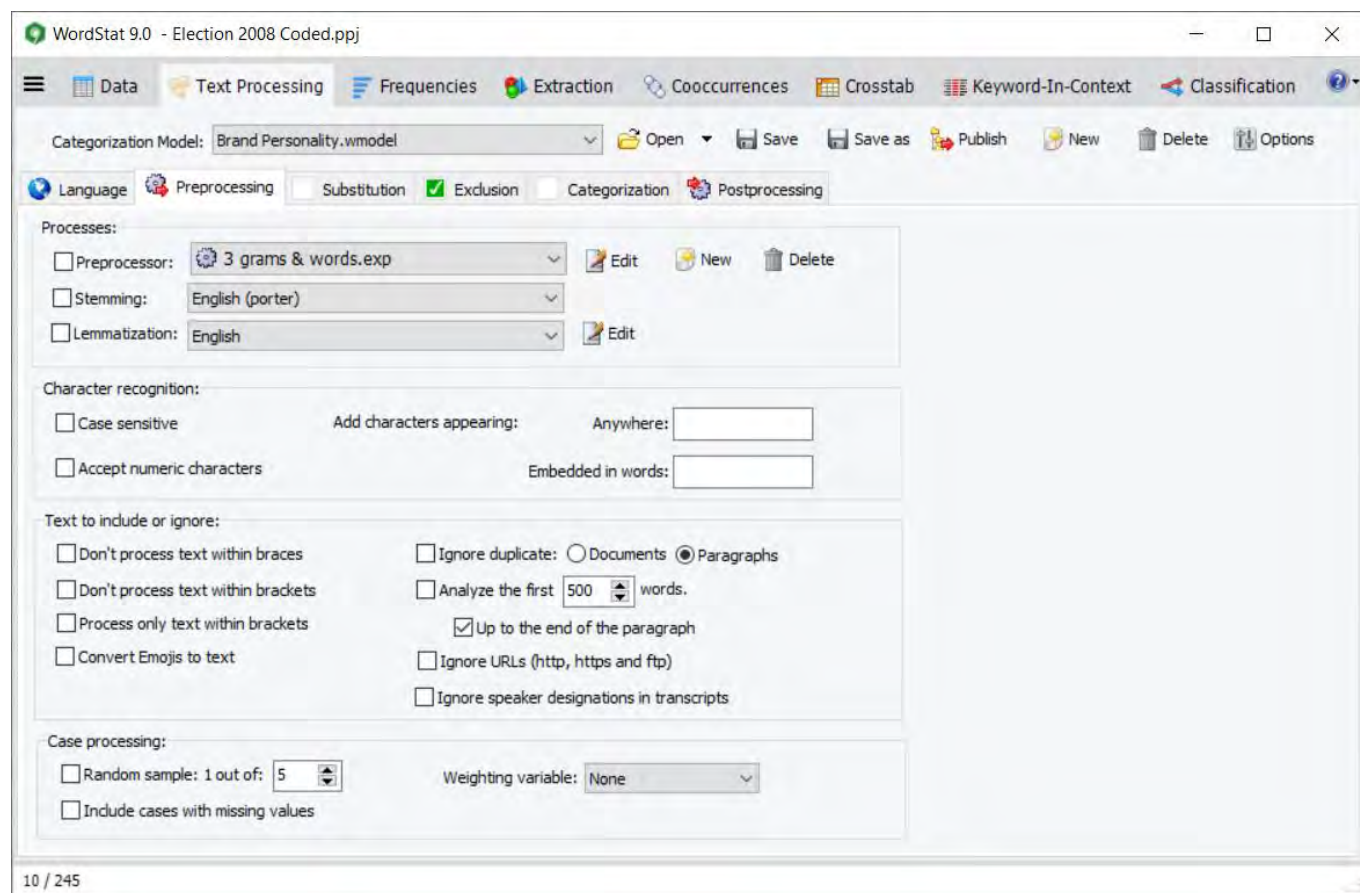
Add or remove language resources: WordStat may use various resources (such as spelling dictionaries, thesauri, and lemmatization routines) to process some natural languages more efficiently. By default, WordStat installs resources for processing English documents. Clicking this button allows you to download and install resource files for other languages or remove language resources previously installed.

Active spelling dictionary: WordStat makes use of language dictionaries to spell-check existing textual data and to suggest inflected forms of words found in the user dictionary. This group of options lets you specify which dictionary to use with the current data file.

Active thesaurus: WordStat's **Suggest** feature can use a thesaurus to suggest synonyms of existing words in the text collection or in the user categorization dictionary. This option allows you to select the language of the thesaurus. For more information, see [Basic Dictionary-Building Tools](#)

Preprocessing

The following section provides a description of the preprocessing steps involved in the transformation of textual data into keywords or content categories.



The Preprocessing tab contains two tabs: **Options** and **Languages**

The **Processes** section of the dialog box offers the following options:

Preprocessor: This option allows for the custom transformation of the text to be analyzed prior to, or in place of the execution of the other three standard processes provided by WordStat: lemmatization, exclusion and categorization. This transformation is accomplished by the execution of specially designed external routines accessible in the form of Python script, an external EXE file or a function in a DLL library. This feature is provided to offer greater flexibility by allowing any user with programming skills or resources to customize the processing of textual information. For more information on this feature see [Notes on Preprocessing](#).

Stemming: This is a natural language processing routine that reduces inflected and derived forms of words to a common root form or word stem. The English stemmer, for example, returns "write" for "write", "writer", "writing" and "writings". Stemming can be especially useful in some exploratory text-mining tasks or when developing automatic document classification models by grouping related words together, and, thus reducing the total number of word forms. However, it may also decrease the precision in the measurement of some topics associated with specific inflected forms. Plural and singular forms of some nouns are often used to refer to different concepts or ideas. The same is true for various tenses of verbs. For example, in a sentiment analysis project, we found that the verb "improve" was often associated with negative comments, while its past-tense form "improved" was generally associated with positive comments. Since stemming is based on a limited number of morphological rules, stemming algorithms are also prone to errors. For example, the English Porter stemmer will group words like "universal", "universe" and "university" into the single word root: "univers". Stemming may also fail to group related words that do not follow typical grammatical rules.

Lemmatization: WordStat provides predefined substitution processes to perform lemmatization on documents. Lemmatization is a process by which various forms of words are reduced to a more limited number of canonical forms. A typical example of lemmatization would be the conversion of plurals to singulars and past tense verbs to present tense verbs. The lemmatization algorithm implemented in WordStat is a dictionary-moderated method, partly inspired by Krovetz's KSTEM suffix substitution algorithm. Since the lemmatization algorithm does not rely on a prior part-of-speech tagging of words, it is much faster than traditional lemmatization routines. It may, however, result in a few invalid word substitutions, but usually, those errors will have no major consequences on the result of an analysis.

You may override specific substitutions by creating custom word substitutions. It is important to remember that lemmatization, like stemming, may decrease the measurement precision of some concepts or topics.

Upon installation, WordStat provides lemmatization for English. It is also possible to install additional language modules, some of which contain lemmatization routines. Seven of those modules currently offer lemmatization: French, German, Italian, Norwegian, Polish, Swedish and Spanish. Additional lemmatization modules may later be available.

The **Character recognition** section of the dialog box offers the following options:

Case sensitive: By default, WordStat internally converts all text to uppercase letters so that processing of words is case insensitive. This may be inappropriate if you want to identify proper nouns or analyze text written in some European languages like German where differences in letter cases may denote different meaning. Enabling this option prevents the internal conversion to uppercase letters and will treat two instances of the same word in different cases (lower or upper case) as two distinct words.

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 these words.

Add characters appearing: This set of options allows you to specify which characters, besides 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 ("ex" and "wife") if the hyphen is included in the list of valid characters. Two edit boxes may be used to specify additional characters. The **Anywhere** option is used to specify special characters that will be considered as part of a word, no matter where they appear. The **Embedded in words** option should be used to specify characters that should be enclosed within other valid characters and not at the beginning or at the end of a word. For example, adding a period and a comma to the list of characters embedded in words will allow you to retrieve numeric values such as 97.5 or 1,000,000 or domain names like www.google.com as a single unit without the risk of retrieving words immediately followed by commas or periods.

The **Text to Include or Ignore** section of the dialog box offers the following options:

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 to insert comments or annotations in the text variable without affecting the analysis of its content. It can also be used to ignore all questions, prompts, and other verbal interventions in an interview transcript made by the interviewer.

Don't process text within brackets: This option can be used to instruct the program to skip all text found between brackets (i.e. [and]). Since WordStat can also be configured to analyze only text found between such brackets (see option below), these two options may be used to toggle between an analysis of keywords entered manually between those brackets and of the surrounding text.

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 is especially useful to perform an analysis on keywords entered manually in the text by one or more coders.

Ignore duplicates: This option can be used to prevent identical **documents** or **paragraphs** to be processed more than once. It may be useful to remove from documents some text segments that appear more than once such as the boilerplate of some forms, ads in magazines, questions in structured interviews, copyright notices, headers and footers, etc.. To identify and either tag or permanently remove identical documents in a project, see [Identifying Duplicate Cases](#).

Analyze the first n words: In some situations, analyzing the full document is not essential for identifying the main topic and one may even get a better representation of those topics by focusing on just a few pages or a few paragraphs. A good example would be a dataset of papers from an academic journal. The first few pages which often include the title, the abstract and part of the introduction and ignore the remaining parts, including the reference section, allowing one to extract more easily and much more quickly the main topic of each paper. Another reason to use such an option may be to standardize the length of documents being analyzed. To enable this option, simply put a check mark and set the number of words you want to include in the analysis. You may also prevent WordStat from stopping in the middle of a paragraph by enabling the **Up to the End of the Paragraph** option.

Ignore URLs: Selecting this option will prevent the text analysis of URLs starting with either HTTP, HTTPS or FTP by removing those links from the text data prior to the analysis.

Ignore speaker designations in transcripts: In interview transcripts, news transcripts, or debates, the persons speaking are often identified by their initials or by name in uppercase letters followed by a colon. Analyzing the text of those documents without removing those turn-taking indicators often results in a word frequency list containing all those indicators as the most frequent words. Enabling this option will remove those speech turn indicators allowing one to identify more easily what people are talking about rather than who is talking.

The **Case Processing** section of the dialog box offers the following options:

Random sample: When this option is activated, the program will randomly select a fraction of all cases and perform the content analysis on this subsample. The proportion of cases 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 you to preview the kind of results that would be obtained on very large data files.

Include cases with missing values: When examining the relationship between textual data and categorical or numerical variables, WordStat will skip any cases with a missing value on any one of these variables. Enabling this option instructs WordStat to include all cases, whether or not values are missing. All missing values are assigned to an additional class labeled as "MISSING." Any analysis involving comparisons between classes of categorical variables (cross-tabulation, correspondence analysis, etc.) will include this additional class.

Weighting variable: This option allows the selection of a variable that will be used to apply weight to the cases. When the program reads a case, the value of the weighting variable for this case is truncated to an integer. This integer value specifies how many times the case will be duplicated. If the value is less than one, the case is excluded from the analysis. This option is especially useful when the textual data to be analyzed has already been reduced to a frequency list, such as when analyzing a list of the most frequent queries on a search engine.

Notes on Preprocessing

The **Preprocessor** option allows users to access external text preprocessing routines that are not part of the WordStat program. This option is useful to perform custom transformation on the text to be analyzed. For example, a routine may be created to remove all foreign accents, to segment a document in a particular way, to perform part-of-speech tagging, word disambiguation, stemming or transforming words into letter n-grams (sequences of letters). These transformations are not applied to the original documents stored in the database but are instead performed live immediately after the textual information has been read into memory and prior to any text processing available in WordStat (lemmatization, exclusion of words, categorization, etc.). The same routines may also be applied permanently on text stored in WordStat using the [Transform Document](#) feature.

Such external routines may be written in R, Python, or in any programming language that can create or be called from a stand-alone EXE file or a DLL. The programmer is responsible for matching the formal parameters and result type of the conventions used by WordStat for data interchanges. Technical information on these programming conventions is available on request from Provalis Research.

To create a new preprocessing routine, see:

[Writing a Preprocessing Routine in R or Python](#)
[Calling an Executable program](#)
[Calling a Function in a DLL](#)

To edit the settings of an external routine:

- Select the preprocessing routine you would like to edit from the drop-down list.
- Click the  icon to edit the external routine settings.

To remove an external routine:

- Select the preprocessing routine you would like to remove from the drop-down list.

- Click the  button.

Writing a Preprocessing Routine with Python

WordStat allows you to use a preprocessing script created in R or Python. This offers you a wide range of text processing options that may not be available in WordStat. There are certain guidelines that must be followed when creating your function:

- The function must be called "preprocessText".
- The function can only have one string argument.
- The return value must be a string.

Other than the restrictions above you are free to compose the functions of your choice.

To add a preprocessing routine:

- Go to the **Text Processing** tab > **Preprocessing** tab > **Options** tab
- To the left of the **Preprocessor** option select the  button.
- Choose **R or Python function** from the list. A dialog like the one below will appear.



The image shows a dialog box titled "External Process". It has a label "Name" above a text input field. At the bottom, there are two buttons: "OK" with a green checkmark icon and "Cancel" with a red X icon.

- Enter the name of your process in the **Name** field.
- Select **OK**. A Python editor dialog will open.



The example above performs a lemmatization using NLTK WordNet lemmatizer using Python. This function tokenizes the text and applies Part-of-Speech (POS) tags to each token. It then lemmatizes each token in relation to its POS tag.

- Enter your source code in the **Script** window on the left of the dialog box. Be careful to follow the guidelines listed above.
- Enter sample text in the sample text window.
- Select the **Test** button
- Ensure the results in the **Results** window are as desired.
- Select **OK**.

To apply your R or Python function:

- Select the check box beside the **Preprocessing** option.
- Choose the R or the Python function from the drop-down menu.

Calling an Executable Program

The second method by which an external routine may be integrated within WordStat is through the calling of an executable program (typically a console application with an EXE file extension). With such a method, information is transferred between the two programs by way of temporary files generated on the fly and stored in a default temporary folder. WordStat first creates a text file (WORDSTAT.IN) in the temporary folder containing the text to be processed. It then calls the external routine, with optional parameters (which may or may not include the input and output file name and their locations). Once the external program ends, WordStat retrieves another text file (WORDSTAT.OUT) created by the external program and containing the text to be processed in place of the original one. The two temporary files are then deleted.

To configure WordStat to call an EXE file:

- Click the  button located on the right of the **Preprocessor** list box.
- Type the name you would like to give to this preprocessing routine and click **OK**. This name will be added to the list box of available routines.
- Enter the name of the program file including the full path or click the  button to display a dialog box that allows browsing through folders and then select the appropriate program file.
- In the **Working Dir** edit box, specify the working directory for the program if necessary. Specifying \$TEMP as the working directory instructs the program to set the working directory to the temporary folder.
- Enter the **parameters** to transfer to the program at start-up. Typically, you will transfer the input and output file names, as well as any command line options needed for the external routine, to perform the required transformation. You can specify multiple parameters and can use any one of these three string constants:

CONSTANT	STANDS FOR
\$TEMP	The system temporary folder.
\$IN	The temporary text file created by WordStat and to be processed by the external routine (the actual file name used is WORDSTAT.IN)
\$OUT	The name of the text file created by the external routine and retrieved by WordStat. (the actual file name used is WORDSTAT.OUT).

To apply your new function:

- Select the checkbox beside the **Preprocessor** option.
- Choose the new function from the dropdown menu.

Calling a Function in a DLL

The third method used to call a preprocessing routine is to execute an external function stored in a dynamic library (DLL). This will manipulate the text stored in the computer memory at a specific memory address. Since no file input and output operations are necessary to transfer information, this method is often much faster than running an EXE file. However, because this external routine has access to the same memory space as WordStat, great care should be taken when running or creating such a routine. To minimize the risks involved in calling external functions, a programming convention has been imposed where the name of the function to be called **must** begin with the two uppercase letters 'WS', thus reducing the risk of calling a function that has not been written specifically for WordStat.

To configure WordStat to call a DLL function:

- Click the  button located on the right of the **Preprocessor** list box.

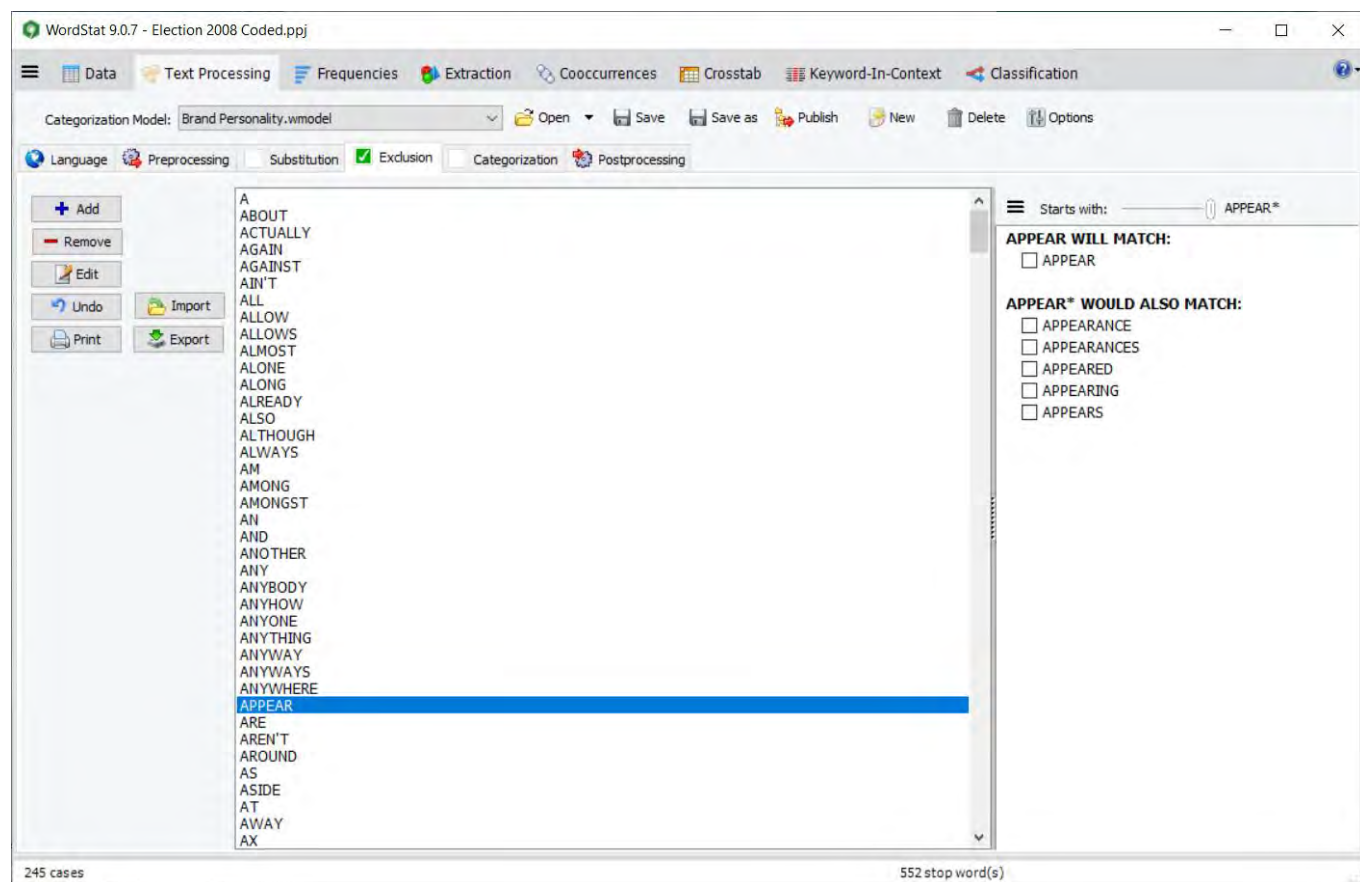
- Type the desired name for this preprocessing routine and click **OK**. This name will be added to the list box of available routines.
- Enter the name of the DLL file including the full path or click the  button to display a dialog box that allows browsing through folders and then select a DLL.
- Once a DLL file has been entered, the dialog box will provide a list of functions that are likely to be compatible with WordStat (with names starting with "WS"). Select the function containing the transformation routine needed.
- WordStat must set apart, in advance a "buffer size" that will contain the transformed text. By default, the memory space is equal to the length of the original document. For many text transformation routines, such as stemming or lemmatization, which often result in shorter text, this space should be large enough. However, for other types of text preprocessing (e.g., part-of-speech tagging or transformation of words into n-grams), the size of the transformed text may be twice or three times larger than the original. The **Buffer Size** option allows you to specify how much larger the memory space should be in order to hold the transformed text. A numerical value between 1 and 10 can be used to represent the value by which the original text should be multiplied. For example, if this option is set to 3 and the size of the original text into memory is 10 kilobytes, then 30 kilobytes of memory will be reserved for holding the transformed text. The calling routine should run a test to see whether the reserved space is sufficient and if not should return an error message, allowing WordStat to respond with an error message to the user.
- Once all the options have been set, click the **OK** button.

To apply your new function:

- Select the checkbox beside the **Preprocessor** option.
- Choose the new function from the dropdown menu.

Exclusion

The **Exclusion** tab contains an exclusion dictionary (also known as a stop list). It is used to remove all words that are not to be included in the frequency analysis. It is used mainly to remove words with little semantic value, such as pronouns and conjunctions. It may also be used to remove phrases.



The currently opened exclusion dictionary may be deactivated by removing the check mark in the check box in the tab. Wildcard symbols - such as *, ?, # and [] are supported. One may also specify case sensitive entries.

The * character (also call the "asterisk" or "star"): matches zero or more characters.

For example, the following expression:

REPORT*

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

while this expression:

COLO*R

will match both COLOR and COLOUR.

The question mark character (?): matches exactly one character.

For example:

WOM"N

will match both WOMEN and WOMAN.

The number sign (#) wildcard: stands for a single numerical digit.

For example:

B##

will match words like B12 (the vitamin) as well as B52 (the aircraft), while:

#####

will match any five digit number, typical of US zip codes. Please note that the use of the # wildcard will work as long as you set the Accept Numeric Characters check box on the **Preprocessing** tab.

The square brackets "[" and "]": are used to match a single character out of a list of characters. For example, [AEIOU] will match any one of these vowels, while [A-E] will match any letter between A and E.

The following pattern:

[A-Z][0-9][A-Z]_[0-9][A-Z][0-9]

will match typical Canadian postal codes, like H3B 1W9. Note that the underscore character in the above example represents a space.

An expression or a phrase 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 ^ character allows you to type a case sensitive entry:

For example, the following two entries:

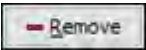
^it
^It

Will match the pronoun "it" whether or not it appears at the beginning of the sentence. These two entries will still allow WordStat to recognize "IT", which stands for "information technology".

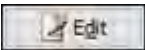
To add words to the exclusion list:

- Press on the  button and select **Words / Phrases...**
- Add the words or phrases into the box on the **Add words** dialog.
- Select **OK**.

To remove a word from the exclusion list:

- Select the words you would like to remove.
- Click the  button.

To edit an entry in the exclusion list:

- Select the words you would like to edit.
- Click the  button.
- Modify the entry in the **Edit selected words** dialog box.

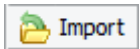
To search for an entry in the exclusion list:

- Click anywhere in the exclusion list.
- Press **Ctrl F**. A search dialog box will appear.
- Type the desired search string in the **Find What** edit box. To restrict the search to items starting with the search string, enable the **Match Beginning of Item** option and to restrict it to whole words matching the search string, enable the **Match Whole Word Only**.
- Click the **Find** button to search the first item matching the entry. Clicking this button again finds the next occurrence of the search string, starting at the currently selected item.

To undo a change in the exclusion list:

- Select the  button. A dialog box appears listing changes.
- Select the operation you would like to undo.
- Select the  button in the dialog box.

To import and exclusion list:

- Select the  button. A dialog box containing files available for import will open. WordStat can import from three file formats: An ANSI exclusion list file with a .exc file extension, a UNICODE stop word list file with a .stop file extension, or an existing categorization file (.wmodel).
- Select the file you would like to import and select **Open** button.
- You will be asked if you want to **append the words to the exclusion list**.
- Select **Yes**.

To export an exclusion list:

- Click the  button. A **Save File** dialog box will appear
- Select the proper file format under which you would like to save the file. WordStat can export to an ANSI exclusion list file with a .exc file extension or a UNICODE stop word list file with a .stop file extension.
- Type the name of the file, and click **OK** to create it.

Adding Words to an Exclusion List from other Places in WordStat

From the Frequencies page:

- Select the rows containing the words you would like to add.
- Right click your mouse, a menu will appear.
- Select **TO EXCLUSION LIST** option.

Note: You may also drag and drop a word into the dictionary panel to the right of the frequency table (see [Using the Dictionary Panel](#)).

From the Crosstab page:

- Select the rows containing the words you would like to add.
- Right click your mouse, a menu will appear.
- Select the **TO EXCLUSION LIST** option.

Note: You may also drag and drop a word into the dictionary panel to the right of the frequency table (see [Using the Dictionary Panel](#)).

From the text editor:

- Select the word you would like to add.
- Right click your mouse, a menu will appear.
- Select the **TO EXCLUSION LIST** option.

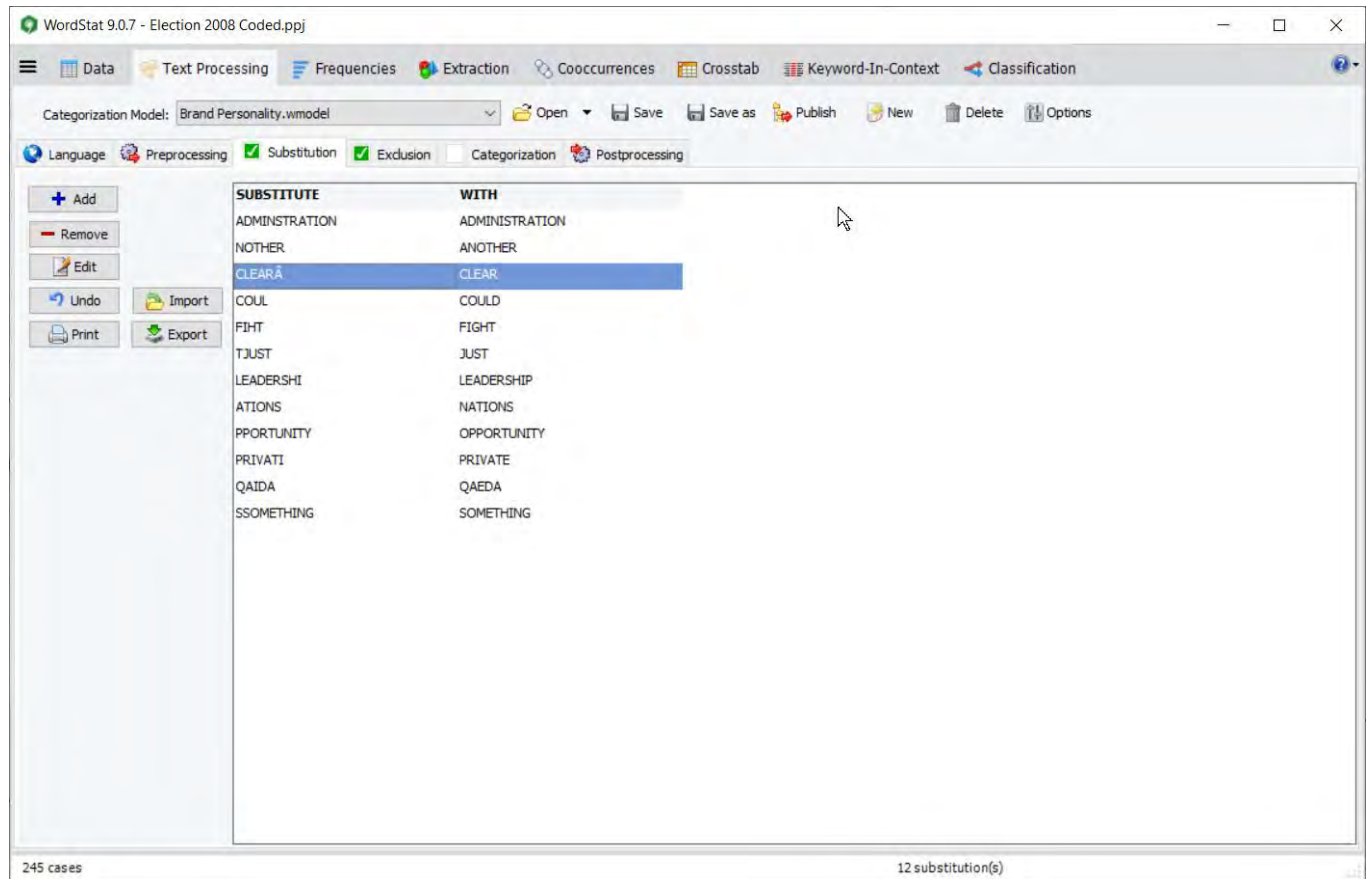
If you choose to add a word to the exclusion list, the word will automatically be stored in this file without any dialog box.

Using the Dictionary Assistance panel

Taking into account the impact of words in the exclusion list on the recognition of other entries in the categorization dictionary may be difficult especially when some items contain a wildcard. A Dictionary Assistance panel displayed on the right can provide some guidance about possible interaction between the currently selected entry and other entries in the exclusion list and the categorization dictionary. For more information on how to use such a feature, see [Using the Dictionary Assistance panel](#).

Substitution

The **Substitution** tab may be used to automatically replace specific words with other word forms. It may also be used to substitute common misspellings. You can also use this process to perform a simple type of categorization where specific words are replaced with keywords.



The currently opened substitution list may be deactivated by removing the check mark in the check box in the tab.

To add a new substitution:

- Click the  button. The following dialog box appears:

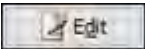
- Type in the **Original** edit box the word you would like to replace, then type the replacement word in the **Replace with** edit box and click **OK** to create the new substitution rule. This new rule is automatically added to the substitution process.

To remove a substitution:

- Select the rule and click the  button.

To edit an entry in the substitution list:

- Select the words you would like to edit.

- Click the  button.
- Modify the entry in the dialog box.

To undo a change in the substitution list:

- Select the  button. A dialog box appears listing changes.
- Select the operation you would like to undo.
- Select the  button in the dialog box.

To import and add entries to the substitution list:

- Select the  button. A dialog box listing categorization files (.wmodel) available for import will open.
- Select the file from which you would like to import substitution entries and select **Open** button.
- You will be asked if you want to append the words to the substitution list.
- Select **Yes**.

To search for an entry in the substitution list:

- Click anywhere in the substitution list
- Press **Ctrl-F**. A search dialog box will appear.
- Type the desired search string in the **Find What** edit box. To restrict the search to items starting with the search string, enable the **Match Beginning of Item** option and to restrict it to whole words matching the search string, enable the **Match Whole Word Only**.
- Click the **Find** button to search the first item matching the entry. Clicking this button again finds the next occurrence of the search string, starting at the currently selected item.

Adding Words to an Substitution List from other Places in WordStat

From the Frequencies page:

- Select the rows containing the words you would like to add.
 - Right click your mouse, a menu will appear.
 - Select **SUBSTITUTE WITH** option.

Note: You may also drag and drop a word into the dictionary panel to the right of the frequency table (see [Using the Dictionary Panel](#)).

From the Crosstab page:

- Select the rows containing the words you would like to add.

- Right click your mouse, a menu will appear.
- Select the **SUBSTITUTE WITH** option.

Note: You may also drag and drop a word into the dictionary panel to the right of the frequency table (see [Using the Dictionary Panel](#))

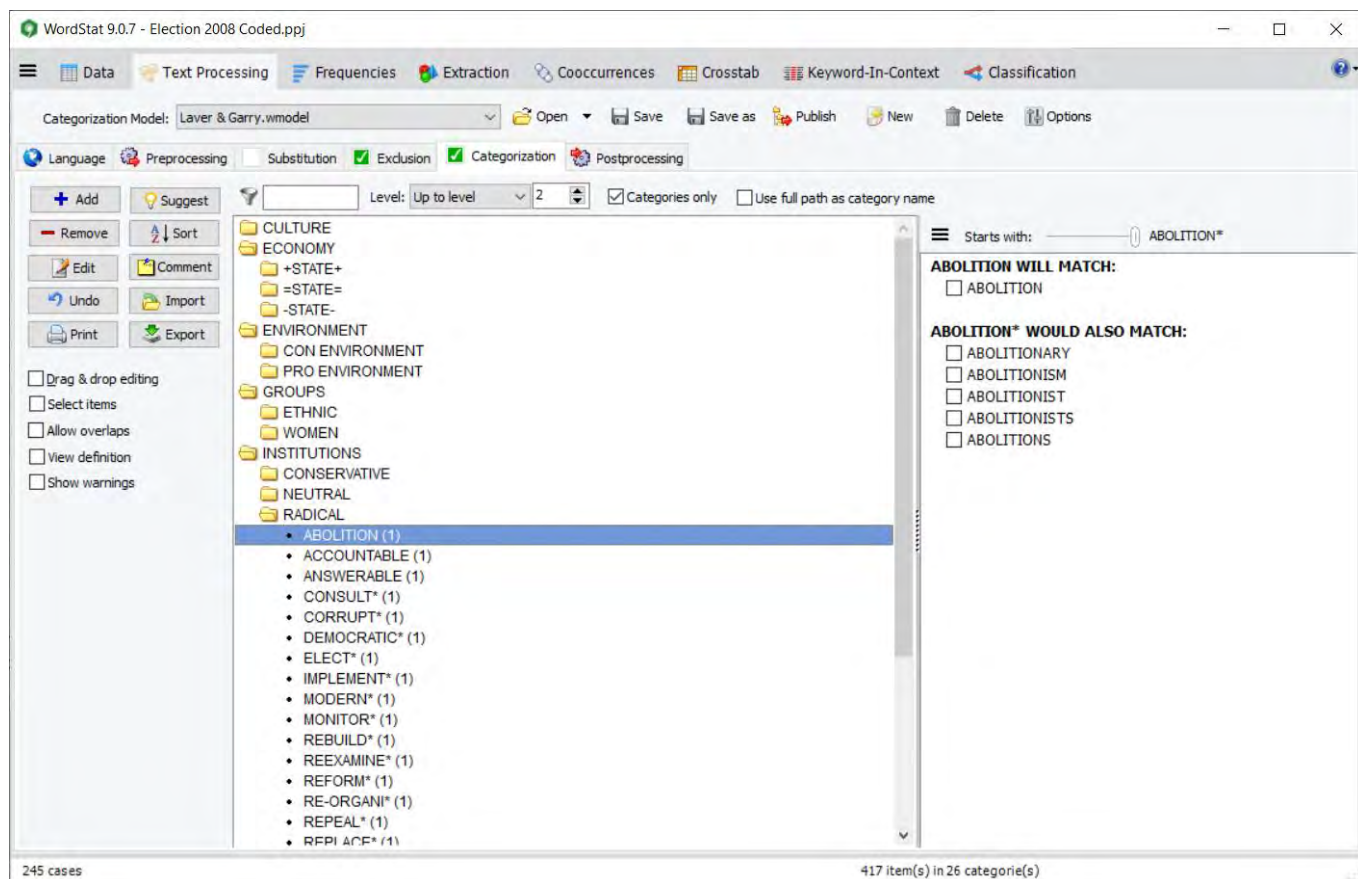
Related Topics:

[Misspellings & Unknowns](#)

Categorization

The categorization process allows you to change specific words, word patterns, or phrases to other words, keywords or content categories and/or to extract a list or specific words or codes. This process requires the specification of a categorization 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".

A categorization dictionary may also contain rules delineating the conditions under which specific words or phrases should be categorized. The rules may consist of complex expressions involving Boolean (AND, OR, NOT) and proximity operators (NEAR, BEFORE, AFTER). These kinds of rules allow you to eliminate basic ambiguity in words by taking into account the presence of other words that may alter the meaning. A good example would be the presence of a negative word form (such as "rarely" or "never") close to an adjective. Another example would be the differentiation of the various meanings of the word "bank" by identifying other words like "river", "money" and "deposit" surrounding "bank". For more information on rules, see section [Working with Rules](#).



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

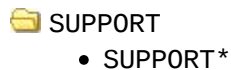


In the above example, words like CANADA, USA or MEXICO may be coded as either NORTH_AMERICA or COUNTRY, depending on whether the categorization is performed up to the first or second level of the dictionary (see [Level of Analysis](#) below).

Wildcards

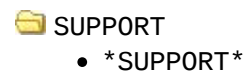
Wildcards such as *, ?, # are supported.

For example, the following item under the support category:



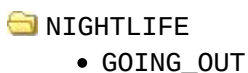
will change SUPPORT, SUPPORTS, SUPPORTING, SUPPORTIVE, SUPPORTER, etc. into a single word SUPPORT,

while the following word pattern:

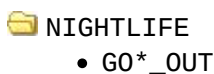


will also substitute all words with the substring "SUPPORT" in it, such as UNSUPPORTEDLY, UNSUPPORTED, etc.

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:



You may also use wildcards in expressions such as:



to substitute several forms of an expression at once.

Integer weights can 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*.

Case Sensitive Entries

While by default, WordStat text analysis is case insensitive, sometimes you need to perform a case sensitive match. For example when you want to differentiate the pronoun "us" from the abbreviation of United States "US", or identify references to "Information technology" using the "IT" abbreviation. To create a case sensitive entry, simply type the ^ character as the beginning of the entry.

For example:

^Bill

will match the given name Bill, but will not match this same word without the uppercase letter B. It also won't match this word if in all uppercase ("BILL") or bill as in "dollar bill".

Regular Expressions

Beside the above wildcards and case sensitive entries, WordStat categorization dictionaries may contain entries consisting of regular expressions, allowing you to detect specific patterns such as zip codes, phone numbers, or email addresses. For more information on how to enter, edit and test regular expression, see [Working with Regular Expressions](#).

Categorization Settings

Level: The level option allows you to specify up to which level the categorization process should be performed. Two ways of setting levels is available. Setting this option to **Up to level** allows you to specify a fixed level up to which the coding should be performed. For example, in the following dictionary:



if a level of 1 is specified, all words that are stored at a higher level than 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 result 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 will remain unchanged. In the above example, increasing the level of analysis to 3 will return the frequency of each country name.

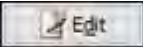
Setting the **Level** option to **As shown** instructs WordStat to match the level of categorization performed to the level of details currently displayed in the tree view of the categorization dictionary. This option allows one to set different levels of categorization by expanding broad categories that should be broken down and by collapsing categories for which finer details are not needed. For example, if we modify the above tree by collapsing the NORTH-AMERICA category, WordStat will display it the following way:



- 📁 NORTH_AMERICA
- 📁 SOUTH_AMERICA
 - BRAZIL (1)
 - CHILI (1)

The program will report frequencies of individual countries like BRAZIL or CHILI but will categorize every instance of CANADA, UNITED-STATES, USA and MEXICO as NORTH-AMERICA.

Please note that it is possible to prevent a category from being broken down into subcategories or items, even if the level of analysis is set to a higher setting or if it is set to AS SHOWN and the items contained in this category are visible. Such a feature is useful when the content of a category consists of different ways of referring to the exact same thing (for example UNITED_STATES, UNITED_STATES_OF_AMERICA, US and USA) or consists of various misspellings.

To make a category unbreakable, select the category in the dictionary tree, click the  button, and put a check mark in the Unbreakable box. The folder icon normally used to represent categories will be transformed into a folder icon  with a key inside. You may also select the category, right click, and then select UNBREAKABLE | YES from the pop-up menu. To unlock the folder, follow the previously described steps for editing the category and remove the check mark in the **Unbreakable** box.

Categories only: When the **Level** 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
 - TREMOR (1)
 - AFRAID (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 ensures that individual words won't be included but will be coded as their parent category.

Use full path as category name: When the **Level** 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 (/) character is used to separate the various levels. For example, in the above example, enabling this option and setting the level of 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 overlaps: By default, categories are mutually exclusive such that a word can only be entered in a single category. Enabling this option allows you 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 violates this assumption and may yield dubious results.

Show warnings: Some items in an exclusion list or categorization dictionary may remain undetected in documents because of their incompatibility with some analysis options. This occurs, for example, when an item is found both in the categorization dictionary and the exclusion list, or when this item includes non-alphabetic characters that have

not been specified as valid. The following table displays the various types of problems that may be identified by WordStat:

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

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

Matching Priority in a Dictionary

Dictionaries or lexicons consisting of words or word patterns only often fail to achieve accurate measurements of the topics or concepts they are supposed to assess. This results from their inability to disambiguate words with multiple meanings. WordStat offer a way to disambiguate words and achieve higher accuracy by using an established priority when matching items in the dictionary. The evaluation order is based on a single principle that consists of giving priority to more specific items over less specific ones. Since phrases are more specific than words, and the longer the phrase, the more specific it is, and since whole words are more specific than word patterns (words with wildcards), the evaluation priority has been set this way:

1. Long phrases over short phrases
2. Phrases over words
3. Case sensitive words over non case sensitive ones
4. Words over word patterns
5. Longer word patterns over shorter ones

Such a matching priority list, not only allows one to measure topics more accurately, but it also prevents double-counting items when more than one candidate item may be matched by a specific text segment, such as when a phrase in the text correspond to a phrase in the dictionary as well as a word or a word patterns also in the dictionary.

For example if your dictionary contains the words and phrases:

- 📁 ACCESSIBILITY
 - HEALTH_CARE_COST (1)
 - COST_OF_HEALTH_CARE (1)
- 📁 TOPICS
 - HEALTH_CARE (1)
 - CHILD_CARE (1)
 - CARE
 - HEALTH

and your document contains the sentence "If I'm elected, I promise to reduce the health care cost for middle-class families", WordStat will only match the item HEALTH_CARE_COST, and won't match entries such as HEALTH_CARE, CARE, or HEALTH, since those are less specific than the longest phrase. This will allow you to disambiguate words such as "bank", and remove false positives by excluding or categorizing elsewhere phrases such as "river bank" or "Gaza bank", or making sure that the dictionary entry JOBS won't be triggered when someone mentions "Steve Jobs". If you use the word pattern TAX* to identify references to "tax", "taxes" or "taxation", it becomes possible to restrict hits to only those words by adding entries like TAXI, TAXONOMY or TAXIDERM to the exclusion list or to another category. Finally, since case sensitive entries are more specific than words, it becomes possible to differentiate IT (information technology), US (country), or WHO (World Health Organization), from the pronouns "it" or "us", or "who".

Using the Dictionary Assistance Panel

Taking into account the above matching priority and envisioning possible interactions between items in a large dictionary may be challenging. It may be difficult to identify for a specific dictionary item whether it will be superseded by another item or, on the contrary, prevent some other items from being matched.

Another challenge when building categorization dictionaries is to assess the impact of using a * wildcard at the end of a word. Using such a wildcard may generate false positives, consisting of unrelated words matching the item. One may also be unaware of the full potential benefit of using such a wildcard resulting in some false negatives, or relevant items that could have been matched if a wildcard had been used. Also, if one decided to use such a strategy to match multiple words, how far can one go in the truncation of a word to get more relevant words without also matching irrelevant ones. For example, if you have a dictionary item COMPASSIONATE, changing this entry to COMPASSIONAT*, will bring three additional items: COMPASSIONATED, COMPASSIONATELY, and COMPASSIONATING, but truncating four more letters to COMPASSI* will also bring COMPASSION and COMPASSING. However, truncating a single additional character will also match COMPASS and COMPASSES, two false negatives.

The Dictionary Assistance panel located on the right of the Categorization and the Exclusion sub-pages provides a way to identify potential interaction between the currently selected item and other dictionary entries. For single word entries it will also allow one to identify the potential impact of using a * wildcard at the end of the word or at different levels of truncation by displaying new entries that would be matched if such a word pattern were used. The following example provides an example of the type of information one may obtain when selecting a single word entry. For this example, we selected the item ENVIRONMENTAL in the Forest Value Dictionary.

Starts with:

ENVIRONMENTAL*

ENVIRONMENTAL WILL MATCH:

☐ ENVIRONMENTAL

BUT NOT IF IT ALSO MATCHES:

ENVIRONMENTAL_CONCERN in category LIFE SUPPORT

ENVIRONMENTAL_COST in category LIFE SUPPORT

ENVIRONMENTAL_DEGRADATION in category LIFE SUPPORT

ENVIRONMENTAL_VALUE in category LIFE SUPPORT

ENVIRONMENTAL_QUALITY in category LIFE SUPPORT

ENVIRONMENTAL_IMPACT in category LIFE SUPPORT

IT WILL SUPERSEDE:

ENVIRON* in category LIFE SUPPORT

ENVIRONMENTAL* WOULD ALSO MATCH:

☐ ENVIRONMENTALISM
☐ ENVIRONMENTALIST
☐ ENVIRONMENTALISTS
☐ ENVIRONMENTALLY
☐ ENVIRONMENTS

BUT NOT IF IT ALSO MATCHES:

ENVIRONMENTALLY_SUSTAINABLE in category LIFE SUPPORT

ENVIRONMENTALLY_SENSITIVE in category LIFE SUPPORT

The panel allows us to see that ENVIRONMENTAL in the LIFE SUPPORT category would not be matched in the situation where it is followed by any one of the following words: CONCERN, COST, DEGRATION, VALUE, QUALITY or IMPACT. It will also prevent the matching of the item ENVIRON* since it is more specific that this word pattern. Now, the last two groups of items on this panel allow us to see that by adding a * wildcard at the end of ENVIRONMENTAL, we would also match five additional words. However, phrases such as "ENVIRONMENTALLY SUSTAINABLE" and "ENVIRONMENTALLY SENSITIVE" would have precedence over this word pattern or any of the suggested words (phrases being more specific than words and word patterns). Moving the **Start With** cursor located on the top of the panel tool bar will allow one to interactively identify the impact of various level of truncation on the matching items.

Several operations are available from such a list.

To add those words to a content category

- Put a check mark beside the words you want to add to your categorization dictionary.
- Click the  button or right-click to display a contextual menu.
- Choose **Add Selected to Categorization**
- Select the category in which you want those words to appear and click **OK**.

To treat those words as exceptions by adding them to the exclusion list

- Put a check mark beside the words you want to exclude
- Click the  button or right-click to display a contextual menu.
- Choose **Add Selected to Exclusion List**

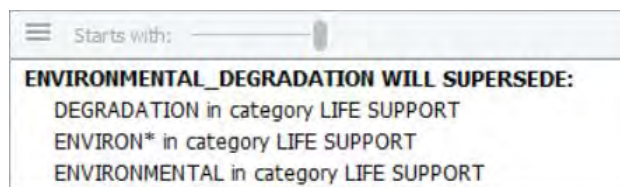
To replace the currently selected dictionary item with the word pattern

- Click the ≡ button or right-click to display a contextual menu.
- Select the **Replace** menu item and then select the displayed word pattern.

To replace the currently selected dictionary item with other words

- Put a check mark beside the words you want to replace the current entry with. You may select the original entry if you wish.
- Click the ≡ button or right-click to display a contextual menu.
- Select the **Replace** menu item and then choose **With Selected Words**.

When a phrase is selected, the panel will either be empty, indicating the absence of interaction with other dictionary entries, or it will contain a list of entries which the selected phrase may interact. In the example below, the phrase "ENVIRONMENTAL DEGRADATION" will have precedence over the words "ENVIRONMENTAL" and "DEGRADATION" as well as over the word pattern ENVIRON*, no matter to which content category they belong.



Related Topics:

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](#).

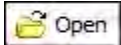
Or jump directly to one of the following topics:

- [Opening an existing dictionary](#)
- [Creating or copying a dictionary](#)
- [Adding a word](#)
- [Adding a category](#)
- [Removing words or categories](#)
- [Editing a word or a category](#)
- [Moving words or categories](#)
- [Using lexical tools for dictionary-building](#)
- [Working with rules](#)

Creating and Maintaining Dictionaries

Opening a Dictionary

To open an existing dictionary:

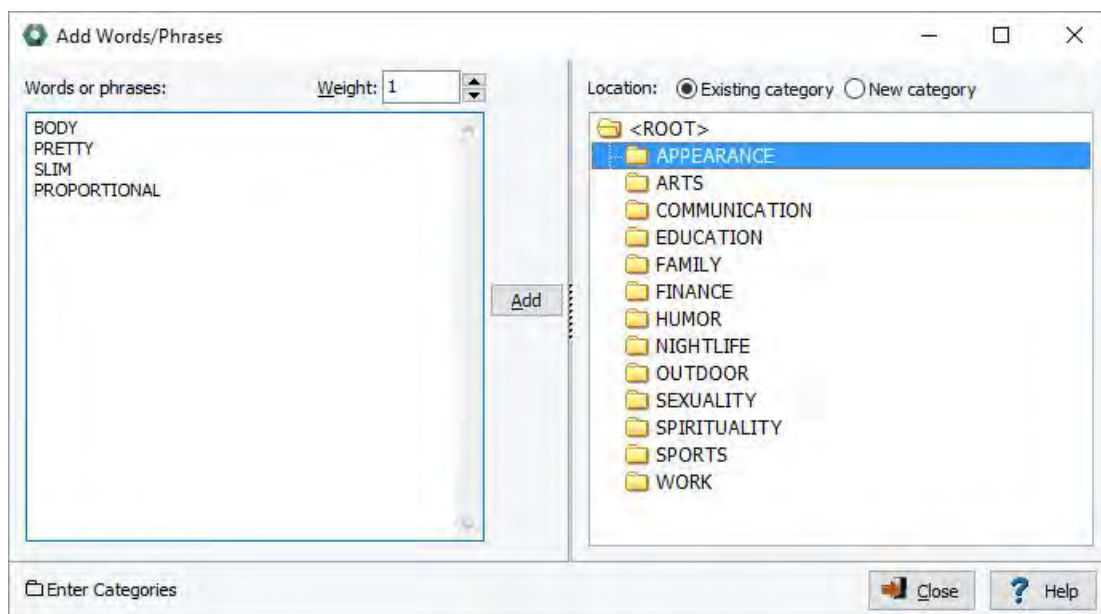
- Select the dictionary from the **Categorization Model** drop-down list. If the dictionary is not listed, click the  button to display a dialog box that will let you browse through folders and select a dictionary.

To create or copy a dictionary:

- Click the  button located on the right of the exclusion list or of the categorization dictionary. A dialog box enables specifying the name and location of the new dictionary file. If a dictionary file is already active, it will ask whether existing entries should be copied 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 words to a dictionary:

- Press on the  button and select **Words / Phrases...** The program will display a dialog box similar to the following:



To add new items to an existing content category:

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

To add new items to a new content category:

- Type the words or phrases you would like to add in the edit box, one item per line. Spaces are automatically replaced by an underscore character.
- Set the check box on the right panel to **New category**.
- Type the name of the new category and select the existing category under which this new category should be created. To create a main category, select the **<ROOT>** item.
- Click the **Add** button.

Wildcards such as * and ? are supported.

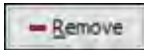
Weights can also be assigned to specific categorization, so that a specific word, word pattern, or expression may count for more than one instance of the concept. The default value for this option is 1. To use a lower or higher value, edit the Weight option either by entering a new numeric value in the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

If you want to add a word to a non-existing category, you first need to create a category (see below) and then follow the above steps to add the word to this new category.

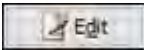
To add categories to the categorization dictionary:

- Press on the  button and select the Categories command.
- 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 names you would like to add in the edit box, one category per line, and click the **Add** button located on the right of this edit box.

To remove a keyword or a content category from a dictionary:

- 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 the  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 be erased.

To edit an entry in a dictionary:

- Select the item you want to modify and click the  button.

To search for an entry in a dictionary:

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

Moving Words or Categories

The easiest way to change the structure of a categorization dictionary is by using drag-and-drop operations. Using the mouse, you can move a word to a different category, or move an existing category or sub-category to another location on the main level or below an existing category. There are two ways to perform such a drag-and-drop operation:

For single-move operations:

- Press and hold the **Alt** key.
- Click the item in the dictionary you would like to move and hold down the mouse.
- Drag the item to the desired location and release the mouse button.

Another way to achieve multiple drag-and-drop operations is by enabling the **Drag & Drop Editing** option located to the left of the dictionary.

For multiple-move operations:

- Check the **Drag & Drop Editing** check box.
- Click the item you want to move, and hold down the mouse button.
- Drag the item to its new location, and release the mouse button.

By default, the dragged item is stored in the category at the cursor's position. To move a word or a category to the main level or to the same level as the category at the cursor's position, simply hold the Alt key while dropping the dragged item.

To move a word to a distant category:

- Select the item you would like to move.
- Click the right button of the mouse to display the contextual menu.
- Select the **MOVE TO** command.
- Choose the category where the selected item should be moved and click **OK**.

Adding Words to an Exclusion List from other Places in WordStat

From the Frequencies page:

- Select the rows containing the words you would like to add.
- Right click your mouse, a menu will appear.
- Select **TO CATEGORIZATION DICTIONARY** option.

Note: You may also drag and drop a word into the dictionary panel to the right of the frequency table (see [Using the Dictionary Panel](#)).

From the Crosstab page:

- Select the rows containing the words you would like to add.
- Right click your mouse, a menu will appear.
- Select **TO CATEGORIZATION DICTIONARY** option.

Note: You may also drag and drop a word into the dictionary panel to the right of the frequency table (see [Using the Dictionary Panel](#)).

From the text editor:

- Select the word you would like to add.
- Right click your mouse, a menu will appear.
- Select **TO CATEGORIZATION DICTIONARY** option.

Related Topics:

[Using Lexical tools for dictionary-building](#)

[Merging categorization dictionaries](#)

[Printing the categorization dictionary](#)

[Using the Dictionary Panel](#)

Working with Rules

The WordStat Rules editor may be used to define complex coding rules allowing you to specify under which conditions a particular item or category of items should be coded. Such a feature may be useful to differentiate between numerous meanings of a single word (disambiguation). For example, you may limit the coding of the word "bank" to situations where the word refers to the financial institution. This can be done by restricting the coding of "bank" to documents containing vocabulary related to monetary or financial transactions ("cash," "money," "mortgage," "investment," etc.) or by excluding alternate meanings such as when "bank" appears in close proximity to words like "river" or "canoe." Rules may also be used to measure various forms of a phrase. For example, the idiom "TURN OFF" may be expressed in many different ways ("turn it off," "turned off," "turned this off," "turned his radio off"). While figuring out all the possible forms of such an idiom may be very difficult, if not impossible, a single coding rule to look for the word pattern "TURN*" followed by "OFF" within the same sentence could very well cover most of these situations. Rules can also take into account the presence of words that may alter the power of an adjective, such as negations or qualifiers like "rarely," "numerous," "few," etc. Rules may even be used to identify sequences of events or complex actions.

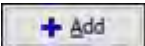
In WordStat, a rule can refer to individual words, word patterns, or phrases. It may also refer to several items belonging to a content category of the current dictionary. A reference to a category is always preceded by the number sign character ('#'). For example, in the rule:

SATISFIED NEAR #PROFESSOR

the first item, SATISFIED, refers to a single word while #PROFESSOR will match any item found in the PROFESSOR content category.

Just like words or phrases, rules may be stored anywhere in a categorization dictionary. A rule consists of a target item and from up to 20 conditions, each condition consisting of another item linked to the first item using a Boolean (AND, NOT) or a proximity operator (NEAR, BEFORE, AFTER) or their negative forms (NOT NEAR, NOT BEFORE, NOT AFTER). The context in which these conditions will be tested also needs to be specified, allowing you to either consider the content of the entire document or restrict the test to a single paragraph or a single sentence. When a proximity operation is used, you also have to specify the maximum distance (in number of words) that must separate the two items in order for this proximity condition to be tested as true or false.

To create a rule:

- Click the  button and select the **RULES** menu item. A dialog box similar to the one below will appear:

Rule Editor

Name: WOMENS_RIGHTS

Target word, phrase or category: EQUALITY

Operator	Word, phrase, or category	Within	Distance
AND	WOM?N	Same paragraph	5
NOT AFTER	#NEGATIONS	Same sentence	4
NEAR		Same document	50

☒ Match All ☐ Match Any

Weight: 1

Store in: <ROOT>

- CULTURE
 - CULTURE-HIGH
 - CULTURE-POPULAR
- SPORT
- ENVIRONMENT
 - CON ENVIRONMENT
 - PRO ENVIRONMENT
- GROUPS
- ETHNIC
- WOMEN

+ Add Close

The minimum requirements for a rule to be valid are:

- A unique name,
- At least one statement consisting of a target item, a Boolean or proximity operator and a second item,
- The text unit within which this rule should be tested, and optionally the maximum distance in words.
- The weight given to items meeting these conditions,
- The content category where the rule will be stored be stored.

The ampersand or at sign ("@"") is used as a prefix to denote the presence of a rule and will be added automatically to the rule name. In the above dialog box, the rule @WOMENS_RIGHTS will be stored under the content category WOMEN and will be considered as true if the word EQUALITY occurs in the same paragraph WOMEN or WOMAN and if there is no item in the #NEGATIONS category in the same sentence up to four words before the target word (i.e. EQUALITY).

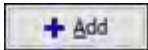
To enter a specific word, word pattern or phase, simply type the desired item. Spaces between words are automatically converted to underscore characters. To enter a content category, type the number or pound sign ('#') immediately followed by the name of the category. An existing category may also be selected from a drop-down list by clicking the down arrow located to the right of the edit box and clicking the appropriate category name, listed in alphabetical order.

The following operators may be used in a rule:

RULE	CONDITION IS TRUE IF...
------	-------------------------

<i>item1</i> AND <i>item2</i>	...both items occur in the same document, paragraph or sentence.
<i>item1</i> NOT <i>item2</i>	...the first item occurs in the document, paragraph or sentence but not the second one.
<i>item1</i> NEAR <i>item2</i>	...both items occur in the same document, paragraph or sentence, and are no more than n words apart.
<i>item1</i> BEFORE <i>item2</i>	...both items occur in the same document, paragraph or sentence, and the second item appears after the first one within the next n words.
<i>item1</i> AFTER <i>item2</i>	...both items occur in the same document, paragraph or sentence and the first item appears after the second one within the next n words.
<i>item1</i> NOT NEAR <i>item2</i>	...the first item occurs in a document, paragraph or sentence, and is not found within n words of the second item.
<i>item1</i> NOT BEFORE <i>item2</i>	...the first item occurs in a document, paragraph or sentence, and is not followed within n words by the second item.
<i>item1</i> NOT AFTER <i>item2</i>	...the first item occurs in a document, paragraph or sentence, and does not occur within n words after the second item.

To add an additional condition, click the **Add** button located on the left of the first condition. To remove a condition, click the **X** button located on its right. When more than one condition is set, you will be asked to specify whether you want to match all criteria or match any one of them..

Once the rule has been properly defined, click the  button located in the lower right corner of the dialog box to append the rule definition to the selected content category and to clear the form. Once you have finished entering rules, click the close button to quit this dialog box and return to the WordStat main screen.

Please note, in order to prevent any recursive or cross-reference problems in rules, content categories can only refer to words, word patterns or phrases stored in categories and will ignore the presence of other rules. For example, if a category named #NEGATION contains 10 word patterns and three rules, any reference to this category in a rule will take into account those 10 words and will ignore instances where any one of the three rules have been found to be true.

Working with Regular Expressions

A regular expression (RegEx) is a formal language made up of a sequence of characters or a text string for describing a search pattern that can be used to extract information from text. Using RegEx is much more powerful than using wildcards because of the number or types of patterns that can be extracted. Regular expressions can be used in WordStat categorization dictionaries.

RegEx can be used to find the following patterns (this list is not exhaustive):

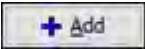
- floating point number
- password
- username
- email address
- IP Address
- credit card numbers in documents for a security audit
- duplicated words
- HTML tag

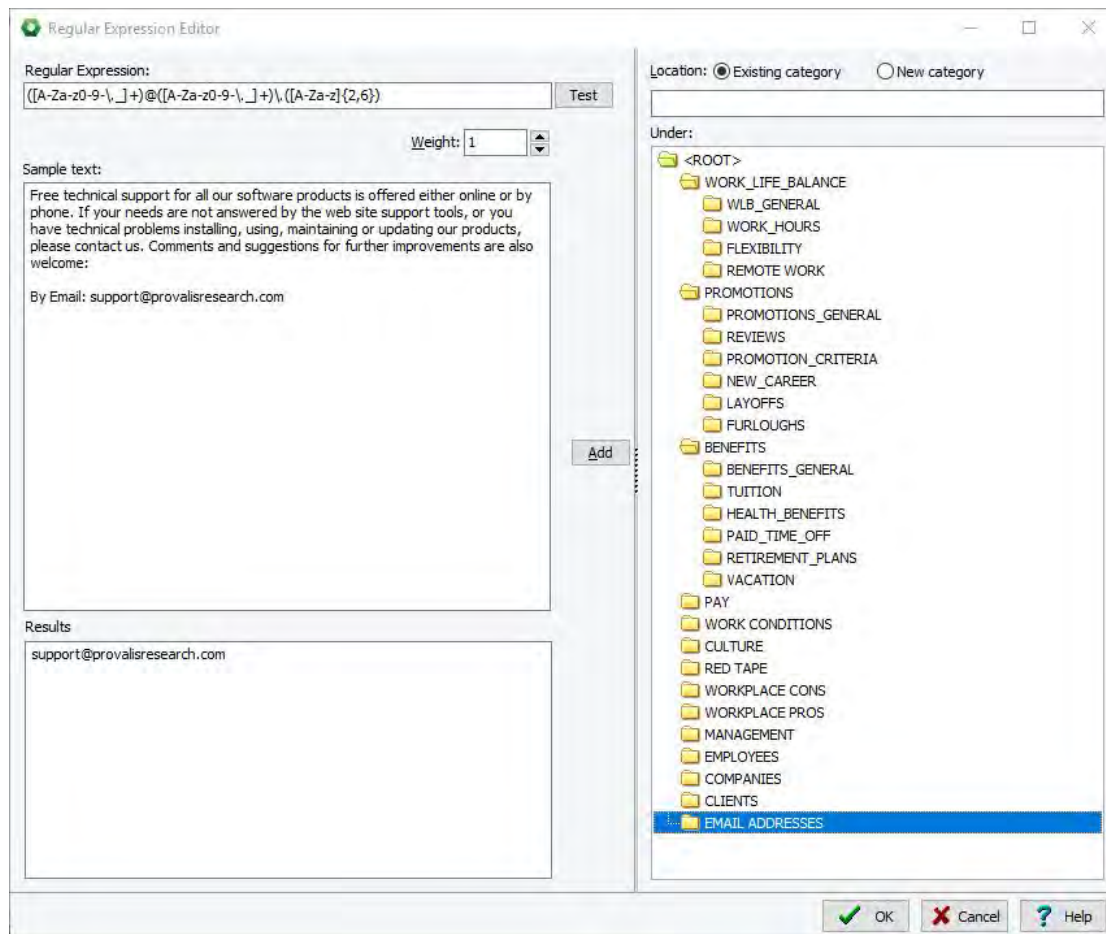
- numeric ranges
- same word with different spelling
- two words near each other
- URL
- valid dates

RegEx can range from simplistic to complicated. You have probably used *.doc to find all word files in a file manager. The regular expression (regex) equivalent is [^]+\.doc. A more elaborate example would be a pattern for matching an email address. The following example is not complete but will convey the idea ([A-Za-z0-9-\._]+)@([A-Za-z0-9-\._]+)\.([A-Za-z]{2,6}).

A free and useful tool for writing RegEx is called can be found at <https://regex101.com>. It is a PCRE-based regular expression debugger with real time explanation, error detection and highlighting.

To add RegEx expressions:

- Click the  button and select the **RegEx** menu item. A dialog box similar to the one below will appear:



In the example above we are looking for email addresses

- Enter your regular expression in the **Regular Expression** field.
- Enter sample text in the sample text window.
- Select the **Test** button

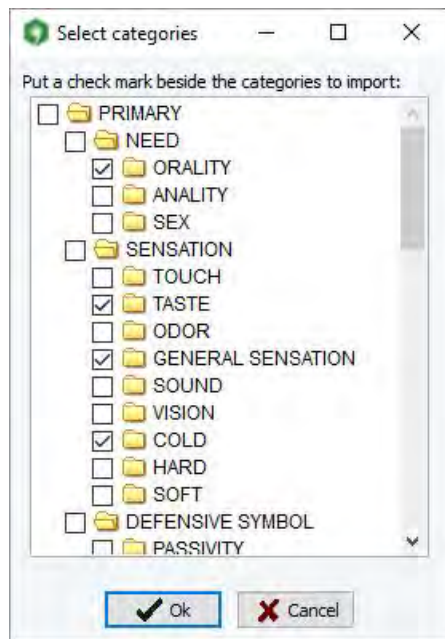
- Ensure the results in the **Results** window are as desired.
- Choose the location in the dictionary panel on the left where you would like the extracted text to be stored.
- Select **OK**.

Importing Dictionaries

The **Import** feature allows you to append categories and items contained in one categorization dictionary into another dictionary. WordStat categorization dictionaries may be imported from WordStat .CAT and .WMODEL files, Excel, CSV or tab-delimited files, as well as from XML files. The exportation process supports multilevel dictionaries as well as weighting.

To import from WordStat .cat and .wmodel files:

- From the dictionary page of WordStat, open the dictionary into which you would like to import new categories.
- Click the  **Import** button. An **Open** dialog box is displayed.
- Locate the dictionary containing the items you would like to import and click **Open**. A dialog box similar to the one below is displayed:



- Select the categories you would like to import by clicking in the box beside the desired category(ies) and clicking **OK**. To select all items, right click anywhere on the list of categories and select **Check All**. Choosing **Uncheck All** removes all check marks previously entered.

If an imported category already exists in the currently active dictionary, WordStat ignores duplicate items and only imports new items not already found in the original category. New categories are appended to the existing structure along with all their items.

Other File Types

To import a dictionary stored in Excel, CSV, Tab-delimited and XML files the data file has to be formatted such that the names of the categories are listed in one-to-four columns and the items such as words and phrases are listed in an

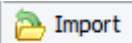
additional column. The data file may also contain an additional column containing weights to be used for each item, this weight being represented by a positive integer or floating-point numerical value. The first row must contain a header that will help you identify the content of each column. A typical dictionary table consisting of items stored in a hierarchical dictionary with two levels (main categories and subcategories) along with individual weights may look like this:

MAIN CATEGORY	SUBCATEGORY	ITEMS	WEIGHT
EMOTION	ANXIETY	NERVOUS	0.6
EMOTION	ANXIETY	EMBARRASS*	2.1
EMOTION	ANXIETY	DISCOMFORT*	2.7
EMOTION	SADNESS	ABANDON*	1.0
EMOTION	SADNESS	ALONE	1.3
BEHAVIOR	AGGRESSION	ABUS*	1.2
BEHAVIOR	AGGRESSION	HUMILIAT*	0.8
BEHAVIOR	AGGRESSION	BEAT*	2.5
BEHAVIOR	ASSERTION	CLAIM*	1.0
BEHAVIOR	ASSERTION	REQUEST*	3.2

Another version that omits the repetition of category names may look like this:

MAIN CATEGORY	SUBCATEGORY	ITEMS	WEIGHT
EMOTION	ANXIETY	NERVOUS	0.6
		EMBARRASS*	2.1
		DISCOMFORT*	2.7
	SADNESS	ABANDON*	1.0
		ALONE	1.3
BEHAVIOR	AGGRESSION	ABUS*	1.2
		HUMILIAT*	0.8
		BEAT*	2.5
	ASSERTION	CLAIM*	1.0
		REQUEST*	3.2

To import from Excel, CSV, Tab-delimited and XML files:

- Click the  button.
- An **Open** dialog box will appear, allowing you to select the file containing the dictionary you want to import.
- Once selected, click **Open**. A dialog box similar to this one will appear:

The drop-down lists (to the left in this dialog box) allow you to identify the columns containing the category and subcategory names. Up to four levels of category can be specified this way. It is, however, possible to import dictionaries with higher levels of category by separating the deeper levels (fifth or more) with a backslash character and storing these at the fourth level as part of this subcategory name. For example, a fourth level subcategory that would be stored as: SENTIMENT\NEGATIVE\UNFAIR would create two additional levels underneath SENTIMENT, with NEGATIVE as a fifth level subcategory of SENTIMENT and UNFAIR as the sixth level (after NEGATIVE).

If the names of the categories are not fully specified, as in our first example, but instead resemble the second table where only the first occurrence of the category name has been entered, then you must select the **Sparse Entry of Category** check box to make sure the items will be imported and stored in their proper category.

- Once all the options have been set, click the **OK** button.
- A **Save File** dialog box will ask you to enter a dictionary file name. WordStat will point to the default folder where other dictionaries are stored. You may browse to a different folder to store the new dictionary at another location. Once saved, the newly created dictionary will become the default one.

Exporting Dictionaries

WordStat categorization dictionaries may be exported to WordStat .CAT , Excel, CSV or tab-delimited files, as well as from XML files. The exportation process supports multi-level dictionaries as well as weighting.

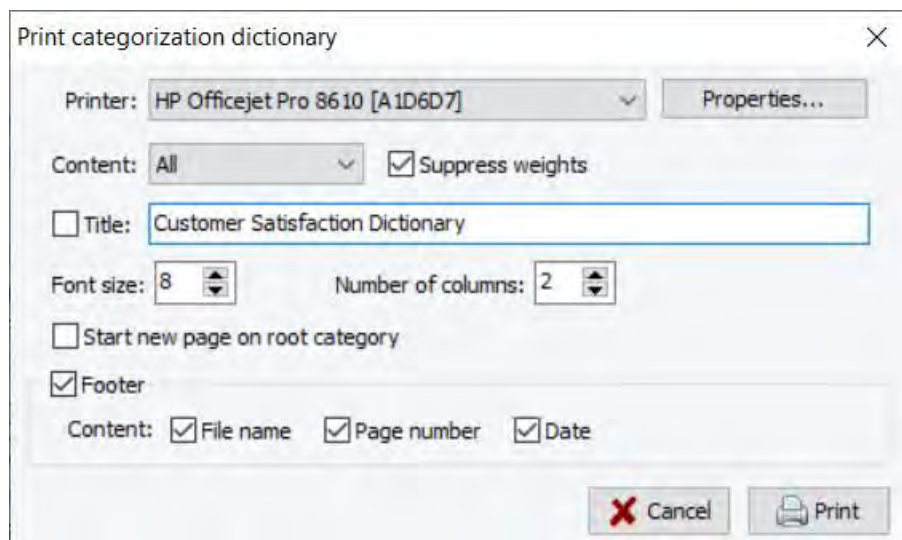
To export a classification dictionary:

- Click the  **Export** button. A **Save File** dialog box will appear.
- Select the proper file format under which you would like to save the file, type the name of the file.
- Click **OK** to create it.

Printing Dictionaries

To create a printed version of the categorization dictionary:

- Click the  button. A print dialog box like the one below will appear, allowing you to set various options of what and how the dictionary should be printed.



The dialog box is titled "Print categorization dictionary" and has a close button (X) in the top right corner. It contains the following options:

- Printer:** A dropdown menu showing "HP Officejet Pro 8610 [A1D6D7]" and a "Properties..." button.
- Content:** A dropdown menu showing "All" and a checked checkbox for "Suppress weights".
- Title:** A checkbox labeled "Title:" followed by a text field containing "Customer Satisfaction Dictionary".
- Font size:** A spinner box set to "8".
- Number of columns:** A spinner box set to "2".
- Start new page on root category:** An unchecked checkbox.
- Footer:** A checked checkbox.
- Footer Content:** Three checked checkboxes: "File name", "Page number", and "Date".

At the bottom right are "Cancel" and "Print" buttons.

Printer: Select the desired printer. To adjust the printer settings, such as the page size and orientation or printer resolution, select the **PROPERTIES** button.

Content: This option allows you to specify what should be printed. Selecting **All** prints the entire dictionary along with all its items. Selecting **As Shown** will print only currently visible items. When the **Select Items** option on the [Categorization](#) tab is enabled, a third option **Selected Items** becomes available allowing you to restrict the printing to previously selected categories and items.

Suppress weights: When this option is disabled, weights of each item are printed within parentheses. Enabling this option prevents those weights from being printed.

Title: Use this option to specify a particular line of text that will appear at the top of each page.

Font size: This option may be used to adjust the font size used to print dictionary items.

Number of columns: Dictionaries may be printed with up to seven columns per page, allowing you to print large dictionaries on fewer pages. Please note that when increasing the number of columns per page, it may be necessary to decrease the font size to prevent the overlapping of items in adjacent columns.

Start new page on root category: Root categories are the dictionary categories still visible when the dictionary is fully collapsed. Selecting this option instructs the program to start the printing of all items starting from this root category at the top of a new page.

Footer: Enable this option to print a footer at the bottom of each page. A footer can consist of up to three items: The **Filename** (printed on the left margin of the footer), the **Page Number** (located at the bottom center of the page), and the **Date** (printed on the right margin of the footer).

Using Lexical Tools 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 several 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.
- Two English thesauri are also used to propose synonyms of words already in your dictionary.
- 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 150,000 root words (including many proper nouns) and offers over 120,000 synonym sets. The availability of word sense definitions allows for manual as well as automatic filtering of proper word senses.

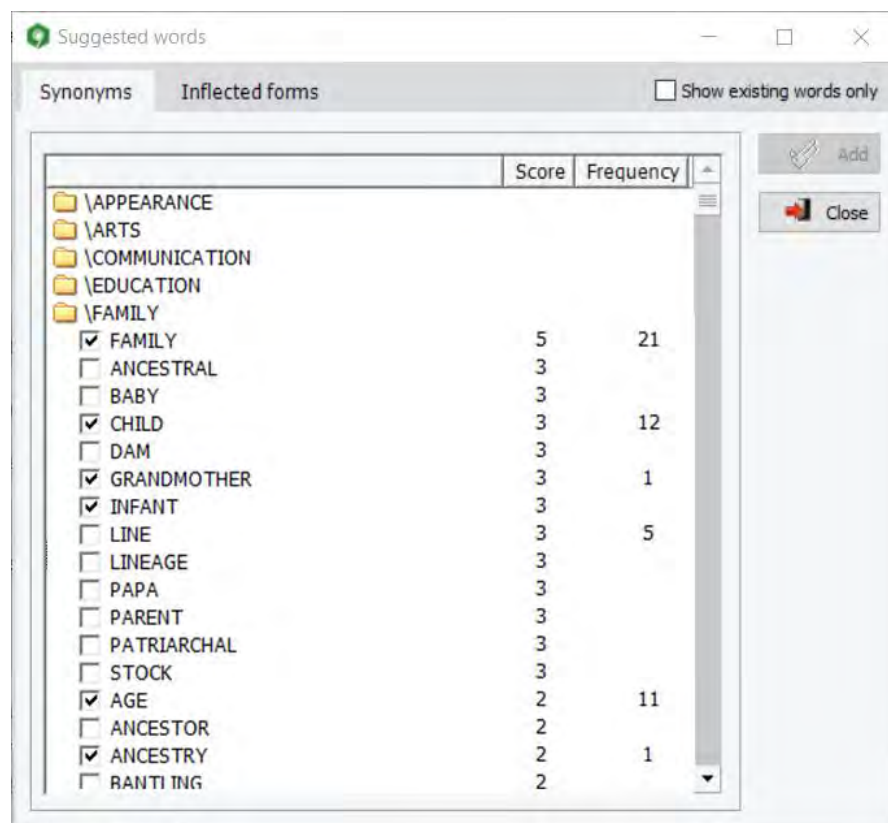
These three tools are available through the auto suggest panel on the frequency list, as well as through two dictionary-building commands.

- The [Basic](#) command uses the selected spelling dictionaries and any available thesauri 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, not only synonyms, but all related words such as hypernyms, hyponyms, holonyms, meronyms, coordinate terms as well as the selected spell-checking dictionaries to find inflected forms of those words.

Basic Dictionary-Building Tools

To access the basic dictionary-building tool:

- Select the **Categorization** tab.
- Press on the  button.
- Select the **Basic** command. WordStat will immediately start looking for synonyms and inflected forms of all words in your categorization dictionary and will report them in a dialog box like this one:



This dialog box 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 under the **Score** column indicates the number of existing dictionary entries to which it was related, while the value under the **Frequency** column indicates how often this word has been found in the current text collection.

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", "understand", "understandable", and "understandably". The frequency score indicates how often this word has been found in the current text collection.

To display only words existing in the current text collection, select the **Show existing words only** option, in the upper right-hand corner of the dialog box. Please note that if this dialog is accessed prior to any WordStat text analysis, this option will be grayed out. Running a simple frequency analysis on the current text collection will collect the frequency information needed to allow this option to be used.

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

Click the **Close** button to return to WordStat.

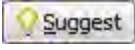
Advanced Dictionary-Building Tools

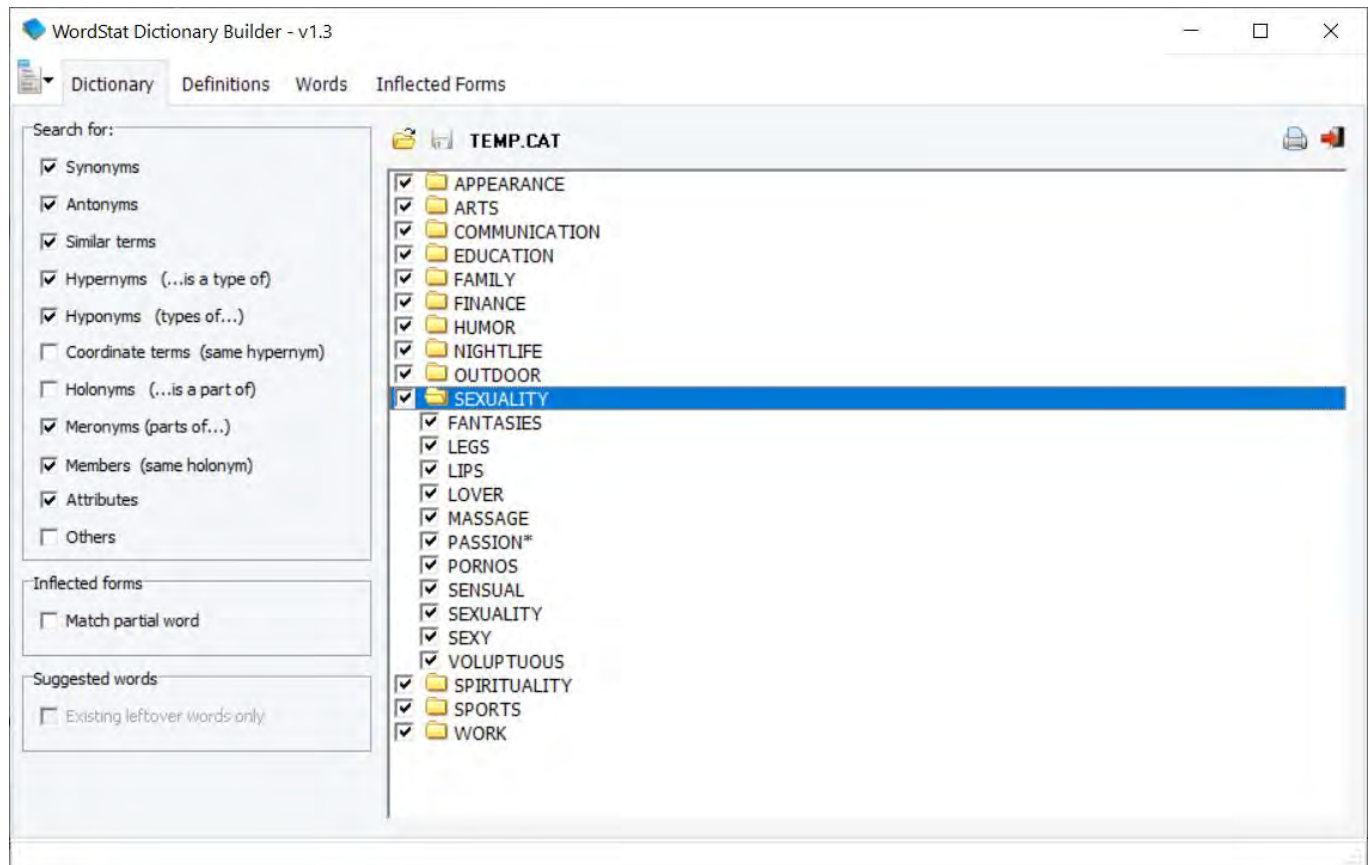
The advanced dictionary-building tool can be accessed either as a stand-alone application or from within WordStat.

To run the stand-alone version:

- Point to the Programs folder in the Windows' Start menu, then select Provalis Research and then click Dictionary Builder.

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

- Select the **Categorization** tab.
- Press on the  button.
- Select the **Advanced** command. A dialog box like this one will appear:



The first tab 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 forms of those entries.

Dictionary Tab

The first tab 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 the  button. A standard Open dialog box 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 mark beside them. Clicking a category check box to change its state also changes the check box state of all words and subcategories within this category.

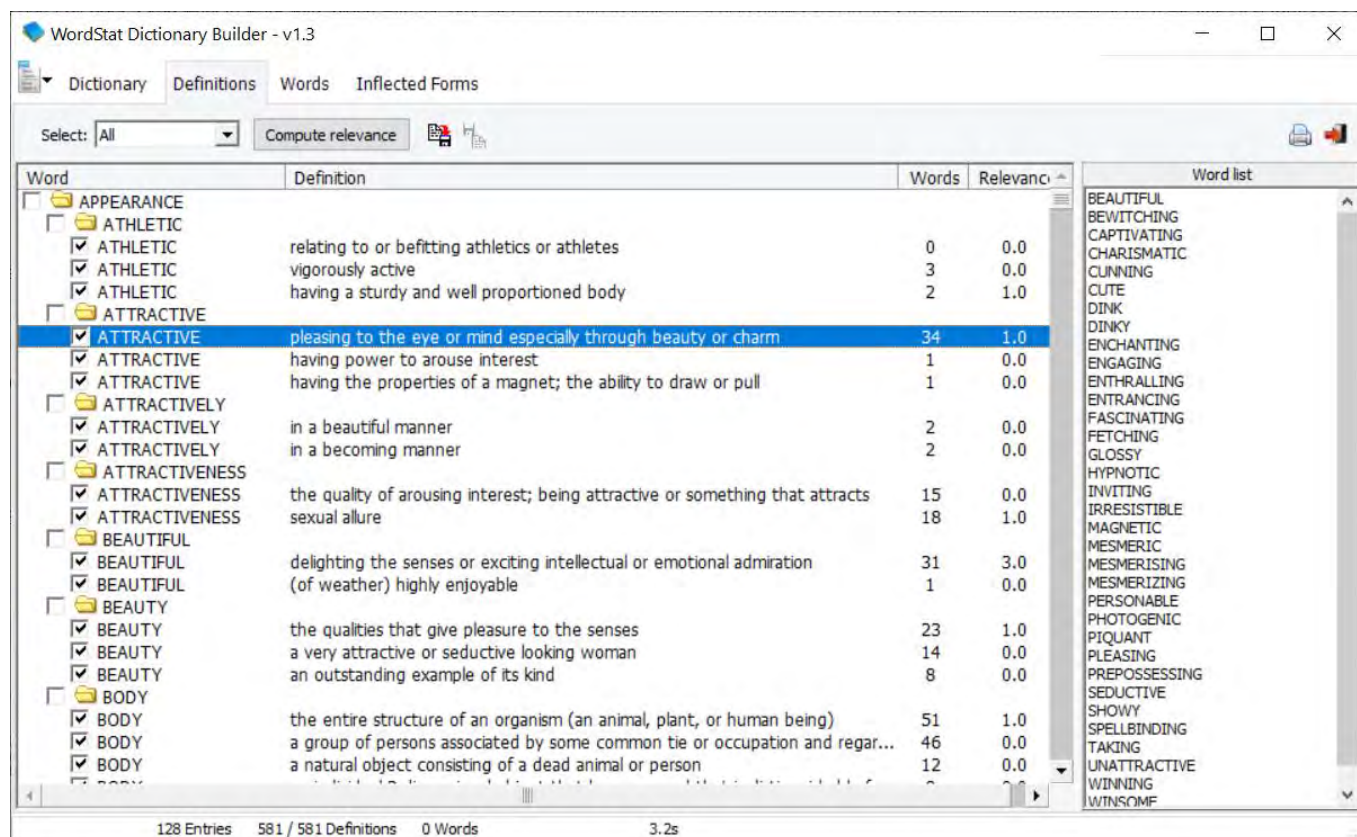
Specifying the type of relationship to look for: The **Search for** group box allows you to specify what type of relationship the program will look for. For example, you may choose to search only for synonyms and similar terms or decide to also search for hypernyms, hyponyms, coordinate terms, etc.

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.

Showing existing words only: By default, suggested words and phrases are presented whether or not they were found in the current text collection. Selecting the **Existing Leftover Words Only** option restricts the list of suggestions to those present in the text collection and not yet captured by the dictionary. This option will be grayed out if no text processing has been done yet in WordStat. Running a simple frequency analysis in WordStat prior to using running this program will collect the frequency information needed to allow this option to be available.

Definitions Tab

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** tab was created to 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 you 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 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 obtained the highest relevance score. When selecting the highest score, a 20% tolerance is used so that, sometimes, more than one word sense will be selected. This selection rule is the most conservative one and ensures 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 many false positives).

Relevance > 0.1: This rule is slightly more conservative than the previous one, in that it also rejects all word senses that have obtained a score of 0.1. Besides a score of zero, 0.1 is the lowest score that may be obtained. Experience has 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 the  button, and later retrieved by clicking the  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 Tab

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** tab includes a list of all words and idioms that were suggested, irrespective of their relationship with the existing entries. The remaining pages allow you to examine those same words by the nature of their relationship with existing entries.

WordStat Dictionary Builder - v1.3

Dictionary Definitions Words Inflected Forms

Compute specificity Add words (8) View definition

All words Synonyms Antonyms Is a type of Types Coordinates Parts Similar

Word	Releva...	Specificity	Other categories
EDUCATION	1046 words		
TRAIN	5.00	1.00	
SCHOLAR	4.08	1.00	
MIND	4.08	0.80	COMMUNICATION 1.00
STUDY	4.00	1.00	
PREPARE	4.00	0.80	WORK 1.00
SCHOLARLY PERSON	3.08	1.00	
CREATIVE THINKER	3.08	1.00	
BOOKMAN	3.08	1.00	
HIGHBROW	3.07	1.00	
MENTOR	3.06	1.00	
EXPONENT	3.06	1.00	
RATIONAL	3.01	1.00	
GRADE	3.00	1.00	
EDUCATION	3.00	0.75	WORK 1.02
DRILL	3.00	1.00	
ASSEMBLAGE	3.00	1.00	
WONDERER	2.06	1.00	
WISE MAN	2.06	1.00	
VISIONARY	2.06	1.00	

STUDY (noun) - a detailed critical inspection
 STUDY (noun) - applying the mind to learning and understanding a subject (especially by reading)
 STUDY (noun) - a written document describing the findings of some individual or group
 STUDY (noun) - a state of deep mental absorption
 STUDY (noun) - a room used for reading and writing and studying
 STUDY (noun) - a branch of knowledge

128 Entries 581 / 581 Definitions 5496 Words 2.8s

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. These 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 the list.

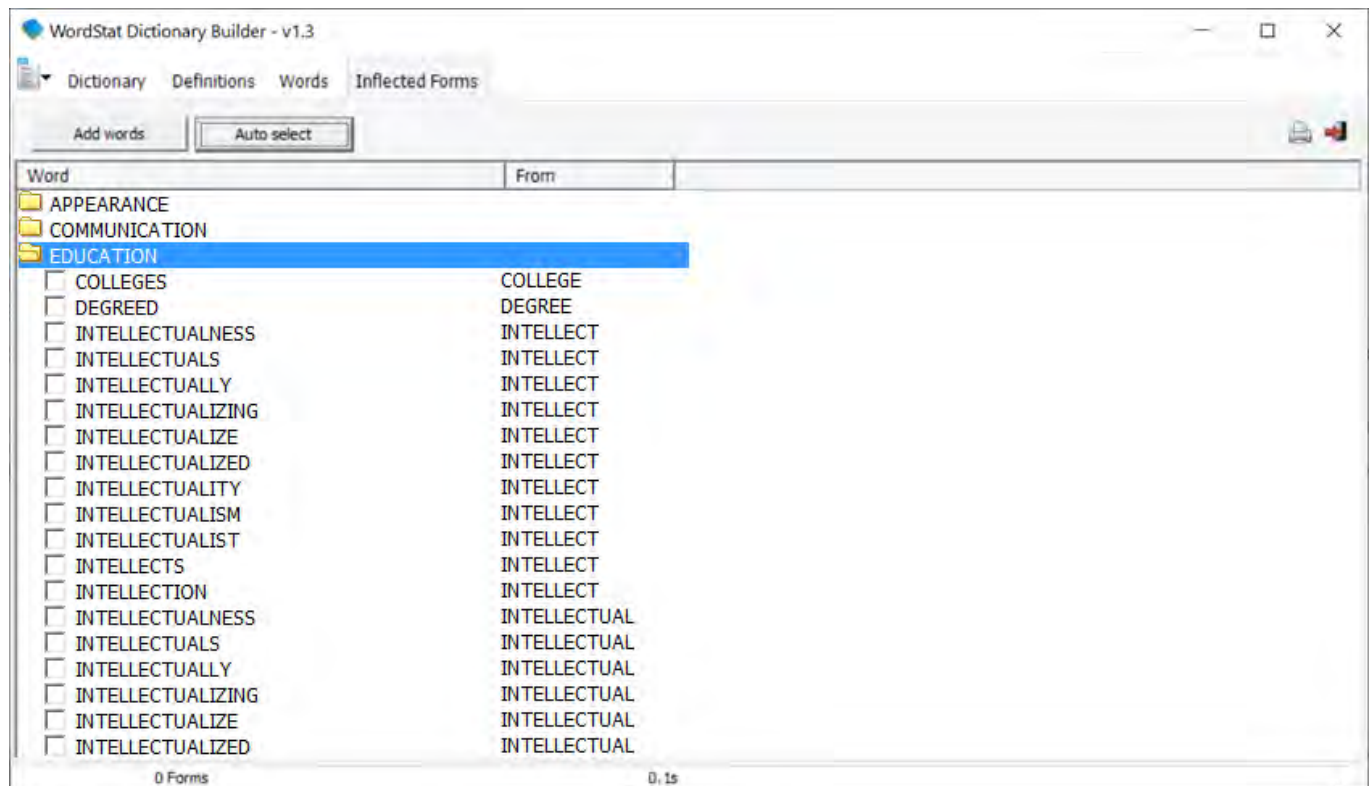
Specificity index: Very often, a word is suggested in more than one category. This is especially true when the dictionary includes categories that are semantically close to each other. One good example of such a categorization system is the Lasswell dictionary that tries to differentiate ten different 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 you to obtain a specificity index as well as a list of all the other categories in which this item 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 the **Add** button.

Inflected Forms Tab

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 **Match Partial Word** option is enabled (see [Categorization](#)), 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:

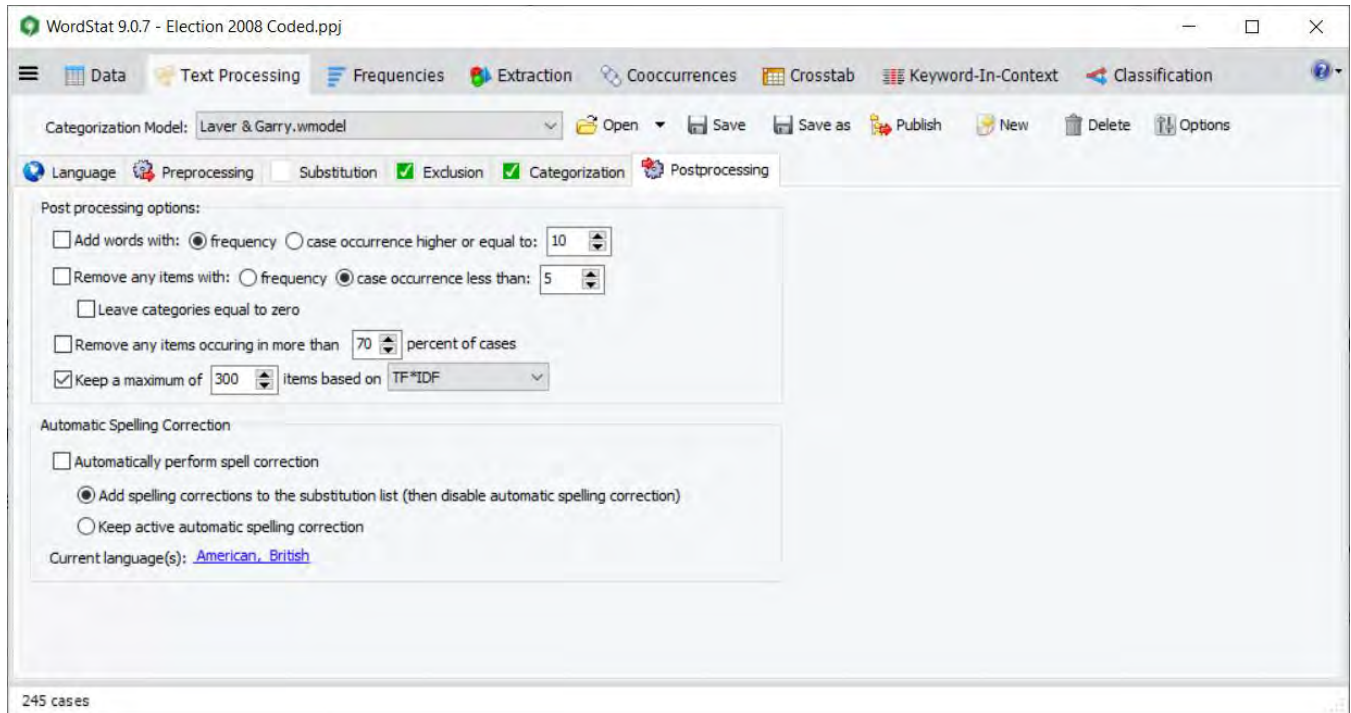
- Place a check mark beside the words you would like to add.
- Click the **Add** button.

The **Auto Select** button allows you to automatically select from all the suggested forms with specific suffixes such as all suggested forms ending with 's' or 'ed'. When searching inflected forms of English words, it is also possible to use WordNet to automatically select words that share the same meaning as the original word from which it was derived. As an example, the program will automatically select words like BEHAVIORS and BEHAVIOURS as valid forms derived from BEHAVIOR since all three forms will yield the same WordNet definitions. You can also set this feature to accept any new word form for which there is at least one WordNet definition containing the original word. For example, when enabling this option the word COMPETING would be automatically selected as a valid inflected or derived form of COMPETITION since one of WordNet definitions associated with COMPETING (i.e. "Being in competition") contains the original word from which it was derived.

- Click the  button to return to quit the dictionary builder program and return to WordStat.

Postprocessing

The **Postprocessing** tab offers different options that control how the textual information should be processed once all other processes (such as stemming, substitution, exclusion, categorization, etc.) have been applied.



Post-Processing Options:

Add Words: When the categorization dictionary is disabled, all words that are not found in the exclusion list will be included in the final keyword frequency analysis. This option allows you to restrict the number of words included to the most frequent ones by setting a minimum **Frequency** or **Case Occurrence** criterion for inclusion. This option may also be used while the categorization dictionary 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 categorization dictionary and cannot be used to remove any items. To remove items in this dictionary based on a frequency or case occurrence criterion see the **Remove Words** option below.

Remove Words: This option allows you to restrict the number of included words or categories to the most frequent ones by setting a minimum **Frequency** or **Case Occurrence** criterion for inclusion. This criterion is applied both to items in the categorization dictionary and words that meet the criterion specified with the **Add Words** option.

Examples:

- If no categorization dictionary is used and you want to include any word that appears at least 10 times, but in no less than 5 different cases, 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 **Case Occurrence** of 5. Only words that meet both criteria will be included.
- When a categorization dictionary 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 dictionary and set the minimum frequency criterion of both the **Add Words** and **Remove Words** options to the required frequency.

- When a categorization dictionary is used to categorize words, but you only want to analyze the most frequent categories, you have to activate the dictionary and set the Remove Words option to the required frequency. In this situation, the Add Words option should be deactivated.

Leave Categories Equal to Zero: By default, WordStat removes from the frequency table any keyword or category in the categorization dictionary that was not found in the analyzed text. Enabling this option instructs the program to leave those items with a zero frequency in the table. This option is especially useful when comparing obtained frequencies to normative data or to other samples. This option should also be enabled when creating norm files (see [Creating and Using Norm Files](#)).

Remove Items Occurring in More than n Percent of Cases: This option allows you to remove keywords or categories appearing in more than a specified percentage of cases. This criterion is applied both to items in the categorization dictionary and to words that meet the criterion specified in the **Add Words** option. This option is especially useful to remove words that are too common to have any informative or discriminative value.

Keep a Maximum of n Items: This option allows you to restrict the number of included words or categories to a maximum number of items, based either on their **total frequency**, number of **case occurrences**, or on the computed **TFxIDF** index. This selection occurs only after all the previous frequency options have been assessed and only if the total number of remaining items is higher than the specified maximum. If the cutting point falls on a frequency or a case occurrence shared by many items, those with the highest **TFxIDF** values will be selected.

Automatic Spelling Correction:

Enabling the automatic spelling correction feature will instruct WordStat to automatically identify misspellings of words or dictionary entries that have met prior selection criteria to be included in the frequency analysis. It will automatically identify misspelling of known words in the currently active spell-checking dictionaries. It will also combine frequency information along with phonetic and string similarities to correct unknown words such as technical terms or proper names. When this automatic spelling correction is enabled, one can choose between two types of outcomes:

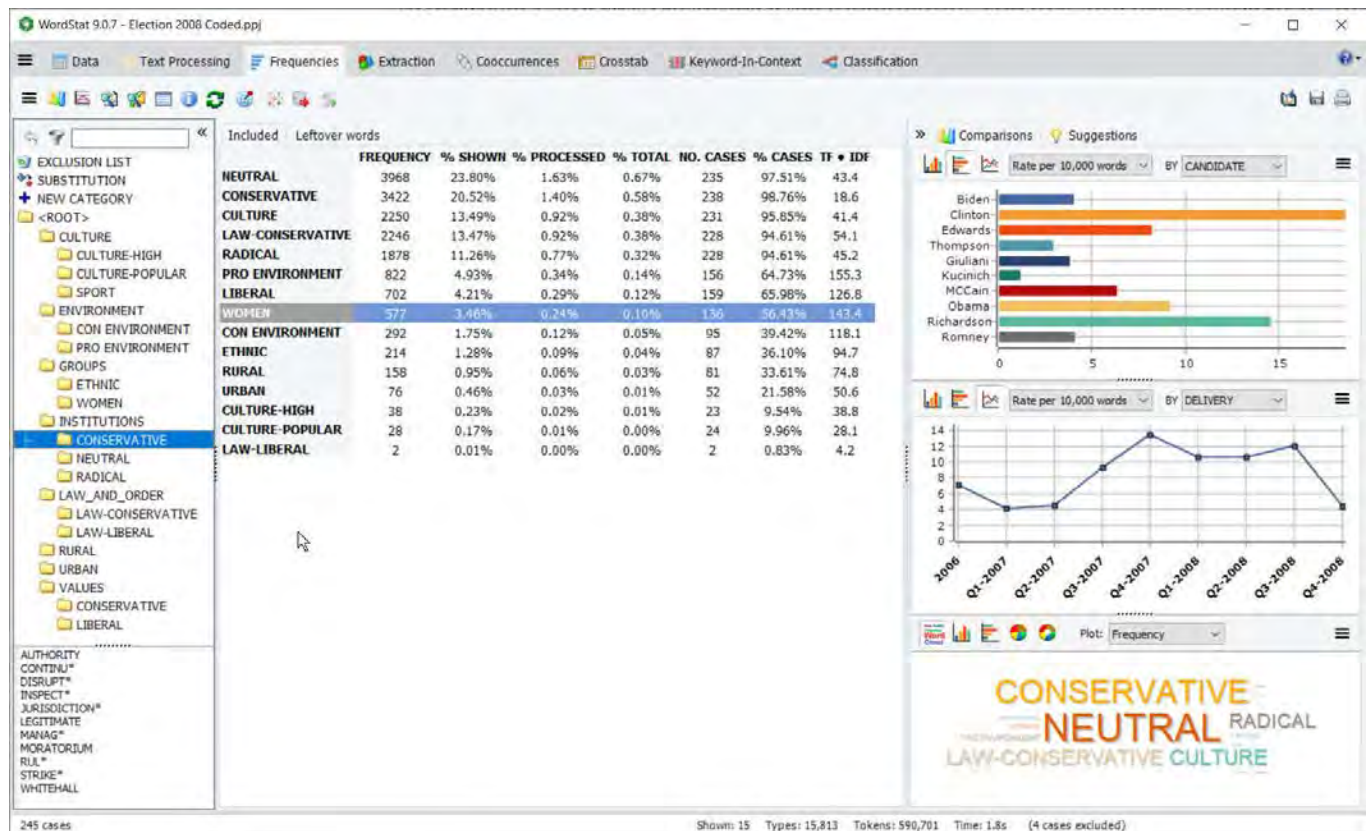
Add Spelling Corrections to the Substitution List - The first option will add all spelling replacements to the substitution list, allowing one to review all replacements made during the last text analysis. Once finished, the automatic spelling correction is disabled, preventing the software from applying spelling corrections on the same set of words. Please note that if the number of words to be analyzed is increased, if additional entries are added to the current categorization model or if one changes the categorization model, it is recommended to re-enable this automatic spelling correction feature in order to identify and correct additional misspellings.

Keep Active Automatic Spelling Correction - The second option will perform automatic spelling correction every time WordStat will need to process the original text documents. Please note that this option will not store spelling corrections in the substitution list. Consequently, it will not be possible to review, override, or modify corrections that have been made. Such an option is especially useful when frequent changes are made to the text processing options.

The **currently language(s)** option lists the active spell-checking dictionaries. To change the active dictionaries, click on dictionary names displayed on the right to display a dialog box that will allow you to enable or disable dictionaries. One may also move back to the [Language](#) page to select different dictionaries.

The Frequencies Tab

The **Frequency** tab is used to display a frequency table of words or content categories. This tab can be used to perform a univariate frequency analysis on words or categories and to modify any of the active dictionaries or word lists.



By default, the table shows the included keywords or categories in descending order of frequency.

The table includes the following statistics:

FREQUENCY	Number of occurrences of the keyword
% SHOWN	Percentage based on the total number of keywords displayed in the table
% PROCESSED	Percentage based on the total number of words encountered during the analysis
% TOTAL	Percentage based on the total number of words that have not been excluded
NO OF CASES	Number of cases where this keyword appears
% CASES	Percentage of cases where this keyword appears
TF*IDF	Term frequency weighted by inverse document frequency. Such a weighting is based on the assumption that the more often a term occurs in a document, the more it is representative of its content yet, the more documents in which the term occurs, the less discriminating it is.

Tabs at the top of the table allow you to access (1) a frequency table of all **Included** content categories or keywords or (2) a frequency table of **Leftover words** consisting of individual words that have not been categorized or included in the analysis.

To the left of the main grid, a panel lists the content categories of the current dictionary following additional entries representing the substitution and exclusion processes. This panel may be used to quickly assign items in the table to any one of these locations using the drag-and-drop operation. For more information on how to use this panel, see [Using the Dictionary Panel](#).

The  button can be used to move one or several words to the exclusion list or to add or remove a word from the categorization dictionary. The permitted moves depend on the items currently displayed. This button may also be used to display a Keyword-In-Context table of the selected word.

It is also possible to quickly access the pop-up menu invoked by this button by pressing the right button of the mouse anywhere on the grid.

To the right of the main grid is a panel containing two tabs: **Comparisons** and **Suggestions**.

The Comparisons Tab

The **Comparisons** tab allows you to look at the distribution of the selected words or categories among values of up to two structured variables. You may display this distribution using either a vertical bar chart, a horizontal bar chart or a line chart, by clicking on the corresponding button.

Four statistics may also be represented on the charts:

Case Occurrence	Number of cases in this subgroup containing at least one of these words.
Category Percent	Percentage of cases in this subgroup containing at least one of these words.
Word Frequency	Total number of these words in this subgroup.
Rate per 10,000 Words	Rate of words in this subgroup per 10,000 words.

The bottom chart contains the distribution of keywords or categories that can be represented as a word cloud, a vertical or horizontal bar chart, a pie chart and a donut chart. You can display the distribution with the same statistics seen on the frequency table: Frequency, % Shown, % Processed etc.

Right-clicking anywhere in the chart areas displays a pop-up menu that allows you to edit the chart, save it to disk or in the **Report Manager**, or copy it to the clipboard. Clicking a specific bar or a data point of a line chart also allows one to retrieve text segments associated with the selected class and containing words of the selected topic.

The Suggestions Tab

The **Suggestions** tab enables you to view an automatic suggestion panel displaying leftover words potentially related to the currently selected item. For more information on the auto-suggest panel, see [Working with the Auto-Suggest Panel](#).

The Frequencies Toolbar

The  button allows you to produce bar charts or pie charts to visually display the distribution of specific keywords or categories.

To produce charts:

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

For further information see [Bar Charts, Pie Charts, and Word Clouds](#)

The  button allows you to create normative frequency data from the current file, to store them on disk and to compare currently displayed frequencies with previously saved norms. See [Creating and Using Norm Files](#) for more information on this topic.

The  button allows you to access a keyword retrieval feature to retrieve all documents, paragraphs or sentences containing a specific keyword or a combination of keywords. See [Keyword Retrieval](#) for more information on this topic.

The  button allows you to automatically attach QDA Miner codings to all paragraphs or sentences associated with currently displayed content categories or keywords. When clicked, a dialog box asks whether the coding should be applied to whole paragraphs or to individual sentences. If some WordStat keywords have no corresponding QDA Miner codes, new codes will be created under a special codebook category before the autocoding process begins.

The  button is used to draw color guidelines on alternate rows in order to facilitate the reading of large tables. When clicking this button color guidelines are shown. Clicking this button again removes the color guidelines.

The  button displays various statistics on the text categorization process, such as the total number of words processed, the number and proportion of words that have been excluded and of those that have been categorized. The dialog box also displays document statistics - such as the average length in words of sentences, paragraphs and documents - as well as several coverage statistics, including the percentage of cases, paragraphs and sentences containing at least one keyword and the proportion of words that have been categorized or included. The coverage statistics are especially useful when you apply a content analysis dictionary developed to describe a specific data set on new data sets. A significant decrease in coverage may indicate the need to update a dictionary in order to better reflect changes over time or specific differences in this new data set.

The  button is used to reapply the content analysis process on the current data set. This button is disabled by default, and it becomes enabled when changes are made to any one of the currently active text analysis processes (such as the categorization dictionary, the exclusion list, or the substitution process). Clicking this button will instruct WordStat to reprocess the text collection and update the current table.

The  button is used to access the mapping function, which allows you to analyze the spatial distribution of various terms.

The  button is used to export the selected data to Tableau, a software used for interactive data visualization.

The  button may be used to append frequency information to the current data file or save to disk a matrix of word or keyword frequency by cases. For more information on one of these topics see [Exporting Frequency Data](#).

The  button is used to apply post-processing scripts on WordStat data files, For more information on how run an existing script or create your own script, see [Running and Creating Post-Processing Scripts](#).

To append a copy of the table in the Report Manager:

- Click the  button. A descriptive title will be provided automatically for the table. To edit this title or to enter a new one, hold down the **Shift** keyboard key while clicking this button.

(for more information on the Report Manager, see the [Report Management Feature](#) topic).

To export the frequency table to disk:

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

- Type a valid file name with the proper file extension.
- Click the **Save** button.

To copy the entire table to the clipboard:

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

To copy selected rows to the clipboard:

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

To search for a specific item:

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

Using the Dictionary Panel

The dictionary panel provides an easy way to assign words or phrases to the current categorization dictionary and to the exclusion or substitution lists. It may also be used to remove or edit existing items. This panel is located to the left of the **Frequencies**, **Crosstab** and **Phrase Finder** pages and looks like the one shown below.

WordStat 9.0.7 - Election 2008 Coded.ppt

Menu: Data, Text Processing, Frequencies, Extraction, Cooccurrences, Crosstab, Keyword-In-Context, Classification

Left Panel: Tree structure showing categories like EXCLUSION LIST, SUBSTITUTION, NEW CATEGORY, <ROOT>, CULTURE, ENVIRONMENT, GROUPS, INSTITUTIONS, CONSERVATIVE, NEUTRAL, RADICAL, LAW, RURAL, URBAN, VALUES, CONSERVATIVE, LIBERAL.

Center Panel: Frequency table

	FREQUENCY	% SHOWN	% PROCESSED	% TOTAL	NO. CASES	% CASES	TF * IDF
NEUTRAL	3968	23.80%	1.63%	0.67%	235	97.51%	43.4
CONSERVATIVE	3422	20.52%	1.40%	0.58%	238	98.76%	18.6
CULTURE	2250	13.49%	0.92%	0.38%	231	95.85%	41.4
LAW-CONSERVATIVE	2246	13.47%	0.92%	0.38%	228	94.61%	54.1
RADICAL	1878	11.26%	0.77%	0.32%	228	94.61%	45.2
PRO ENVIRONMENT	822	4.93%	0.34%	0.14%	156	64.73%	155.3
LIBERAL	702	4.21%	0.29%	0.12%	159	65.98%	126.8
WOMEN	577	3.46%	0.24%	0.10%	136	56.43%	143.4
CON ENVIRONMENT	292	1.75%	0.12%	0.05%	95	39.42%	118.1
ETHNIC	214	1.28%	0.09%	0.04%	87	36.10%	94.7
RURAL	158	0.95%	0.06%	0.03%	81	33.61%	74.8
URBAN	76	0.46%	0.03%	0.01%	52	21.58%	50.6
CULTURE-HIGH	38	0.23%	0.02%	0.01%	23	9.54%	38.8
CULTURE-POPULAR	28	0.17%	0.01%	0.00%	24	9.96%	28.1
LAW-LIBERAL	2	0.01%	0.00%	0.00%	2	0.83%	4.2

Right Panel: Comparisons, Suggestions. Suggests: Less, More. Lists: SYNONYMS (BROAD, BROADER, etc.), ANTONYMS (BOYS, MAN, etc.), RELATED WORDS (ADULT, ADULTS, etc.).

Status: 245 cases, Shown: 15, Types: 15,813, Tokens: 590,701, Time: 1.8s, (4 cases excluded)

The main section of this panel is a tree representing the structure of the current content analysis along with the substitution and exclusion processes (if active). When an item on this list is selected, its content is listed in a resizable window below the tree structure. At the top of the panel, a button allows you to undo the last change made to any one of the lists.

To move items from the frequency list to a list or to an existing content category:

- In the table to the right of this panel, click the word or phrase you wish to assign to a list or content category. To select a group of adjacent entries, move the mouse cursor over the first item in the list, click and hold the mouse button, drag the mouse to the last entry to highlight the block of rows you want to assign, and then release the mouse button. To select disjointed items, hold down the **Ctrl** key while clicking each one of the items.
- Once the words or phrases have been highlighted, drag and drop them on the category of your choice. To add an item to the root folder of a categorization dictionary, simply drop it into the **<ROOT>** folder.

To move items to a new category:

- Select the words or phrases you would like to assign to this new category.
- Drag and drop the items on the **+ NEW CATEGORY** item. An [Add Word/Category dialog box](#) will appear, allowing you to type the name of this new category.

To rename or delete an existing content category:

- In the tree representation of the dictionary, right-click the category you would like to rename or delete.
- Select the appropriate command.

To rename, delete or move an item in a list or in a content category:

- In the window below the tree display, right-click the item you would like to modify.
- Select the appropriate command.

To cancel a modification:

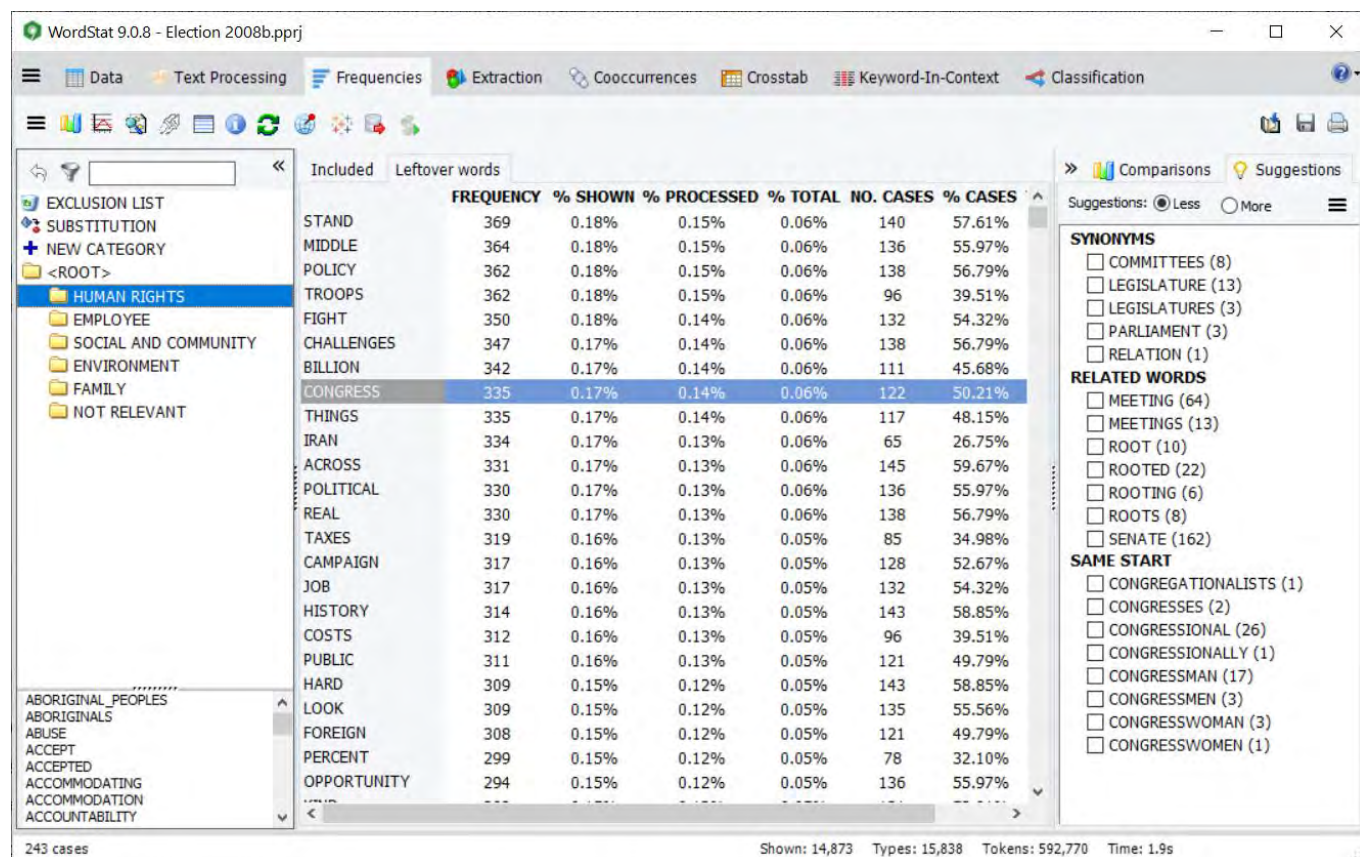
- If you want to undo the last assignment, deletion or modification made using this panel, click the  **Undo** button.

Note: If you leave the mouse cursor over this button a hint window will appear showing you which modification will be canceled by clicking this button.

Working with the Suggestions Tab

One of the biggest challenges of quantitative content analysis lies in the fact that a single idea may be expressed in many different ways, like using synonyms, paraphrases, or idioms. When you need to identify all instances of such an idea, a critical task becomes identifying these numerous forms. WordStat provides several tools to support this task such as the [Suggest](#) feature on the **Text Processing** tab that can be used to retrieve a list of all known synonyms, related words and inflected forms of the items already in a dictionary from another source such as a thesaurus or a lexical database. The **Suggestions** tab on the **Frequencies** tab also lists suggestions. It differs, however, from the **Suggest** feature in several ways. First, it only displays suggested words that were found in the current text collection and that are not already in the categorization dictionary (leftover words). It may also be used not only on content categories but also on any words extracted by WordStat and displayed in the **Included Words** and **Leftover Words** lists. It may thus be used from the very beginning of the dictionary construction process to quickly identify potential groupings of words and assign the relevant ones to existing or new content categories. To display this panel, select the **Suggestions** tab to the right of the frequency table. To display suggestions, simply select the appropriate row in the frequency grid. When the selected item is a word, the panel will display synonyms, antonyms, related items and words with a similar beginning (potentially related items, inflected and misspelled forms). When the selected row consists of a content category, then the panel displays the same information but for all words currently in this content category. Selecting more than one row will also result in a compound view of suggestions of all selected items.

At the top of the panel, radio buttons allow you to choose the extent of the suggestions. By default, the panel returns the most likely synonyms, while related words consist of hypernyms, hyponyms, holonyms and meronyms. Choosing the **More** option retrieves other potential synonyms, as well as coordinate terms and possible attributes of the selected item.



To assigning suggestions to a dictionary:

There are two methods to assign suggested words to the categorization dictionary or to the exclusion list. The first method will perform these operations on items in this panel only, while the second method will also include selected items in the main frequency list.

- To perform one of the above-mentioned operations on the suggested items only, select the check boxes of one or several suggestions, right-click your mouse, and choose the appropriate command.
- To include with the suggestions, currently selected words in the frequency table, click the  button on the toolbar and choose the appropriate command. You may also drag and drop words from the frequency table to the appropriate location in the **Dictionary** panel on the left. When dropping items from the frequency table, all selected suggestions will also be moved to the selected location.

Barcharts, Pie Charts, and Word Clouds

WordStat allows you to produce bar charts or pie charts to visually display the distribution of specific keywords or categories.

To produce charts:

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

- Click the  button.

Types of Charts

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



The word cloud is an image composed of keywords or content categories, sometimes phrases in which the size of each item indicates its frequency.



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



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



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



The donut chart, similar to the pie chart, is useful to display the relative frequency of each value and compare individual values to other values and to the whole. Donut charts are considered easier to read as you can focus on reading the length of the arcs, rather than comparing the proportions between slices.

The **Plot** option allows you to select the values that will be used as the scale for the length of bars in bar charts or as the percentage base for pie charts.

For bar charts the options are:

FREQUENCY	Number of occurrences of the keyword
% SHOWN	Percentage based on the total number of keywords displayed in the table
% PROCESSED	Percentage based on the total number of words encountered during the analysis
% TOTAL	Percentage based on the total number of words that have not been excluded
NO OF CASES	Number of cases where this keyword appears
% CASES	Percentage of cases where this keyword appears

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

FREQUENCY	Percentage based on the total frequency of keywords
NO OF CASES	Percentage based on the total number of case occurrences

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

The Toolbar

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

Control: **Description:**



Press this button to vertically display the labels on the bottom axis.



Press this button to append a copy of the graphic in the Report Manager. A descriptive title will be provided automatically. To edit this title or to enter a new one, hold down the SHIFT keyboard key while clicking this button (for more information on the Report Manager, see the [Report Management Feature](#) topic).



Press this button to retrieve a chart previously saved on disk.



Press this button to save a chart on disk. Charts may be saved in BMP, JPG or PNG graphic file format or may be saved in a proprietary format (.WSX file extension) that may later be edited and customized using the Chart Editor.



Pressing this button prints a copy of the displayed chart.



Click this button to turn on/off the 3-D perspective for the current chart.



This button allows the editing of various features of the chart such as the left and bottom axis, the chart and axis titles, the location of the legend, etc. (see below for available options)



This button is used to create a copy of the chart to the clipboard. When this button is clicked, a pop-up 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 box and returns to WordStat's main screen.

Customizing the Chart

Clicking the  button displays a dialog box that allows you to customize the appearance of bar charts and pie charts. The options available in this dialog box represent only a small portion of all settings available.

Left or Bottom Axis

Minimum / Maximum: WordStat automatically adjusts the vertical axis scale to fit the range of values plotted against it. To manually set these values, type the desired minimum and maximum.

Increment: Increasing or decreasing this value affects the distance between numbers as well as tick marks. Horizontal grid lines are also affected by the modification of this value.

Grid: This option turns horizontal grid lines on and off. Grid lines extend from each tick mark on an axis to the opposite side of the graph. To increase or decrease the number of grid lines or the distance between these lines, change the increment value of the axis. A list box also allows a choice among five different line styles to draw these grid lines.

Titles

Proper titles and axis labels are of utmost importance when describing the information displayed in a chart. By default, WordStat uses variable names and labels as well as other predefined settings to provide such descriptions.

The title page allows you to modify the **top** title, as well as the labels on the **left**, **bottom** and **right** axis. To edit the title, select the proper radio button. Enter several lines of text for each title by pressing the <Enter> key at the end of a line before entering the next line.

The **Font** button on the right side of the edit box allows changing the font size or style of the related title.

3D View

Orthogonal: Turning this option off disables the free elevation and rotation of the 3-D chart.

Zoom: This option zooms the whole chart. Expressed as a percentage, increasing the value positively will bring the chart towards the viewer, increasing the overall chart size as the Zoom value increases.

3-D Percent: The 3-D Percent property indicates the size ratio between chart dimensions and chart depth by specifying a percent number from 1-to-100.

Perspective: Use this property with Orthogonal unchecked to modify the 3-D perspective of the Chart. Larger values add more depth perspective.

Bar shadow: Enabling this option adds a shadow to the sides of 3-D bars. Turning it off will color the sides of the bar the same as the front.

Bar width: This option determines the percent of total bar width used. Setting this value to 100 makes joined bars.

Bar depth: Use this property to limit the depth that each bar series uses. By default, bars will take up the part proportional to the number of bar series in the chart so that the back of a bar will join the front of the bar immediately behind it. To insert a gap between a series of bars, decrease this value.

Pie depth: Use this property to change the thickness of the pie chart.

To further customize the chart, modify data points or value labels, click the  button located on the right side of the dialog box.

Keyword Retrieval

The **Keyword Retrieval** feature can retrieve any document, paragraph or sentence containing a specific keyword, a combination of keywords or no keyword at all. Text units corresponding to the search criteria are returned in a table on the **Results** tab. This table may then be printed or saved to disk. It may also be used to create tabular or text reports as well as to attach QDA Miner codes to the retrieved text segments.

To start the Keyword Retrieval feature:

- Go to the Frequencies page and click the  button. This dialog box can also be accessed from other parts of the program to retrieve text units associated with a cluster or with a specific association. It may also be accessed by selecting an item in the frequency page, right-clicking and selecting **Keyword Retrieval** from the pop-up menu. When calling this function, a dialog box similar to this one appears, allowing you to specify the desired search criteria:

Retrieve: This option determines the text unit on which the search will be performed as well as what will be retrieved. You can select three different text units. The **Documents** search unit allows WordStat to apply the search expression on each document associated with a specific case and, if a specific document meets the search condition, its location will be displayed. The **Paragraphs** search unit allows WordStat to display any paragraph meeting the search condition. The **Sentences** search unit will instruct WordStat to return sentences meeting the search condition.

Keyword filtering: This group of options allows you to select the keywords on which the retrieval will be based. Setting the first list box to **No Keyword** will retrieve all text units for which no keyword has been found. This option is especially useful to identify topics or themes that have not been covered by the current categorization dictionary or new keywords that should be added to enhance the coverage of existing categories in the dictionary. Setting the filtering to **Include at least** allows you to retrieve text units containing a minimum number of keywords from a selected list. To select the keywords, click the arrow button to show all available keywords and then click the desired items. The minimum number of keywords a specific unit must contain in order to be retrieved is specified using a small edit box with spin buttons on the right. Setting this numerical value to the total number of keywords selected will force the program to retrieve only those units containing all those keywords. Setting this number to a lower value will retrieve all text units containing at least this number of keywords from the selected list. In the example shown above, any unit containing keywords from at least one of the three categories AMENDMENT, CIVIL, or DIPLOMACY will be retrieved.

To enter a second filtering condition, click the  button. You can choose to link the two filtering conditions using either one of the three Boolean expressions: AND, OR or NOT. Choosing AND will retrieve all text units fulfilling both criteria; selecting OR will result in a retrieval of text units meeting either the first or the second condition, or both, while choosing the NOT Boolean operator will retrieve text units meeting the first condition but not the second one.

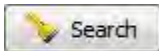
To limit the filtering conditions to a single statement, click the  button.

Variable filtering: The second group of options allows you to restrict the retrieved text units to specific cases selected according to some logical condition. This filtering condition may consist of a simple expression, or may include up to two expressions joined by a logical operator (i.e., AND, OR). In the above screen shot, only text units from cases where the variable CANDIDATE is equal to PARTY will be retrieved.

The following table shows the various operators available for each data type:

DATA TYPE	AVAILABLE OPERATORS
NOMINAL / ORDINAL	Equals Does not equal Is empty Is not empty
NUMERIC and DATE	Equals Does not equal Is greater than Is lesser than Is greater than or equal to Is lesser than or equal to Is empty Is not empty
BOOLEAN	Is true Is false
STRING	Contains Does not contain Is empty Is not empty

Add variables: This drop-down checklist box may be optionally used to add the values stored in one or more variables to the table of retrieved segments for the specific case from which a text segment originated.

- Once all the search options have been set properly, simply click the  button to retrieve the selected text units.

Working with the Retrieved Text Units

The retrieved text units are displayed in a table found on the **Results** tab.

Keyword Retrieval

Criterion Results

CODE: Multilines grid

Case #	Text	Matching	Case	CANDIDATE	DELIVERY	Variable	Paragraph	Word Count
63	Most importantly, this plan will ensure that we control the energy we use with resources and technology that are available today. The steps I just spoke about are not far-off, pie-in-the-sky solutions, they are now. Today, there are waiting lists for fuel-efficient cars. There's an old steel mill in Pennsylvania that has become the home of a new wind turbine factory. I've seen a small business in Nevada powered entirely by solar power. Across the planet, countries like Germany and the United Kingdom have already implemented clean energy policies that are reducing their carbon emissions right now, and leaders like Tony Blair and Angela Merkel have done a great job of raising the visibility of climate change within the G8. Now it's our turn to lead - to show that this future is possible for America.	POWER	1-Obama20080711	Obama	Q3-2008	DOCUMENT	23	143
64	Qaeda, the Taliban, and all of the terrorists responsible for 9/11, while supporting real security in Afghanistan.	SECURITY	1-Obama20080715	Obama	Q3-2008	DOCUMENT	2	16
64	Our men and women in uniform have accomplished every mission we have given them. What's missing in our debate about Iraq - what has been missing since before the war began - is a discussion of the strategic consequences of Iraq and its dominance of our foreign policy. This war distracts us from every threat that we face and so many opportunities we could seize. This war diminishes our security, our standing in the world, our military, our economy, and the resources that we need to confront the challenges of the 21st century. By any measure, our single-minded and open-ended focus on Iraq is not a sound strategy for keeping America safe.	MILITARY; SECURITY	1-Obama20080715	Obama	Q3-2008	DOCUMENT	11	113
64	I am running for President of the United States to lead this country in a new direction - to seize this moment's promise. Instead of being distracted from the most pressing threats that we face, I want to overcome them. Instead of pushing the entire burden of our foreign policy on to the brave men and women of our military, I want to use all elements of American power to keep us safe, and prosperous, and free. Instead of alienating ourselves from the world, I want America - once again - to lead.	POWER; MILITARY	1-Obama20080715	Obama	Q3-2008	DOCUMENT	12	91

To achieve that success, I will give our **military** a new mission on my first day in office: ending this war. Let me be clear: we must be as careful getting out of Iraq as we were careless getting in. We can safely redeploy our combat brigades at a pace that would remove them in 16 months. That would be the summer of 2010 - one year after Iraqi **Security** Forces will be prepared to stand up; two years from now, and more than seven years after the war began. After this redeployment, we'll keep a residual force to perform specific missions in Iraq: targeting any remnants of al Qaeda; protecting our service members and diplomats; and training and supporting Iraq's **Security** Forces, so long as the Iraqis make political progress.

123 Paragraphs (0.2s)

This table contains the case number and the variable from which the segment originates, as well as the value of all additional variables selected by the user (see the **Add variables** option above). When searching for paragraphs or sentences, the table also displays the text associated with the retrieved unit and its location (its paragraph and sentence number). By using arrow keys or by clicking a row the associated text is displayed in a separate window at the bottom of the screen with all keywords in bold. Selecting specific words or phrases in this text window by right-clicking displays a pop-up menu to assign them to a content category, the exclusion list or to obtain a keyword-in-context list.

To sort the table of retrieved units in ascending order in any column, simply click the column header. Clicking the same column header a second time sorts the rows in descending order. Tables may also be printed, stored as a text report, or exported to disk in various file formats such as Excel, ASCII, or HTML.

To remove a search hit from the hit list:

- Select its row and then click the button.

To assign a QDA Miner code to a specific search hit:

- In the table of search hits, select the row corresponding to the text segment you want to code.
- Use the **CODE** drop-down list located above this table to select the code to assign.
- Click the button to assign the selected code to the highlighted text segment.

To assign a QDA Miner code to all search hits:

- Use the **CODE** drop-down list located above this table to select the code to assign.
- Click the button to assign the selected code to all text segments matching the search expression.

NOTE: To automatically attach QDA Miner tags to all paragraphs or sentences associated with all currently displayed content categories or keywords, you may use the autocoding feature by clicking the  button on the **Frequencies** tab.

To create a new QDA Miner code:

- Click the  button. A dialog box will appear allowing you to create a new QDA Miner code (for more information see Adding a QDA Miner codes).

To obtain a word cloud and perform a word frequency analysis on retrieved text segments:

- Click the  button. The analysis will be performed on all text segments retrieved and displayed in the table. (See [Word Frequency Analysis](#) for more information on this feature).

To save the table to the Report Manager:

- Select the  button. A copy of the table will be appended in the **Report Manager**.

To create a report of coded segments:

- Click the  button. The sort order of the current table is used to determine the display order in the report. This report is displayed in a text-editing dialog box and may be modified, stored on disk (in RTF, HTML or plain text format), printed, or cut-and-pasted into another application. Graphics and tables may also be inserted anywhere in this report.

To export the table to disk:

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

To print the table:

- Click the  button.

To close the keyword retrieval dialog box:

- Click the  button.

Running and Creating Post-Processing Scripts

The post-processing scripting feature allows one to extend the capabilities of WordStat by the writing of additional computation or data transformation scripts. Those scripts may be written in Python or R using custom codes or existing libraries for NLP, machine learning, statistical processing, visualization, and more. Such a feature allows data scientists to focus on the crucial processing step of interest, leaving to WordStat the various tasks associated with the data preparation and preprocessing such as document importation, stemming, lemmatization, spelling correction, removal of stop words, and so on. It also allows one to combine WordStat's powerful categorization features (with word patterns, phrases, disambiguation rules, etc.) with new analytics techniques not available in WordStat.

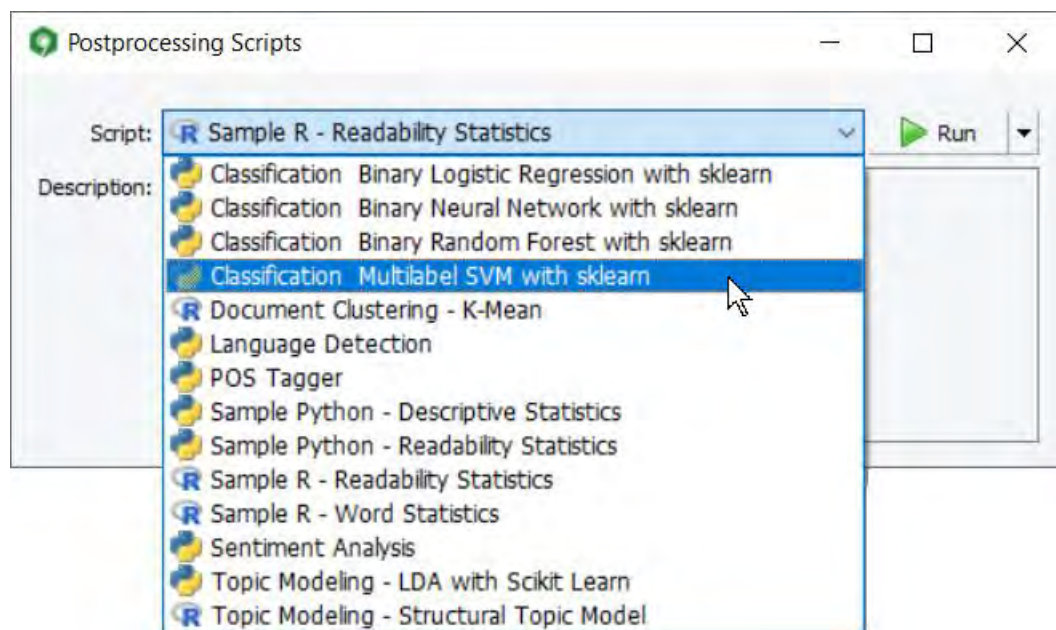
The additional possibility to design dialog boxes allows data analysts to create simple graphic users' interfaces that may be used for customizing script execution and facilitate their use by other WordStat users with no programming skills.

Running an existing Python or R post-processing script:

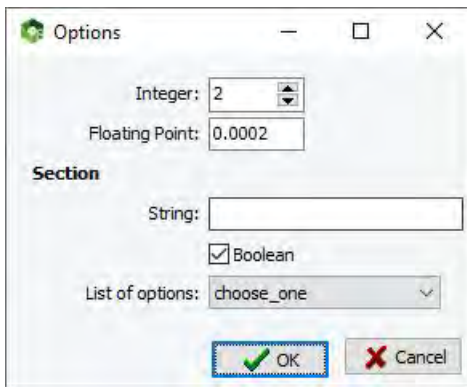
On the Frequencies page, applying an existing post-processing script can be as simple as these steps:

- Click the  button to open the main post-processing script window.

The **Script** dropdown menu lists all existing Python and R scripts. Select the desired script. Beneath the **Script** dropdown menu is the **Description** section. Any text describing the script will be displayed here. Resizing the post-processing script window may be required to display all the description text,



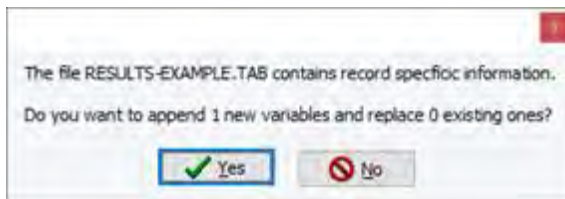
- Click the  button. For some scripts, no other steps are required, and the Report Manager will display the output of your script. Running other scripts will result in the appearance of a dialog box like the one below, allowing you to set some analysis options.



- While the script is running, a Python console or R console will appear, displaying what is logged or printed as output (stdout) or any error messages.

Output

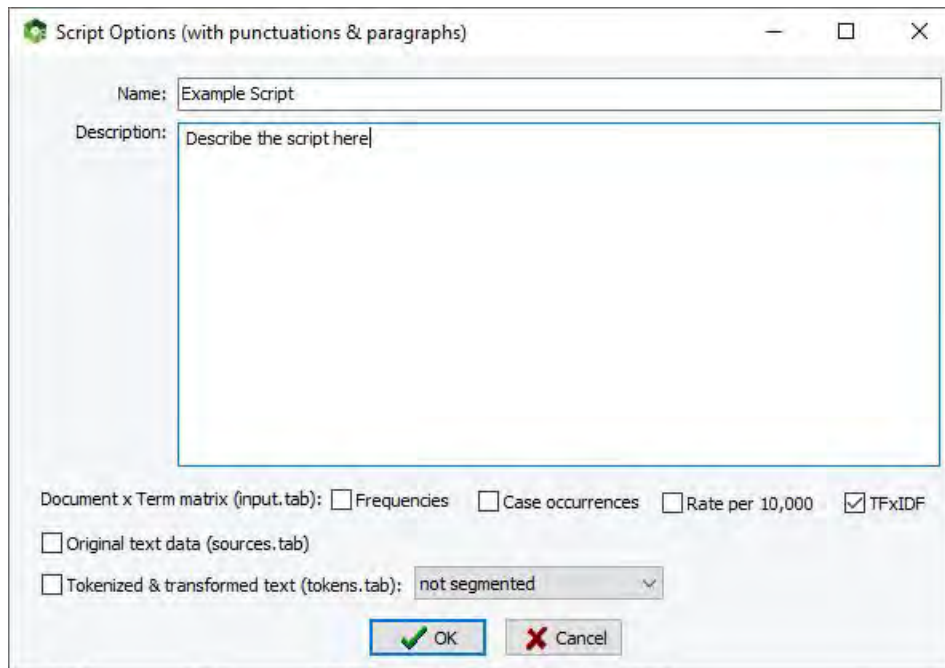
- If a script is executed successfully, WordStat will import any output stored in any supported file format and display them in the Report Manager, which will open automatically. WordStat will import text output stored in any file with a .TXT file extension, assuming an UTF-8 text encoding, or documents stored in .DOCX, .RTF or HTML files. It will also import table output files with either .CSV, .TAB, .TSV, or .XLS file extension, and any graphic produced with a .JPG, .JPEG, .PNG, .GIF, or BMP file extension. Once imported, those files are deleted.
- If a table created by the script has a column named **RECNO** containing record numbers, a dialog box will appear, giving the possibility to append data contained in this table to the current project data.



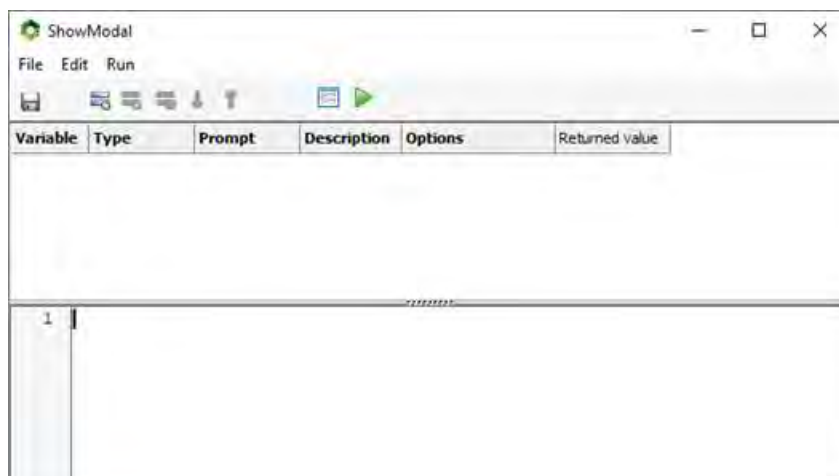
To append data in this table, click **Yes**. If the table's column names correspond to existing variables, data into those variables will be overwritten, while new column names will result in new variables being appended. Clicking **No** will append the table to the report manager.

Writing a new Python or R post-processing script

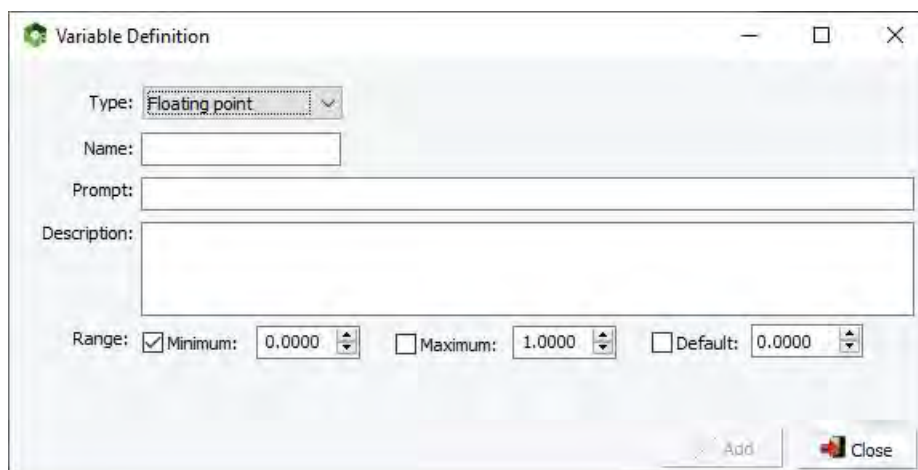
- To create a new post-processing script, open the main post-processing script window by clicking the  button on the **Frequencies** page.
- Then click the  button next to the **Run** button and select **New script** from the dropdown menu.
- Select the desired programming language for this script (Python or R)
- A **Script Options** window will appear. Type the **Name** of the new script.
- Optionally, add a **Description**. This description will appear when the script is selected from the script selection dropdown box.



- The options below allow you to select the types of input files that will be generated and processed by the script. Up to three separate data files can be generated from a choice of seven options.
 - The **input.tab** data file contains numerical values resulting from the quantification of text data by WordStat where each row represents a document, and columns consist of frequency information for every item displayed in the Frequencies tab (i.e., either words or content categories). When no categorization process is applied, such a data file corresponds to a Document x Term matrix. A choice of one of four statistics may be stored in this data file: the term's frequencies, the case occurrences (i.e., either 0 or 1), The rate of this term per 10,000 words, or its TF-IDF score. If more than one of these metrics is checked, the user will be asked to select from these options upon execution.
 - The **sources.tab** is the original text data stored in a single file, one document per line. To store a document in a single line, carriage return (ASCII #13) and line feed characters (ASCII #10) have to be replaced with other characters (ASCII 30 and 31).
 - The **tokens.tab** will hold the result of the project's text processing steps including preprocessing, word replacements, stop word removal and categorization. By default, each document is stored on a single line. One may also segment this file by paragraphs or by sentences.
- Once the script options have been set, click **OK** to open the script editor window.



- The **Variables** section at the top allows you to define variables that can be used in the script to customize its execution. Upon execution of a script, if variables have been defined, a dialog box will be presented to set those options.
- To add a new variable to the script, click the  button. The variable definition window will appear.



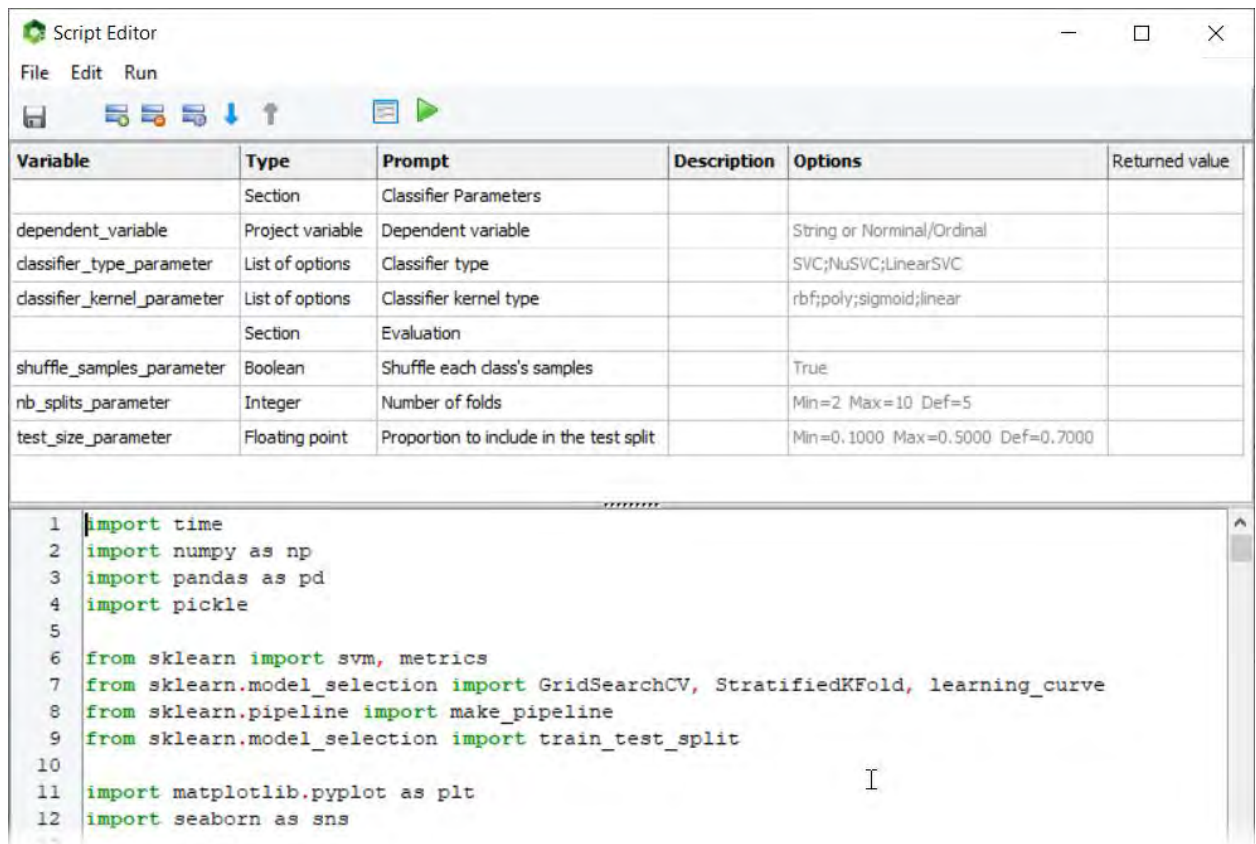
The Variable Definition dialog box is a window with a title bar containing a green icon and the text 'Variable Definition'. It has standard window controls (minimize, maximize, close). The dialog contains the following fields and controls:

- Type:** A dropdown menu currently showing 'Floating point'.
- Name:** An empty text input field.
- Prompt:** An empty text input field.
- Description:** A larger empty text input field.
- Range:** A section with three options:
 - ☒ **Minimum:** A spinner box set to 0.0000.
 - ☐ **Maximum:** A spinner box set to 1.0000.
 - ☐ **Default:** A spinner box set to 0.0000.
- Buttons:** 'Add' and 'Close' buttons at the bottom right.

- From the **Type** dropdown menu, select what will be the type of the variable. You may select among six types of variables: an integer, a floating point, a string, a Boolean, a list of options or a project variable. An additional **Section** type may also be used to group various options into distinct sections. When such an option is selected, you will be asked to enter a string that will be displayed in bold character as the title of the section.
- In the **Name** edit box, type the name of this variable. This is the name that should be inserted into the script source code to customize its execution.
- The **prompt** edit box is the text that will be displayed on the left of the data editing control. The **Description** edit box may also be used to provide additional information about this option. If a description or instructions are provided, a hint window will display this information upon hovering over the variable data entry control.
- Depending on the variable type, different specifications can be added:
 - For a **Floating Point** or an **Integer** variable, the range section allows you to set a **Minimum**, **Maximum** and/or **Default** value. First check the box for the desired range specification, then increase or decrease the value via the up/down arrows, or by typing in the numerical value.
 - A Boolean variable can be enabled by default, by checking the default box under the **Description** textbox.
 - For a **List of options**, the strings that will appear as options to be selected must be specified in the **Options** textbox. Each option should be added on a separate line without quotation marks.
 - If a script includes a **Project variable**, choose the type of data that will be expected via the **Data type** dropdown list. Choosing a data type will restrict the list presented to the selected data type. Selecting **Any type** will present a list of all variables on the project. Checking the **Optional** checkbox allows the Project variable to be left blank.
- To remove an existing variable, first select the variable to be deleted and then click the  button.

- Select a variable and click the  edit variable button to open the variable definition window and make any changes to the existing specifications.
- Click the  down arrow to change the order of the variables and bring the selected variable down in the list of variables, or the  up arrow to move the variable up.
- The  preview button shows the **Options** dialog box as it would display when the script is run, without having to run the script. The dialog box can also be previewed by selecting **Test Dialog Box** under the **Run** menu.

In the example below, the dialog box consists of six variables grouped under two sections: Classifier Parameters and Evaluation.



The Script Editor window displays a table with the following variables:

Variable	Type	Prompt	Description	Options	Returned value
	Section	Classifier Parameters			
dependent_variable	Project variable	Dependent variable		String or Nominal/Ordinal	
classifier_type_parameter	List of options	Classifier type		SVC;NuSVC;LinearSVC	
classifier_kernel_parameter	List of options	Classifier kernel type		rbf;poly;sigmoid;linear	
	Section	Evaluation			
shuffle_samples_parameter	Boolean	Shuffle each class's samples		True	
nb_splits_parameter	Integer	Number of folds		Min=2 Max=10 Def=5	
test_size_parameter	Floating point	Proportion to include in the test split		Min=0.1000 Max=0.5000 Def=0.7000	

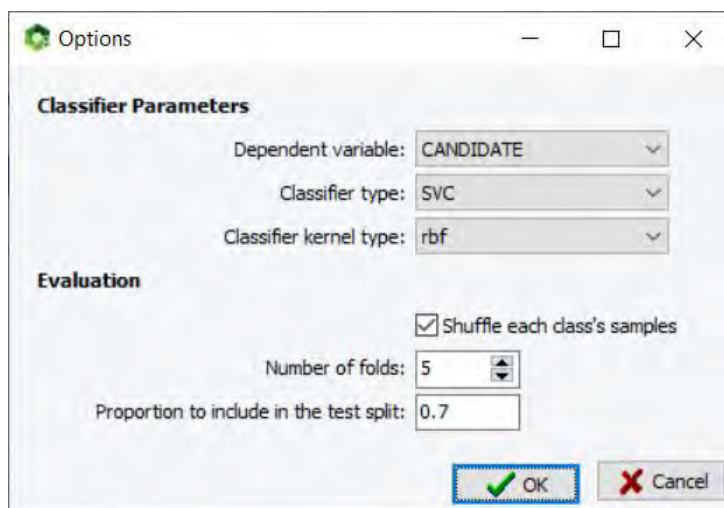
Below the table, the following Python script is visible:

```

1 import time
2 import numpy as np
3 import pandas as pd
4 import pickle
5
6 from sklearn import svm, metrics
7 from sklearn.model_selection import GridSearchCV, StratifiedKFold, learning_curve
8 from sklearn.pipeline import make_pipeline
9 from sklearn.model_selection import train_test_split
10
11 import matplotlib.pyplot as plt
12 import seaborn as sns

```

Running the script will cause the following dialog box to appear:



The Options dialog box contains the following settings:

Classifier Parameters

- Dependent variable: CANDIDATE
- Classifier type: SVC
- Classifier kernel type: rbf

Evaluation

- ☒ Shuffle each class's samples
- Number of folds: 5
- Proportion to include in the test split: 0.7

Buttons: OK, Cancel

- The bottom section of the script editor window should contain valid Python or R code, and the formatting will reflect the syntax of the script's language. Proper commands for importing needed libraries should be specified at the beginning of the script. Names of the specified variables may be used in the script to customize its execution. The Python script below illustrates how one may use the defined variables in a script (highlighted in yellow) and their associated values set by the user using the dialog box.

```

201 data = pd.read_csv(file, sep='\t', encoding='ISO-8859-1')
202 data_clean = data.copy()
203 data_clean = data_clean.drop(columns=[entry_number])
204 if data[dependent_variable].dtype == 'object':
205     # Convert categorical variable to numeric
206     tags_list = data[dependent_variable].unique()
207     tags = {x: i for i, x in enumerate(tags_list)}
208     data_clean['cleaned_tag'] = data_clean[dependent_variable].apply(lambda x: tags[x])
209 else:
210     data_clean['cleaned_tag'] = data_clean[dependent_variable]
211 data_clean = data_clean.drop(columns=dependent_variable)
212 data_X = data_clean.drop(columns='cleaned_tag')
213 data_y = data_clean['cleaned_tag']
214
215 # Split the data into training and testing sets
216 return train_test_split(data_X, data_y,
217                         test_size=test_size_parameter,
218                         random_state=1)
219
220 def main():
221     """Train SVM classifier
222     """
223     try:
224         # read and split data
225         X_train, X_test, y_train, y_test = prepare_data(input_file)
226
227         # cross-validation object
228         kfolds = StratifiedKFold(n_splits=nb_splits_parameter, shuffle=shuffle_samples_parameter, random_state=1)
229
230         # Training
231         np.random.seed(1)
232         svc_param_grid = {}
233         # Linear classifier
234         if classifier_type_parameter == 'LinearSVC':
235             pipeline_svm = make_pipeline(
236                 getattr(svm, classifier_type_parameter)(class_weight="balanced"))
237             svc_param_grid = {'linearsvc__C': [0.01, 0.1, 0.5, 1]}
238         else:
239             pipeline_svm = make_pipeline(
240                 getattr(svm, classifier_type_parameter)(probability=True, class_weight="balanced"))
241             if classifier_type_parameter == 'NuSVC':
242                 if not classifier_kernel_parameter == 'linear':

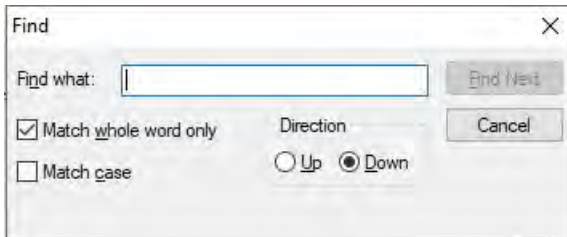
```

- Clicking the  button will run the script. Any changes made will prompt a dialog box asking if changes should be saved before running. Running a script from within the script editor opens a Python or R console. In contrast to running a script from the main post-processing scripts window, the console will also display the script's code in the upper half of the window. The code will include all the variable assignments. Once the script terminates, the console window will have to be closed for the output to be processed. The script can also be run by selecting **Run Script** under the **Run** menu.
- A new script can also be created by selecting **New** under the **File** menu of the script editor. There is an option to choose between Python or R.
- Save a new script with the  button or by selecting **Save** from the **File** menu. Selecting **Save as** will open the **Script Options** window allowing you to enter a new name for the script.

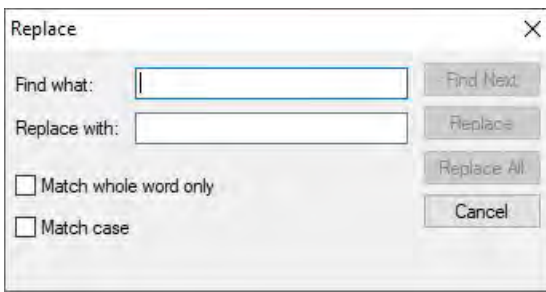
Editing a script

- To edit an existing script, first select it from the **Script** dropdown menu, then click the  button next to the **Run** button and select **Edit Script**.

- Variables can be added, deleted, or edited (See details above)
- Script Options of an existing script can also be edited by selecting **Settings** under the **Edit** menu.
- Changes can be made to the code of the script in the Code section of the script editor. Any edits can be typed into this section or using the **Edit** menu. Under the **Edit** menu, you will find common edits such as **Undo**, **Cut**, **Copy**, **Paste**, **Select all**.
- Click **Find** to search for any particular sequence in the code.



- Select **Replace** in the Edit menu to substitute a sequence in the code.



Importing/Exporting a script

Post-processing scripts created in WordStat (whether Python or R) are saved as '.wscr' files in a distant folder not easily accessible. The import and export features have been designed to easily copy an existing script to this folder or create a copy outside of this folder, allowing one to share scripts with other users.

- To import a script, click the  button next to the run button and select Import. Select an .wscr script from the file explorer and click Open. A script can also be imported by selecting the **Import** command from the **File** menu in the script editor.
- To export a script, first select it, click the  button next to the RUN button and select Edit. Then choose the Export command from the file menu of the script editor. Choose its destination in the file explorer and click Save

Console

Any libraries that are imported in a post-processing script will automatically be installed as the script is run. In some cases, a package requires extra steps for installation, which Wordstat cannot perform under-the-hood. A Python or R console can be opened to perform these installations or any other tasks.

- From the Script Editor window, select **Console** from the **Run** menu,
- Depending on the language of the script, a Python or R console will open

The Extraction Tab

The **Extraction** tab groups tools that are used to extract useful features from the text collection.

The [Topic](#) modeling tool will automatically extract the most important topics from a text collection using factor analysis. Results may be saved as a content analysis dictionary or may be further examined using cooccurrence analysis or crosstabulation.

The [Phrase](#) extraction feature will identify idioms and common phrases and will allow you to add them to a content analysis dictionary as well as perform cooccurrence analysis and comparison analysis of the phrases.

The [Named Entities](#) extraction feature can identify proper nouns, names of people, locations or organizations as well as acronyms. You can select relevant items and move them to the categorization dictionary.

The [Misspellings & Unknowns](#) extraction feature provides a tool that identifies misspellings and some technical terms by comparing the list of word forms encountered in the entire text collection against a list of common words. Extracted words may be added to the current categorization dictionary or to a substitution process. They may also be replaced in the original documents with properly spelled words.

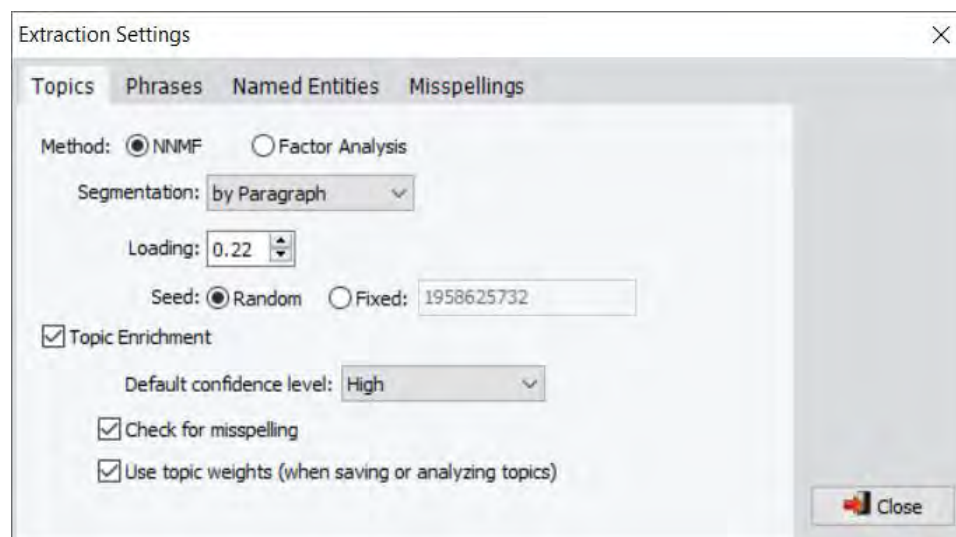
Topics

The **Topics** extraction feature, the first tab on the **Extraction** tab, attempts to uncover the hidden thematic structure of a text collection by applying a combination of natural language processing and statistical analysis. The main statistical procedure used for topic extraction in WordStat is a factor analysis. Technically speaking, the extraction is achieved by computing a word by document frequency matrix, or alternatively by segmenting documents into smaller chunks and computing a word by segment frequency matrix. Once this matrix is obtained, Wordstat performs either a non-negative matrix factorization (or NNMF) or a factor analysis with Varimax rotation, in order to extract a small number of factors. All words with a loading higher than a specific criterion are then retrieved as part of the extracted topic. While in hierarchical cluster analysis, a word may only appear in one cluster, topic modeling using factor analysis may result in a word being associated with more than one factor, a characteristic that more realistically represents the polysemic nature of some words as well as the multiple contexts of word usage.

To ensure the stability of the factoring solution, low frequency items should preferably be excluded. It is thus strongly recommended to remove any word occurring less than 10 times on smaller data sets, ideally less than 30-to-50 times on larger ones. Stemming, lemmatization or the creation of a categorization dictionary may also be used to group words or phrases, including less frequent ones, prior to the topic extraction.

To extract topics:

- Select the  button. An Extraction Setting dialog opens similar to the one below.



The  button is present on all of the tabs on the **Extraction** tab. Each tab of this dialog contains extractions options for the currently select tab.

Method: WordStat implements two methods for topic extraction. The **NNMF** method uses non-negative matrix factorization to extract topics from a word x word correlation matrix computed by WordStat, while **Factor Analysis** perform a similar extraction using a principal component analysis with VARIMAX rotation. The NNMF method is faster and can handle larger matrices than factor analysis. It is however probabilistic in nature, yielding different, yet likely similar, solutions every time you run it, while factor analysis will always produce the exact same results.

Segmentation: This option allows you to specify whether the data to be used for topic modeling will be based on the cooccurrence of words in the same **document**, or whether it will be based on cooccurrence within **paragraphs**, within **sentences**, or within a **windows of words**. The choice of segmentation should ideally reflect how topics are being distributed in a typical document and across documents, as well as the objective of the analysis. When the text collection consists of long documents containing multiple topics (such as long political speeches) and you need to identify all topics in order to compare their relative frequencies, then performing a segmentation by paragraph or by sentence may be more sensitive than computing cooccurrences by documents. Alternatively, if you attempt to differentiate documents by identifying domains or disciplines, or to identify the dominant issue of documents, then performing the analysis at the document level may be more appropriate. When analyzing responses to open-ended questions, which may include several topics listed in a single paragraph, segmenting by sentence may also result in a more precise extraction of the various topics they contain. Finally, if long documents have been imported without any paragraph delimiters, setting a windows of words may be needed to extract meaningful topics.

Loading: This option allows you to set a minimum topic loading a word should reach in order to be retained in the factor solution. By default, this value is set to 0.3. Increasing the cutoff value will reduce the number of words, keeping only the more representative ones, while reducing it may include words that are somewhat less characteristic of the extracted topic.

Seed: This option, which is available only when NNMF method if selected, allows you to set a specific seed value in the pseudo random number generator of WordStat, allowing you to reproduce the exact same results on every run, overriding the probabilistic nature of NNMF. By default, it is set to **Random**, so that repeated analysis on the same data set with the same settings will yield different results. You can achieve replicability of the results by specifying a

Fixed seed value in the WordStat's random number generator. Clicking the  button generates a random number that will become the new fixed seed value.

Pruning: Factor analysis on large correlation matrices can be time consuming. The factor analysis routine in WordStat is limited to a maximum of 2500 words. To speed up the extraction of topics using factor analysis or to analyze number of words beyond this limit, WordStat implements a pruning technique by which words that are less likely to be included in a factor solution will be removed from the matrix prior to the analysis. Disabling this function prevents the removal of those words and attempts a topic extraction on all words, up to the upper limit of 2500 words.

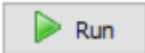
Topic Enrichment: Topic modeling solutions usually consists of a series of words listed in descending order of topic specificity. The combined presence of those words in a text is used to identify whether a topic is present or not. This "bag-of-word" approach may result in imprecise topic identification and measurement since words often have multiple meanings and are often used in very different contexts. The topic enrichment feature allows you to go beyond this "bag-of-word" approach by attempting to identify some phrases highly associated with the extracted topic, as well as other phrases that may represent exceptions and help disambiguate the various meanings of words. WordStat topic enrichment feature also attempts to reduce false negative that may result from the presence of misspellings by identifying misspelled forms of words in the original topic solution or that are part of any of the suggested phrases. Enabling the **Topic Enrichment** feature gives access to the following options:

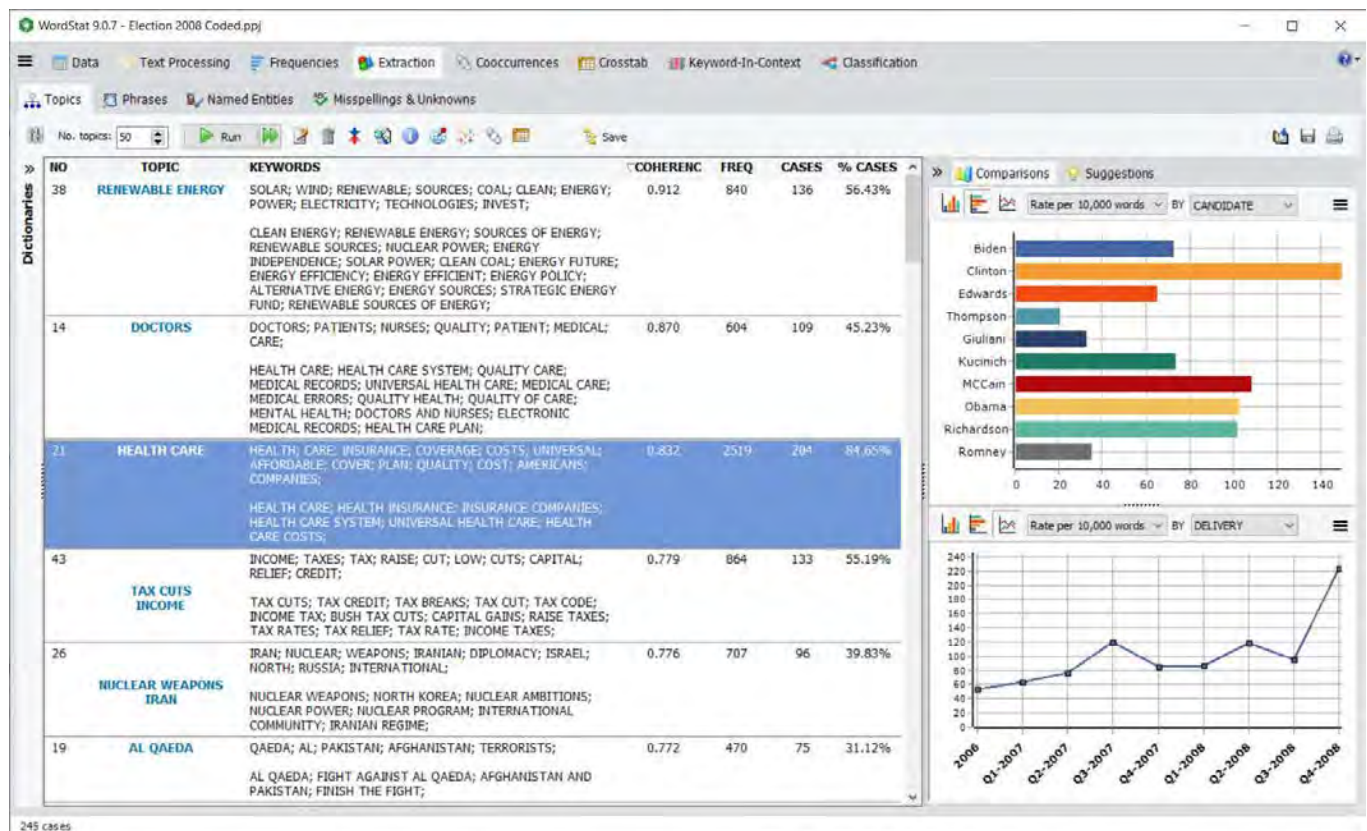
Default confidence level: This option allows you to set a confidence level that will be used when adding phrases to topics. Selecting a lower setting (such as **Low** or **Moderate**) tends to bring more phrases than if this option is set to a higher setting. Phrases that are not automatically added to the topics will still be accessible in the **Suggestions** panel on the left from which they can be manually selected and added to the topic.

Check for misspelling: When this option is enabled, WordStat will identify potential misspellings of words in the topic or that are part of any added or suggested phrases. This feature helps reduce the number of false negatives resulting from spelling errors in the original data set.

Use topic weights: Words in extracted topics are presented in descending order of relevance of specificity to the topic. The first words have higher weights than the last ones. When using factor analysis for topic extraction,

the weight of each word in a topic corresponds to its factor loading, while in NMF it corresponds to a coefficient obtained from the product of the W and H matrices. By default, when running a cooccurrence analysis or performing a crosstab on the extracted topics, or when saving topics to a categorization dictionary, each word will be saved along with its weight, while phrases will get a weight equal to one. Disabling this feature will assign a weight of one to all entries.

- Set the extraction settings and click on the **Close** button.
- Set the number of topics you would like to extract by typing the number in the **No. topics** field or by using the up or down arrows.
- Once the options have been set, click the  button to perform the analysis. Please note that extracting topics on more than a few thousand words can take several minutes. Once extracted, the **Topics** tab should look like this:



The table to the left contains the following information:

TOPIC	WordStat uses an algorithm to automatically provide a label for the extracted topic.
KEYWORDS	Lists all words meeting the factor loading cutoff criteria in descending order of factor loading, along with associated phrases.
COHERENCE	The coherence is a weighted average of the correlations of words associated with the topic.
EIGENVALUE	When performing topic modeling using factor analysis, this column contains the eigenvalue of each factor.
FREQ	Displays the total frequency of all items listed in the keywords column.
CASES	Shows the number of cases containing at least one of the items listed in the keywords column.

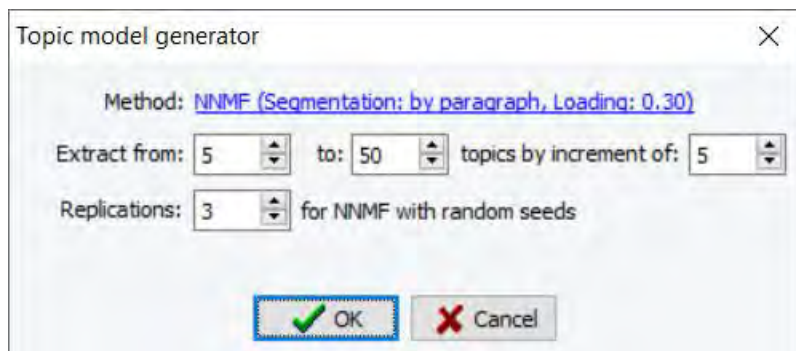
% CASES

Displays the percentage of cases with at least one of the items listed in the keywords column.

Running multiple topic models

Choosing the number of topics to extract using topic modeling techniques remains a question for which there is, to our knowledge, no definitive answer. We may even raise doubts about the fact that such an optimal number exists. In fact, one may even suggest that information obtained using different settings may well serve different purposes or reveal different aspects of a reality. In such a context of uncertainty, researchers often want to compare various solutions. The batch processing feature allows one to compute multiple topic models by systematically varying the number of topics to extract, and to perform several runs using the same settings in order to assess the stability of the results.

To run multiple topics, click the  button, a dialog box similar to this one will appear:



The **Method** option displays the current topic extraction technique and its associated settings. To change any of those values, click this description to display the Topic Extraction Settings dialog box.

The next line of options allows one to set the smallest and largest number of topics to extract as well as the increment to be used. In the example above, the settings will produce topics of 10 different sizes from 5 topics up to 50 topics, incrementing the number of topics by five up until it reaches 50 topics.

Topic extraction using Non-negative matrix factorization (NNMF) is probabilistic in nature and will never give the exact same results unless a fixed random seed is provided. For this reason, one may be interested in assessing the stability of some topic solutions by performing several runs with the same settings and the same number of topics. The **Replications** option allows one to specify how many replications will be performed for each setting. This option is available only when the NNMF method is selected and when no fixed random seed has been set. In the above example, since the user requested topic solutions with 10 different numbers of topics, the total number of topic solutions computed will be 30.

Once the desired options have been set, click the **OK** button to start the computation of all topic models. All topic model solutions will be aggregated in the report manager allowing one to compare solutions obtained in multiple runs using different settings. A summary table with various coherence statistics will be displayed. Those statistics are currently experimental measures of coherence and will later be either dropped or documented. For NNMF, a seed value is also printed, allowing one to replicate a specific topic solution by entering its associated seed value.

Topic Modeling Buttons



Select to edit the topic name and to remove or exclude words and phrases from the topic.



Allows you to delete the topic on the selected row.



Click to merge a topic into another one. You first need to select the row containing the first topic you would like to merge, and then click this button. A dialog box will appear with a list of all other topics. Select the second topic and click **OK**.

-  To retrieve segments associated with a topic, select it and click this button. All text segments containing at least two keywords of the selected topic will be retrieved and presented in a table format. You may however change both the type of segments retrieved (paragraphs, sentences or full documents) or the minimum number of topic words needed for retrieval.
-  When using Factor Analysis as an extraction method, this button open a dialog box containing a summary of the statistical calculations used in the topic extraction including a scree plot, percentage communalities, percentage of trace and factor pattern.
-  Accesses the mapping function, which allows you to analyze the spatial distribution of various topics.
-  Stores the extracted topics currently displayed into a new categorization dictionary where folders at the first level correspond to different topics, and where each of those folders contains the associated words. A dialog box allows you to save
-  Click to export the selected data to Tableau, a software used for interactive data visualization.
-  Allows you to perform cooccurrence analysis of all the extracted topics including clustering and multidimensional scaling, and to create proximity plots as well as link charts. For more information on the various features available, see the [Cooccurrence Tab](#) topic.
-  Allows you to perform full crosstabulation analysis of all the displayed topics with structured data, to apply statistical analysis, and to create various charts such as correspondence plots, heatmaps, bubble charts, and bar charts. For more information on the various features available for crosstabulation analysis, see the [Crosstab Tab](#) topic.
-  Press this button to append a copy of the topic table in the Report Manager. A descriptive title will be provided automatically. To edit this title or to enter a new one, hold down the SHIFT keyboard key while clicking this button (for more information on the Report Manager, see the [Report Management Feature](#) topic).
-  Allows you to store the topic table to disk in various formats, including Excel, tab and comma delimited files, plain text, HTML, XML, SPSS or Stata files.
-  Allows you to print a copy of the displayed chart

The Comparisons Tab

To the right side of the table is a panel containing two tabs. The **Comparisons** tab allows you to look at the distribution of the selected topic among values of up to two structured variables. You may display this distribution using either a vertical bar chart, a horizontal bar chart or a line chart, by clicking on the corresponding button. Four statistics may also be represented on the charts:

- **Case Occurrence** - number of cases in this subgroup containing at least one of these words.
- **Category Percent** - percentage of cases in this subgroup containing at least one of these words.
- **Word Frequency** - total number of these words in this subgroup.
- **Rate per 10,000 Words** - rate of words in this subgroup per 10,000 words.

Right-clicking anywhere in the chart areas displays a popup menu that allows you to edit the chart, save it to disk or in the Report Manager, or to copy it to the clipboard. Clicking a specific bar or a data point of a line chart also allows one to retrieve text segments associated with the selected class and containing words of the selected topic.

The Suggestions Tab

The **Suggestions** tab displays in its top panel suggested phrases, potential exceptions and spelling corrections that can be added to currently selected topic. The bottom panel contains segments related to the selected suggestion, allowing you to

judge its relevance by looking at the context in which it appears. Suggestions are highlighted in yellow, while keywords from the topic are highlighted in green.

To add any of those to the selected topic, check the boxes beside the entries you want to add, right-click the mouse and choose the appropriate destination on the adjacent menu.

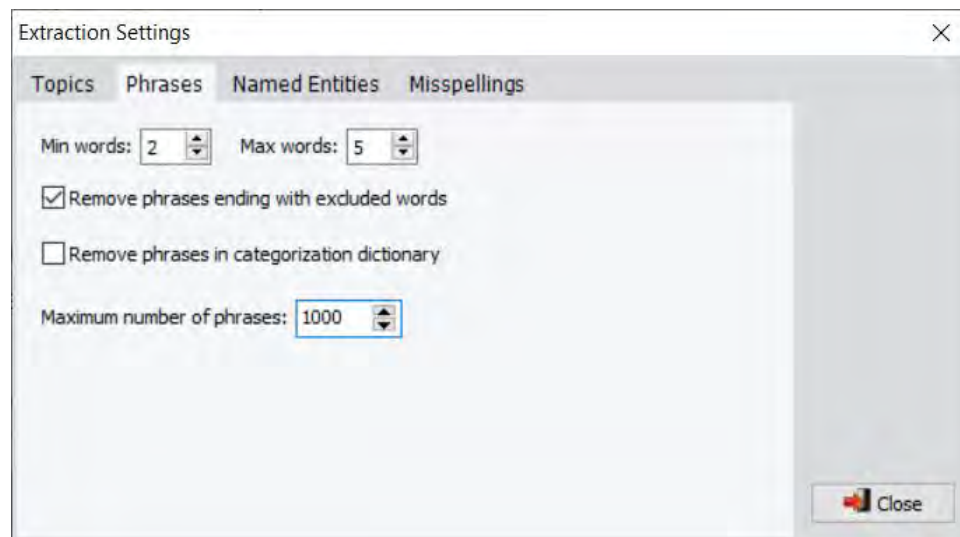
Items listed as **Exceptions** result from an attempt to perform automatic disambiguation of words in the topic by identifying phrases containing topic words that are statistically unrelated to the other words. Automatic disambiguation is a very difficult task and often results in incorrect classification of relevant phrases as exceptions. For this reason, right-clicking an phrase classified as a exception gives the possibility to move it either to the list of exceptions or to the list of related phrases. Since WordStat gives precedence to phrases over single words, adding a phrase representing an exception to a topic will ensure that the word, when part of this phrase, won't be considered as an indication of the topic. When a topic containing exceptions is saved, phrases representing exceptions will be stored in the topic itself but with a weight equal to zero.

Phrases

To accurately represent the meaning of a document, it is sometimes not good enough to rely on words alone but you should look at idioms and phrases. While obtaining a comprehensive list of words is easy, finding common phrases in a specific text corpus is often much more difficult. The phrase finder feature, on the second tab (**Phrases**) of the **Extraction** tab, provides such a tool. It will scan an entire text corpus and identify the most frequent phrases and idioms and allow you to easily add them to the currently active categorization dictionary. In order to reduce redundancy in such a table, short phrases that are part of larger ones are automatically removed from the list, provided that their frequency is lower than or equal to the frequency of a longer version.

To extract phrases:

- Select the  button. An Extraction Setting dialog opens similar to the one below.

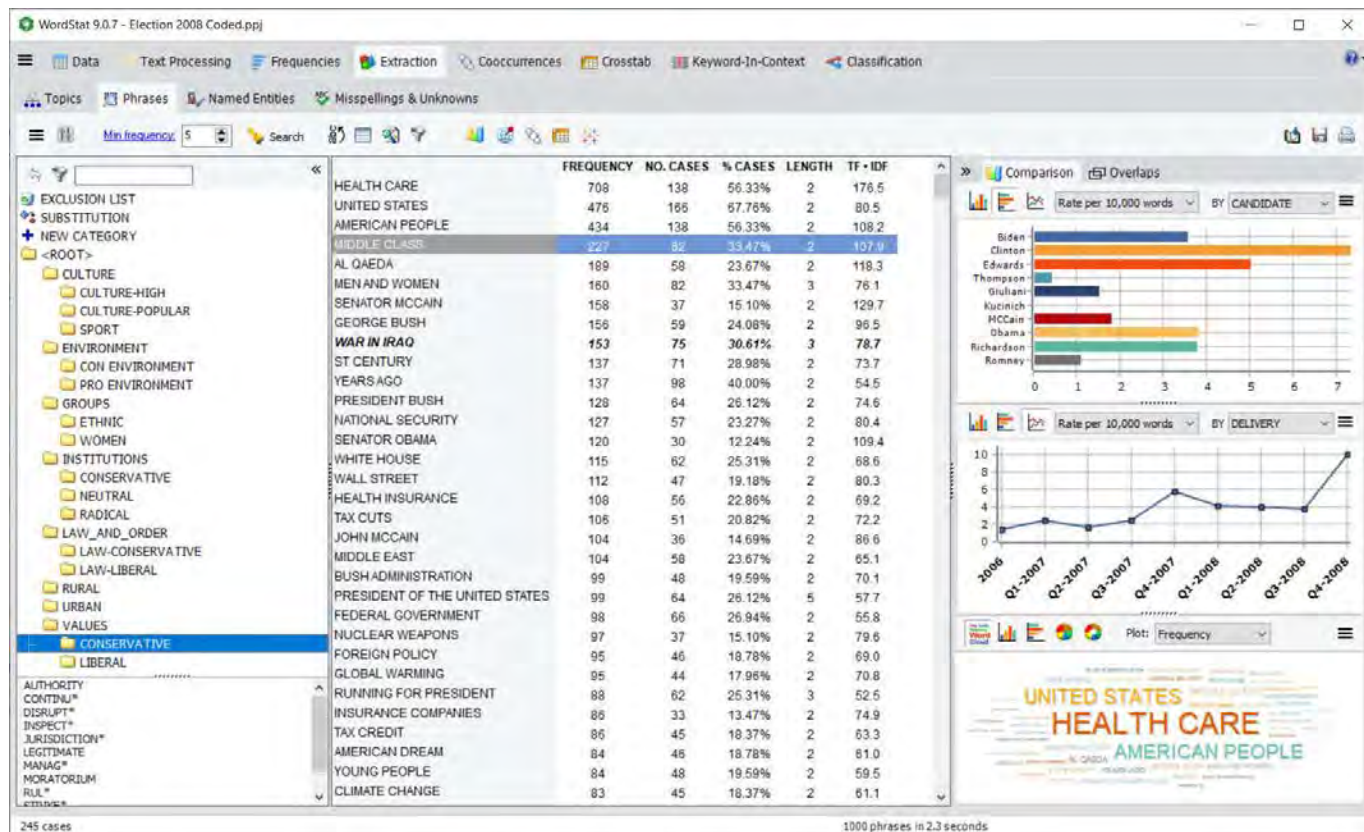


The  button is present on all of the tabs on the **Extraction** tab. Each tab of this dialog contains extraction options for the currently select tab. Before scanning for phrases, you have to set various options that will be used to determine the extent of the scanning process. The first two options that need to be set are the minimum and maximum number of words a phrase can have (**Min words** and **Max words**). These two values determine both the processing time, the memory requirement as well as the number of resulting phrases. The larger the range between these minimum and maximum values, the longer it will take to collect all possible sequences of words. You can also **remove phrases ending with excluded words** or **phrases in categorization dictionary** by selecting the checkboxes beside the options.

- Set the extraction settings and click on the **Close** button.

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

- When these options have been set properly, click the  button to perform the search. Once extracted, the **Phrases** tab should look like this:



By default, found phrases are presented in descending order of frequency. You can sort on any column of the table by clicking its header once to sort the rows in ascending order and a second time to sort the rows on the same column in descending order.

To add a phrase or idiom to the currently selected categorization dictionary or to the exclusion list, simply drag it to the proper location in the **Dictionary panel** located to the left of the screen (see [Working with the Dictionary Panel](#)). You may also select a phrase, right click your mouse and select the desired location.

You should keep in mind that phrases in a dictionary are always treated as a single unit and always have precedence over single words or word patterns. This means if a word is included in one content category ("A") and some phrases containing this word have already been added to another category ("B"), the word will only be categorized into "A" if it is not already part of a phrase found in "B". This feature is essential to perform disambiguation of words, since it allows you to remove some false positives associated with a word by identifying phrases associated with those false positives. For example, if a content category measuring references to money contains the word "BILL," then adding phrases like "BILL OF RIGHTS" or "BILL CLINTON" to another category will prevent those instances of "BILL" being categorized as "money." Note: When two phrases start with the same words, longer phrases have precedence over shorter ones, since they are likely more specific. However, when two phrases partially overlap, the first one encountered in the text will be categorized, preventing the second one from being recognized.

The Comparisons Tab

To the right side of the table is a panel containing two tabs. The **Comparisons** tab allows you to look at the distribution of the selected phrase among values of up to two structured variables. You may display this distribution using either a vertical bar chart, a horizontal bar chart or a line chart, by clicking on the corresponding button. Four statistics may also be represented on the charts:

- **Case Occurrence** - number of cases in this subgroup containing at least one of these words.
- **Category Percent** - percentage of cases in this subgroup containing at least one of these words.
- **Word Frequency** - total number of these words in this subgroup.
- **Rate per 10,000 Words** - rate of words in this subgroup per 10,000 words.

The bottom chart contains the distribution of phrases that can be represented as a word cloud, a vertical or horizontal bar chart, a pie chart and a donut chart. You can display the distribution with the same statistics seen on the frequency table: Frequency, No. of cases, % of cases.

Right-clicking anywhere in the chart displays a popup menu that allows you to edit the chart, save it to disk or in the Report Manager, or copy it to the clipboard. Clicking a specific bar or a data point of a line chart also allows you to retrieve text segments associated with the selected class and containing words of the selected topic.

The Overlaps Tab

While WordStat tries to reduce redundancy in the list of phrases by automatically removing short phrases that are part of longer ones, the resulting list may still contain items that are not independent of each other such as phrases that sometimes overlap. In order to allow users to take into account potential overlaps when selecting phrases, WordStat provides a display option that allows you to see when a selected phrase includes a shorter one, is part of a longer one, or sometimes overlaps other phrases. Such information is especially useful when you need to identify idioms that are more specific, often found in longer phrases or more generic ones usually composed of shorter phrases.

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

Assigning overlapping phrases to a dictionary or obtaining a KWIC table:

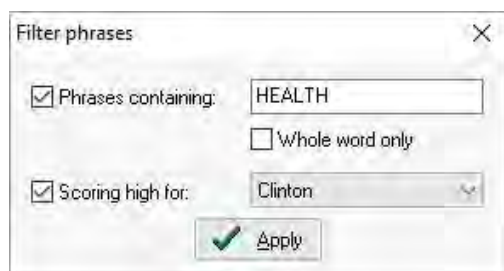
There are two methods of assigning phrases listed in the overlap panel to the categorization dictionary or the exclusion list or of producing a keyword-in-context table. The first method performs these operations on items in this panel only, while the second method includes selected items in the main phrase list.

- To perform one of the above-mentioned operations on the overlapping items only, select the checkboxes of one or several overlapping phrases, right-click your mouse and choose the appropriate command.
- To include phrases in the main table, select one or several phrases in the main table, then the overlapping phrases listed in the overlap panel, and right click the mouse and choose the appropriate command. You may also drag and drop phrases from the main table to the appropriate location in the Dictionary panel to the left. All selected overlapping phrases will be added.

Filtering the Table

Extracting phrases from a large collection of documents can result in a very large table containing thousands of phrases.

- Clicking the  button brings a dialog box offering filtering options that allow you to view only phrases containing either a key word or phrases that are characteristic of a specific class. Filtering conditions are specified in a dialog box similar to this one:



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

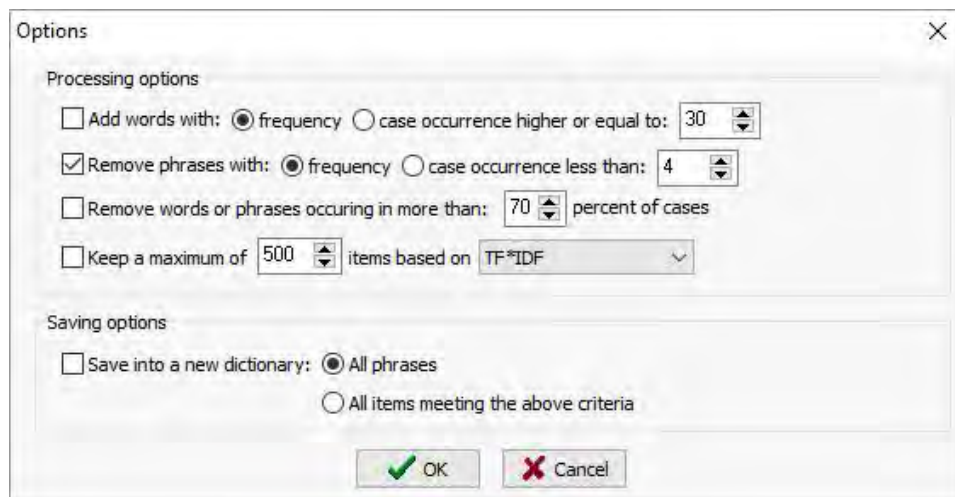
- To apply the filtering condition, click the **Apply** button. When a filtering condition is active, the  button is down. To remove filtering conditions and display all extracted phrases, click this button again.

Phrase cooccurrence Analysis and Crosstabulation

WordStat offers the possibility to perform cooccurrence analysis (clustering, multidimensional scaling, proximity plot) and crosstabulation on defined content categories as well as on words meeting specific frequency criteria. To apply these operations to phrases, you could add them to a user-defined dictionary, enable this dictionary and then move to the **Cooccurrences** or **Crosstab** tabs in order to access the appropriate command. The **Phrases** tab allows you to perform these operations on extracted phrases without the need to assign them to a dictionary. Special dialog boxes also allow you to add frequent words, define additional selection criteria or analysis options, and save the resulting list of items into a new content-analysis dictionary.

To perform a cooccurrence analysis on phrases:

- Click the  button. A dialog box similar to this one will appear:



The options are almost identical to those found on the [Postprocessing](#) tab, allowing you to add to the extracted phrases, single words occurring more than a specific number of times or in a specific number of cases, or to remove items too frequent or not frequent enough. You may also restrict the analysis to a specific number of items.

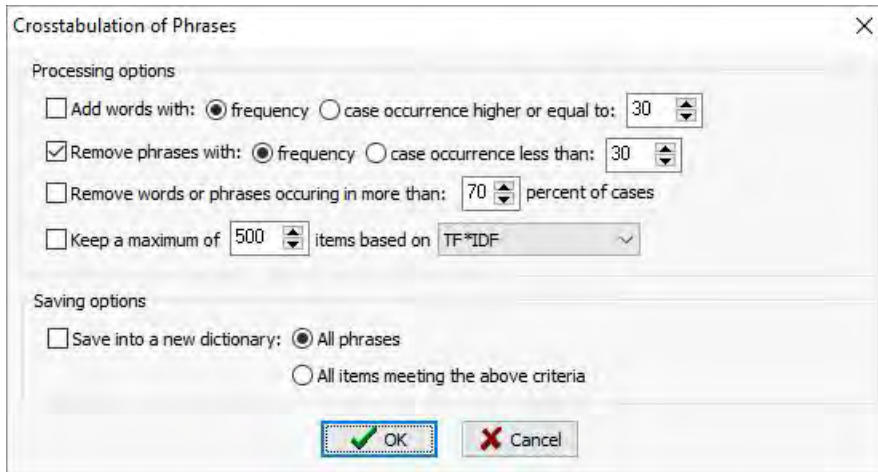
The **Saving into a new dictionary** options may be used to store phrases and words into a new dictionary file. Enabling this option and selecting **All Phrases** will give you the opportunity to store all phrases extracted using the **Phrases** tab, whether or not they meet the specified frequency criteria. Selecting the other option will store only phrases meeting these criteria as well as all words that were also included in the analysis.

The **Saving into a new dictionary** options may be used to store phrases and words into a new dictionary file. Enabling this option and selecting **All Phrases** will allow you the opportunity to store all phrases extracted using the **Phrases** tab, whether or not they meet the specified frequency criteria. Selecting the other option will store only phrases meeting these criteria as well as all words that were included in the analysis.

For further information on the various tools available for annualizing cooccurrences, see [Hierarchical Clustering and Multidimensional Scaling](#).

To perform a crosstabulation analysis:

- Click the  button. A dialog box similar to this one will appear:



The dialog box titled "Crosstabulation of Phrases" contains two sections: "Processing options" and "Saving options".

Processing options:

- ☐ Add words with: ☒ frequency ☐ case occurrence higher or equal to: 30
- ☒ Remove phrases with: ☒ frequency ☐ case occurrence less than: 30
- ☐ Remove words or phrases occurring in more than: 70 percent of cases
- ☐ Keep a maximum of 500 items based on TF*IDF

Saving options:

- ☐ Save into a new dictionary: ☒ All phrases ☐ All items meeting the above criteria

At the bottom are "OK" and "Cancel" buttons.

The **Processing** and **Saving** options are identical to the ones available when performing cooccurrence analysis on phrases (see above).

For more information on crosstabulation, see [Crosstab Tab](#).

Other Table Operations

To copy the entire table to the clipboard:

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

To copy selected rows to the clipboard:

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

To search for a specific item:

- Right-click anywhere in the table.
- Select the **FIND** command from the pop-up menu. A search dialog box will appear.
- Type the desired search string in the **Find What** edit box. To restrict the search to items starting with the search string, enable the **Match Beginning of Item** option and to restrict it to whole words matching the search string, enable the **Match Whole Word Only** option.

Click the **Find** button to search the first item matching the typed string. Clicking this button again finds the next occurrence of the search string, starting at the currently selected item.

Named Entities

The **Named Entities** extraction tool uses pattern-based algorithms to identify people, locations, organizations and acronyms. This feature relies on the existence of mixed-case words and all upper case words in text, and will ignore sentences or paragraphs in full upper case. This approach to named-entity recognition has the benefit of working in many languages but may not be appropriate for some languages, such as German, where upper case letters are used in common nouns. It may also miss some named entities if some of their significant words are not capitalized. For example, "Ministry of education" will not be recognized, while "Ministry of Education" will. The phrases extraction tool may, however, be able to retrieve these named entities if they occur more than once.

To extract named entities:

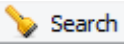
- Clicking the  **Settings** button displays a dialog box similar to the one below.

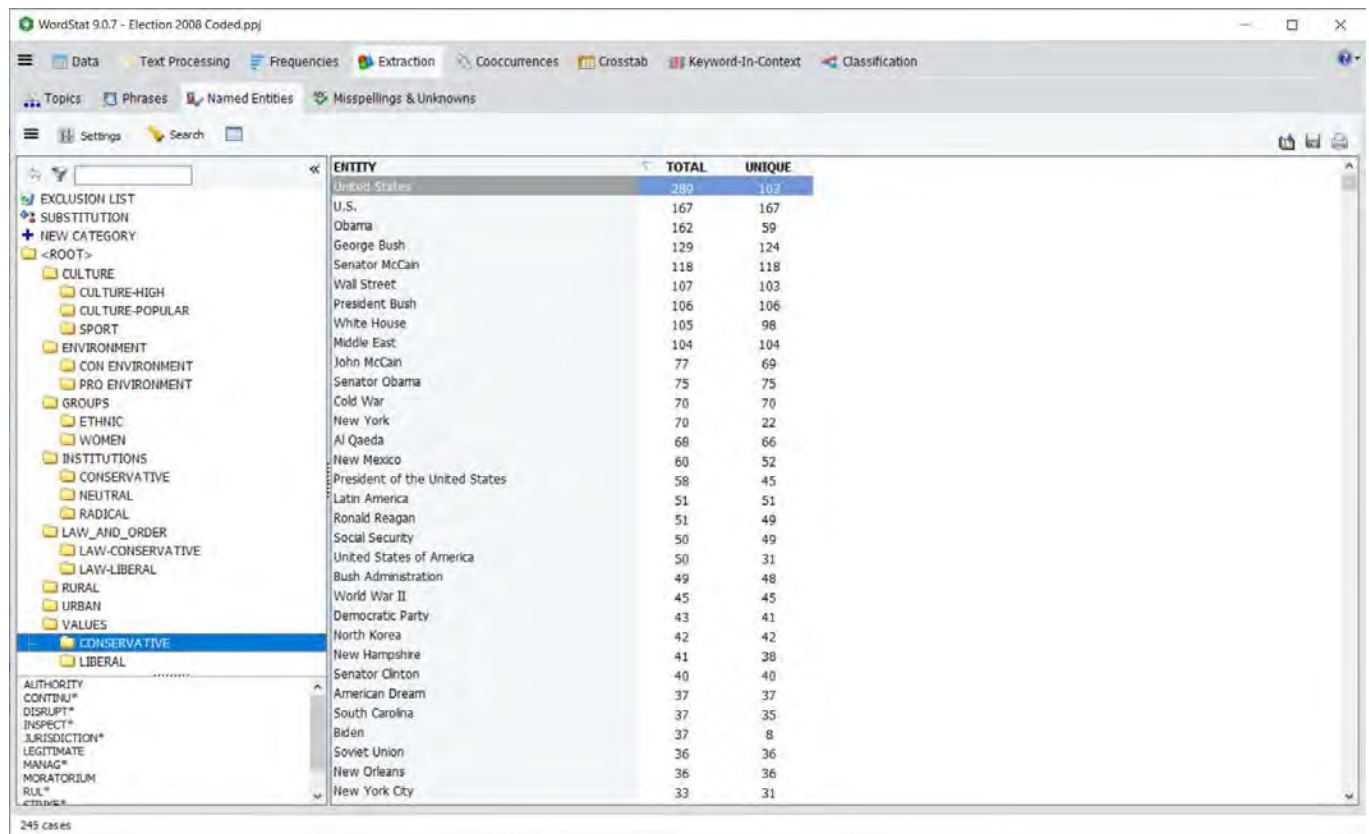


Remove items in categorization dictionary: This option removes, from the final list, named entities that are already in the active categorization dictionary. By default, this option is enabled.

Remove known common words: Enabling this option removes extracted single word entities that are also common words. This option is especially useful to remove capitalized words in titles or words in social-media data that have been put in full upper case, typically to stress their importance. It may, however, miss the identification of named entities that are also common nouns such as "Apple", "Windows" or some acronyms.

Minimum total frequency: This option allows you to eliminate, from the list, named entities that appear only a few times.

- When these options have been set properly, select the 
- Click the  button to perform the search.



ENTITY	TOTAL	UNIQUE
United States	289	102
U.S.	167	167
Obama	162	59
George Bush	129	124
Senator McCain	118	118
Wall Street	107	103
President Bush	106	106
White House	105	98
Middle East	104	104
John McCain	77	69
Senator Obama	75	75
Cold War	70	70
New York	70	22
Al Qaeda	68	66
New Mexico	60	52
President of the United States	58	45
Latin America	51	51
Ronald Reagan	51	49
Social Security	50	49
United States of America	50	31
Bush Administration	49	48
World War II	45	45
Democratic Party	43	41
North Korea	42	42
New Hampshire	41	38
Senator Clinton	40	40
American Dream	37	37
South Carolina	37	35
Baden	37	8
Soviet Union	36	36
New Orleans	36	36
New York City	33	31

Retrieved items are listed in descending order of frequency. The **Total** column indicates the total frequency of this named entity, while the **Unique** column reports how many do not overlap with others.

Adding Named Entities to a Dictionary or Obtaining a KWIC Table

To add a named entity to the currently selected categorization dictionary or to the exclusion dictionary simply drag it to the proper location in the dictionary panel left of the screen (see [Working With the Dictionary Panel](#)). To put them into a new category, drop them into the **New Category** item near the top of the dictionary panel. You may also right click your mouse and select the desired location.

To produce a keyword-in-context table of a specific named entity, first select it and then right click your mouse and select **Keyword-in-Context** list.

Misspellings & Unknowns

The **Misspellings & Unknowns** feature of WordStat offers a tool that identifies common misspellings by comparing the list of word forms encountered in the entire text collection against a list of common words. This feature may also retrieve single words that represent or are part of technical terms, company and product names, as well as abbreviations. Thus, it partly

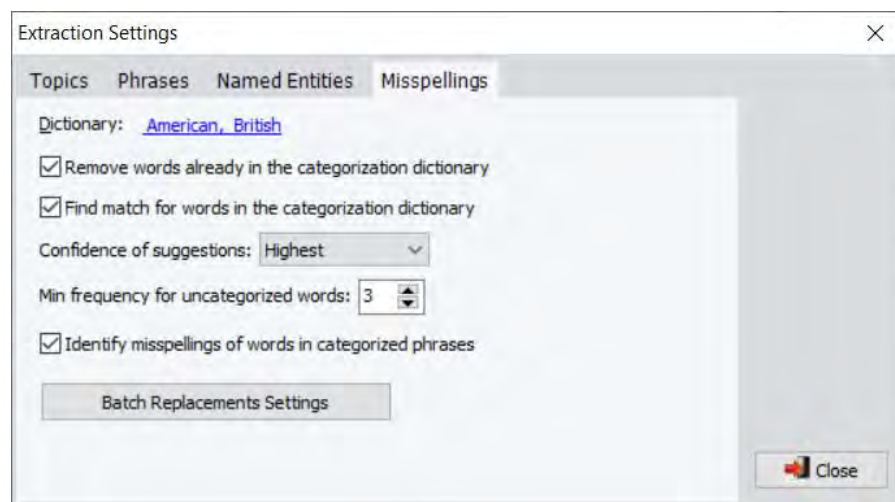
overlaps the [Named Entity](#) extraction tool that will often provide a more complete list of the entities by adding multi-word terms representing the entities as well as common words with irregular capitalization.

When a categorization dictionary is active, this feature will also attempt to match unknown words to existing items in this dictionary and will present the extracted words in two separate lists: the **Categories** list will contain all unknown words that are potentially related to existing items in your dictionary, presented in a tree view organized by categories while all other unknown words will be listed on the **Others** tab.

Once extracted, identified misspellings may be added to the categorization dictionary, to the exclusion list, or a substitution process. You may also replace these words in the original documents with the proper spelling.

To extract misspelled and unknown words:

- Clicking the  **Settings** button displays a dialog box similar to this one:



Several options are available to control how unknown words are extracted and how they are matched to potential replacements.

Dictionary: By default, the extraction is performed in reference to common English words (British and American English). To identify unknown words in documents written in another language or to exclude technical terms from a specific domain, click the language name(s) listed beside the **Dictionary** option or modify the settings on the [Languages](#) section located on the main **Text Processing** tab of WordStat.

Remove words already in the categorization dictionary: Some misspellings or uncommon words may have already been added to an exclusion list or a categorization dictionary. This option automatically removes these items from the lists of retrieved words.

Find match for words in the categorization dictionary: WordStat will attempt to match unknown words with existing ones in the language spelling dictionary and identify the most likely replacements. Enabling this option will also attempt to match uncommon words in the categorization dictionary, such as technical terms, proper names, or even other misspelled words.

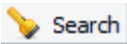
Confidence of suggestions: This option allows you to adjust the confidence level used by WordStat when identifying potential replacements. Setting it to **Highest** will result in less suggestions by applying a stricter matching criterion, while setting it to **Moderate** will provide more suggested replacements at the risk of increasing the number of irrelevant suggestions.

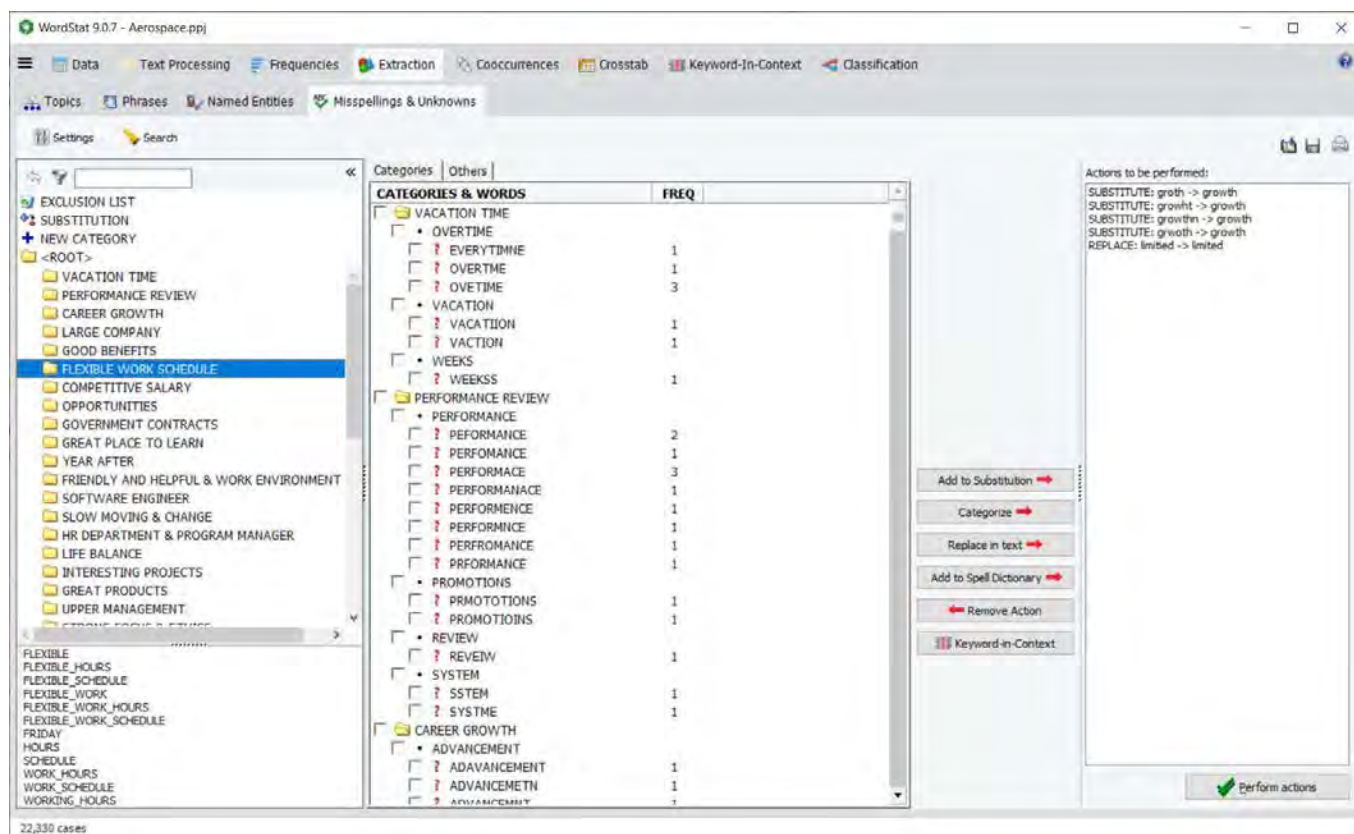
Minimum frequency for uncategorized words: This option allows you to eliminate, from the **Others** list, items that appear only a few times, thus allowing you to focus on the most frequent items. Please note that this setting has no effect on the **Categories** list, since matches will be reported even for words appearing only once.

Identify misspelling of words in categorized phrases: For a phrase in a categorization dictionary to be recognized, all its words must be properly spelled. It is thus important to pay close attention to those misspellings that

may result in a failure to recognize such a phrase. Setting this option will add an additional column in the **Others** list to indicate whether it matches a word that is part of one of the phrases in the categorization dictionary.

Batch replacements settings: WordStat remembers text replacements made in prior sessions and can reapply them in batches. Clicking this button brings a dialog box that allows you to edit or export this list of text replacements as well as to import a list created by another user or on another computer. (see [Managing the Batch Replacement List](#) for more information).

- Once your settings have been established, click the  **Search** button. A dialog similar to this one will appear.



WordStat will retrieve all unknown words and display them in two lists located in the middle panel of the dialog box: the **Categories** list and the **Others** list. You can move from one list to the other by selecting the corresponding tab.

The Categories List

If a content analysis dictionary is available and active, WordStat will attempt to match unknown words with existing items in this dictionary and present them in a tree list view, with the unknown forms stored below the original dictionary form with which it was matched. The suggestions are identified with a "?" symbol and are followed by a numerical value representing their frequency in the current text collection. Existing items and suggested forms are also grouped under their containing content category. Check boxes on the left are used to select multiple items in order to perform some operations on them. To select a single item, click its check box. Click again to remove this check mark. Selecting a content category, subcategory or an existing dictionary item, automatically selects all items below it. If there is no active categorization dictionary, the **Categories** tab will be disabled.

The Others List

The second list contains all words that could not be matched to an existing item of the content analysis dictionary, either because no dictionary was currently active or because they were not similar enough to be matched confidently. On the right of the word, WordStat shows the most likely word replacement that has been found. This column will remain empty if no replacement was similar enough. A third column contains a numerical value representing the frequency of the unknown

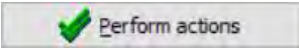
form in the current text collection. Finally, a fourth column will contain the word "Yes" if the word may be matched to a word that is part of a phrase in the content dictionary. Such an indication may be useful to identify situations where a phrase may go undetected because one of its words has been misspelled.

Available Operations

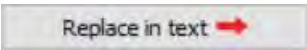
Up to four types of transformation or processing are allowed on the words: 1) You can replace all instances of a selected word in the original document by another word or phrase; 2) You can assign the words to existing content categories; 3) you can instruct WordStat to automatically replace this word with another one by adding it to a live substitution process; 4) you can add this word to a custom list of valid words causing the program to ignore these words the next time there is a search for vocabulary words. You may also obtain a keyword-in-context list associated with a specific word in order to decide how that word should be treated.

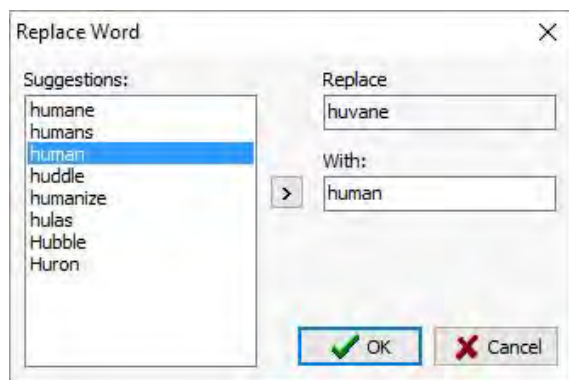
Except for the Keyword-in-context list, none of the other four operations are performed immediately. Instead they are added to an action list allowing you to review, modify or cancel previously defined actions prior to the application of all the specified changes.

To add a word to the substitution process:

- Select the word you want to add.
- Click the  button.
- The selected word followed by the word it will be substituted for will be added to the **Actions to be performed** panel on the right.
- Once all desired actions are present in the **Actions to be performed** list, select the  button. The substitution will now be listed on the **Substitution** tab on the **Text Processing** tab.

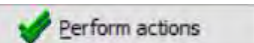
To replace words in the original documents:

- Select the word to be replaced.
- Click the  button. A dialog box similar to this one will appear.

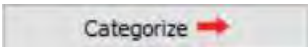


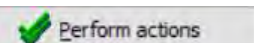
- Type the new replacement word or phrase or choose from the **Suggestions** list box on the left side of the dialog box.
- Click the **OK** button to confirm this replacement and add this operation to the list of **Actions to be performed**.

- Once all desired actions are present in the **Actions to be performed list**, select the button.



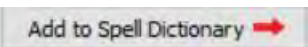
To add a word to a category of your content analysis dictionary:

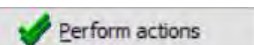
- Select the word you want to categorize.
- Click the  button. A dialog box will appear with a list of all categories and subcategories in the current content analysis dictionary.
- Select the category under which this item should be stored and then click **OK**.
- Once all desired actions are present in the **Actions to be performed list**, select the button.



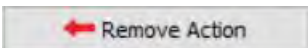
You may add words from this list to the current categorization dictionary or to the exclusion list, or assign it to the substitution process in order to have it replaced automatically by another word. To perform any one of these assignments, simply select and drag the item into the proper location in the dictionary panel to the left of the table (see [Using the Dictionary Panel](#)).

To add a word to the custom list of words to ignore:

- Select the word you would like to add to the custom dictionary.
- Click the  button.
- The selected word to be added will appear the **Actions to be performed** panel on the right.
- Once all desired actions are present in the **Actions to be performed list**, select the button.



To remove operations previously defined:

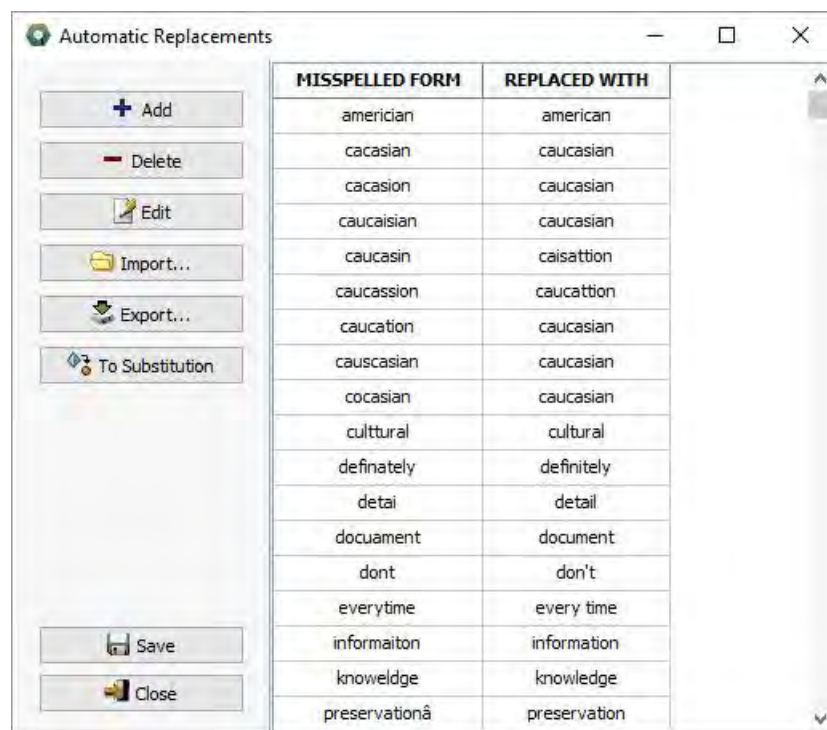
- Select the operations that you want to remove.
- Click the  button. All words associated with the removed actions are moved back to the list of unknown words and positioned at the bottom of the list.

Managing the Batch Replacement List

WordStat remembers text replacements made in prior sessions and can reapply them in batches. You can edit entries in this list of replacements, remove items, merge lists by importing lists created by other users, or you can export this list to a file.

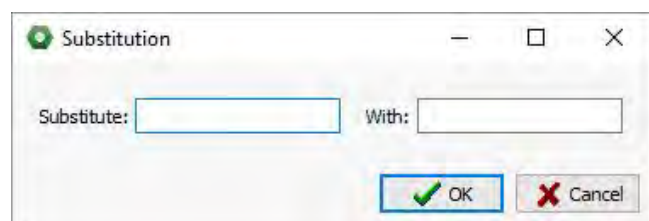
To manage the Batch Replace list:

- Click the **Settings** button. The Extraction Settings dialog will appear.
- Select the **Batch Replacements Settings** button. A dialog box similar to this one will appear:



To manually add a replacement:

- Click the **Add** button. A dialog box like this one will appear:



- Type the word to be replaced in the **Substitute** edit box.
- Type the word that will replace it in the **With** edit box.
- Click the **OK** button to add the replacement to the existing list.

To edit a replacement:

- Select the row containing the item you want to edit.
- Click the **Edit** button. A dialog box similar to the one shown above will appear.
- Edit any one of the entries in the **Substitute** or the **With** edit box.
- Click the **OK** button to confirm the change.

To delete replacements:

- Select the rows containing the items you want to delete. You can hold down the Ctrl key to select disjointed rows, or the Shift key to select a range of successive rows.

- Click the **Delete** button.

To import a replacement list:

- Click the **Import** button. An **Open File** dialog box will appear.
- Select the file containing the replacements you want to import, and click **Open**. All replacements in the imported file will be appended to the current replacement list. Duplicate items won't be imported, while conflicting items will be shown, allowing you to either keep the existing replacement or overwrite it with the one in the imported file.

To export a replacement list:

- Click the **Export** button. A **Save File** dialog box will appear.
- Type a valid file name and select the location where the file should be stored.
- Click the **Save** button.

To move automatic replacements to the substitution process:

- Select the words you would like to add to the substitution process.
- Select the **To Substitution** button. The substitution will now be listed on the **Substitution** tab on the **Text Processing** tab.

The Cooccurrences Tab

The **Cooccurrences** tab allows you to perform hierarchical cluster analysis and multidimensional scaling on all keywords. Results are displayed in the form of dendrograms, concept maps and proximity plots, based on item cooccurrence. It also allows you to compute the similarity of cases based on keyword use. For further information see [Hierarchical Clustering and Multidimensional Scaling](#).

The **Cooccurrences** tab consist of six interior tabs.

The [Options tab](#) allows you to specify whether the clustering should be performed on keywords or on cases and to set various analysis and display options.

The [Dendrogram tab](#) uses average-linkage hierarchical clustering method to create clusters from a similarity matrix.

The [Mapping tab](#) provides a graphic representation of the proximity values computed on all included keywords using multidimensional scaling.

The [Link Analysis tab](#) allows you to visualize the connections between keywords or dictionary items using a network graph.

The [Proximity Plot tab](#) is the most accurate way to graphically represent the distance between objects by displaying the measured distance from one or several target objects to all other objects.

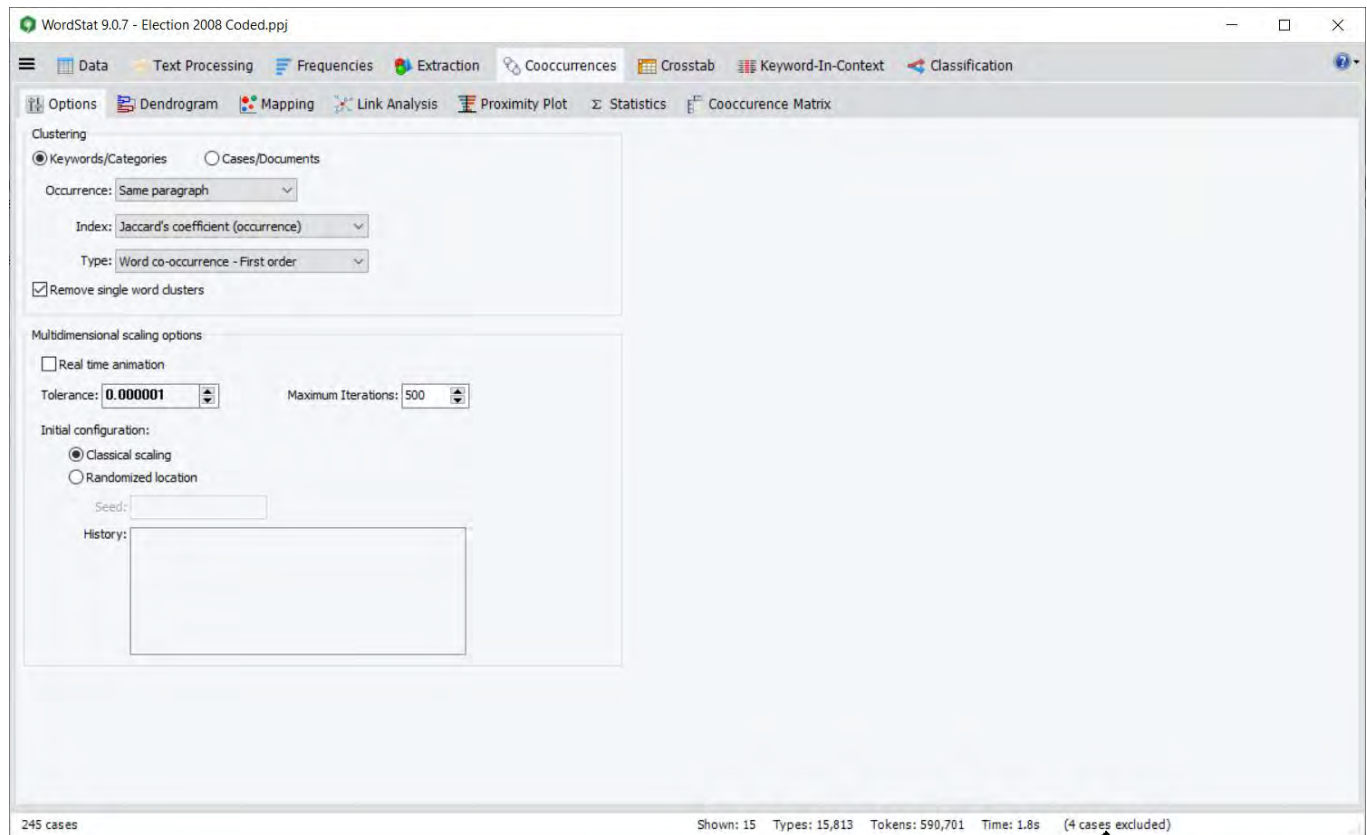
The [Statistics tab](#) displays the cooccurrence and similarity matrices used for building the dendrogram, MDS plots, as well as the proximity plot.

The [Cooccurrence matrix tab](#) provides an interactive matrix that allows one to focus on specific co-occurrences, retrieve associated text segments, and see how they are associated with other variables.

Hierarchical Clustering and Multidimensional Scaling Options

WordStat allows you to further develop categorization by providing various graphic tools to assist in the identification of related words or categories. The 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. Cases or documents may also be clustered based on their content similarity using the same statistical and graphic tools.

The first tab, the **Options** tab, is used to specify whether the clustering should be performed on keywords or on cases and to set various analysis and display options.



Clustering Cases/Documents

When the clustering is set to be performed on cases or documents, the distance matrix used for clustering and multidimensional scaling consists of cosine coefficients computed on the relative frequency of the various keywords. The more similar two documents are in terms of the distribution of keywords, the higher the coefficient. The case label that is used to identify the various cases can be set by choosing the [Edit Case Descriptors](#) command from the WordStat main menu.

Clustering Keywords

When clustering keywords or content categories, several options are available to define cooccurrence and choose which similarity index will be computed from the observed cooccurrences.

Cooccurrence: This option allows you to specify how a cooccurrence will be defined. By default, a cooccurrence is said to happen every time two words or two categories appear in the same case (**by case** option). You may also restrict the definition of cooccurrence 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 cooccurrences by limiting the cooccurrence 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 identifying idioms or phrases that may need to be added to the categorization dictionary. cooccurrence on larger text segments such as cases or paragraphs may be more appropriate to identify the cooccurrence of themes in individual subjects.

Index: The Index option allows the selection of 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 case and do not take into account their frequency. In all these indexes, joint absences are excluded from consideration.

Jaccard's Coefficient: This coefficient is computed from a fourfold table as $a/(a+b+c)$ 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 (also known as the Dice coefficient) is similar to Jaccard's but matches are weighted double. Its computing formula is $2a/(2a+b+c)$ 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.

Phi Coefficient: This is a measure of association for two binary variables. It is similar to the Pearson correlation coefficient in its interpretation.

Cosine Theta: This measures the cosine of the angle between two vectors of values. It ranges from -1 to +1. This coefficient takes into account not only the presence of a word or category in a case, but also how often it appears in this case.

Inclusion Index; This index measures the conditional probability that a document that contains an item X will also contain an item Y. It will take the maximum value of 1 when one of these items always appears when the second one appears, even if the reverse is not necessarily true. The Inclusion Index is optimal for analyzing fields that are organized hierarchically.

Association Strength: This measures the cooccurrence of items taking into account the possibility that two items will sometimes cooccur by chance.

Clustering type: Two broad types of keyword clustering are available. The first method is based on **keyword cooccurrences (First Order Clustering)** and will group together words appearing near each other or in the same document (depending on the selected cooccurrence window). The second clustering method is based on **cooccurrence profiles (Second Order Clustering)** and will consider that two keywords are close to each other, not necessarily because they cooccur but because they both occur in similar environments. One of the benefits of this clustering method is its ability to group words that are synonyms or alternate forms of the same word. For example, while TUMOR and TUMOUR will seldom or never occur together in the same document, second order clustering may find them to be pretty close because they both cooccur with words like BRAIN or CANCER. Second order clustering will also group words that are related semantically such as MILK, JUICE, and WINE because of their propensity to be associated with similar verbs like DRINK or POUR or nouns like GLASS (for more information, see [Grefenstette, 1994](#)).

Remove single word clusters: One way to extract potentially interesting knowledge from dendrograms is to focus on the aggregation of items at an early stage of the clustering process. However, when clustering hundreds or thousands of items, the identification of those items requires the user to scroll through a very long dendrogram which includes many clusters of isolated items. Enabling this option simplifies the use of the dendrogram by hiding all single item clusters and allowing you to concentrate only on the strongest associations. Setting this option also removes isolated items from multi-dimensional scaling plots, greatly enhancing their value when analyzing a large number of items. Please note, however, that when this option is enabled, changing the number of clusters while viewing a 2-D or 3-D MDS plot will cause the program to recompute the distance and location of remaining items.

The options below apply whether clustering cases or keywords. They will affect either the computation or the display of multidimensional scaling charts.

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 you 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 you 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 box. 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 the proper line in the history list box.

For more information on the available options, click the links below:

[Dendrogram](#)

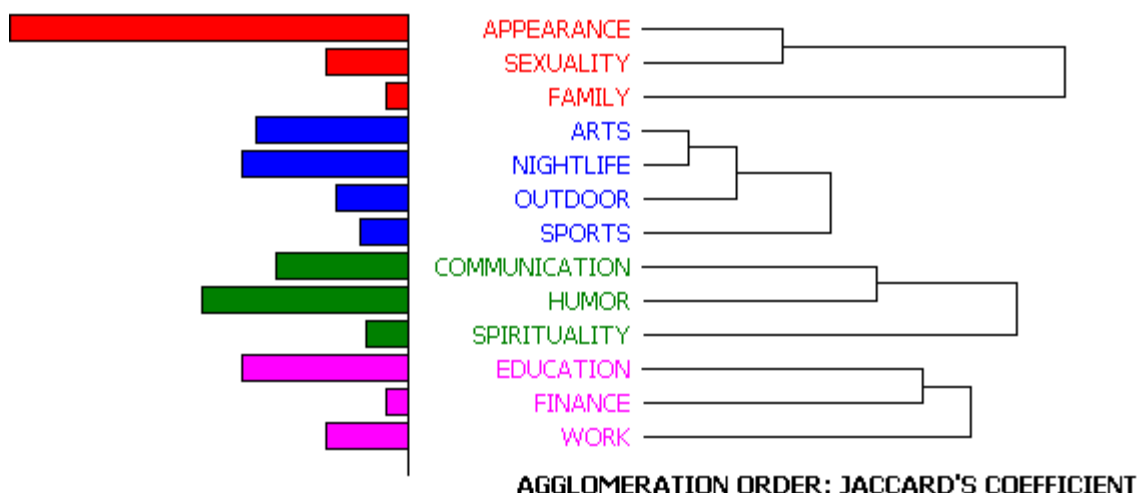
[2-D and 3-D Concept Maps](#)

[Proximity Plot](#)

[Table and Statistics](#)

Dendrogram

WordStat uses 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 known 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.



No clusters This option allows the setting of the number of clusters that the clustering solution should have. Different colors are used both in the dendrogram and in the 2-D and 3-D maps to indicate membership of specific items to different clusters. However, if the option to remove single item clusters is enabled, an increase in the number of clusters may, in fact, result in a decrease in the number of clusters displayed and in the overall height of the dendrogram since all single item clusters will be hidden.

Display This option lets you choose whether the vertical lines of the dendrogram represent the agglomeration schedule or the similarity indices.



When clustering keywords or content categories, clicking this button displays bars beside each dendrogram item to represent their relative frequencies.



Use this button to increase the dendrogram font size and focus on a smaller portion of the tree.



Use this button to reduce the dendrogram font size and view a larger portion of the tree.



This button allows you to perform full crosstabulation analysis with structured data, apply statistical analysis, and create various charts such as correspondence plots, heatmaps, bubble charts and bar charts. A dialog box allows you to restrict the analysis to specific clusters containing a minimum number of items, and cluster names are automatically generated using characteristic words and phrases of each cluster. For more information on the various features available for crosstabulation analysis, see the [Crosstab](#) topic.



This button stores the cluster solution currently displayed into a new categorization dictionary where folders at the first level correspond to different clusters, and where each of those folders contains the associated words or expressions. A dialog box allows you to save only clusters containing a minimum number of items. Cluster names are automatically generated using characteristic words and phrases from each cluster. You may then edit these cluster names and their content from the [Dictionary tab](#).



Press this button to append a copy of the graphic in the Report Manager. A descriptive title will be provided automatically. To edit this title or to enter a new one, hold down the **Shift** key while clicking this button (for more information on the Report Manager, see the [Report Management Feature](#) topic).



This button allows you to store the displayed dendrogram into a graphic file. WordStat supports three different file formats: .BMP (Windows bitmap files), .PNG (Portable Network Graphic compress files) and .JPG (JPEg compressed files).



To retrieve text segments or documents associated with a specific cluster, click anywhere on a cluster to select it (the selected cluster is displayed using thicker black lines), and then click this button to retrieve the associated documents. When performing first order clustering on keywords, this operation retrieves all text segments containing at least two keywords of the selected cluster. When performing second order clustering of keywords, all text segments containing a single one of those keywords will be retrieved.



The slide ruler provides another way of quickly changing the number of clusters included in the clustering solution. Moving the slider to the left increases the minimum distance required to form a cluster and thus produces a dendrogram with more clusters. Moving the slider to the right aggregates smaller clusters into bigger ones. However, if the option to remove single-item clusters is enabled, an increase in the number of clusters may, in fact, result in a decrease in the number of clusters displayed and in the overall height of the dendrogram.

Using the Right Panel

On the right side of this table, a panel allows you to look at the distribution of the selected topic among values of up to two structured variables as well as a network graph representing the relationship between all items in the selected cluster.

For the first two panels, you may choose to display the distribution of this cluster (using either a vertical bar chart, a horizontal bar chart, or a line chart) by clicking the corresponding button. Four statistics may also be represented on the charts:

- **Case Occurrence** - number of cases in this subgroup containing at least one of these words.
- **Category Percent** - percentage of cases in this subgroup containing at least one of these words.
- **Word Frequency** - total number of these words in this subgroup.
- **Rate per 10,000 Words** - rate of words in this subgroup per 10,000 words.

Right-clicking anywhere in the chart area displays a pop-up menu that allows you to edit the chart, save it to disk or in the **Report Manager**, or copy it to the clipboard. Clicking a specific bar or a data point of a line chart allows you to retrieve text segments associated with the selected class and containing words of the selected cluster.

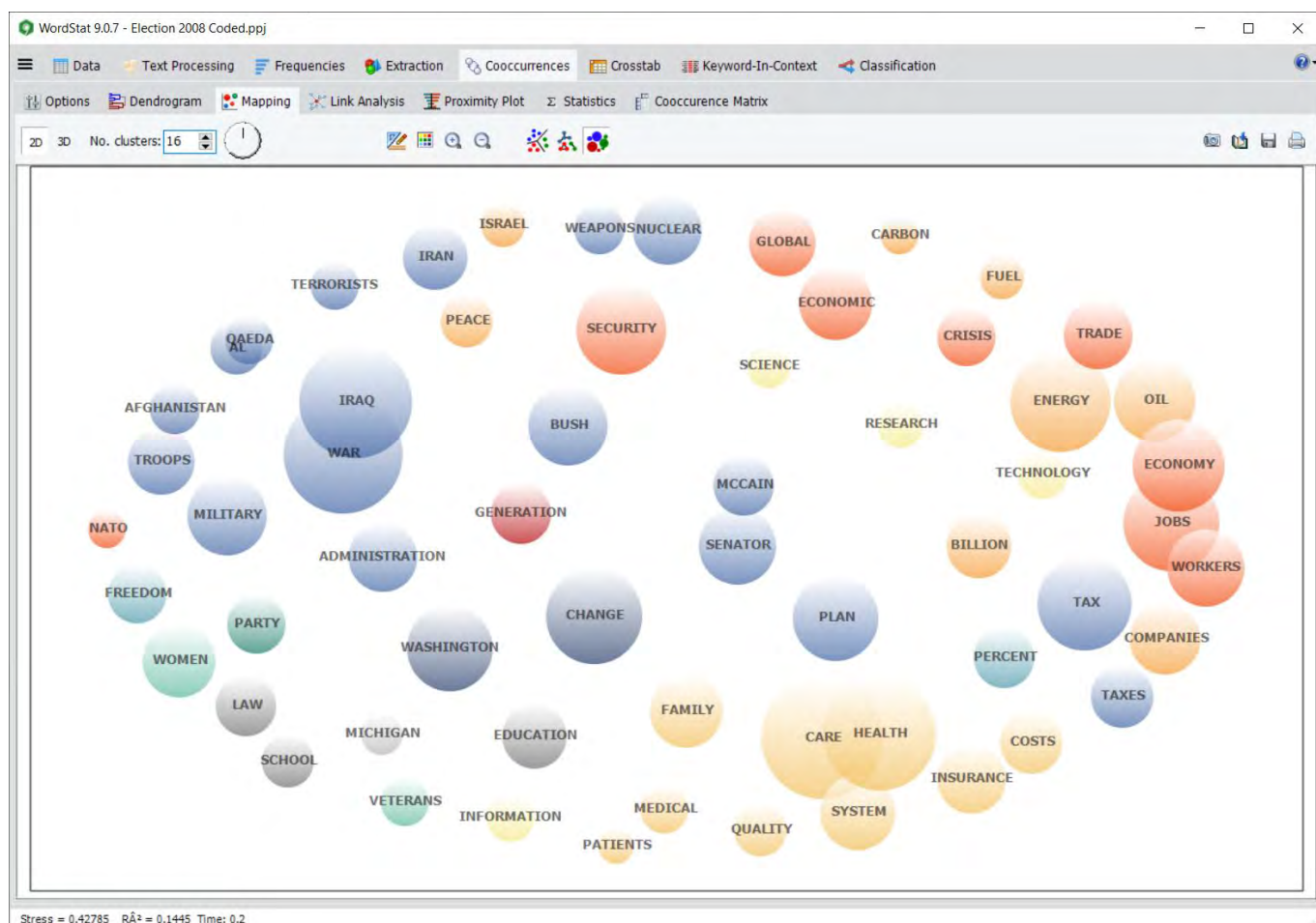
In the lower-right panel, the relationships between words within the selected cluster are represented using a network-type graph. The layout of the nodes can be based on a multidimensional scaling analysis, a force based graph, or may be spread out in a circle. A track bar allows you to hide the weaker connections, while other buttons may be used to zoom in and out.

Clicking the  button allows you to jump to the full-sized [Link Analysis](#) tab.

2D and 3D Concept Maps

The concept maps are graphic representations of the proximity values computed on all included keywords using multidimensional scaling (MDS). In these maps, a point represents an item (keyword or content 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. An option also allows you to vary the size of each data point in order to take into account the observed frequency of each item. 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 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.



2-D and 3-D Map controls:

No. clusters: This option allows the setting of the number of clusters that the clustering solution should have. Different colors are used both in the dendrogram and in the 2-D and 3-D maps to indicate membership of specific items to different clusters. The slide ruler located on the top toolbar of the dialog box may also be used to quickly change the number of clusters. Please note that when the **Remove single word clusters** option is enabled, changing the number of clusters either way often causes the program to recompute MDS maps to take into account the different number of items displayed.



The actual orientation of axes in the final solution is arbitrary. The map may be rotated in any way you want provided 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 that can be most easily interpreted.



Clicking this button enables you to zoom in on a plot. To zoom an area of the plot, hold the left mouse button and drag the mouse down/right. A rectangle indicates 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 button 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 pop-up menu appears, allowing you to select whether the chart should be copied as a bitmap or as a metafile.



This button allows editing of various features of multidimensional scaling plots such as the appearance of value labels and data points, the chart and axis titles, the location of the legend, etc. (see [Multidimensional Scaling Plot Options](#))



Pressing down this button creates a constrained multidimensional scaling. This mapping algorithm allows you to preserve the clustering structure in multidimensional scaling plots, making the interpretation of 2-D and 3-D MDS maps a lot easier and more consistent with the clustering solutions. Enabling this option allows you to use the MDS module to create maps of concepts similar to those suggested by Trochim, in its Concept Mapping procedure.



Pressing down this button displays lines to represent relationships between data points of the multidimensional scaling plot. When the button is down, a cursor will appear in a tool panel below the plot, allowing you to select the minimum association strengths to be displayed.



Clicking this button creates a bubble plot where the areas of data points are proportional to the relative frequency of those items. This type of display is especially useful when you need to take into account a third variable, in this specific case the frequency of items, when interpreting the distance between data points.



Press this button to append a copy of the graphic in the Report Manager. A descriptive title will be provided automatically. To edit this title or to enter a new one, hold down the **Shift** key while clicking this button (for more information on the Report Manager, see the [Report Management Feature](#) topic).



This button allows storing the displayed multidimensional scaling plot into a graphic file. WordStat supports four different file formats: .BMP (Windows bitmap files), .PNG (Portable Network Graphic compress files) and .JPG (Jpeg compressed files) as well as .WSX a proprietary file format (WordStat Chart file). Charts stored in the latter format may be opened, further edited and customized using the Chart Editor external utility program.



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

3-D 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 draws anchor lines from the floor to the data point to better locate data points in all 3 dimensions.



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



Locating a data point on the depth dimension of a 3-D plot can be very difficult, especially when the plot remains static. You often have to rotate this plot constantly on the various axes to get an accurate idea of where the data point is located on this third axis. Clicking this button forces WordStat to rotate the plot automatically. To disable the automatic rotation, click a second time.

Multidimensional Scaling Plot Options Dialog (COPY)

The various options in this dialog box are used to customize the appearance of multidimensional scaling plots. These options represent only a small portion of all settings available.

To further customize the chart, edit data points, value labels, etc., click the  button located at the right side of the dialog box.

Data Points

View data points: By default, multidimensional scaling plots display both the data points and their associated labels. This option toggles on and off the display of the data points.

Style: This option defines the shape used to display the data points. Nine different pointer shapes may be used to represent data points (e.g. square, triangles, dots, cross, etc.).

Shadow: This option toggles on and off the shade on the sides of data points. This option only affects square, triangle and down triangle data points when viewed in three-dimension.

Size: This option specifies the width and height of data points in pixels.

Transparency: This option sets the transparency level of the data points in bubble plots.

Data Labels

Transparent labels: This option allows you to specify whether data labels are to be displayed on an opaque background or whether the background should be invisible.

Background: This option allows you to select the background color of data labels when not transparent

Border: This option enables or disables the rectangle surrounding the label background.

Vertical gap: The vertical gap property determines the vertical distance in pixels between the top of the data point and the bottom of its label.

Font: Click this button to adjust the font properties used to display data labels (i.e., style, size, and color).

Walls and Frame Tab

Wall visible: Use this option to toggle the display of left, back and bottom walls to simulate a 3-D effect.

Transparent: The Transparent property controls whether the wall background will be opaque or transparent.

Dark 3-D: This option shades the sides of the walls.

Title

Show title: By default, multidimensional scaling plots have no title. To display a title, enable the **Show Title** option and enter the desired title in the edit box to the right of this check box. Enter several lines of text by pressing the **<Enter>** key at the end of a line before entering the next line.

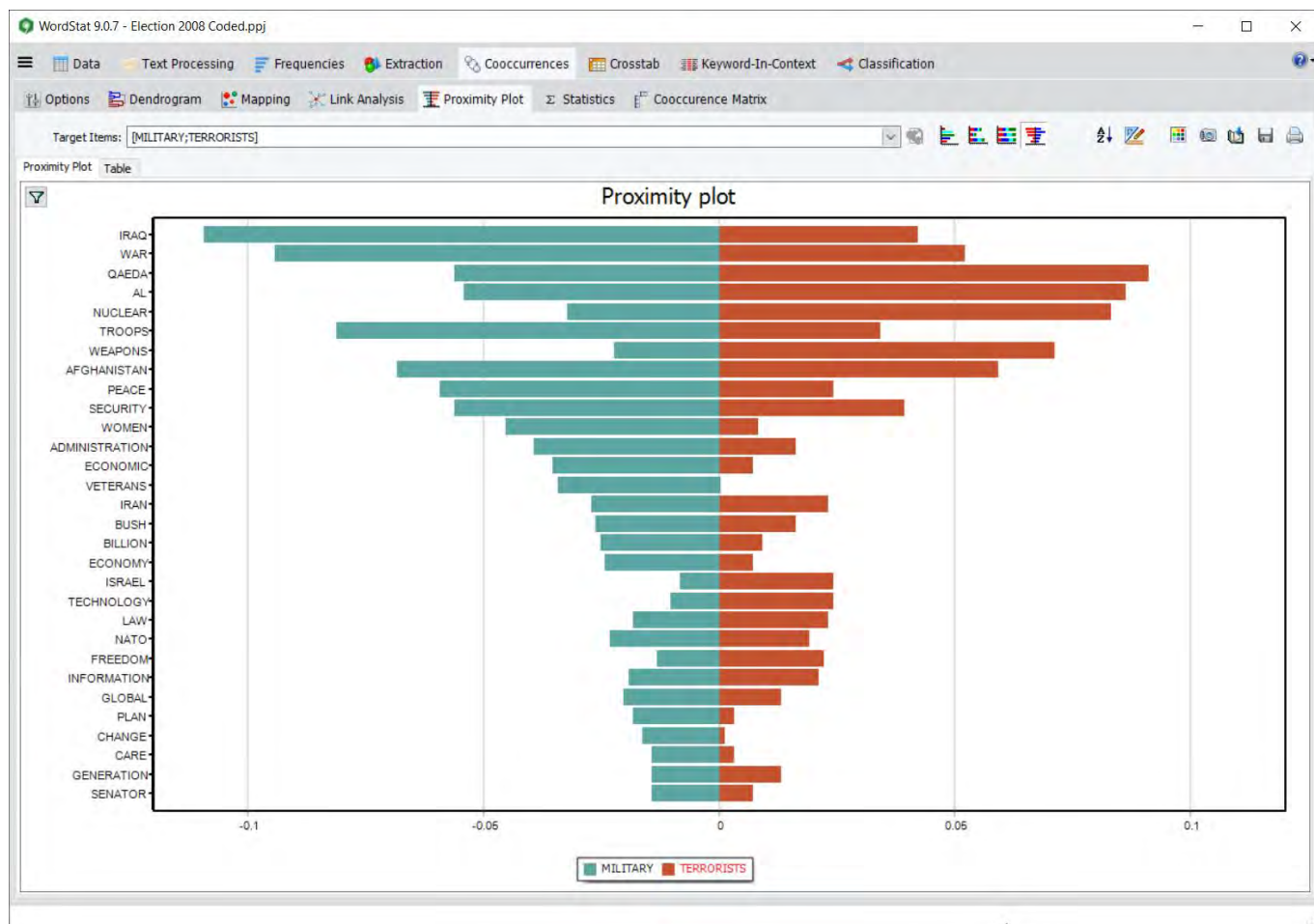
The **Font** button at the bottom of this tab is used to change the font size or style of this title.

Proximity Plot

Cluster analysis and multidimensional scaling are both data reduction techniques and may not accurately represent the true proximity of keywords or cases to each other. In a dendrogram, while keywords that cooccur or cases that are similar tend to appear near each other, you cannot really look at the sequence of keywords as a linear representation of those distances. You have to remember that a dendrogram only specifies the temporal order of the branching sequence. Consequently, any cluster can be rotated around each internal branch on the tree without, in any way, affecting the meaning of the dendrogram. The best analogy here is to think of a Calder mobile. Different photos of such a mobile will yield different distances between hanging objects. While multidimensional scaling is a more accurate representation of the distance between objects, the fact that it attempts to represent the various points into a two- or three-dimensional space may result in distortion. As a consequence, some items that tend to appear together or be very similar may still be plotted far from each other.

Proximity Plot Tab

The **Proximity Plot** is the most accurate way to graphically represent the distance between objects by displaying the measured distance from one or several target objects to all other objects. It is not a data reduction technique but a visualization tool to help you extract information from the huge amount of data stored in the distance matrix at the origin of the dendrogram and the multidimensional scaling plots. In this plot, all measured distances are represented by bars, the closer an object is to the selected one, the longer the bar will be.



To select a keyword or a case that will be used as the point of reference, you can choose from the **Target Items:** drop-down checklist located at the top of the tab. You can also freely browse through different keywords or cases by double-clicking its bar on the **Proximity Plot**. The cooccurrence or similarity to more than one target item may be displayed in a single chart allowing for easy comparisons. When several target items are selected, the proximity plot may consist of bars clustered side-by-side (clustered bars) or stacked, representing either the total amount (stacked bars) or the relative distribution of scores (100-percent stacked bars). When two target items are selected, it is also possible to display the bars on both sides of a central axis like the sample chart above (mirrored bars).

By default, the chart displays the proximity of the target items to a maximum of 30 related items. Clicking the  button located in the upper left-hand corner of the chart, displays a dialog box that allows you to either manually choose items to be plotted or to automatically select a specific number of items.

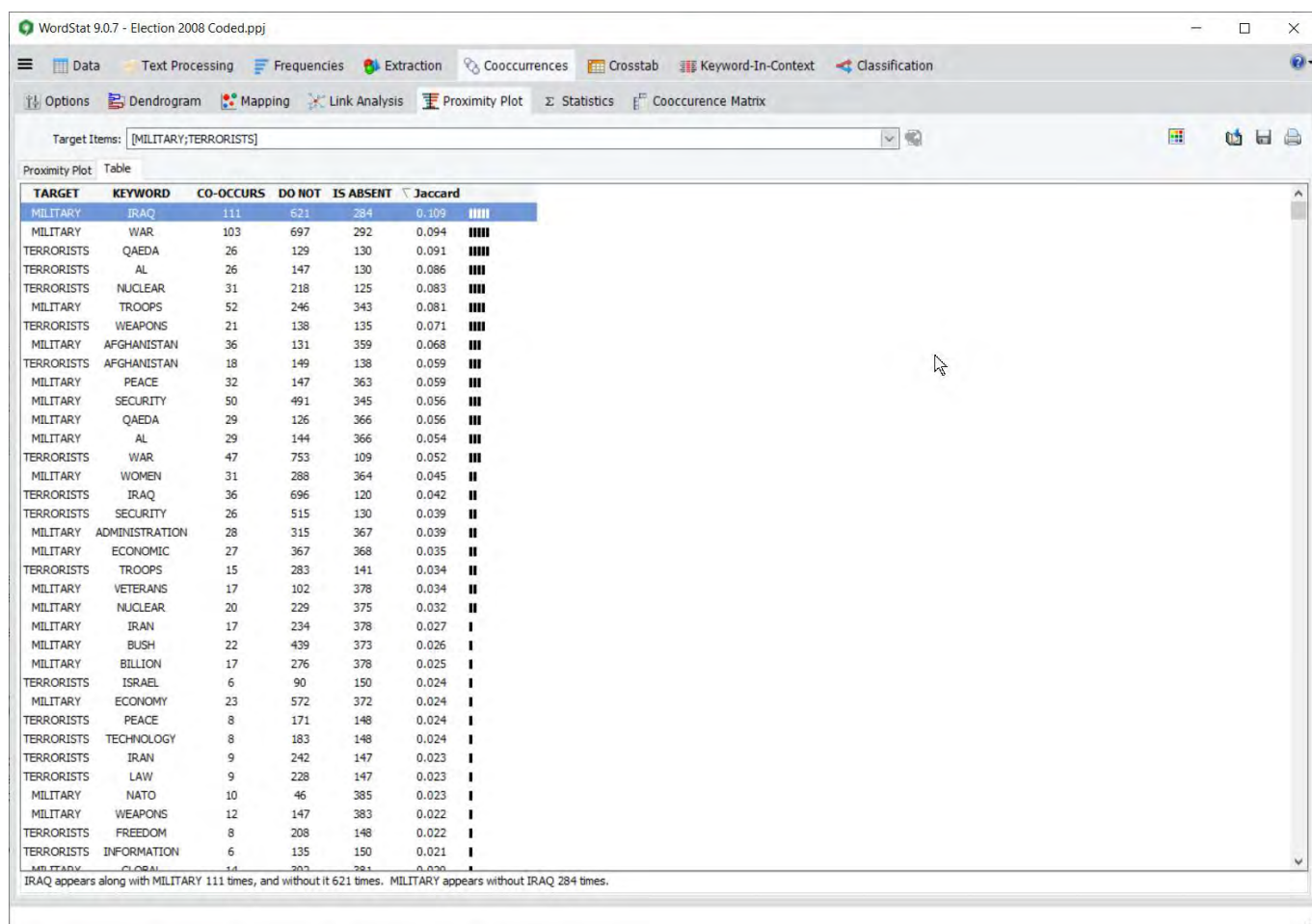
When looking at keyword cooccurrences, selecting a bar enables the  button. Clicking this button retrieves every document or text segment containing both keywords, allowing you to further explore the factors that may explain this cooccurrence. When examining the similarity of documents rather than keywords, clicking this button retrieves both documents and displays them side-by-side in a dialog box.

Right-clicking any existing bar displays a menu that allows you to remove the selected item, move it to the list of target items either by adding it to the existing bars or replacing one of them. You may also retrieve documents or text segments using this popup menu.

Table Tab

The **Table** tab allows you to examine in more detail the numerical values behind the computation of these plots. When the distance measure is based on cooccurrences, the table provides detailed information, such as the number of times a given keyword cooccurs with another one (**COOCCURS**) and the number of times it appears in the absence of this selected keyword (**DO NOT**). Such a table also includes the number of times the selected keyword appears in the absence of the

given keyword (**IS ABSENT**). In the example below (computed using the paragraph as the frequency criteria), we can see on the highlighted line that the category **WORK LIFE BALANCE** cooccurs 115 times with **BENEFITS**, but this word is encountered in 230 paragraphs without the word **BENEFITS**, while **BENEFITS** is found in 914 paragraphs in the absence of **WORK LIFE BALANCE**. The Jaccard coefficient of 0.091 indicates that of all paragraphs that contain either one of these words, only 10.9 percent contains both words. Note, however, that not all proximity measures can be interpreted this easily. To facilitate the interpretation of this table, the status bar provides a textual interpretation of some of the statistics.



TARGET	KEYWORD	CO-OCCURS	DO NOT	IS ABSENT	Jaccard
MILITARY	IRAQ	111	621	284	0.109
MILITARY	WAR	103	697	292	0.094
TERRORISTS	QAEDA	26	129	130	0.091
TERRORISTS	AL	26	147	130	0.086
TERRORISTS	NUCLEAR	31	218	125	0.083
MILITARY	TROOPS	52	246	343	0.081
TERRORISTS	WEAPONS	21	138	135	0.071
MILITARY	AFGHANISTAN	36	131	359	0.068
TERRORISTS	AFGHANISTAN	18	149	138	0.059
MILITARY	PEACE	32	147	363	0.059
MILITARY	SECURITY	50	491	345	0.056
MILITARY	QAEDA	29	126	366	0.056
MILITARY	AL	29	144	366	0.054
TERRORISTS	WAR	47	753	109	0.052
MILITARY	WOMEN	31	288	364	0.045
TERRORISTS	IRAQ	36	696	120	0.042
TERRORISTS	SECURITY	26	515	130	0.039
MILITARY	ADMINISTRATION	28	315	367	0.039
MILITARY	ECONOMIC	27	367	368	0.035
TERRORISTS	TROOPS	15	283	141	0.034
MILITARY	VETERANS	17	102	378	0.034
MILITARY	NUCLEAR	20	229	375	0.032
MILITARY	IRAN	17	234	378	0.027
MILITARY	BUSH	22	439	373	0.026
MILITARY	BILLION	17	276	378	0.025
TERRORISTS	ISRAEL	6	90	150	0.024
MILITARY	ECONOMY	23	572	372	0.024
TERRORISTS	PEACE	8	171	148	0.024
TERRORISTS	TECHNOLOGY	8	183	148	0.024
TERRORISTS	IRAN	9	242	147	0.023
TERRORISTS	LAW	9	228	147	0.023
MILITARY	NATO	10	46	385	0.023
MILITARY	WEAPONS	12	147	383	0.022
TERRORISTS	FREEDOM	8	208	148	0.022
TERRORISTS	INFORMATION	6	135	150	0.021
MILITARY	GLOBAL	14	202	381	0.020

IRAQ appears along with MILITARY 111 times, and without it 621 times. MILITARY appears without IRAQ 284 times.

The following list provides a brief description of the buttons found on these two tabs

Proximity plot controls:



By default items in the proximity plot are sorted in descending order of proximity scores. Clicking this button sorts items in alphabetical order.



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



This button allows editing of various features of the proximity plot, such as the appearance of value labels and bars, the chart axis titles, and the location of the legend.

Proximity plot and proximity table controls:



Press this button to append a copy of the chart or the proximity table in the Report Manager. A descriptive title will be provided automatically. To edit this title or to enter a new one, hold down the SHIFT keyboard key while clicking this button (for more information on the Report Manager, see the [Report Management Feature](#) topic).



When the proximity plot is displayed, this button allows storing the chart on disk in one of the supported graphic file formats. When the proximity table is shown, the table can be saved to disk in an Excel, text delimited, XML, HTML or SPSS file.

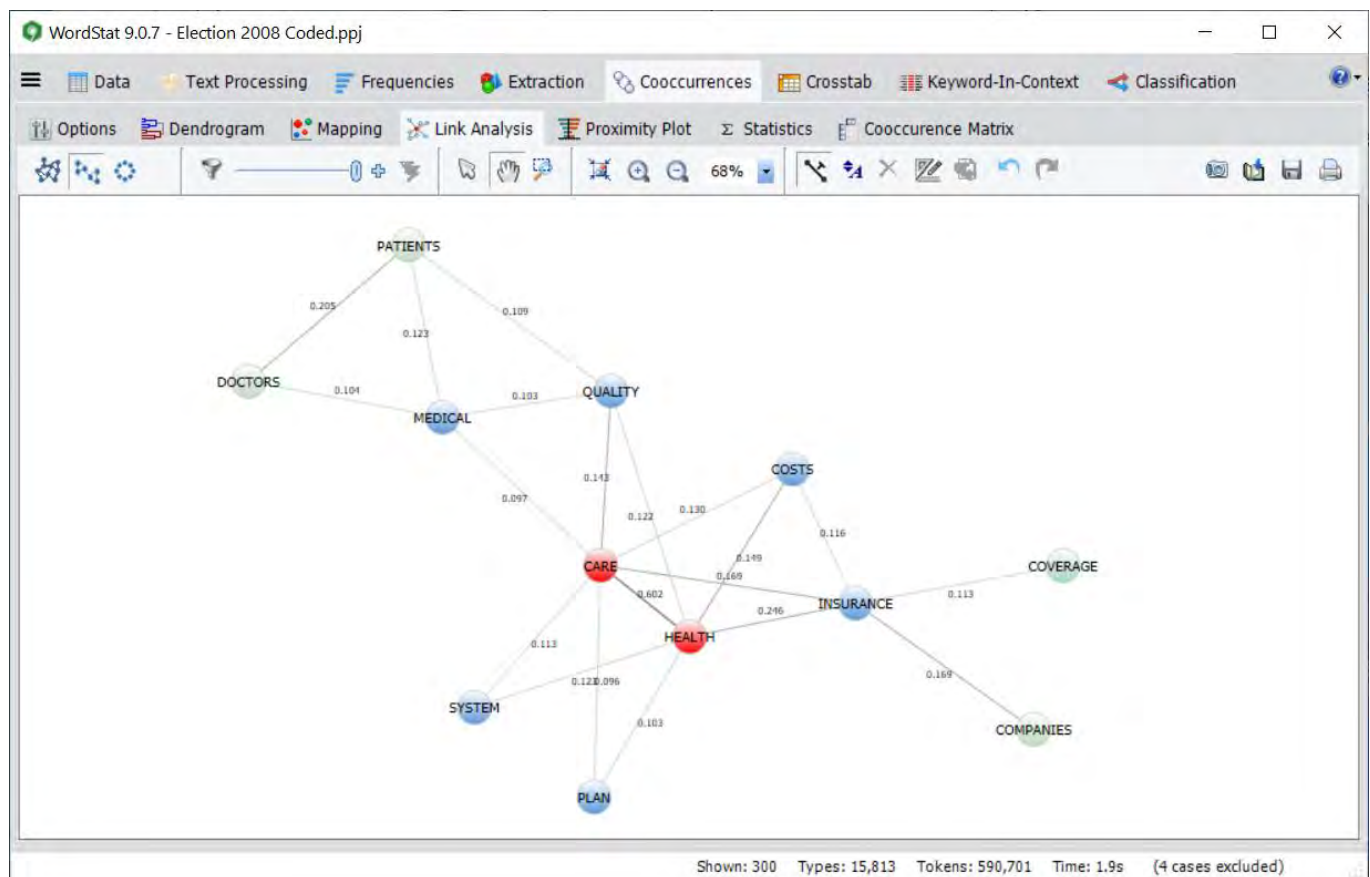


Clicking this button allows you to print a copy of the displayed chart or table.

Link Analysis

The **Link Analysis** tab allows you to visualize the connections between keywords or dictionary items using a network graph. It offers a high level of interactivity, allowing you to explore relationships as well as detect underlying patterns and structures of cooccurrences using three layout types: a multidimensional scaling, a force-based graph, and a circular layout.

This tab is initially associated with the clustering features and the **Dendrogram** tab such that selecting a specific cluster in the dendrogram will result in a network view of its elements - where each element is represented as a node, while their relationship will be represented as a line connecting these nodes (also called an "edge"), the thickness of this line representing the strength of this relationship. Nodes and edges can be edited, deleted or moved, and new nodes and edges may also be added, either manually or using statistical criteria.



Layout Selection

The following three buttons allow you to select one of three layout types.



Clicking this button assigns the location of nodes in a multidimensional space such that nodes that cooccur more often are plotted close together, while those cooccurring less often are plotted far from each other.



Clicking this button draws a force-directed graph in which links are more or less of equal length and there are as few crossed links as possible.



Clicking this button draws nodes in a circle. Nodes that cooccur often are plotted close together.

Nodes and Links Selection



Clicking this button allows you to select nodes and links, either manually or based on statistical criteria. See the Selecting Nodes and Links section below for more information on this feature



This track bar can be used to hide links. By default, the cursor is positioned to the right. Moving it to the left gradually removes the weakest links, allowing you to more clearly identify the strongest ones.



By default, a graph includes links reaching a specific statistical threshold. Clicking this button adds additional links by gradually reducing this statistical threshold.



Clicking this button removes all links that have been hidden using the track bar control (see above), deletes all nodes that are no longer connected, and refreshes the display to take into account the lower number of elements in the graph.

Navigation Mode

The following three buttons allow you to select the default mouse behavior.



Click this button to select specific nodes and links. To select multiple items, hold the Ctrl key while clicking additional items or select the rectangular region of the graph containing the items to select.

To select a node along with all nodes connected to it (neighbors), hold the Alt key while clicking the node. You may also select a node, right-click and choose GET NEIGHBORS.

Once items are selected, they can then be moved, deleted or edited. You may also search for associated text segments either by right-clicking to call up a contextual menu or by clicking the  button on the toolbar.



Clicking this button lets you control which part of the image is visible in the image window.



Clicking this button allows you to zoom into a specific region. Once activated, click the upper-left corner of the rectangular region you want to zoom in on, drag the mouse down the lower-right corner of this region and then release the mouse button.

Zooming In and Out



Clicking this button zooms in or out so you can see the entire map.



Clicking this button increases the zoom level of the map.



Clicking this button decreases the zoom level of the map.



This list box allows you to set the zoom level to predefined levels. You can also type in the desired zoom level.

Editing Buttons



This button allows you to display or hide the numerical value associated with links.

-  This button allows you to change the size of the nodes, links and their associated text. When clicked, a dialog box with four track bars appears. Moving the cursor to the right increases the size of its associated elements while moving it to the left decreases its size.
-  Clicking this button deletes the selected nodes and links.
-  Click this button to display a dialog box to change the style, color and size of the selected nodes and links.
-  To retrieve segments associated with a node, select the node and click this button. When more than one node or when a link is selected, all text segments containing at least two keywords will be retrieved and presented in a table format. You may, however, change the type of segments retrieved (paragraphs, sentences or full documents) or the minimum number of topic items needed for retrieval.
-  This button is used to send a copy of the graph to the clipboard. When this button is clicked, a pop-up menu appears, allowing you to select whether the graph should be copied as a bitmap or as a metafile.
-  Click this button to append a copy of the graph into the Report Manager. A descriptive title will be provided automatically. To edit this title or to enter a new one, hold down the Shift key while clicking this button. (For more information on the Report Manager, see the [Report Management Feature](#) topic.)
-  This button allows you to store the displayed network graph in a graphic file. WordStat supports three different file formats: .BMP (Windows bitmap files), .PNG (Portable Network Graphic compressed files) and .JPG (Jpeg compressed files).
-  Clicking this button allows you to print a copy of the displayed graph.

Selecting Nodes and Links

There are four methods available to select nodes to be plotted:

1. Selecting a Cluster on the Dendrogram tab

The link analysis graph is initially connected to the **Dendrogram** tab, so that selecting one of its clusters results in its items being represented as nodes on the link graph. The similarity criterion is automatically set to the value connecting the last item of this cluster to the other elements. To map connections between items of another cluster, simply move back to the **Dendrogram** tab and click anywhere on the new cluster. A smaller version of the link map should appear in the lower-right panel of the **Dendrogram** tab. You may then move back to the **Link Analysis** tab, or click the  button in the lower-right panel.

2. Double-Clicking a Node

When the graph is in selection mode, double-clicking a node will set the node as the target item and will display all nodes associated with it according to the last criterion settings (primary links). All currently displayed nodes not associated with this new target item will be removed.

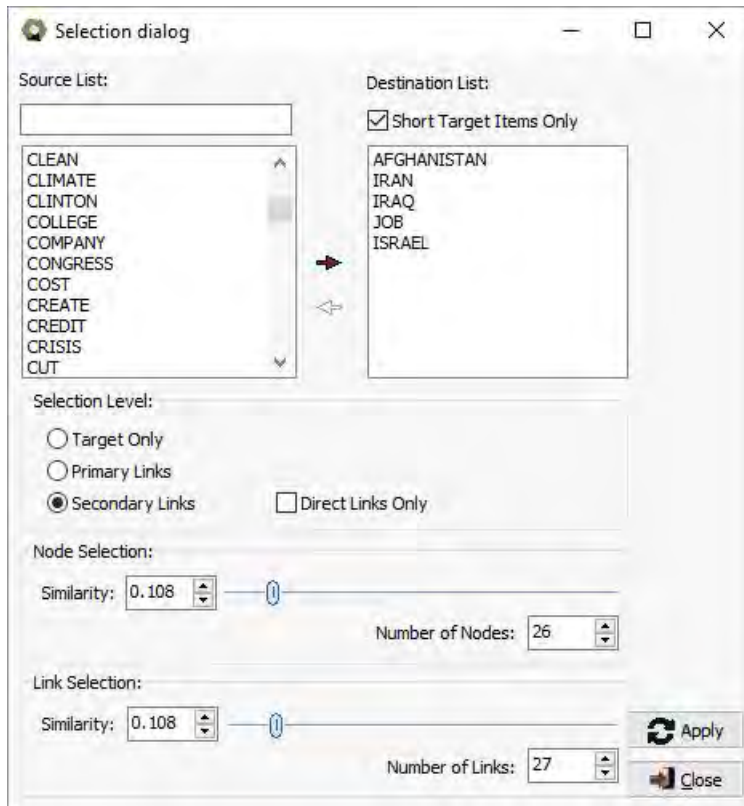
If the Shift key is held down while double-clicking a node, the graph will also include nodes associated by a second degree to the target node (secondary links) as long as they also match the same criterion.

3. Selecting Nodes and Right-Clicking

When one or several nodes are selected, right-clicking brings a popup menu that allows you to either **ADD** to the current graph, primary links and secondary links or **REPLACE** the currently displayed nodes with those associated directly (one level) or indirectly (second level) with the selected nodes.

4. Using the Nodes and Link Filtering Dialog Box

- Clicking the  button on the toolbar displays a dialog box that allows you to manually select items to be graphed. Such a dialog box looks like this:



The following groups of options are available for selecting nodes and links to be plotted:

Source List This list box contains all items that are available to be plotted. To include them in a graph, simply move them to the **Destination List** using the arrow button. A **Search** edit box at the top of this list allows you to quickly find specific items.

Destination List This list box contains items that are currently plotted. When the **Show Target Items Only** option is clicked, the dialog box will offer the possibility of selecting additional nodes and links based on similarity or proximity criteria. When not checked, the graph will display the connection of only the nodes currently listed. Arrow buttons may be used to remove or add additional items.

Selection Level When the destination list is set to include only target items, you may choose whether additional nodes should be plotted and how those will be selected. When set to **Target Only**, no additional items will be plotted. Clicking **Primary Links** will allow the selection of additional nodes directly linked to target items based on a similarity criterion. When the **Secondary Links** option is chosen, additional items may be included if they are indirectly connected to the target nodes through the primary links, as long as they reach the same similarity criteria. By default, all links reaching such a criteria will be plotted, including links among the primary and secondary linked nodes. Selecting **Direct Links Only** will omit most of those links keeping only the ones that connect them to the target items.

Node Selection This set of options allows you to adjust the similarity criterion to be used for selecting additional items. The similarity criteria displayed is the one currently set on the [Options](#) tab of the cooccurrence section. You can type a specific value (typically between 0 and 1) in the edit box, increase or decrease the value using either the **UP** or **DOWN** arrow buttons, or by sliding the cursor on the track bar on the right side of the edit box. The **Number of Nodes** option is automatically adjusted to indicate how many items meet the current similarity criterion. Editing its value will reciprocally adjust the similarity criterion.

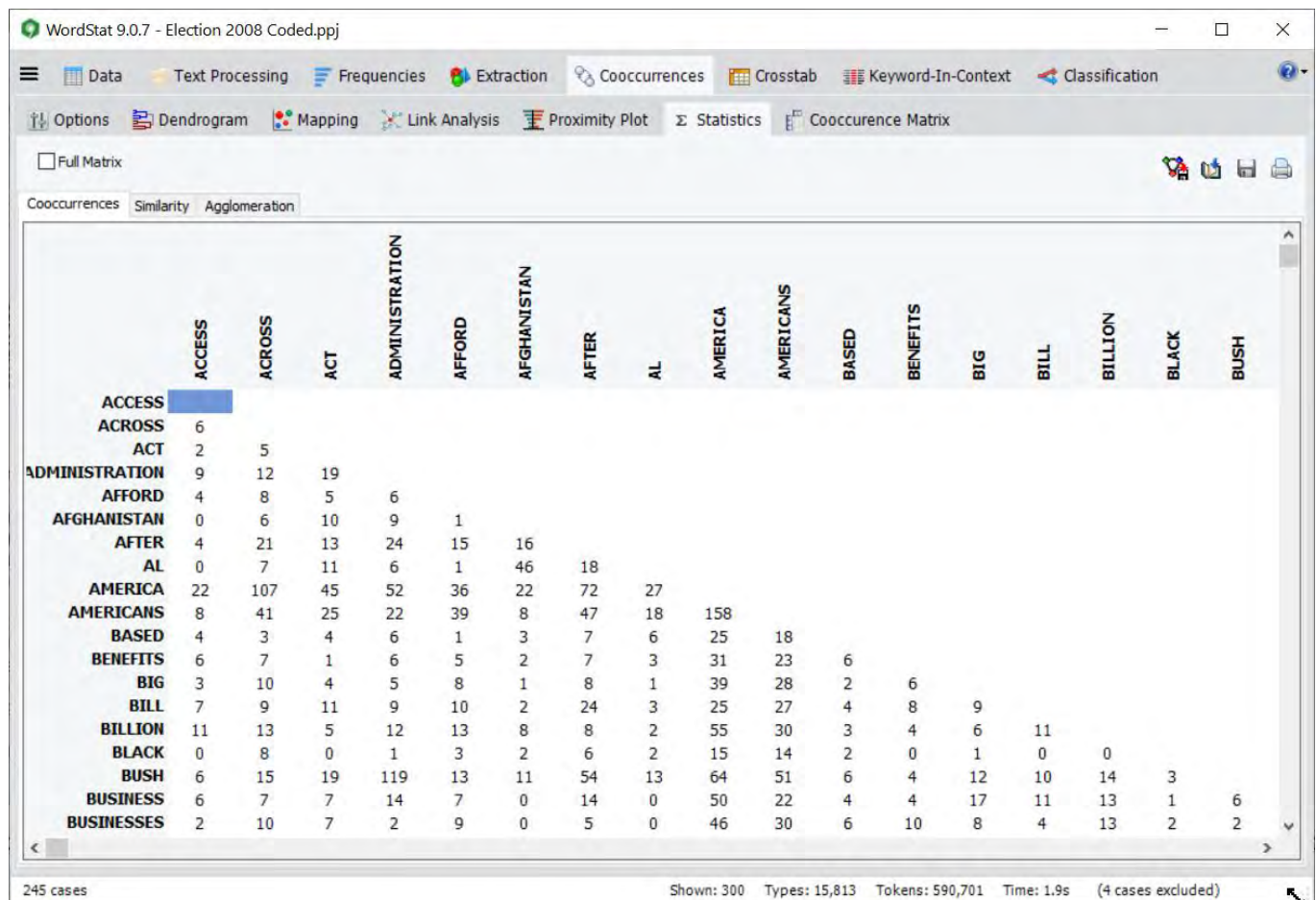
Link Selection This set of options allows you to adjust the similarity criterion to be used for selecting links to be plotted. The similarity criterion for displaying links should always be equal to or less than the one used for selecting nodes. Reducing this criterion will plot additional links. You can type a specific value in

the edit box, and increase or decrease this value using either the **UP** or **DOWN** arrow buttons, or the slide track bar on the right of the edit box. The **Number of Links** option is automatically adjusted to indicate how many links will be plotted. Editing its value will reciprocally adjust the similarity criterion.

- Once the options have been set, click  to update the graph and display the selected nodes and links.

Statistics

The **Statistics** tab displays the cooccurrence and similarity matrices used for building the dendrogram, MDS plots, as well as the proximity plot and table.



The first two tabs display by default, only the lower triangular part of the matrix of cooccurrence and similarity. Selecting the **Full Matrix** option will display data on both sides of the diagonal.

To export a matrix to social network analysis software tools:

- Select the matrix you would like to import by activating either the cooccurrences or the similarity tab.
- Click the  button.
- In the **Save As Type** list box, select the file format under which to save the table. The following formats are supported: UCINET file (*.DL), Pajek Network file (*.NET), NetDraw file (*.VNA) and NetMiner file (*.SNF).

- Type a valid file name with the proper file extension.
- Click the **SAVE** button.

To append a copy of the table in the Report Manager:

- Click the  button. A descriptive title will be provided automatically for the table. To edit this title or to enter a new one, hold down the **Shift** key while clicking this button.

For more information on the Report Manager, see the [Report Manager](#) topic.

To export the table to disk:

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

To print the table:

- Click the  button.

Cooccurrence Matrix

The co-occurrence matrix feature allows one to focus on specific co-occurrences. The main results consist of a table displaying a choice from various co-occurrence statistics. Such a matrix is also highly interactive allowing one to transform specific rows into new columns or vice versa using simple drag-and-drop operations. A charting panel on the left also allows one to assess the distribution of a specific co-occurrence across other variables. One may also obtain a quick view of all text segments associated with a specific co-occurrence. This feature may also be called from the frequency list by selecting target items (words or content categories) that should be displayed as columns, right-clicking, and selecting Co-Occurrence Matrix.

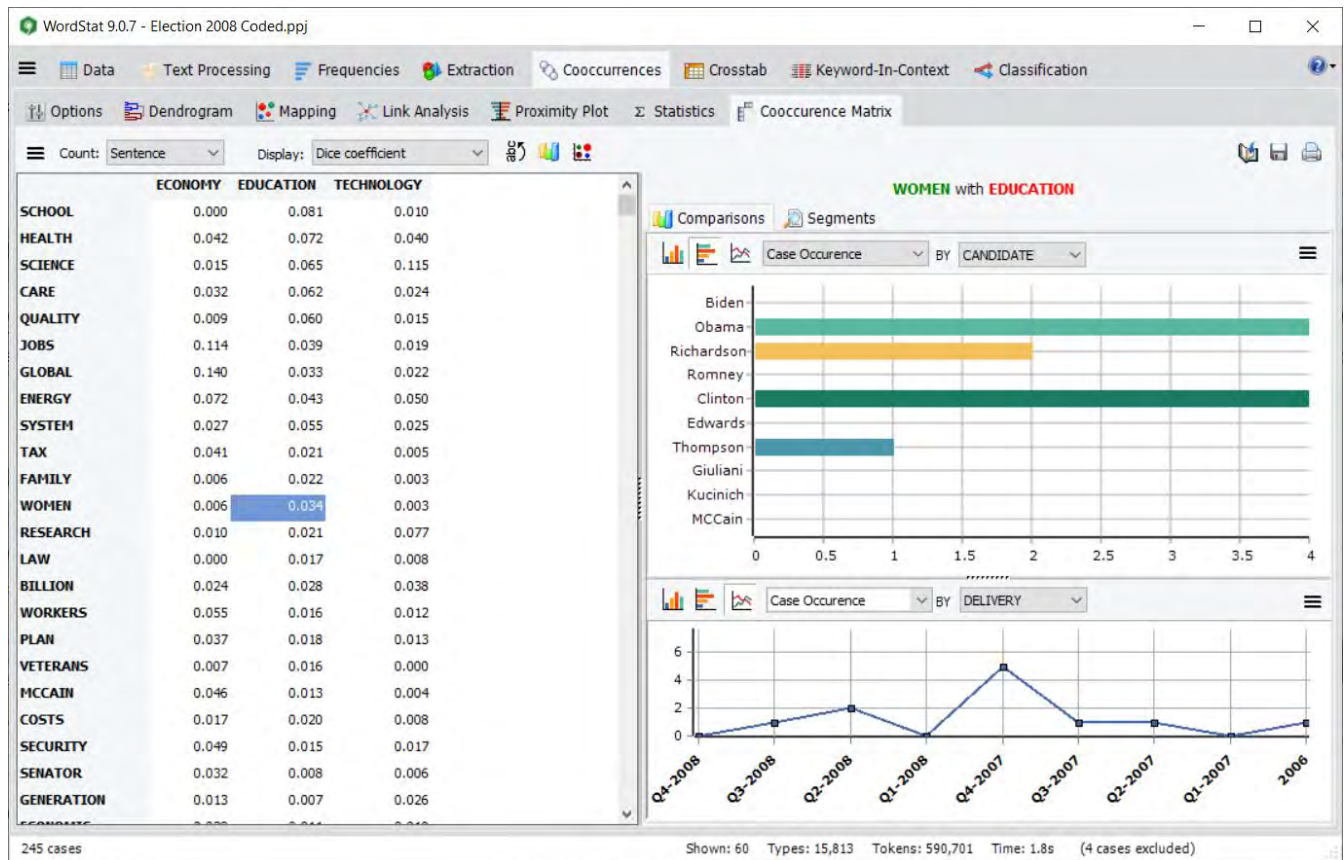
Calling the cooccurrence matrix from the Frequency table

- Select the rows containing the words or content category you would like to use as references. You can select multiple disjunct rows by keeping the CTRL keyboard key down while clicking on the desired rows.
- Right-click to display the popup menu, or click the  button on the left of the Frequencies page toolbar and select the **co-occur with...** command. A dialog box will appear with the selected items as column items and all other words or content categories of the frequency table as row items.

Accessing the co-occurrence matrix feature from the cooccurrences page.

The cooccurrence matrix is also accessible as a separate tab from the main **Cooccurrences** page. When accessing this feature from this location, the unit of text on which cooccurrences are computed is automatically adjusted to the general setting used for clustering, multidimensional scaling and other related features. Moving directly to the **Cooccurrence Matrix** page will result in all clustered items to be displayed as rows without any target item. Selecting a cluster on the **Dendrogram**

page prior to moving to this page, will assign all words of the selected cluster as column items allowing you to see how all items of the selected clusters are related to the remaining items.



Adjusting rows and column items

To transform a row item as a column item:

- Move the mouse over the item name (first column), click on it, and keep the mouse down
- Move the mouse cursor up to the first row at the location where you want this item to appear and release the mouse button.

or

- Click on any cell on the row of the item you want to move to a column. You may also select a range of cells.
- Right-click to display the contextual menu, select **Move** and then choose the **Selected Rows as Columns** command.

To transform a column item as a row item:

- Move the mouse over the item name (first row), click on it, and keep the mouse down
- Move the mouse cursor down to the first column at the location where you want this item to appear and release the mouse button.

or

- Click on any cell on the column of the item(s) you want to move to rows..
- Right-click to display the contextual menu, select **Move** and then choose the **Selected Columns as Rows** command.

To adjust the order of column or row items:

- Move the mouse over the item name you want to move, click on it, and keep the mouse down and drag it to its new location. Green arrows will indicate the location where the item will be inserted.
- One over the desired location, release the mouse button.

To sort rows:

- Double-click on the column header containing the values on which you want rows to be sorted. When clicking on the first column header, items will be sorted alphabetically, while clicking on any other column will sort the rows in ascending order of the numerical value in this column.
- Double-clicking again on the same column header will sort the rows in descending order.

Adjusting display options

The **Count** listbox allows you to specify the size of the text segment that will be used for computing co-occurrences. You can obtain cooccurrence statistics computed on the sentence, the paragraph, the full document, and when there is more than one document selected, on the cases. When only one document variable is analyzed, computed statistics for documents and cases will be identical.

The **Display** list box allow you to create a table using any one of the following statistics:

- Frequency
- % of column item
- % of row item
- Jaccard coefficient
- Dice coefficient
- Adjusted Phi
- Inclusion index

Charting cooccurrences of selected cells:

The **Comparisons** tab on the right panel allows one to compare how co-occurrences of two items are distributed among the selected independent variables using a bar chart or a line chart. Please note that while the charting is linked to the table selection, only one pair of items is being plotted at a time even if multiple cells are selected. The elements of this pair appear at the top of the right panel. One also must remember that what is being charted is not how often those two words are being used, but the absolute or relative frequency by which those two words are being used together. The **Case Occurrence** statistics simply plots how many cases such a cooccurrence is present, irrespective of its frequency while the **Category Percent** computes the percentage of cases for this specific value of the independent variable containing this cooccurrence. All other statistics are computed to represent either the absolute frequency or the relative frequency of one word relative to the other word, or, for the **Jaccard** or **Dice Coefficient**, relative to the presence of either one of them.

One may also represent graphically the co-occurrence statistics of several items by selecting the range of cells to display and by clicking the  button. Alternatively, one may select cells, right-click and select the **Chart Selected Cells** command. For more information on the charting options of this dialog box, see [Bar Chart and Line Chart Options](#).

Creating a bubble chart

Bubble charts are graphic representations of statistical tables where the displayed statistics are represented by circles of different diameters. One can create a bubble chart for representing the entire grid or only selected cells. To create a bubble charts for selected cells only:

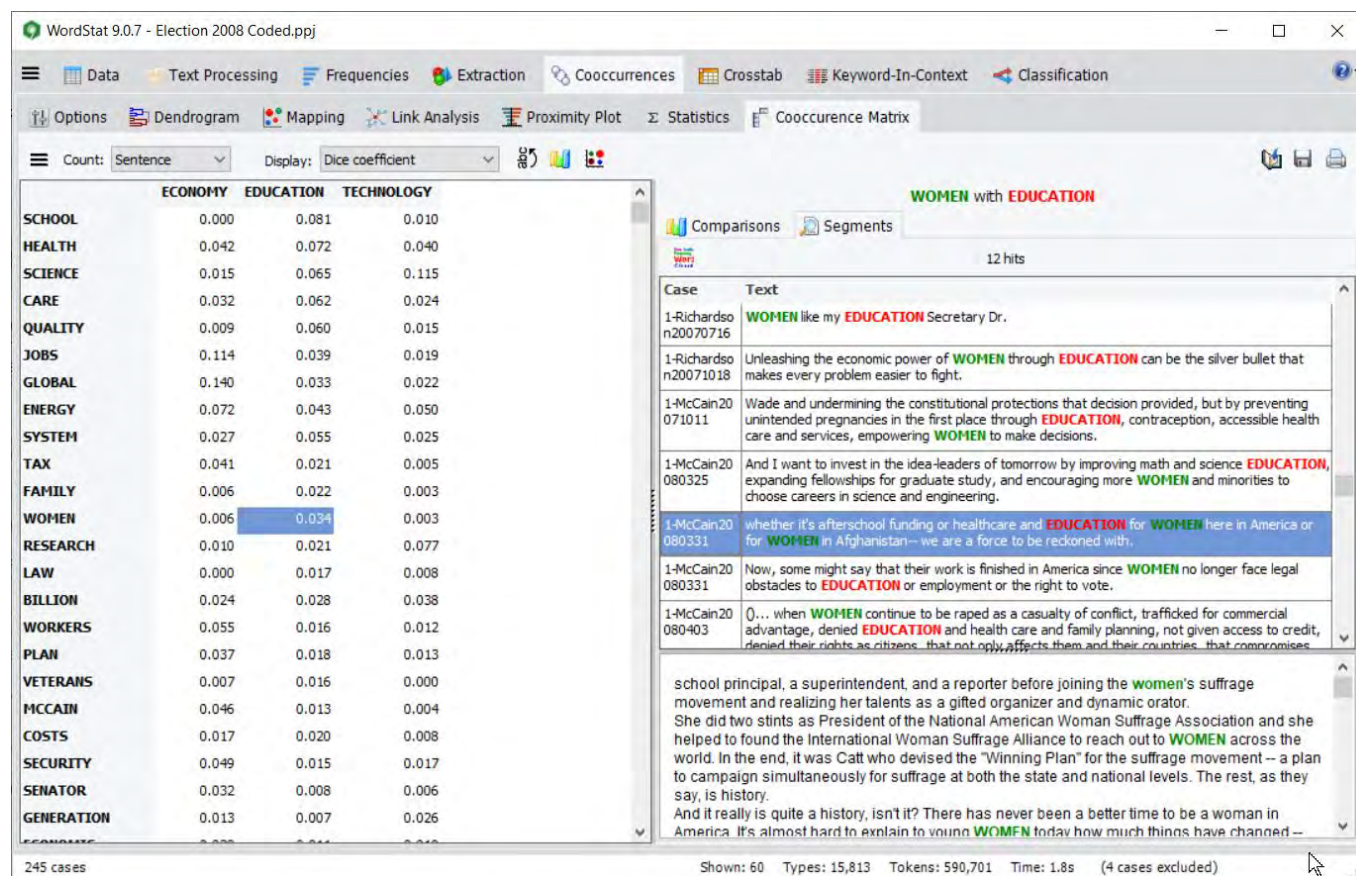
- Select the cells you want to plot.
- Right click and choose **Create Bubble Chart** or click the  button, and then chose **Selected Cells..**
- Select **Entire grid...** instead to create a bubble chart to represent the values contained in the entire table.

For more information, see [Bubble Charts](#).

Retrieving associated text segments

The **Segments** tab on the right panel allows one to quickly retrieve all text segments where both items of the selected cooccurrence appear. Please note that while this feature is linked to the table selection, only one pair of items is being searched for at the time even if multiple cells are selected. The elements of this pair appear above the right panel. The top panel displays the extracted text segment while the bottom part displays the entire document from which this segment come from.

One may further explore the context of such a cooccurrence by clicking the  button to produce a word frequency table and a word cloud from all the retrieved text segments. (See [Word Frequency Analysis](#) for more information on this feature).



The screenshot displays the WordStat 9.0.7 interface for the file "Election 2008 Coded.ppj". The main window shows a table of cooccurrences between various categories. The categories listed are ECONOMY, EDUCATION, and TECHNOLOGY. The rows represent different topics, with "WOMEN" highlighted in blue. The table shows Dice coefficients for each pair of categories.

	ECONOMY	EDUCATION	TECHNOLOGY
SCHOOL	0.000	0.081	0.010
HEALTH	0.042	0.072	0.040
SCIENCE	0.015	0.065	0.115
CARE	0.032	0.062	0.024
QUALITY	0.009	0.060	0.015
JOBS	0.114	0.039	0.019
GLOBAL	0.140	0.033	0.022
ENERGY	0.072	0.043	0.050
SYSTEM	0.027	0.055	0.025
TAX	0.041	0.021	0.005
FAMILY	0.006	0.022	0.003
WOMEN	0.006	0.034	0.003
RESEARCH	0.010	0.021	0.077
LAW	0.000	0.017	0.008
BILLION	0.024	0.028	0.038
WORKERS	0.055	0.016	0.012
PLAN	0.037	0.018	0.013
VETERANS	0.007	0.016	0.000
MCCAIN	0.046	0.013	0.004
COSTS	0.017	0.020	0.008
SECURITY	0.049	0.015	0.017
SENATOR	0.032	0.008	0.006
GENERATION	0.013	0.007	0.026

The right panel shows the "Segments" tab for the selected cooccurrence "WOMEN with EDUCATION". It displays 12 hits. The top section shows the extracted text segment, and the bottom section shows the full document text from which the segment was extracted.

WOMEN with EDUCATION

12 hits

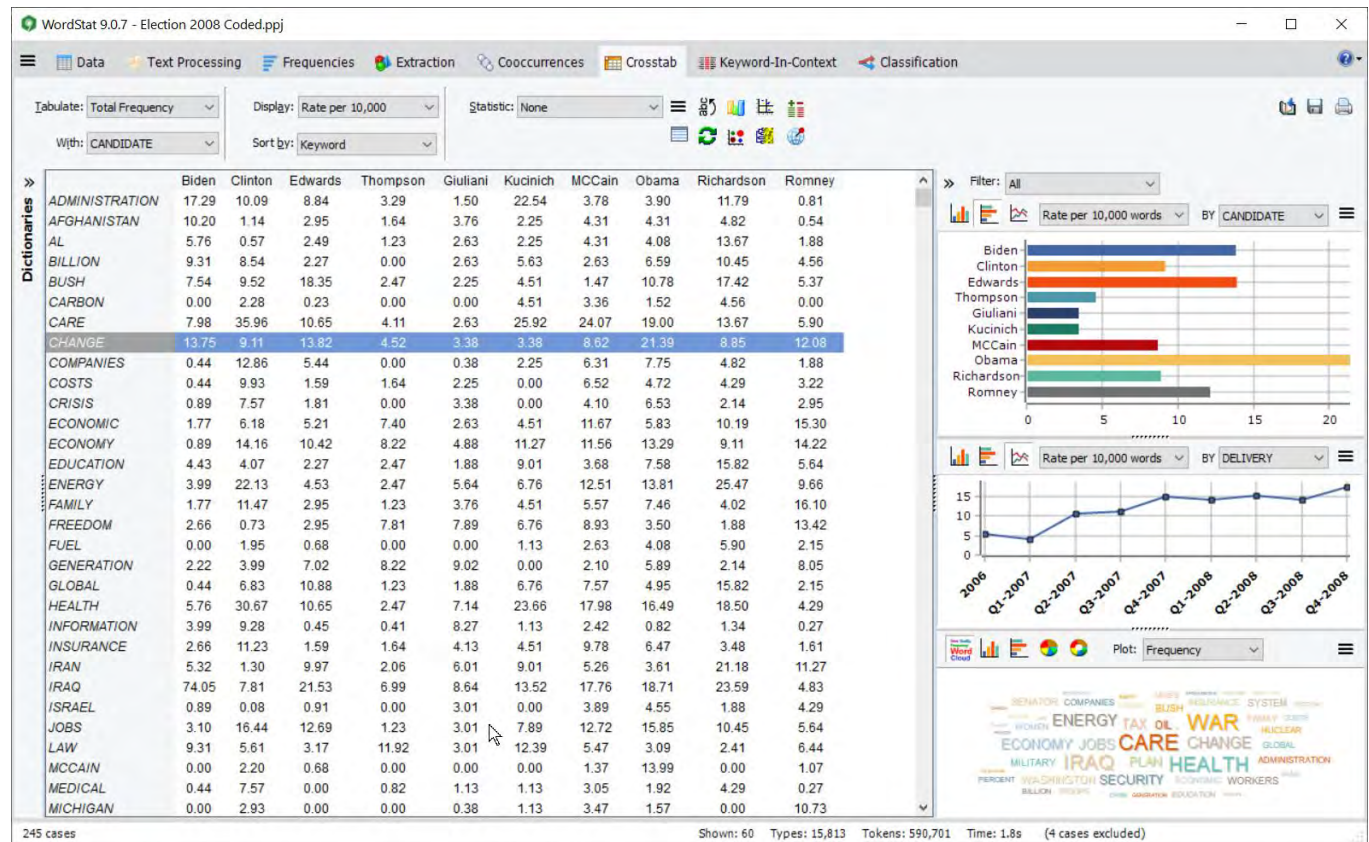
Case	Text
1-Richardson20070716	WOMEN like my EDUCATION Secretary Dr.
1-Richardson20071018	Unleashing the economic power of WOMEN through EDUCATION can be the silver bullet that makes every problem easier to fight.
1-McCain20071011	Wade and undermining the constitutional protections that decision provided, but by preventing unintended pregnancies in the first place through EDUCATION, contraception, accessible health care and services, empowering WOMEN to make decisions.
1-McCain20080325	And I want to invest in the idea-leaders of tomorrow by improving math and science EDUCATION, expanding fellowships for graduate study, and encouraging more WOMEN and minorities to choose careers in science and engineering.
1-McCain20080331	whether it's afterschool funding or healthcare and EDUCATION for WOMEN here in America or for WOMEN in Afghanistan-- we are a force to be reckoned with.
1-McCain20080331	Now, some might say that their work is finished in America since WOMEN no longer face legal obstacles to EDUCATION or employment or the right to vote.
1-McCain20080403	Q... when WOMEN continue to be raped as a casualty of conflict, trafficked for commercial advantage, denied EDUCATION and health care and family planning, not given access to credit, denied their rights as citizens... that not only affects them and their countries... that compromises

school principal, a superintendent, and a reporter before joining the women's suffrage movement and realizing her talents as a gifted organizer and dynamic orator. She did two stints as President of the National American Woman Suffrage Association and she helped to found the International Woman Suffrage Alliance to reach out to WOMEN across the world. In the end, it was Catt who devised the "Winning Plan" for the suffrage movement -- a plan to campaign simultaneously for suffrage at both the state and national levels. The rest, as they say, is history. And it really is quite a history, isn't it? There has never been a better time to be a woman in America. It's almost hard to explain to young WOMEN today how much things have changed --

245 cases Shown: 60 Types: 15,813 Tokens: 590,701 Time: 1.8s (4 cases excluded)

The Crosstab Tab

The **Crosstab** tab is used to display a contingency table of words or categories. This contingency table is computed only on items that have been included. If a categorization dictionary has been specified, this grid will display only the words or keywords in this list. If no dictionary has been specified, the grid will display all words that have not been explicitly excluded. Along with absolute and relative frequency of keyword occurrence or keyword 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.



Tabulate: This option allows you to choose whether the values in the table should be based on the total frequency of keywords or the number of cases containing the keywords.

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

- **<other keywords>:** display a square table showing the number of cooccurrence of keywords in the same case.
- **<case number>:** display the keyword occurrence or frequency for each individual case.
- **Any independent variable:** If numeric, categorical or date variables were selected as independent variables, their names will appear in this list box. Selecting any of those variable names will display a contingency table allowing for the assessment of the relationship between this variable and the keywords or content categories. When a date variable is selected, a dialog box like the one below will appear allowing you to automatically recode those dates into various time periods or date units such as week days or months.



- **<variables>**: When content analysis is performed on several alphanumeric variables, this option will allow for comparison of the occurrence of keywords according to variable name.
- **Select variables...**: By default, the variables listed in the With list box are the independent variables that have been selected when calling WordStat. Choosing this option displays a dialog box that allows you to select other numeric, categorical or date variables.
- **Combine variables...**: This option allows you to compare the frequency of keywords or content categories among the combined values of two variables. When this item is selected, a dialog box appears allowing you to choose the two variables whose values will be combined.

Sort by: This option presents the opportunity to sort the table by word or category names (alphabetical order) or by descending order of frequency or case occurrence. When a statistic is displayed (see option **Statistic**), 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 this column heading. Clicking several times on the same column heading toggles between ascending and descending orders.

Display: This list box allows you 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 **Case Occurrence**, two additional statistics are also available:

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

Statistics: When keyword frequency or occurrence is broken down by an independent variable (see **With** option), a drop-down list will appear. This list box allows you 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 you 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.

To display column labels vertically, click the  button located to the right of the **Tabulate With** list box.

The  button is used to reapply the content analysis process to the current data set. This button is disabled by default and becomes enabled when changes are made to any one of the currently active text-analysis processes, such as the categorization dictionary, the exclusion list or the substitution process. Clicking this button will instruct WordStat to reprocess the text collection and update the current table.

The Comparisons Panel

The **Comparisons** panel on the right allows you to look at the distribution of the selected words or content categories among values of up to two structured variables. It may be used to get a quick graphic representation of the distribution of the variable currently used in the crosstabulation table but may also be adjusted to display the distribution on other variables. You may display such distributions using either a vertical bar chart, a horizontal bar chart or a line chart, by clicking on the corresponding button. Four statistics may also be represented on those charts:

- Case Occurrence** Number of cases in this subgroup containing at least one of these words or category items.
- Category Percent** Percentage of cases in this subgroup containing at least one of these words or category items.
- Word Frequency** Total number of selected items in the subgroup.
- Rate per 10,000 Words** Rate of words in this subgroup per 10,000 words.

The bottom chart contains the distribution of selected items in the crosstab page using either a word cloud, a vertical of horizontal bar chart, a pie chart, or a donut chart. Right-clicking anywhere in the chart areas displays a pop-up menu that allows you to edit the chart, save it to disk or in the **Report Manager**, or to copy it to the clipboard. Clicking a specific bar or a data point of a line chart also allows one to retrieve text segments associated with the selected class and containing words of the selected topic.

The comparison panel may also be used to filter data on one or several values of the variables currently used for comparison in the crosstab page.

To filter cases and compare distribution on other variables:

- Adjust the **Filter** list box at the top of the comparison panel to the value you want to filter on. The two graphs below will be automatically adjusted to display the distribution of the selected items on the chosen variables for only cases corresponding to the selected value. The word cloud at the bottom will represent the distribution of all items for cases corresponding to the selected value. Changing from one value to another may allow one to identify meaningful differences in distributions associated with those values.

- It is also possible to examine the distribution on a combination of values by selecting **Multiple...** and then putting check marks beside all values that you want to combine and filter on.
- To disable the filter on values, simply set the **Filter** list box to **All**.

Creating Bar Charts or Line Charts

The Crosstab tab also allows you 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 subgroups of individuals (male vs. female) or time periods.

To produce such charts:

- Set the **Tabulate** and **Display** options so that the information to visualize is displayed in the table.
- Using the mouse, select the rows you would like to display. Multiple but separate rows can be selected by clicking while holding down the **Ctrl** key.
- Click the  button or right-click the mouse and select the **Chart Selected Rows** command.

Creating Bubble Charts

Bubble charts are graphic representations of contingency tables where relative frequencies are represented by circles of different diameters.

To create a bubble chart:

- Set the **Tabulate**, **Count** and **Display** options so that the information to view is displayed in the table.
- Click the  button.

For more information, see [Bubble Charts](#).

Creating Heat Maps with Clustering of Rows and Columns

A heatmap plot is a graphic representation of crosstab tables where cell frequencies are represented by different color brightness or tones. When combined with clustering of rows and/or columns, this exploratory tool allows you to identify functional relationships between specific keywords and subgroups defined by values of the independent variable.

To create a heatmap:

- Set the **With** option to an independent variable.
- Set the **Tabulate** option to either **Case Occurrence** or **Keyword Frequency**.
- Click the  button to access the heatmap dialog box.

For more information on heatmaps, see [Heatmap plot](#).

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.
- Set the **Tabulate** option to either **Case Occurrence** or **Keyword Frequency**.
- Click the  button to access the correspondence analysis dialog box.

For more information on correspondence analysis, see [Correspondence Analysis](#).

Other Tasks

To export the table to disk:

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

To append the table to the Report Manager:

- Click the  button. A descriptive title will be provided automatically for the table.
 - To edit this title or to enter a new one, hold down the **Shift** key while clicking this button.
- (for more information, see the [Report Manager](#) topic).

To copy the entire table to the clipboard:

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

To copy selected rows to the clipboard:

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

To search for a specific item:

- Right-click anywhere in the table.
- Select the **FIND** command from the pop-up menu. A search dialog box will appear.

- Type the desired search string in the **Find What** edit box. To restrict the search to items starting with the search string, enable the **Match Beginning of Item** option and to restrict it to whole words matching the search string, enable the **Match Whole Word Only** option.
- Click the **Find** button to search the first item matching the typed string. Clicking this button again finds the next occurrence of the search string, starting at the currently selected item.

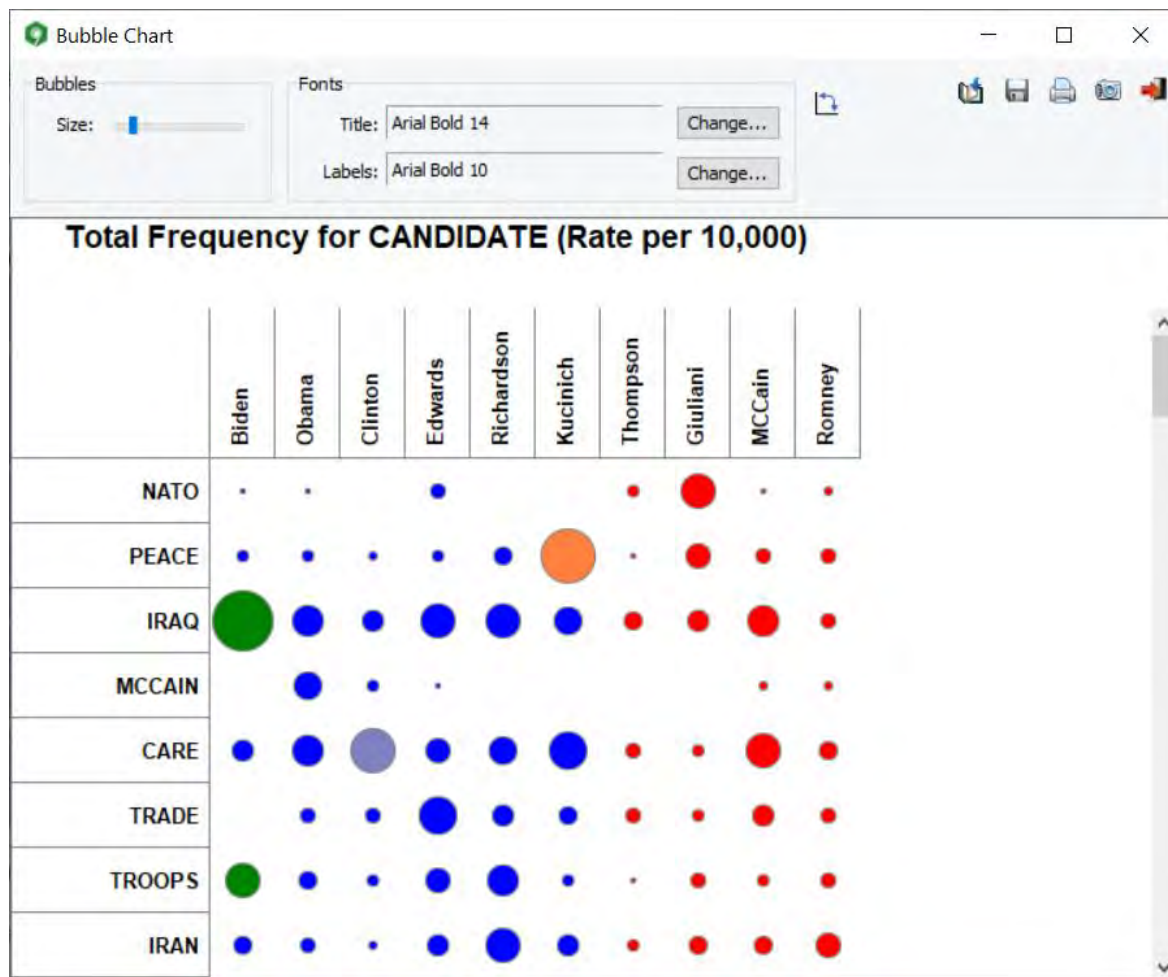
Bubble Charts

Bubble charts are graphic representations of contingency tables where relative frequencies are represented by circles of different diameters. This type of graphic chart allows you to quickly identify high-and low-frequency cells and is especially useful for presentation purposes. Many features of the chart can be customized to highlight specific findings. Rows and columns can be moved freely or deleted, and you can adjust the color of each cell as well as the fonts used in the chart.

The bubble chart represents graphically the underlying table and displays the corresponding measure used (code occurrences or frequencies, word counts or percentages) and will also reflect the currently selected display settings (such as count, percentage of rows or columns, etc.).

To create a bubble chart:

- Move to the **Crosstab** tab.
- Set the **Tabulate** option to either **Case Occurrence** or **Total Frequency**.
- Set the **With** option to the desired independent variable.
- Set the **Display** option to specify how this information will be displayed.
- Click the  button. A dialog box similar to this one will appear:



To adjust the size of the bubbles:

- In the **Bubbles** group box at the top of the window, adjust the **Size** option to Small, Medium or Large.

To adjust the color of the bubbles:

- Select the cell or group of cells you would like to alter.
- Right click and select the **SET COLOR** command. A dialog box will appear letting you choose a specific color value.

To move a row or a column:

- Click anywhere on the cell header containing the row or column label and hold the mouse down.
- Drag the mouse cursor over the desired new location and release the mouse button.

To delete selected rows or columns:

- Select a cell range covering the rows or columns you would like to delete.
- Right-click to display a popup menu.
- Select the **REMOVE** command and then choose **SELECTED ROWS** or **SELECTED COLUMNS** depending on your desired action.

To adjust the font size and style of the title and labels:

- Click the **Change** button beside the **Title** or **Labels** options. A Font setting dialog box will appear, letting you change the font, the font size, style, and color.

To edit the title:

- Click anywhere in the title region and type in the new title. The height of the title region is automatically adjusted to the number of lines in the title.

The following table provides a short description of additional buttons:

Control	Description
	Pressing this button transpose the grid so that rows become columns while columns are transformed into rows. Such a feature may be useful to flip rectangular charts and choose between a portrait or a landscape display.
	Press this button to append a copy of the graphic in the Report Manager. A descriptive title will be provided automatically. To edit this title or to enter a new one, hold down the Shift key while clicking this button (for more information on the Report Manager, see the Report Management Feature topic).
	Click this button to save a chart on disk. Charts may be saved in BMP, JPG or PNG graphic file format or may be saved in a proprietary format (.WSX file extension) that can later be edited and customized using the Chart Editor.
	Clicking this button prints a copy of the displayed chart.
	This button creates a copy of the chart to the clipboard. When this button is clicked, a shortcut menu appears allowing you to select whether the chart should be copied as a bitmap or as a metafile.

Heatmap Plot

Heatmap plots are graphic representations of crosstab tables where relative frequencies are represented by different color brightness or tones and on which a clustering is applied to reorder rows and/or columns. This type of plot is commonly used in biomedical research to identify gene expressions. When used for text mining, such an exploratory data analysis tool facilitates the identification of functional relationships between related keywords and a group of values of an independent variable by allowing the perception of cell clumps of relatively high or low frequencies or of outlier values.

The heatmap plot used in WordStat allows you to graphically examine the relationship between keywords (rows) and values of an independent variable (columns). While the clustering available through the WordStat Frequency tab (see [Hierarchical clustering and multidimensional scaling](#)) is performed on individual cases and documents, the clustering analyses used in the heatmap plot are performed directly on the crosstabulation tables. As a consequence, the similarity index computed for two keywords and used for clustering does not represent their cooccurrences within cases but measures the similarity of their distribution among the various groups of the independent variable. Likewise, two subgroups defined by values on the independent variable will be considered near to each other if the distributions of keywords in those two groups are similar.

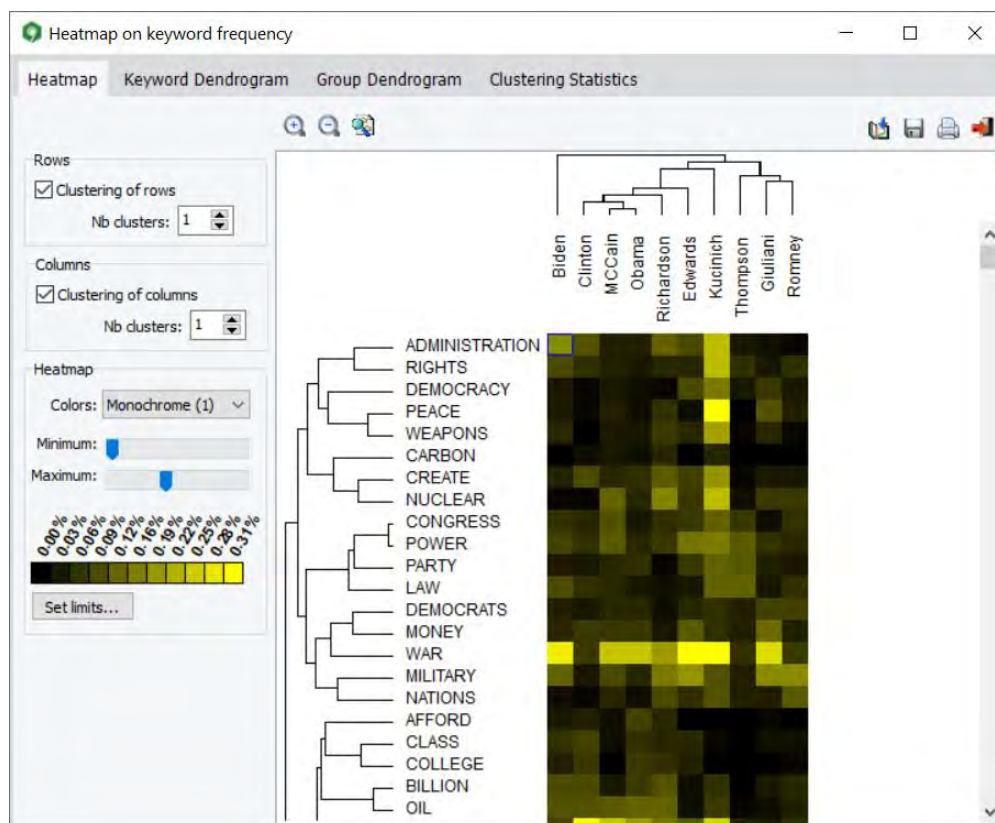
The heatmap plot can be performed either on keyword frequency or occurrences within cases (present or absent). For both types of analysis, the observed frequency is transformed into a percentage by dividing the frequency by either the total number of words in a specific subgroup (keyword frequency) or by the total number of cases in this subgroup (case occurrences).

To create a heatmap:

- Select the Crosstab tab.
- Set the **Tabulate** option to either **Total Frequency** or **Case Occurrence**.
- Set the **With** option to the desired independent variable.
- Click the  button.

Heatmap Tab

The main section on the right of this tab as shown in the screen shot below, displays the heatmap grid representing the relative frequencies of each cell (row and column intersection) using different brightness or color tones. Optional dendrograms are displayed at the top and to the left margin of this grid. The size of these dendrograms may be adjusted by moving the mouse cursor over the bottom edge (upper dendrogram) or the right edge (dendrogram on the left) of the dendrogram and dragging its limit to the desired size.



The font size used to display the row and column values may also be adjusted by clicking the  or  buttons located on the top toolbar. The size of cells and the distance between dendrogram leaves are automatically adjusted to the new font size.

To identify which specific cases or text documents are associated with a cell or group of cells, simply select the rectangular area of the list you would like to obtain, click the  button and select **documents**, **paragraphs** or **sentences**. WordStat locates the documents or text segments associated with these cells and displays them in the [Keyword Retrieval](#) dialog box.

Rows option group:

Clustering of rows: Enabling this option reorders the rows according to the result of a cluster analysis and displays a dendrogram at the left of the heatmap plot. Keywords that are distributed across the various subgroups in a similar way will tend to be grouped under the same cluster.

Sort by: When the clustering of keywords is disabled, rows may be sorted in alphabetical order, in descending order of frequency or case occurrence or displayed in their original categorization dictionary order.

No clusters: This option allows setting how many clusters the clustering solution should have. When two clusters or more are selected, horizontal red lines are drawn in the heatmap to delineate those clusters.

Columns options group:

Clustering of columns: Enabling this option reorders the columns according to the result of a cluster analysis and displays a dendrogram at the top of the heatmap. Columns with similar distribution of keywords will tend to be grouped under the same cluster. When the clustering option is disabled, the columns are presented in their ascending order of values. Disabling the clustering of columns is especially useful to preserve the ordinal nature of the values. For example, when looking for a relationship between some words or keywords and publication years, then disabling the clustering of columns will automatically sort the columns in chronological order allowing the identification of temporal trends.

No clusters: This option allows setting how many clusters the clustering solution should contain. When two clusters or more are selected, vertical red lines are drawn in the heatmap to delineate those clusters of keywords.

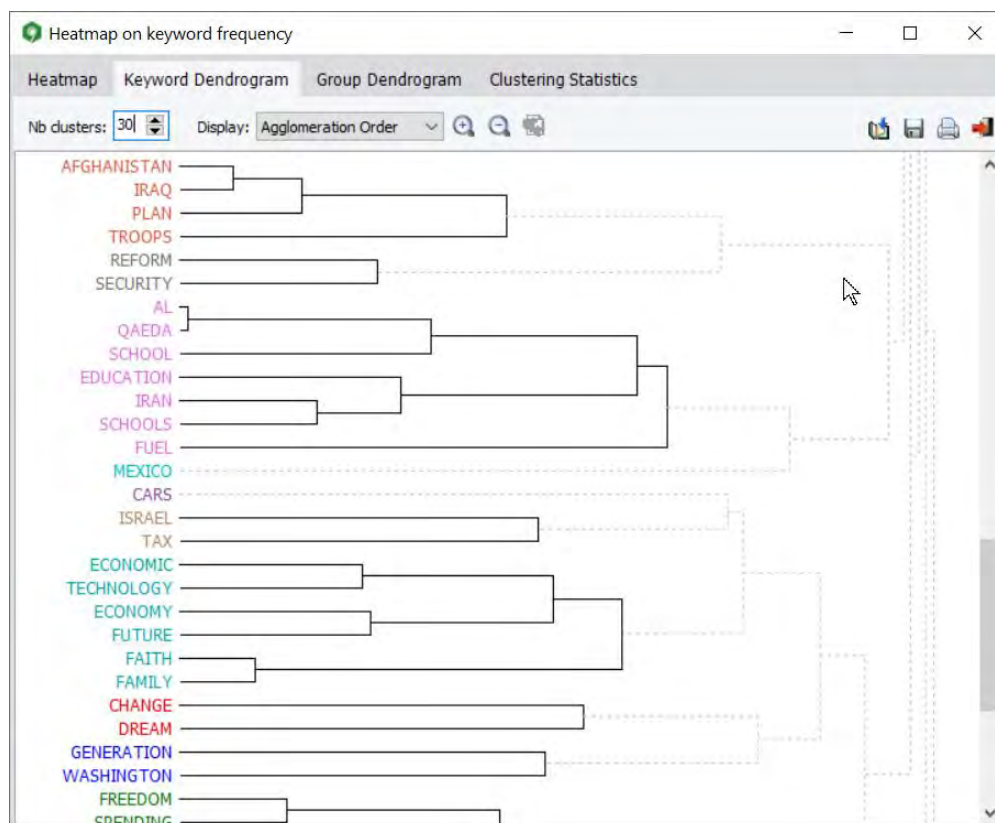
Heatmaps options group:

Colors: The colors list box allows the selection of various color schemes to represent differences in percentages. Monochrome schemes will express differences using levels of brightness of a single color where black always represents the lower limit of the selected range. Multicolor spectrums may be used to represent greater nuances in the selected range of percentages.

Minimum / Maximum: The minimum and maximum slide bars allow the setting of the range of values that will be used to display variations of color tones or brightness. By default, the minimum value is set to zero while the maximum value is set to the highest observed percentage. To increase the contrast at a specific location of this range, minimum and maximum limits may be adjusted. All cells with values lower than the minimum will be represented with the color located on the left end of the selected color spectrum while cells with percentages higher than the maximum limit will be displayed with the color located on the right end of this spectrum. Reducing the range of values increases the contrast in a specific region of percentage values.

Keyword and Group Dendrogram Tabs

The second and third tabs of the heatmap dialog box allow a more detailed examination of the clustering of words or categories (Keyword Dendrogram) and of all values on the independent variable (Group Dendrogram). WordStat uses 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 known 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. Items that tend to be distributed similarly against the other variable appear together and are combined at an early stage while those that have dissimilar distributions tend to be combined at the end of the agglomeration process.



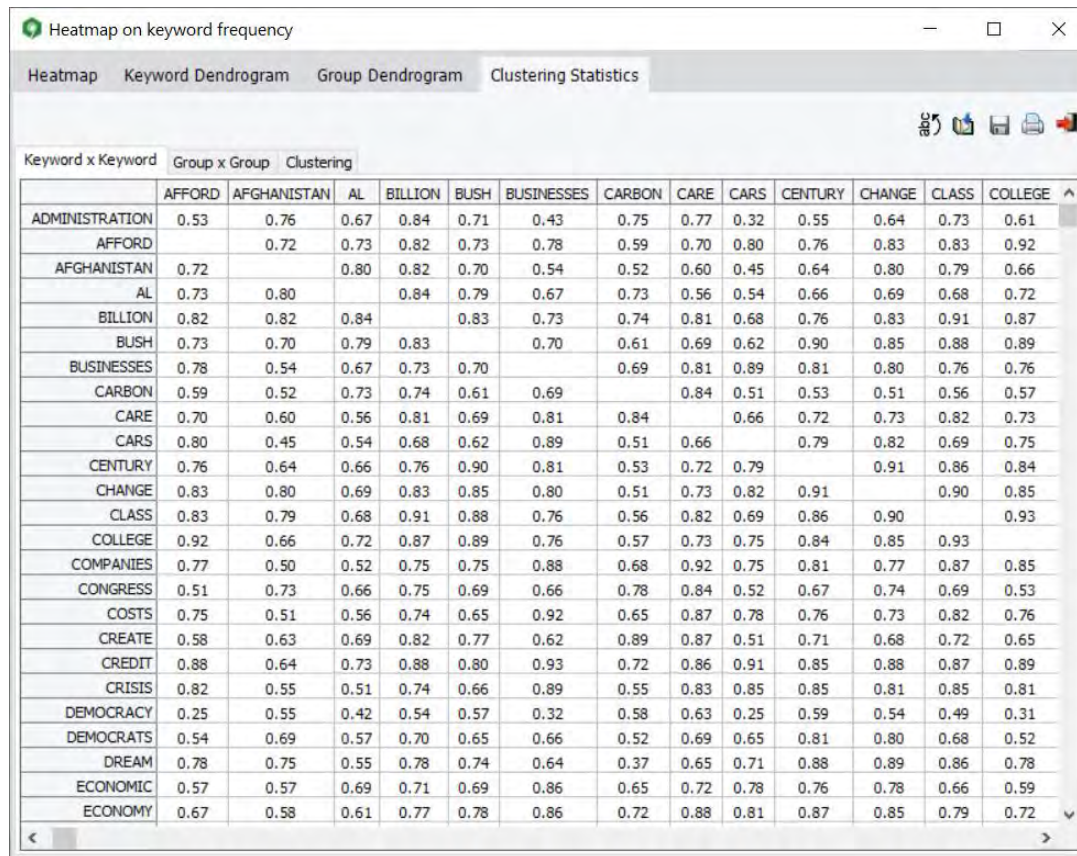
No clusters: This option sets how many clusters the clustering solution should have. Different colors are used in the dendrogram to indicate membership of specific items in different clusters.

Display: This option sets whether the vertical lines of the dendrogram represent the agglomeration schedule or the similarity indices.

At times, it is desirable to see some areas of a dendrogram. The  and  buttons may be used to zoom in or out of the dendrogram.

Clustering Statistics Tab

The fourth tab of the heatmap dialog box allows the close examination of the matrices of similarities among keywords or among groups as well as the statistics associated with the agglomeration processes of both the keywords and the groups.



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 you to graphically examine the relationship between keywords 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 ensures that the results remain easily interpretable. To further differentiate among some values of the independent variable, we recommend applying a filter to restrict the analysis to a subset of values.

The first two tabs of the dialog box, provides graphical displays of correspondence maps. When the number of words or categories or the number of subgroups is less than four, only the 2-D Map tab 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 reasons: plotting all the keywords on a single horizontal axis would have produced a cluttered list of keywords that would have made the graph useless. (The horizontal axis represents, by convention, the first extracted dimension).

The 2-D correspondence plot offers the ability to remove a keyword or a class of the independent variable and to automatically recompute the analysis on the remaining items. It also allows you to obtain a KWIC table or perform a keyword retrieval for a specific keyword. To perform any of these operations, simply click the desired item to display a pop-up menu and choose the proper command. If more than one item is close by the cursor, the menu will offer the possibility to choose on which keyword or class the operation should be performed.

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

2-D Map control

Plot - When more than two axes have been extracted, this control allows the selection of all the possible axis combinations that can be graphed on the two axes of the plot.

2-D and 3-D Map Controls

Keywords This checkbox displays or hides the row points (i.e., words or category names).

Groups This checkbox displays or hides the column points (i.e., subgroup labels)



Clicking this button enables you to zoom in a plot. To zoom an 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.



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



In 2D correspondence plots, there is no easy way to identify where items are located on the third dimension, while in 3D it remains difficult to position an item on the depth axis without applying a rotating motion to the chart or inserting anchor lines. When this button is pressed gradients of colors are applied on fonts in order to differentiate items on the front end of the depth axis from those in the back end and those in between. Up to four colors may be used to create those gradients. To adjust the color of those gradient, use the chart customization button (see below).



Moving this slider allows the adjustment of the scaling used to plot groups or classes of the categorical variable on the correspondence plot. While the positions of keywords and groups relative to the origin of the axes are significant, it is important to remember that the scaling used for keywords and groups are independent of each other. This slider may be used as a reminder of this fact. It may also be used to increase the readability of some plots, by moving group names so they do not overlap keyword names.



Press this button to append a copy of the correspondence plot in the Report Manager. A descriptive title will be provided automatically. To edit this title or to enter a new one, hold down the SHIFT keyboard key while clicking this button (for more information on the Report Manager, see the [Report Management Feature](#) topic).



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



Clicking this button brings up a dialog box to customize the appearance of the correspondence plots (click [here](#) to obtain more information on the various settings that may be changed).



Clicking this button prints a copy of the displayed chart.

3-D Plot controls



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



Clicking this button draws anchor lines from the floor to the data point to better locate data points in all 3 dimensions.



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



Locating a data point on the depth dimension of a 3-D plot can be very difficult, especially when the plot remains static. You often have to rotate this plot constantly on the various axes to get an accurate idea of where the data point is located on this third axis. Clicking this button forces WordStat to rotate the plot automatically. To disable the automatic rotation, click a second time.

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 the distribution of a keyword or a content category among subgroups is 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 far from each others. Again, it is important to remember 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.

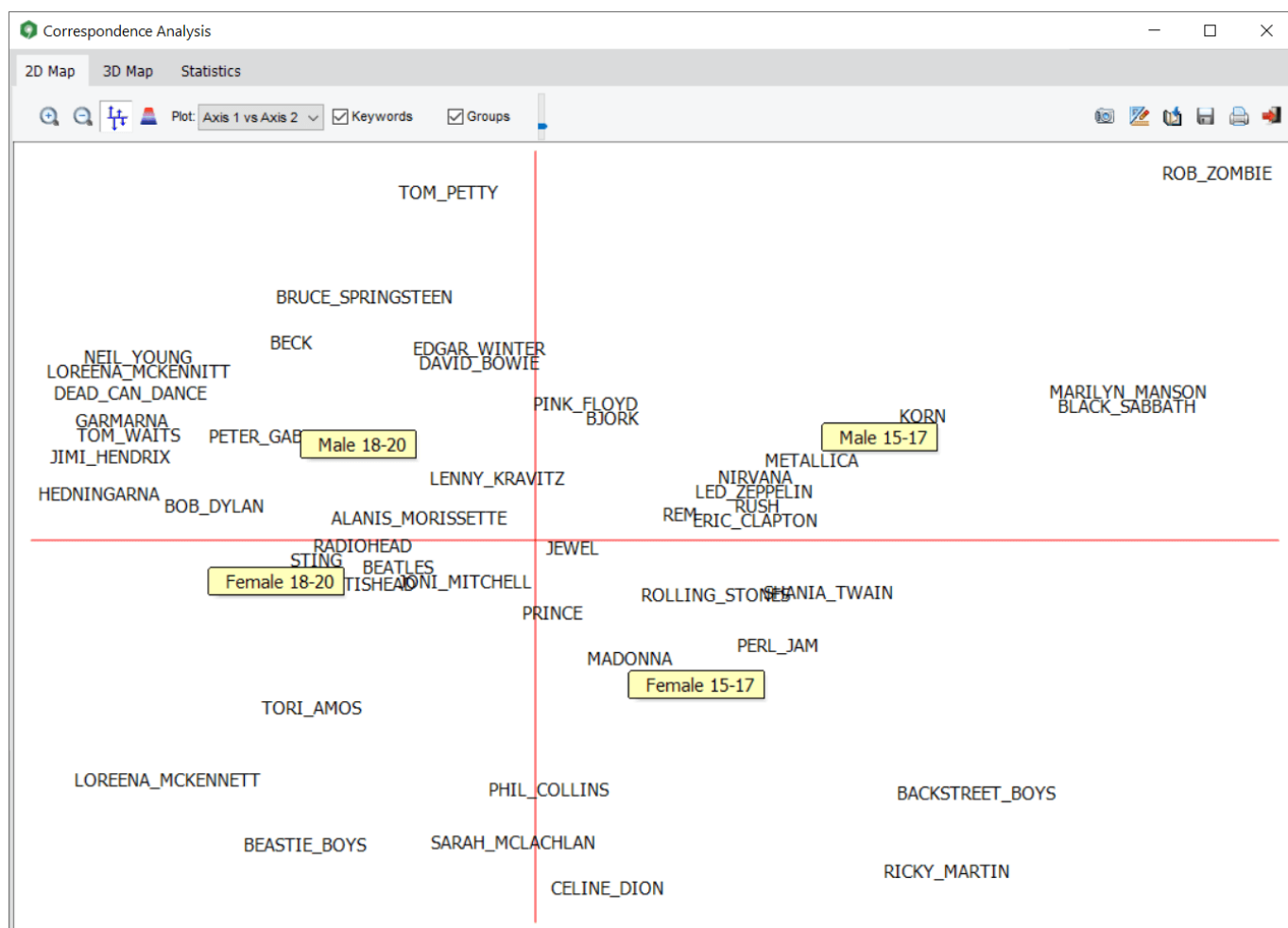
Relationship between words/categories and subgroups (row and column points):

- Great caution should be taken 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 keyword 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 keyword 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, Metallica and Korn could be viewed as characteristic of males between 15 and 17 years old. However, Rob Zombie is much more specific to those listeners since it is much farther from the origin. Marilyn Manson and Black Sabbath would come next, followed by Korn and then Metallica,



- Words or categories closely associated with two subgroups will be plotted at an angle from the origin that will lie between those two groups. In the above example, the Beastie Boys seem to be characteristics of both female age groups.

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 \(1998\)](#).

Keyword-In-Context Page

The **Keyword-In-Context** (KWIC) technique allows you to display in a table the occurrences of either a specific keyword, or of all items 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	few hours before...
He said that he would	kill	if I call the police.
Too much garlic	kills	taste of the meat.
The Black Death was a disease that	killed	millions of people.

My shoes are **killing** me.
The French skier Jean Claude **Killy** 3 gold medals.

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

#SATISFACTION near #TEACHER and not after #NEGATION

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

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

The KWIC technique is also useful to highlight syntactical or semantic differences in word usage between individuals or subgroups 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 **Keyword-In-Context** tab has been designed to facilitate the various tasks involved in content analysis. The tab looks like this:

WordStat 9.0.7 - Election 2008 Coded.ppt

Menu: Data, Text Processing, Frequencies, Extraction, Cooccurrences, Crosstab, **Keyword-In-Context**, Classification

List: Included Sort by: Keyword & After

Keyword: ADMINISTRATION Context delimiter: Paragraph

SHOW ALL ITEMS	FREQ	CASENO	KEYWORD	CANDIDATE	DELIVERY
care	1175	207	ke right away. My administration will create a Veterans'	McCain	Q3-2008
care for	115	206	make right away. My administration will create a Veterans'	McCain	Q3-2008
care and	109	207	est, most straightforward terms that the Veterans Health	McCain	Q3-2008
care of	68	207	anty or exclusively affect women. And here the Veterans	McCain	Q3-2008
care system	63	206	behind in the services it provides. And here the Veterans	McCain	Q3-2008
care costs	46	101	man, woman and child in America spends \$412 on health	Clinton	Q2-2007
care in	36	185	f children from online predators, in helping to make health	McCain	Q1-2008
care plan	35	41	i ill. She needs us to finally come together to make health	Obama	Q1-2008
care they	33	49	dn't. But they believe it's finally time that we make health	Obama	Q2-2008
care that	28	33	ca. Thank you. I'll be a President who finally makes health	Obama	Q1-2008
care to	25	46	etiree making less than \$50,000 per year. To make health	Obama	Q1-2008
care or	24	54	a wealthy we just can't afford. Rather than making health	Obama	Q2-2008
care is	18	57	er costs for ordinary Americans. We need to make health	Obama	Q2-2008
care about	16	57	in average family health care plan, and won't make health	Obama	Q2-2008
care reform	16	42	en invested in job training and child care; in making health	Obama	Q1-2008
care we	15	105	amatic savings which I will use to continue to make health	Clinton	Q3-2007
care the	15	101	study found this model could reduce the cost of diabetes	Clinton	Q2-2007
care crisis	13	81	y've been in a decade, at a time when the cost of health	Obama	Q4-2008
care i	12	73	ue that would make a difference in your lives - on health	Obama	Q3-2008
care coverage	12	109	nored to work throughout my career on issues like foster	Clinton	Q4-2007
care providers	12	176	ity, my Florida neighbors looked after my family with great	McCain	Q4-2007
care it	11	94	g aid from loans to grants. We will focus on primary health	Richardson	Q4-2007
care if	11	188	urse care managers to make sure they receive the proper	McCain	Q2-2008
care by	11	44	Senate Veterans Affairs Committee - working to improve	Obama	Q1-2008

245 cases Number of items: 1175

The upper-right part of the screen provides a list of all instances of keywords associated with a dictionary category or of a specific word or phrase, along with its surrounding text. The panel on the left shows a tree view of items and their context in descending order of frequency, which may be used to browse through and filter through the KWIC list on the right. The text panel at the bottom of the screen displays the full document from which the selected keyword comes from and highlights the selected keyword. The text panel can be used to examine the full context of a keyword, but may also be used to add words and phrases to the current categorization dictionary or to the exclusion list. To assign a word or a phrase to a list or content category, position the text cursor on the word you want to assign, or select one or several words with the mouse or and right-click to display a contextual menu. Select the **To Categorization Dictionary** or the **To Exclusion List** menu item.

List: This option allows for specifying whether the words for display in the KWIC table either should be selected from the list of included keywords or from the list of all remaining words that have not been explicitly excluded. The option **User Specified** allows you to enter a word or word pattern at the keyboard and search for all instances of this expression.

Keyword: This option allows for choosing among all words belonging to the list of **Included** keyword or **Leftover** words (see above). When the **List** option is set to **User Specified**, this option becomes an edit box where you can type a word or word pattern. (Wildcards such as ***** and **?** are supported as well as phrases).

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

- **Case number:** The KWIC table is sorted in ascending order of case 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 variables have been selected, the KWIC table includes a column indicating in which variable the keyword 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 alphabetical order. Lines with identical words are sorted on the variable name from which they come. This display is useful to examine whether specific words are used with the same meaning in different variables.
- **Variable & Keyword:** When several textual variables have been selected, the KWIC table includes a column indicating in which variable the keyword was found. This option displays a KWIC table where all lines are sorted on the variable name from which they originate. Lines extracted from a single variable are sorted by keywords. This display is useful to establish whether different variables contain different information. For a more detailed analysis of difference 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 words, use the **Keyword & VARNAME** sort order.

Context delimiter: This option allows selecting 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.

Once the settings have been set, click the  button to start searching all instances of the selected keyword.

Using the Tree List

The upper left panel displays a tree list view where each keyword is presented in descending order of frequency along with either its context or some associated variable. The content by which keywords are sorted depends on the sort setting, so that when the **Sort By** option is set to **Keyword & Before** or **Keyword and After**, this tree list will be sorted and broken down by up to five words of its surrounding text (similar to the screen shot above). When the **Sort By** option is set to the keyword and a variable, the tree list will be sorted on keywords first, and each keyword will be broken down by the values of the selected variable.

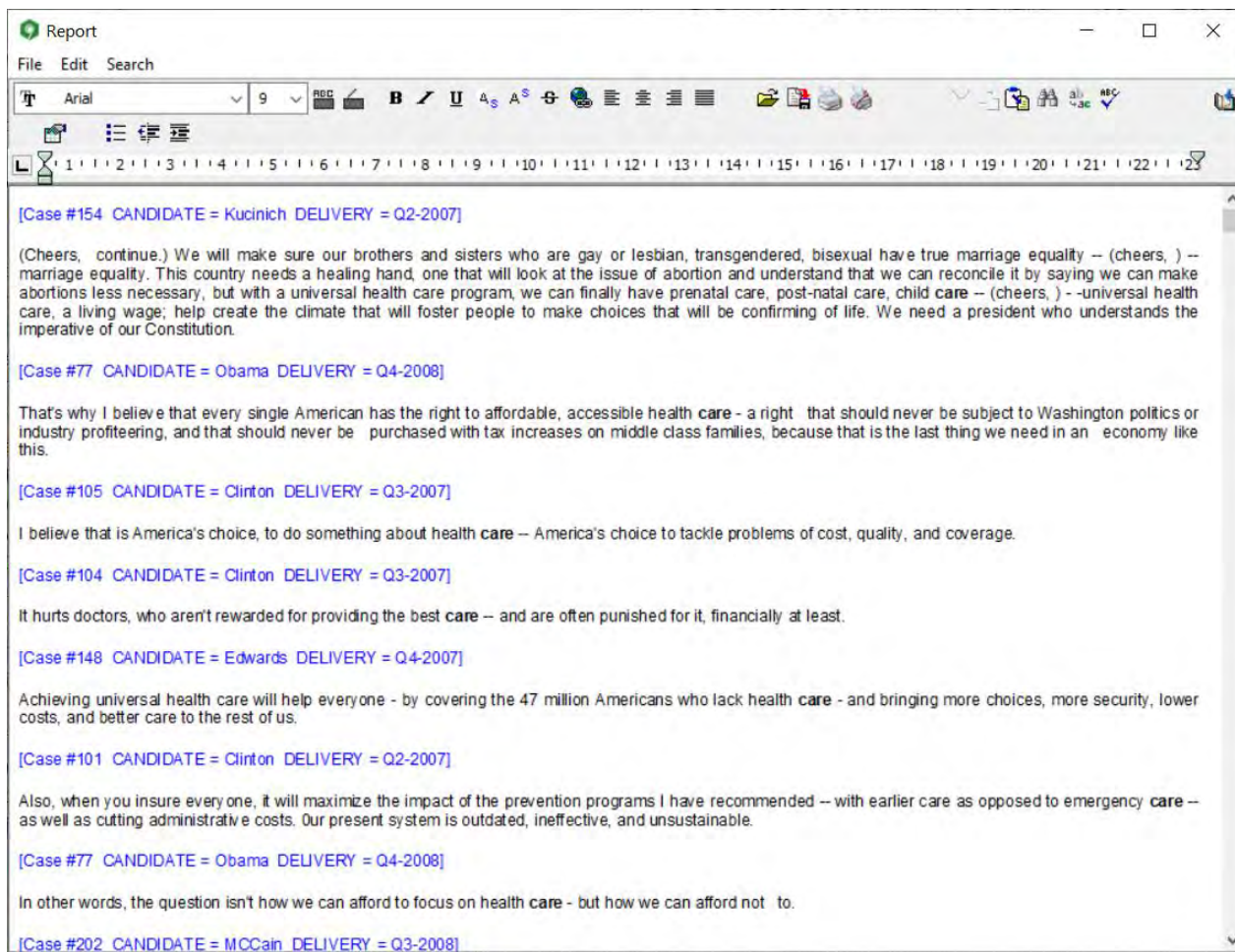
Selecting any item in this tree list will filter the keyword-in-context list on the right to display only the selected item. Moving to **SHOW ALL ITEMS** will remove any filtering conditions and will show all the hits.

By default, the tree list displays all items. You can filter this list and display items meeting a minimum frequency criterion. To adjust this criterion, right-click anywhere in this panel and select **SET MINIMUM**. A submenu allows you to increase the minimum frequency criterion, to reset it to 1, or to adjust to the current value the default minimum frequency criterion to be used for all other KWIC analysis.

Any KWIC table may be saved to disk in Excel, plain ASCII, text delimited, or HTML format by clicking the  button. To export the content of the table to a new Simstat data file, press the **Export** button located at the top of the Frequencies tab.

To print the KWIC table, click the  button.

Clicking the  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 box (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.



A word cloud along with a word frequency analysis may be useful to identify context words associated with the current target item displayed in the KWIC table. To perform such an analysis, click the  button and choose whether you want to analyze context words appearing **before only**, those appearing **after only** or any words appearing **before and after** the target item. The analysis will be performed on text currently displayed in the corresponding grid column(s) and will thus take into account the context delimiter option and any filtering option currently being applied using the tree list. (See [Word Frequency Analysis](#) for more information on this feature).

The Classification Tab

Performing Automated Text Classification

The automated text classification module allows you to apply a machine-learning approach to the existing textual database in order to develop a classification model that can later be used to accurately classify uncategorized documents into predefined classes.

What is Automatic Text Categorization?

Automated text classification is a supervised machine-learning task by which new documents are classified into one or several predefined category labels based on an inductive learning process performed on a set of previously classified documents. This machine-learning approach of classification has been known to achieve comparable if not superior accuracy than classification performed by human coders, yet at a very low cost in manpower. It has been used to automatically classify documents into proper categories or to find relevant keywords describing the content and nature of a document. It has also been used to automatically file or re-route documents or messages to their appropriate destinations, to classify newspaper articles into proper sections or conference papers into relevant sessions, to filter emails or documents (like spam filtering), or to route a specific request in an organization to the appropriate department. Automated text classification may also be used to identify the author of a document of unknown or disputed authorship. For a good overview of automated text classification, see [Sebastiani \(1999\)](#).

The automated text classification module in WordStat allows you to apply either Naive Bayes or K-Nearest Neighbors learning algorithms on an existing textual database in order to develop a classification model (or classifier). The program also provides features to test the accuracy of the classification and to optimize the various parameters. Once optimized, the obtained classification model may be used immediately to classify classification documents or may be saved on disk to be applied later outside WordStat using the [WordStat Document Classifier](#) utility program. The classifier may also be incorporated into a desktop or web application or within a document management system using the [WordStat Software Developer's Kit](#).

The development and application of a text classifier often involve the following steps:

1. **Removal** of function words, words that appear in only a few documents and words that appear too often.
2. **Dimension reduction**, through lemmatization, stemming, categorization, word clustering or other dimension-reduction techniques.
3. **Feature selection**, which consists of a selection of terms based on their capability to discriminate between categories of documents.
4. **Training** the classifier on the train set.
5. **Testing** the accuracy of the classification on a test set.
6. **Applying** the classifier to new documents.

While the basic content analysis features of WordStat may be used to deal with the first two steps, the Automated Text Classification dialog box allows you to accomplish tasks related to the last four steps. This dialog box consists of five tabs:

- The [Settings](#) tab is used to select the variable to predict and the validation method to use to test the accuracy of the classifier.
- The [Select Features](#) tab allows you to apply various feature selection methods to select a subset of terms to be used by the classifier.
- The [Learn & Test](#) tab is the location where machine-learning algorithms are set and tested. This tab also allows the storing of classification models to disk.

- The [History](#) tab keeps track of every learning test performed during a session allowing you to choose the best setting and algorithm for a specific classification task. It also gives access to a batch experiment dialog box that may be used to define numerous tests and perform them all at once.
- The [Apply](#) tab is used to apply a classifier to an external document, a list of documents or to the current data file.

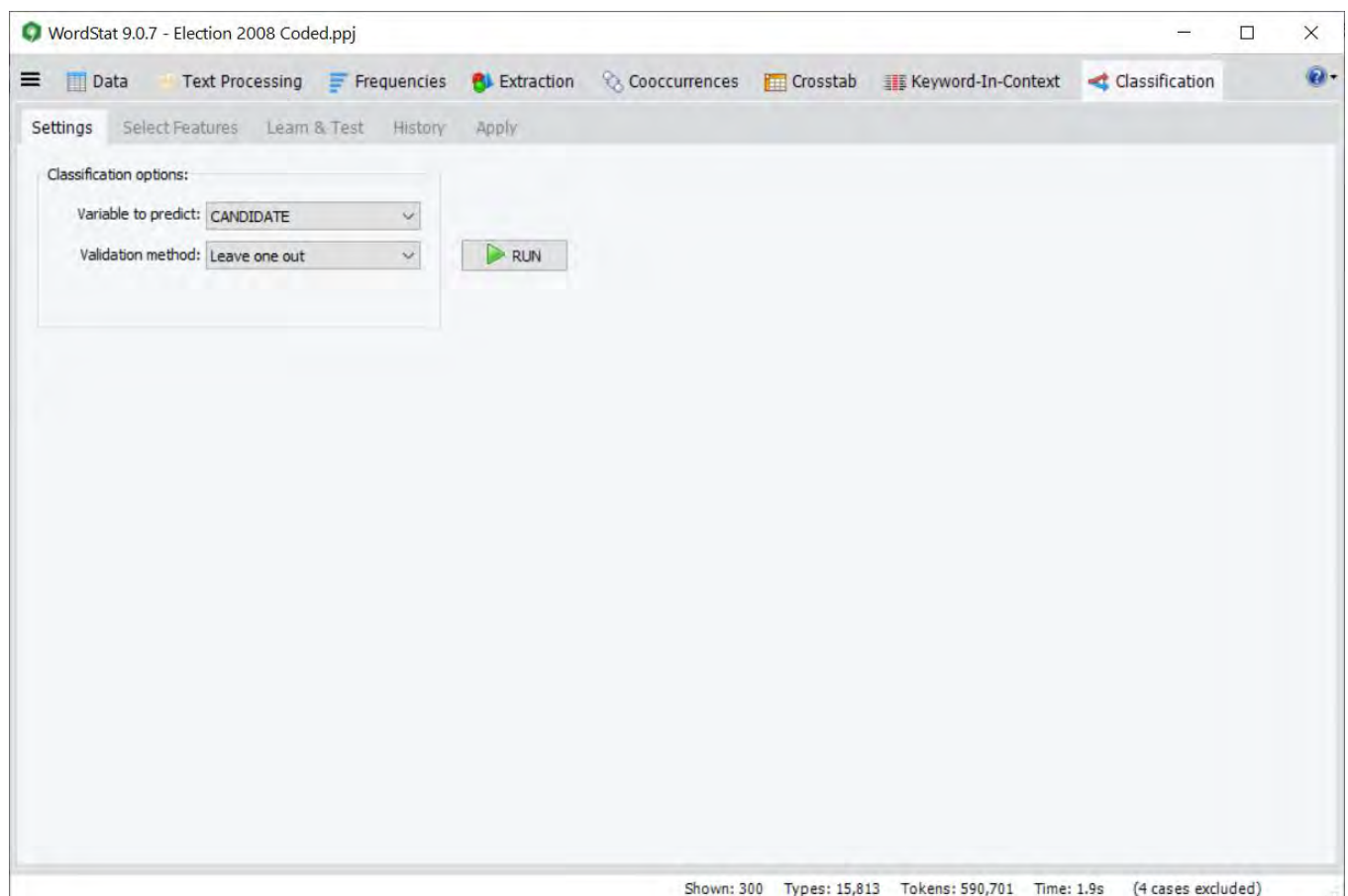
Accessing the Automated Text Classification feature

To develop a classifier for a specific categorical variable, you need to select a categorical variable containing the values you want to predict. It can be done in WordStat and from QDA Miner or SimStat while calling WordStat. In WordStat, when you select the **Analyze** button on the **Data** tab, on the left side of the dialog box are listed the categorical, numeric or date variables that you can analyze in relation to the text variables. Choose the variables and then move to the **Classification** tab. In QDA Miner, you have to choose this categorical variable in the **In Relation With** section. In SimStat, this variable should be assigned to the **Independent** list box, and the text variables on which the prediction should be based should be assigned to the **Dependent** list box. Once in WordStat, set the various text-processing options (such as the lemmatization, the exclusion and categorization lists, and all the other analysis options needed) to obtain the desired list of keywords or content categories. Then move to the **Classification** tab.

Settings

The **Settings** tab allows you to select which variable contains the values to predict and choose a validation method.

Selecting the Variable to Predict



To select the variable containing the values to predict, set the first list box to the name of this variable. If its name is not listed, choose the **<Select Variables>** item to display a list of all available variables, select the variable to predict and click **OK**. Then set the drop-down list to this newly added variable.

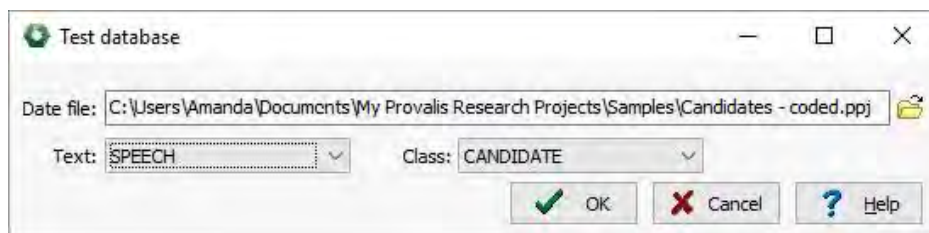
Selecting the Validation Method

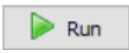
The evaluation of a classifier consists of measuring its effectiveness at classifying documents that have already been classified. Those documents should, however, not be part of the training set used to develop the classification model, since it would likely overestimate the real performance of the classifier. Yet, training a classifier on only a portion of the available training set may result in a less than optimal classifier. Cross-validation methods have been proposed as a compromise solution that allows you to develop a classification model on all the available documents in the training set yet provides a somewhat more realistic estimate of the classifier performance. WordStat offers three broad types of validation methods:

Leave-one-out - This cross-validation method consists of 1) removing a document from the training set, 2) developing a classification model on the remaining documents, 3) applying this model to predict the membership of this single document and 4) comparing the decision made by the classifier to the actual class to which this document belongs. This procedure is then repeated for each document in the training set and the different decisions are combined to estimate the performance of the classifier. While this method logically involves the computation of a large number of models and may seem to be time consuming, in practice the classification model is computed only once but adjusted analytically to remove the contribution of the test document prior to its classification. This cross-validation method will often overestimate the performance of a classifier if the training set includes duplicate documents or if included documents are not totally independent from one another.

n-folds - This method consists of splitting the training set into smaller partitions and testing each partition on the classification performance obtained by a model developed on the remaining ones. For example, when using a five-fold cross-validation method, the training set is divided randomly into five subsets, each containing approximately 20% of the documents. For each subset, the program tests the accuracy obtained by a classification model developed on the remaining 80% of the original training set. The performances obtained on all five classifiers are then used to estimate the performance of the classifier computed on the full training set. WordStat provides a choice between five-fold, 10-fold and 20-fold cross-validation.

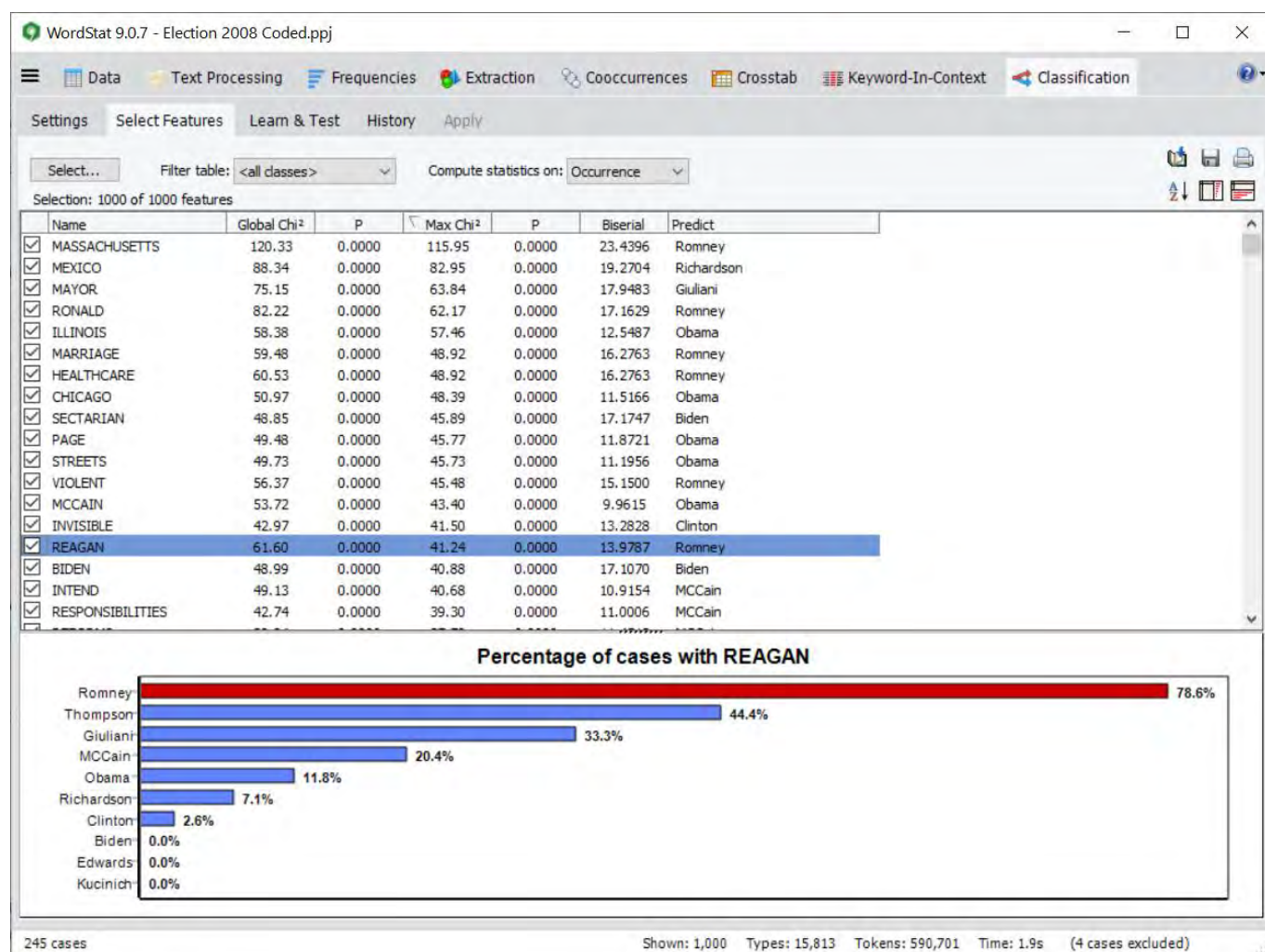
External file - A more conventional method for assessing the performance of a classifier is to test the accuracy of the classifier on an entirely different set of documents that have also been classified but are totally independent of the training set on which the categorization model is based. To perform such a test, WordStat requires the test set to be stored in a different data file. When this option is selected, an Open File dialog box is displayed allowing you to identify the file containing the external set. WordStat then displays a dialog box like the one below allowing you to choose the text variable containing the documents to be used for classification and the numerical variable containing the class to which this document belongs. Once set, click **OK** to return to the classification tab.



- Once the variable and the validation method have been set, click the  button to continue. WordStat will compute all statistics needed, and will then automatically move to the [Select Features](#) tab.

Select Features

The **Select Features** tab allows you to view the strength of the relationship between words, keywords or content categories and classes of the selected categorical variable. It also allows you to select a subset of items based on the statistics.



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

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

- Name** The word, keyword or content category.
- Global Chi²** The overall chi-square value computed on all classes of the categorical variable.
- P** The probability of the above chi-square value.
- Max Chi²** The chi-square value computed on the class with the highest case occurrence or frequency against all the other classes.
- P** The probability of the Max Chi² value.
- Biserial** The biserial correlation computed between the class of the categorical variables with the highest case occurrences and the remaining classes. This coefficient assumes that the presence or absence of a class is determined by a trait normally distributed. Contrary to the standard correlation coefficient, this measure of association may yield a value lower than -1.0 or higher than +1.0.

Predict Indicates the class in which the item most frequently occurs. When the highest case occurrence **or frequency** appears for more than one class, the column includes the labels of all those classes.

Clicking any column header sorts the table in ascending order of the data in that column. Clicking the same column header a second time sorts its content in descending order. The check boxes in the first column show, by default, that all items are to be included in the classification model. To manually remove an item, simply click in the box to remove the check mark.

The **Filter table** option allows you to display only the terms characteristic of a specific class. It may also be set to **<selected>** to display only items that have been selected for inclusion in the categorization model. To display all items set this option to **<all classes>**.

The lower portion of the tab displays a bar chart with the percentage of cases in each class of the categorical variable containing the selected item. Classes are presented in descending order of case occurrence. This graph is synchronized with the above table so that changing the selected item in this table results in the display of the corresponding distribution chart.

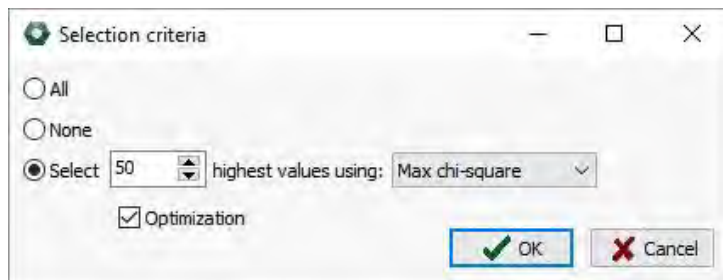
To display the chart on the right side of the table, click the  button. To bring the bar chart back to the bottom of the table, click the  button.

By default, bars in the chart are displayed in descending order of frequency. Moving from one feature to another will cause the order of the values to be rearranged according to the observed frequencies. To display bars associated with specific values at a fixed location, click the  button.

Performing Automatic Feature Selection

It has been shown that reducing the number of terms in classification models can reduce not only the cost of recognition by reducing the number of features that need to be collected, but also, sometimes, can improve the classification accuracy and can reduce the risk of overfitting. Several methods have been proposed to select a subset that yields the best performance for a specific classification system. WordStat allows you to apply a filter on available items in order to select a specified number based on the computed association with the classes to predict.

- To access the feature selection dialog box, click the  button. The following dialog box will appear:



To include all items, set the selection criterion to **All** and click **OK**. To remove all the check marks beside the items, set the selection criterion to **None** and click **OK**. When the **Select** criterion is chosen, specify how many items should be selected and which statistic will be used to select them. Three statistics are available for performing such a selection: The global chi-square value, the max chi-square and the bi-serial correlation (see above for a description of these statistics). By default, the selection is performed by extracting items with the highest values on the selected statistic. Enabling the **Optimization** option instructs the program to apply a special algorithm that will take into account this statistic as well as current contrasts between classes. This option has been found, in most situations, to improve the performance of the classifier for an equal number of selected features.

To close this dialog box and apply the selection criteria, click **OK**. You may also leave this dialog box without affecting the current selection by clicking the **Cancel** button.

Once features have been selected, move to the [Learn & Test](#) tab to select a learning algorithm and test it.

Learn & Test

The **Learn & Test** tab is where machine-learning algorithms are chosen and tested and from which the classifier may be exported to disk.

WordStat 9.0.7 - Election 2008 Coded.ppj *** FILTER: 152 of 245 cases ***

Settings Select Features Learn & Test History Apply

Learning:
Method: Naive Bayes Use: Keyword frequency
☒ Uniform Prior Feature weighting: Max-chi2 Run Publish

Confusion matrix Confusion List Review Errors

Correct = 135 Average precision = 0.9144 Nominal Accuracy = 0.9122 Ordinal Accuracy = 0.9574
Incorrect = 13 Average recall = 0.8817 F1 = 0.8977

Actual	Predicted						
Frequency Row Pct Col Pct Tot Pct	Biden	Clinton	Obama	Edwards	Richardson	TOTAL	PRECISION RECALL
Biden	10 76.92 100.00 6.76	0 0.00 0.00 0.00	3 23.08 4.23 2.03	0 0.00 0.00 0.00	0 0.00 0.00 0.00	13 8.78	1.0000 0.7692
Clinton	0 0.00 0.00 0.00	35 89.74 92.11 23.65	1 2.56 1.41 0.68	1 2.56 7.69 0.68	2 5.13 12.50 1.35	39 26.35	0.9211 0.8974
Obama	0 0.00 0.00 0.00	2 2.94 5.26 1.35	65 95.59 91.55 43.92	0 0.00 0.00 0.00	1 1.47 6.25 0.68	68 45.95	0.9155 0.9559
Edwards	0 0.00 0.00 0.00	0 0.00 0.00 0.00	2 14.29 2.82 1.35	12 85.71 92.31 8.11	0 0.00 0.00 0.00	14 9.46	0.9231 0.8571
Richardson	0 0.00 0.00 0.00	1 7.14 2.63 0.68	0 0.00 0.00 0.00	0 0.00 0.00 0.00	13 92.86 81.25 8.78	14 9.46	0.8125 0.9286
TOTAL	10 6.76	38 25.68	71 47.97	13 8.78	16 10.81	148 100.00	0.9144 0.8817

FILTER: 152 of 245 cases Shown: 2,592 Types: 13,089 Tokens: 398,486 Time: 1.4s (4 cases excluded)

The panel at the top of the tab allows you to select which machine-learning algorithm to use, set various analysis options and choose how the classification model will be tested.

Methods: Two machine-learning algorithms are available for text classification:

The **Naive Bayes** algorithm classifies text by estimating the probability of a class, given the presence or absence of specific words or keywords in the document to be classified. It first computes the probability of each term to occur in documents of specific classes in the training set. It then combines the probabilities associated with words found in the document to classify to estimate the probability that this document belongs to different classes. Finally, it assigns the document to the class with the highest probability. A multinomial Naive Bayes model has been chosen to handle both binomial and multinomial classification tasks as well as binary and numerical weights of items.

The **k-Nearest neighbor** classification method compares a document to be classified to all documents in the training set, retrieves the k most similar documents, and then assigns the new document to the most common classes in this retrieved set. This method is usually known to provide accurate classification when the training set is large enough, yet it can be very time-consuming because of the need to compare and rank the entire training set for similarity with the test document. It also usually requires a larger storage space since it must keep frequency information for all documents in the training set rather than just a few classification rules or mathematical formulas like many other machine-learning methods. However, WordStat uses a very efficient K-NN algorithm that drastically improves the computing speed and reduces

the disk space and memory requirement. When this method is chosen, a **NO** edit box appears below the Method list box, allowing you to set the number of similar documents on which this classification will be based. Values higher than 20 or 30 are typically used in text classification tasks.

Use: This option is used to select the item statistic to be used in training and classification. Choosing **Case Occurrence** results in the use of binary weights, indicating whether or not a word or keyword occurs in the document. Selecting **Keyword Frequency** allows you to use additional information related to how often this item occurs in each document. **Percentage of Words** and **Percentage of Keywords** provide two methods to normalize the obtained frequency to take into account the document length. Such normalization is performed by dividing the frequency either by the total number of words found in the document or the total number of keywords that has been extracted by WordStat.

Feature weighting: Feature weighting has been presented as an alternative to feature selection or as a way to further improve classification accuracy from selected item sets. This method consists of giving more weight to items that are rather good at differentiating documents from distinct classes and negligible weight to those that are distributed evenly among classes. The most frequently used weight in information retrieval is the TF*IDF measure where the frequency of an item is adjusted to take into account the number of documents containing this item. However, such a weighting can be considered to be only a crude approximation of the capacity of the item to differentiate documents from distinct classes. More accurate performance of the classifier can be expected from using a weight based on a more direct indicator of this discriminative capability such as the **Global Chi-Square** or the **Max Chi²** described previously.

- Once the analysis options have been set, click the  button to perform the training and test the performance of the obtained classifier.

Results

A common way of assessing the accuracy of a classifier is by comparing the accuracy of predicted class membership against actual membership. Such information is provided by the **Confusion Matrix** where each predicted class is plotted against the actual class. Accurate predictions are plotted in the diagonal going from the top left to the bottom right of the table. Values in this diagonal are printed in bold characters for easy identification. Values in cells below or above this diagonal represent classification errors. Besides the actual number of documents in each cell, the table shows the row, column and total percentages. Row percentages represent the number of documents in a class that have been classified in a specific way, while column percentages express the percentage of a specific prediction actually belonging to a known class. This table may be used to identify which classes are the easiest or hardest to predict, as well as which classification errors are the most common. To facilitate comparisons across the classes of the categorical variable, two related statistics are printed on the right of the table: **Precision** is the probability that documents identified as belonging to a class are correctly classified and **Recall** is the probability of documents in a class to be correctly identified.

Several statistics are provided to assess the global performance of the classifier. The **Nominal Accuracy** measure is the proportion of documents correctly classified. It is considered a micro-average statistic since it gives equal weight to documents regardless of how they are distributed among classes of the categorical variable. The **Average Precision** and **Average Recall** measures are macro-average statistics obtained by computing the mean precision and recall obtained for every class. The **Ordinal Accuracy** measure weights disagreements so that errors in prediction will be considered higher when the predicted value is far from the original value, while predictions that are closer to the original value will be counted as partial disagreements.

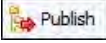
The **Confusion List** tab presents information already found in the confusion matrix but in the form of a single list that allows you to identify more easily the most common errors. The table may be sorted on the actual class of the documents, the predicted classification, the number of times such a classification error occurred, or the proportion of documents that have been misclassified this specific way. By default, the table is sorted in descending order of frequency. To sort the table on values in another column, simply click this column header. Clicking the same column header a second time sorts its content in descending order.

The **Review Errors** window displays a list of all documents that have been misclassified, allowing you to examine for each document the classification error made by the classifier as well as the computed values associated with every class of the categorical variable. A text window in the bottom of the list also allows you to review the text on which the classification has been made and potentially identify some of the reasons why the document had been misclassified.

Exporting a Classifier to Disk

Once developed and optimized, a document classification model may be saved to disk and later be used outside the WordStat main program to categorize new documents. The saved categorization model includes all word exclusion, extraction and categorization settings needed to accurately retrieve items used in the classification model as well as either the classification rules (Naive Bayes) or the keyword indexing of documents in the training set (K-nearest neighbors) needed to perform document classification.

To save the classifier on disk:

- Click the  button. A standard Save File dialog box will appear.
- Enter the file name under which you would like to save the classifier and then click the **Save** button. WordStat will automatically provide a .wclas file extension.

Document classifiers may be retrieved and applied to new documents using the [WordStat Document Classifier](#) utility program. A special [Software Developer's Kit](#) (SDK) is also available upon request from Provalis Research allowing any programmer to integrate WordStat categorization and classification technologies into one's own database or document management system.

History

There is, unfortunately, no rule or rationale to help you chose a classifier and its options and parameters, and to indicate how many or which items should be selected for inclusion in the development of a classification model. For this reason, the selection of classifiers and their optimization must rely on experimentation. In other words, in order to obtain the best classifier, you needs to perform numerous tests involving systematic variations of the various settings and compare the results. The **History** tab offers a way to keep track of previously performed trials, the options used, as well as the obtained performances. From this tab, you can also access an **Experiment** dialog box that provides a convenient way to define and run a large number of classification experiments on the same data set.

WordStat 9.0.7 - Election 2008 Coded.ppj *** FILTER: 152 of 245 cases ***

Settings Select Features Learn & Test History Apply

Experiment Class: <all classes>

Table Graph

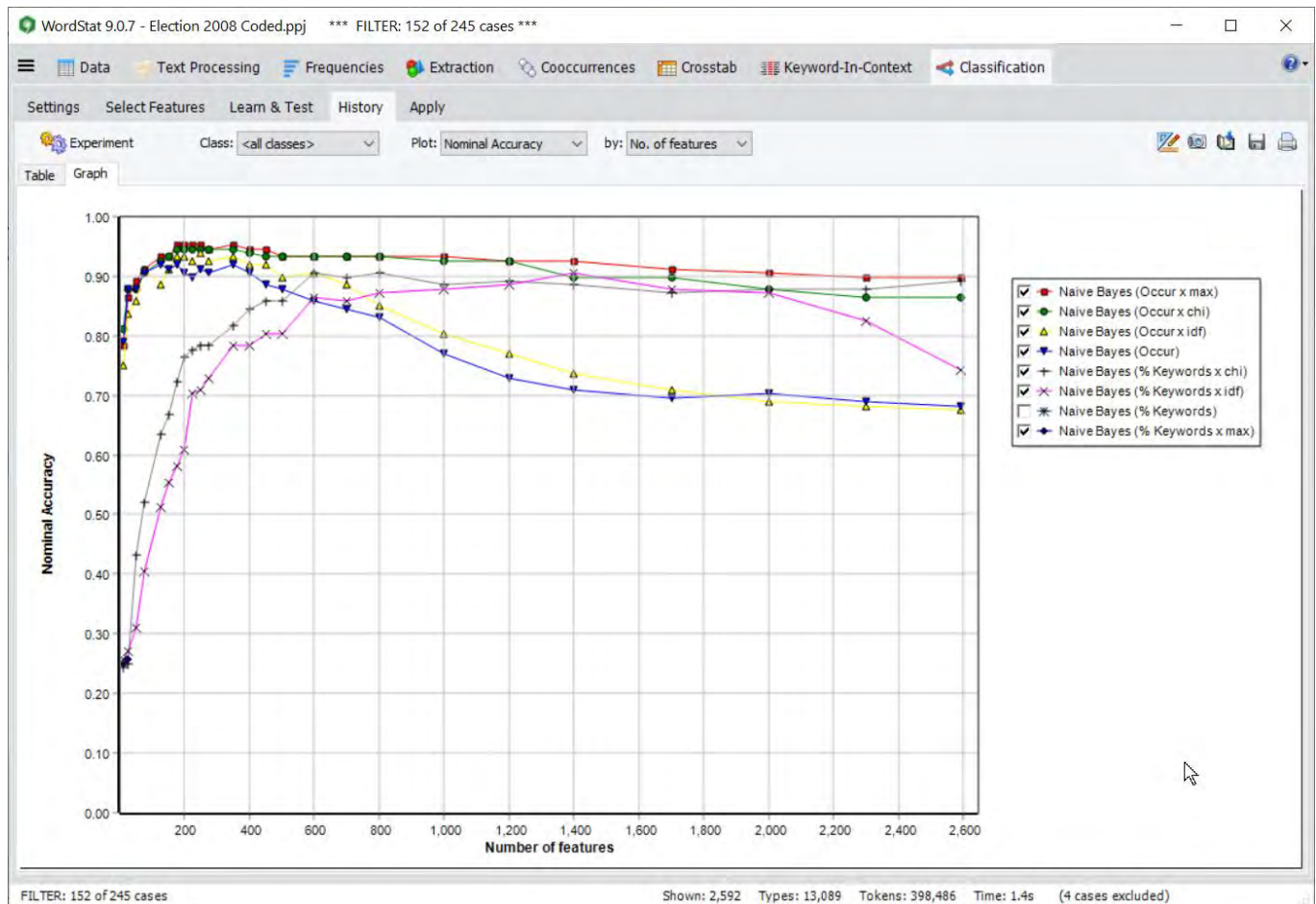
Method	Parameters	No. of	Selection	Validation	Nominal Accuracy	Ordinal Accuracy	Precisio	Recal	F1
Naive	Occur x max	2592	Chi ² -O opt	Leave one	0.8986	0.9568	0.9179	0.840	0.877
Naive	% Keywords x chi	2592	Chi ² -O opt	Leave one	0.8919	0.9583	0.8850	0.818	0.850
Naive	Occur x chi	2592	Chi ² -O opt	Leave one	0.8649	0.9377	0.8999	0.771	0.830
Naive	% Keywords x idf	2592	Chi ² -O opt	Leave one	0.7432	0.8627	0.6849	0.624	0.653
Naive	Occur	2592	Chi ² -O opt	Leave one	0.6824	0.8364	0.4805	0.404	0.439
Naive	Occur x idf	2592	Chi ² -O opt	Leave one	0.6757	0.8368	0.4763	0.401	0.435
Naive	Occur x max	2300	Chi ² -O opt	Leave one	0.8986	0.9568	0.9179	0.840	0.877
Naive	% Keywords x chi	2300	Chi ² -O opt	Leave one	0.8784	0.9524	0.8564	0.812	0.833
Naive	Occur x chi	2300	Chi ² -O opt	Leave one	0.8649	0.9377	0.8999	0.771	0.830
Naive	% Keywords x idf	2300	Chi ² -O opt	Leave one	0.8243	0.9071	0.7720	0.761	0.766
Naive	Occur	2300	Chi ² -O opt	Leave one	0.6892	0.8412	0.4800	0.421	0.449
Naive	Occur x idf	2300	Chi ² -O opt	Leave one	0.6824	0.8417	0.4776	0.416	0.445
Naive	Occur x max	2000	Chi ² -O opt	Leave one	0.9054	0.9589	0.9220	0.854	0.887
Naive	% Keywords x chi	2000	Chi ² -O opt	Leave one	0.8784	0.9524	0.8564	0.812	0.833
Naive	Occur x chi	2000	Chi ² -O opt	Leave one	0.8784	0.9489	0.9112	0.800	0.852
Naive	% Keywords x idf	2000	Chi ² -O opt	Leave one	0.8716	0.9416	0.8221	0.844	0.833
Naive	Occur	2000	Chi ² -O opt	Leave one	0.7027	0.8490	0.4855	0.440	0.461
Naive	Occur x idf	2000	Chi ² -O opt	Leave one	0.6892	0.8447	0.4768	0.432	0.453
Naive	Occur x max	1700	Chi ² -O opt	Leave one	0.9122	0.9626	0.9288	0.869	0.898
Naive	Occur x chi	1700	Chi ² -O opt	Leave one	0.8986	0.9586	0.9255	0.842	0.882
Naive	% Keywords x idf	1700	Chi ² -O opt	Leave one	0.8784	0.9455	0.8306	0.870	0.850

FILTER: 152 of 245 cases Shown: 2,592 Types: 13,089 Tokens: 398,486 Time: 1.4s (4 cases excluded)

Data from prior trials are presented either in the form of a table (**Table** tab) or as a line chart (**Graph** tab). Clicking any column header of the table sorts it in ascending order of the data in this column. Clicking the same column header a second time sorts its content in descending order. By default, the displayed statistics are computed for all classes of the categorical variable. To restrict the display of either the table or the line chart to statistics related to a single class, set the **Class** list box to the desired class. Setting this option to **<all classes>** brings back the micro- and macro-average statistics computed on all classes.

Data from specific trials may be deleted from the table by selecting their rows and pressing the  button.

The **Graph** tab allows you to compare the performance of various settings and the relationship between those settings using a line chart like the one shown below.



By default, the chart displays the relationship between accuracy and the number of features in the classification model. You may, however, examine the relationship between other parameters (such as precision versus recall) by changing the information plotted either on the horizontal or the vertical axis of the chart.

The following table provides a short description of buttons available:

Control	Description
	Press this button to save either the table or the chart on disk. The table may be saved to disk in Excel, plain ASCII, text delimited, or HTML Charts may be saved in BMP, JPG or PNG graphic file format or may be stored on disk in a proprietary format (.WSX file extension) that may later be edited and customized using the Chart Editor.
	Pressing this button prints a copy of the displayed table or chart.
	This button allows the editing of various features of the chart such as the left and bottom axis, the chart and axis titles, the location of the legend, etc.
	This button is used to create a copy of the chart to the clipboard. When this button is clicked, a pop-up menu appears allowing you to select whether the chart should be copied as a bitmap or as a metafile.

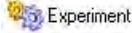
The History tab also gives access to the Experiment feature where you can quickly perform a series of classification experiments.

- To access this feature click the  Experiment button. For more information on this feature, see [Classification Experiment Dialog Box](#)

Classification Experiment Dialog Box

The Classification Experiment feature allows you to quickly perform a series of classification experiments in order to choose among classification methods, analysis settings and selected sets of items (or features).

To access this dialog box:

- Click the  Experiment button. A dialog box similar to the one below will appear:

Experiment

Feature selection

Select: 50 100 150 200 300

Statistic: Max chi-square Based on: Pct of words Optimization

Learning: Method: Naive Bayes Use: Percentage of words Feature weighting: x IDF

Testing: Method: Leave one out

Run

Method	Parameters	No. of Features	Selection method	Testing method	Executed
Naive Bayes	Case occurrence x IDF	50 100 150 200 300	Max Chi ² -O (opt)	Leave one out	Yes
Naive Bayes	Keyword frequency	50 100 150 200 300	Max Chi ² -O (opt)	Leave one out	Yes
Naive Bayes	Keyword frequency x IDF	50 100 150 200 300	Max Chi ² -O (opt)	Leave one out	Yes
Naive Bayes	Percentage of words	50 100 150 200 300	Max Chi ² -O (opt)	Leave one out	Yes
Naive Bayes	Percentage of words x IDF	50 100 150 200 300	Max Chi ² -O (opt)	Leave one out	Yes
Naive Bayes	Case occurrence	50 100 150 200 300	Max Chi ² -O	Leave one out	Yes
Naive Bayes	Case occurrence x IDF	50 100 150 200 300	Max Chi ² -O	Leave one out	Yes
Naive Bayes	Keyword frequency	50 100 150 200 300	Max Chi ² -O	Leave one out	
Naive Bayes	Keyword frequency x IDF	50 100 150 200 300	Max Chi ² -O	Leave one out	

The basic principle of this dialog box is to create a list of classification experiments involving different settings and then to instruct WordStat to perform all those experiments one after the other. Experiments first need to be defined in the upper part of the dialog box and then moved to the table of experiments located at the bottom of the dialog box. After several experiments have been defined and added to the list, you can execute all of them at once.

The first steps involve setting the experiment options. The **Feature Selection** option located at the top of the dialog box provides automatic feature selection options similar to those available from the first tab of the document classification dialog box with only one exception: while the original dialog box allows you to select only one feature set size at a time, the current dialog box allows you to set numerous feature set sizes at once. For example, by entering the following string in the **Select** edit box: 50 100 150 200 300

Five classification experiments will be performed using the same analysis settings but on features set sizes of 50, 100, 150, 200 and 300, picked out using the chosen selection method. For more information on the **Statistics**, **Base On** and **Optimization** options, see [Performing Automatic Feature Selection](#).

The **Learning** and **Testing** groups of options also provide similar settings as those available on the [Learn & Test tab](#). Please refer to this section for information on the available algorithms and options.

To add a classification experiment to the list:

- Set the various classification experiment settings.

- Click the  button to move the defined experiment to the list.
- Clicking the  button displays a dialog box that allows you to quickly define numerous variations of the current classifier and to add all those at once to the list of experiments to be performed.

To remove an experiment from the list:

- In the list of previously added experiments, select the one to delete.
- Click the  button.

To run experiments in the list:

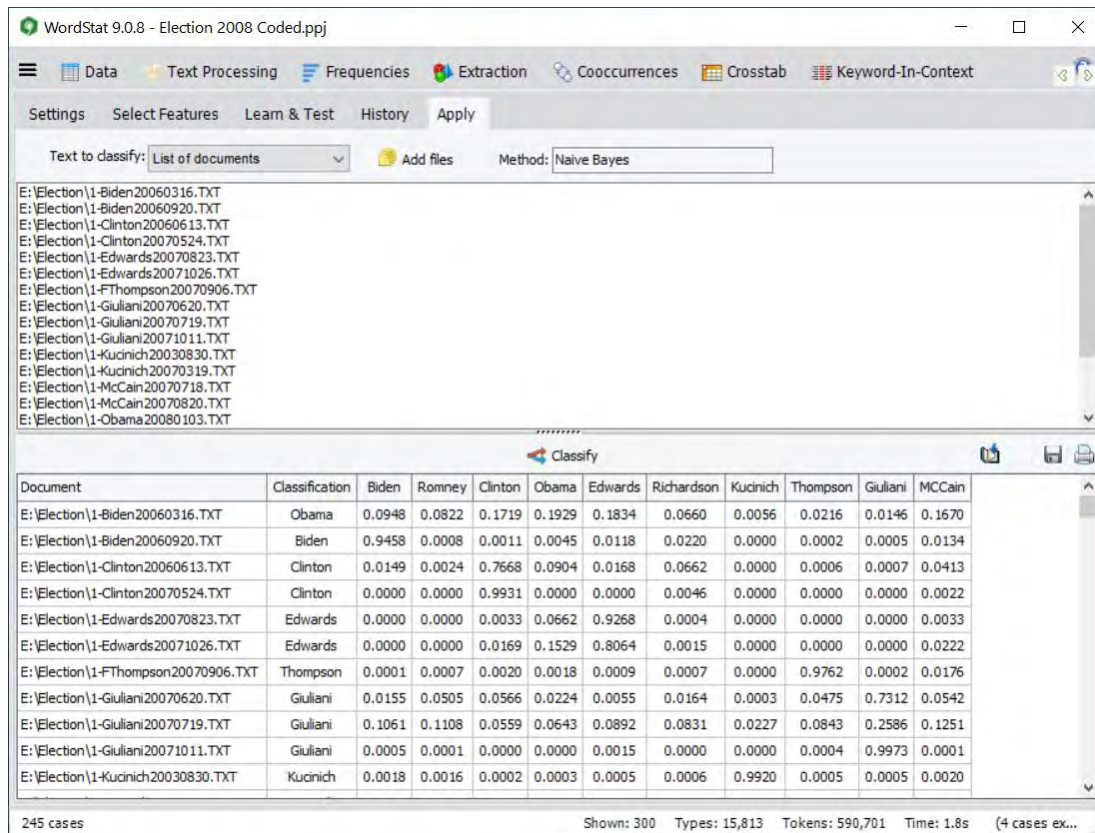
- Click the  button. The program performs all the experiments in the list that had not been executed before. Once an experiment is completed, the **Executed** column for this item is set to **Yes** preventing the program from executing the same experiments twice. Results of every experiment are automatically appended to the History tab.

To exit:

- Click the  button to close the dialog box and return to the History tab of the classification dialog box.

Apply

The **Apply** tab allows you to use the most recently tested classifier to categorize either a single document, a list of files or documents stored in the current data file or another Simstat/QDA Miner data file. To perform similar tasks using previously saved classification models, use the [WordStat Document Classifier](#) utility program.



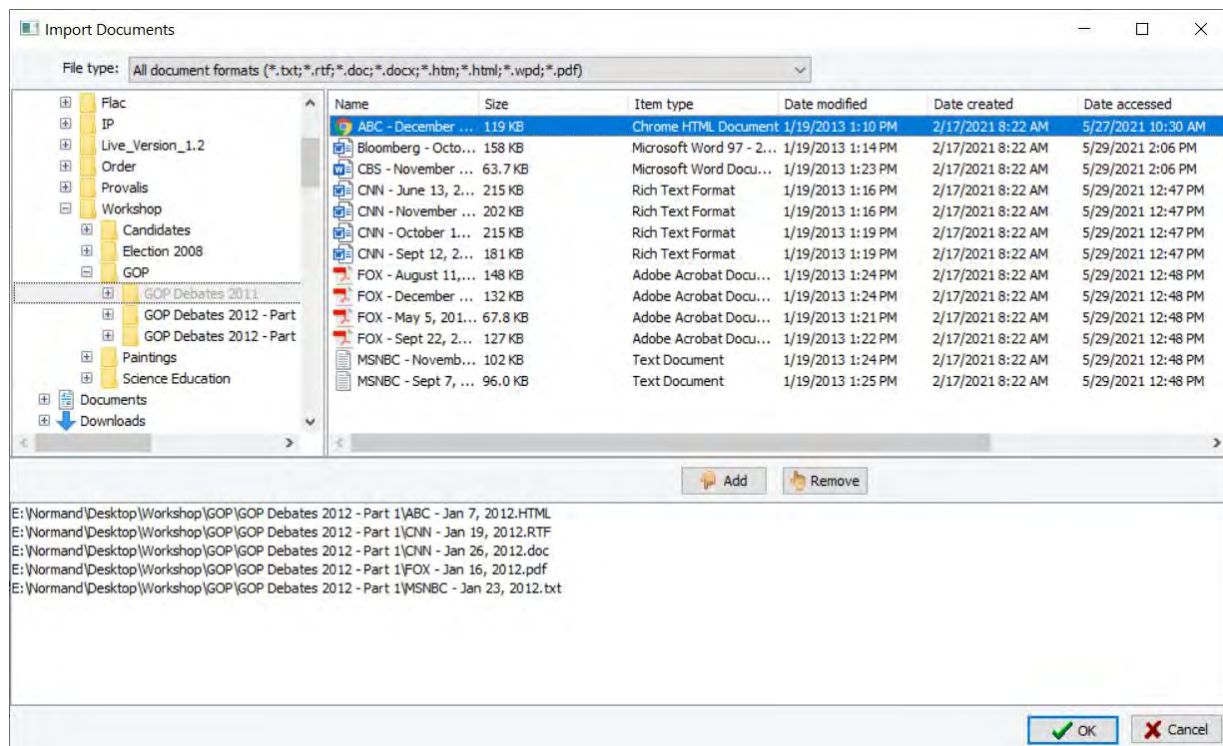
The document classification feature supports numerous file formats such as plain ASCII text files as well as HTML, Rich Text, MS Word, WordPerfect, Acrobat PDF files. Detailed results of classifications are displayed in a table at the bottom of the dialog box and may be either saved to disk or printed. When applied to the current database or another database, the automatic classification feature may be useful to categorize unclassified documents or to review existing classifications based on the results of the new classifier.

To classify a single document:

- Set the **Text To Classify** list box to **Single Document**.
- Click the **Open File** button to locate and import the file containing the text to be classified. You may also type directly in the text-editing window or paste a text previously copied to the clipboard (by moving to the text-editing window and pressing **Ctrl-V**).
- Click the **Classify** button to apply the current classifier to the displayed text.

To classify a list of documents:

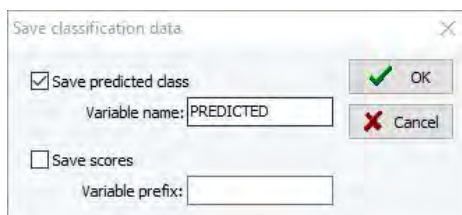
- Set the **Text To Classify** list box to **List of Documents**.
- Click the **Edit List** button to display a dialog box like the one below that allows you to browse through your computer and select certain documents. You may add documents located in different folders by successively adding documents located in a specific folder and then moving to a new location where other documents are located. Click **OK** to confirm the changes to the file list.



- Click the **Classify** button to apply the current classifier to all documents in the list.

To classify documents in the current data file:

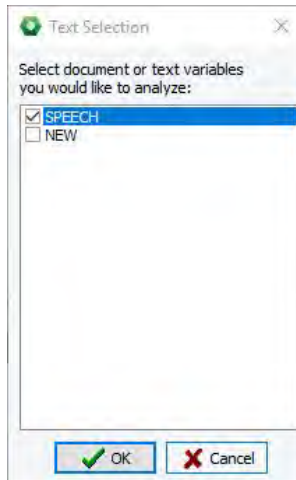
- Set the **Text To Classify** list box to **Current Data File**.
- Click the **Classify** button to apply the current classifier to all documents contained in the selected text variables. Please note that these are the same text variables used for developing the current classifier. However, some documents may have been ignored during the classification phase, either because they had been filtered out or because they had not been classified before and contained missing values in the categorical variable. These documents, as well as all other previously classified documents, will be categorized by the current classifier and the result of this classification will be displayed in a result table.
- To store the predicted class or the computed score obtained for every class, click the  button. The following dialog box will appear:



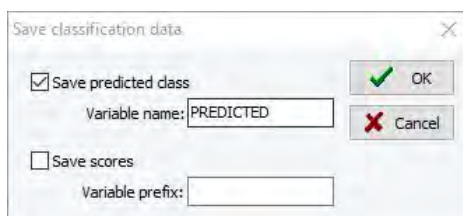
- To save the predicted class put a check mark beside **Save Predicted Class** and enter a variable name.
- To save the scores associated with each class and upon which the classification has been made, put a check mark beside **Save Scores** and enter a variable prefix (up to 7 characters). Variable names are created by adding successive numeric values to this prefix. For example, if the edit box at the right of the Variable Prefix option is set to "CLASS_", the variable names will be CLASS_1, CLASS_2, CLASS_3, etc.
- If any one of the specified variables does not exist, WordStat will create new ones and store the numerical values associated with either the predicted class or the class scores. A confirmation dialog box will ask for confirmation of the creation of these new variables as well as the overwriting of any existing ones.

To classify documents in another data file:

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



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



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

To export the table to disk:

- Click the  button. A Save File dialog box will appear.

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

Miscellaneous

Preparing and Importing Data

This section provides general information on how to prepare textual data for analysis in WordStat and specific instructions on how to import data into QDA Miner and SimStat.

Preliminary Text Preparation

While interview transcripts, responses to open-ended questions, or any other kind of textual information may be typed directly within SimStat or QDA Miner, there are many situations where electronic versions already exist either in the form of text files or as data files accessible only through specific applications such as word processor, spreadsheet or database programs. This information must be transferred into a QDA Miner project or Simstat data file for further processing. Projects can also be created directly in WordStat from documents and spreadsheets etc. Prior to using WordStat for content analysis, some modification or adjustments may need to be made.

Uppercase and lowercase letters

By default, WordStat is case-insensitive and therefore accepts files in either upper-or-lowercase.

Check spelling of documents

The automatic content analysis feature of WordStat involves 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. Two strategies may be used to deal with misspellings:

1. One may run documents through a spell-checker to make sure all words are spelled correctly. WordStat provides spell-checking for more than 20 languages. The spell-checking may be performed in QDA Miner or through [Text Editor](#) feature of WordStat. An even more efficient approach is to use the [Misspellings and Unknown Words](#) feature of WordStat to quickly retrieve all potentially misspelled words and to replace them all at the same time.
2. An alternative approach would be to build a content analysis process that would take into account the misspelling of words. To achieve this, one may use the [Substitution](#) feature to automatically replace misspelled words with their correct forms or add the most commonly misspelled keywords to the content-analysis dictionary.

Remove hyphenation

While WordStat can be configured to accept compound words with dashes, it cannot differentiate between dashes and hyphens. As a consequence, a hyphenated word will often be treated as two separate words. It is recommended to revise the text to ensure no hyphenation is present.

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

Square brackets and braces have special meanings for WordStat. For example, braces are often 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 text. If these symbols are used in a text for other purposes, they should be replaced with other symbols.

If there are specific parts of your text that you do not want to process, such as 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. QDA Miner's coding feature may also be used to restrict the analysis to some sections or exclude specific text segments from the content analysis process. Once the text segments have been manually tagged in QDA Miner, one could then specify, when

calling WordStat, to ignore sections tagged with specific codes or to only analyze segments associated with one or several codes.

Importing Spreadsheet or Database Files

Spreadsheet Data Files

Most spreadsheet programs allow for entry of both numeric and alphanumeric data into cells of a data grid. QDA Miner, SimStat and WordStat can import spreadsheet files produced by EXCEL (*.xls; *.xlsx).

To import an Excel spreadsheet from QDA Miner:

- Choose the **NEW** command from the **PROJECT** menu.
- Click the Import from an **Data File** button.
- Select the file format using the **List File of Type** drop down list.
- Select the file you want to import and click the **OK** button.

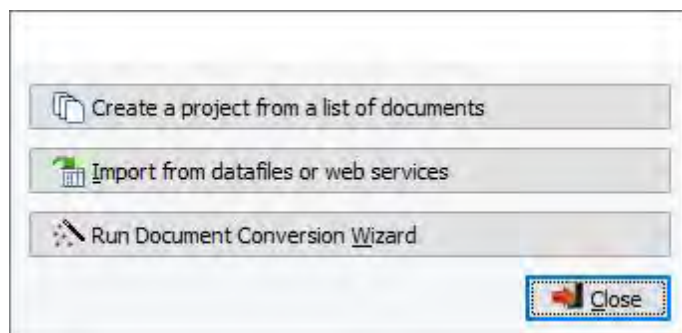
To import an Excel spreadsheet from Simstat:

- 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 the **OK** button.

The program displays a dialog box where you can specify the spreadsheet tab 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 tab.

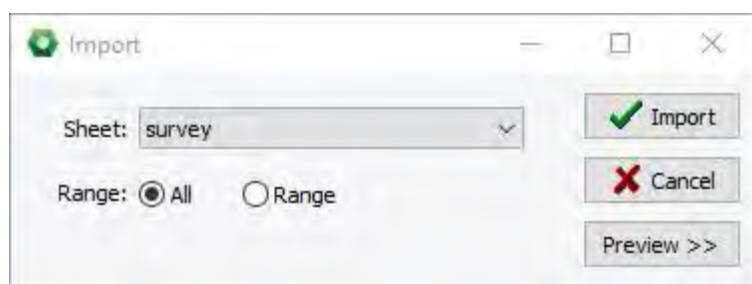
To create project from a Excel spreadsheet in WordStat:

- Select the **New button**. A dialog box similar to the one below will appear.



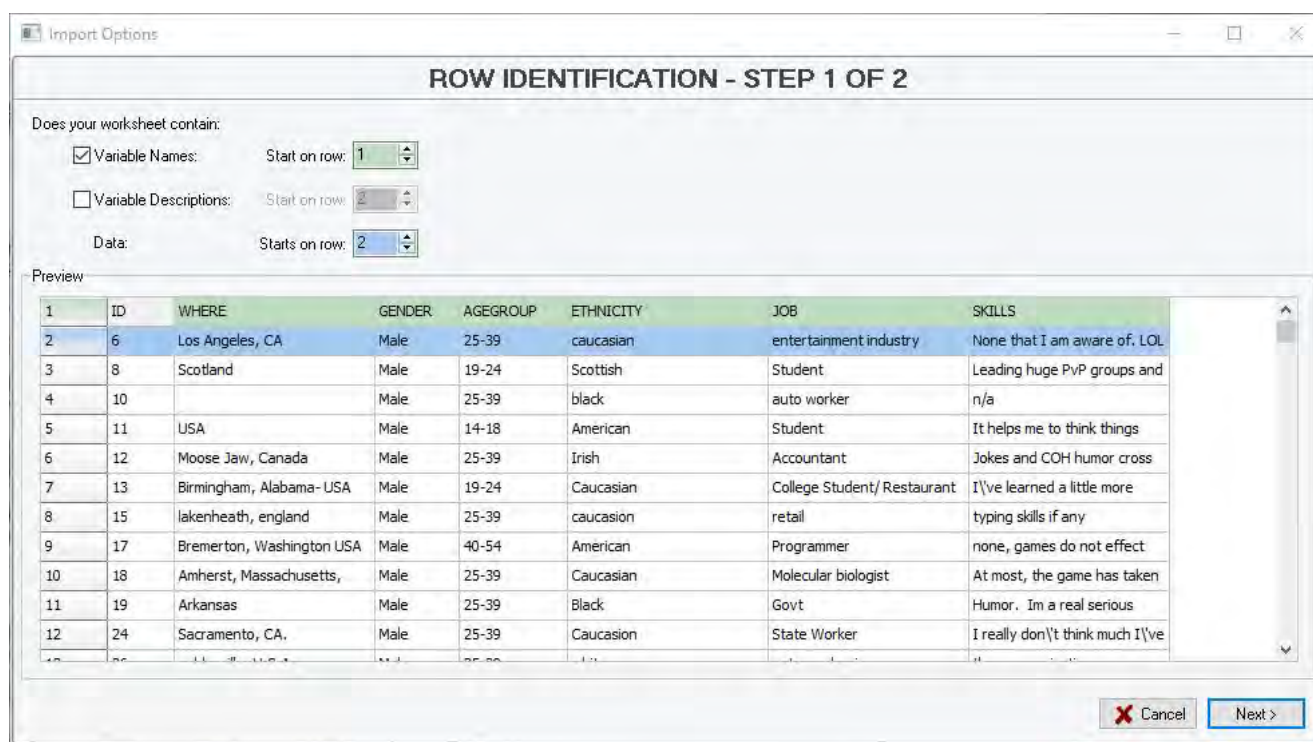
- Click the **Import from an existing data files or web service** button.
- Select the desired file format from the menu. An import dialog box will appear.
- Select the corresponding file format from the **Files of type** drop-down list and select the file you want to import.
- Click **Open**.

- Name your project and save it in the appropriate location.
- Once you have named your project and saved it, a dialog similar to the one below appears.



When you select an Excel file format for importation, the program displays a dialog box in which you can specify the spreadsheet tab 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** radio button to **All**, the program attempts to read the whole tab. You can preview the data before you import by selecting the **Preview** button. Your Excel spreadsheet must be closed before you select **Import**. You can only import one tab at a time. If you have more than one tab to import, containing the same column headers, simply append the other tabs.

- In the **Sheet** drop-down choose the tab you would like to import.
- Select either a **Range** or **All** and set you range, if necessary.
- Select **Import**. An Import Options dialog appears similar to the one below.



This dialog is a two-step process that helps to properly configure your data for import. On the first page you are asked to identify the location of the variable names and variable descriptions, if present. You are also asked to identify the row in which the data starts.

- If your spreadsheet contains variable names, check the checkbox called **Variable Names** and enter the row number in which they are located.

- If your spreadsheet contains variable descriptions, check the checkbox called **Variable Descriptions** and enter the row number in which they are located.
- In the **Data** field enter the row number in which the data starts.
- Select **Next**. The second page appears.

Import Options

SELECTION OF VARIABLES & SETTINGS - STEP 2 OF 2

Select the variables that you would like to import. You can customize the name and description if necessary by typing in the associated cell. Change the variable type by selecting from the dropdown menu.

Variables

Selected	Name	Description	Type
☒	ID	ID	Integer
☒	WHERE	WHERE	Short String
☒	GENDER	GENDER	Nominal
☒	AGEGROUP	AGEGROUP	Nominal
☒	ETHNICITY	ETHNICITY	Document
☒	JOB	JOB	Short String
☒	SKILLS	SKILLS	Document

< Previous Import...

The second page allows you to choose the variables you would like to import. You can change the names and descriptions by typing directly in the cells. To change the data type select from the drop-down list.

- Select the variables you want to import by checking their checkboxes.
- Modify the variable names, descriptions and data types as necessary.
- Select **Import**. The data will be imported and WordStat's **Data** tab will appear containing a table with your imported Excel data. From here you can further tailor your data set if necessary, or start analyzing immediately.

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. Also, the first row should preferably contain the variable names while the remaining rows hold the data, one case per row. QDA Miner and 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 variable names. If no variable name is encountered, QDA Miner as well as 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.

Database Files

MS Access, dBase and Paradox files

QDA Miner and Simstat can directly import MS Access, dBase and Paradox data files. For the last two file formats, the length of alphanumeric variables should not exceed 256 characters and memo variables are not supported. If you do have this 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 Visual FoxPro data file or a tab delimited text file).

Other database files

QDA Miner and SimStat offer ways to import from various database formats by connecting directly to the database system using an ODBC connection. For database systems for which no ODBC driver exists, it is still often possible to import data by exporting the data to a common file format that can be read by QDA Miner or SimStat. The recommended file formats are, in descending order of preference, Excel, Tab-delimited text files, or CSV files.

Importing memo variables

Memo variables 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](#).

Importing Plain Text or Word Processor Files from Simstat

QDA Miner provides an easy way to import documents stored in various formats including MS Word, WordPerfect, Rich Text, HTML, PDF and plain text files. When using SimStat as the base module, such a task is not as obvious. One way 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 variables:

- Using the Document Conversion Wizard program
- Retrieving a text file into a memo variable
- 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, RTF or Acrobat PDF 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 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 Variable

This method should be used to retrieve a single unit of text into a memo variable for a specific case. If textual data for several cases 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 the 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 the Import Text Into Memo button , select the text file you wish to retrieve 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 variables should ideally be removed. If a tab delimited format is chosen, all tab characters already present in a text variable should be removed. Another important limitation is that all the information of a single case 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).

QDA Miner as well as Simstat can read up to 2000 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 text by the use of page delimited memo files. This file format consists of a plain text file which contains the textual data of numerous individuals for a single memo variable. The text for each record must be separated by page break characters (ASCII 12) or by the ^P string. 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 the OK button.

The resulting file consists of a SimStat data file with two variables: RECNO, a numeric variable containing a sequential number going from 1 up to the total number of cases encountered in the input file, and TEXT, a memo variable containing the textual data for this case.

Note: Importation of numerous text variables 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 variables and to assure the case sequence of the various text files is identical.

Creating Comparison Charts

The chart window allows you to graphically examine the relationship between specific items 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 may be preferred for the examination of the relationship between these categories and an ordinal or quantitative variable. Eight types of chart can be obtained from this dialog box.



The **vertical bar chart** can be used to display the distribution of various categories within subgroups as defined by a nominal independent variable. Bar charts are easy to read since one can easily compare the heights of the bars.



The **Horizontal bar chart** has the same advantage as the vertical bar chart. The horizontal orientation provides a way to accommodate longer category labels without the need to change their text orientation.



The **stacked bar chart** allows one to display relative or absolute frequency of items by stacking them for each other class. It allows one to quickly show the relationship of parts to the whole or to emphasize a sum of several codes



The **100% stacked bar chart** is designed to show the relative percentage of multiple data series in stacked bars, where the total of each stacked bar always equals 100%. Like a pie chart, a 100% stacked bar chart shows a part-to-whole relationship.



The **line chart** may be preferred for the examination of the relationships between these categories and an ordinal or quantitative variable. It is especially useful when looking for trends over time.



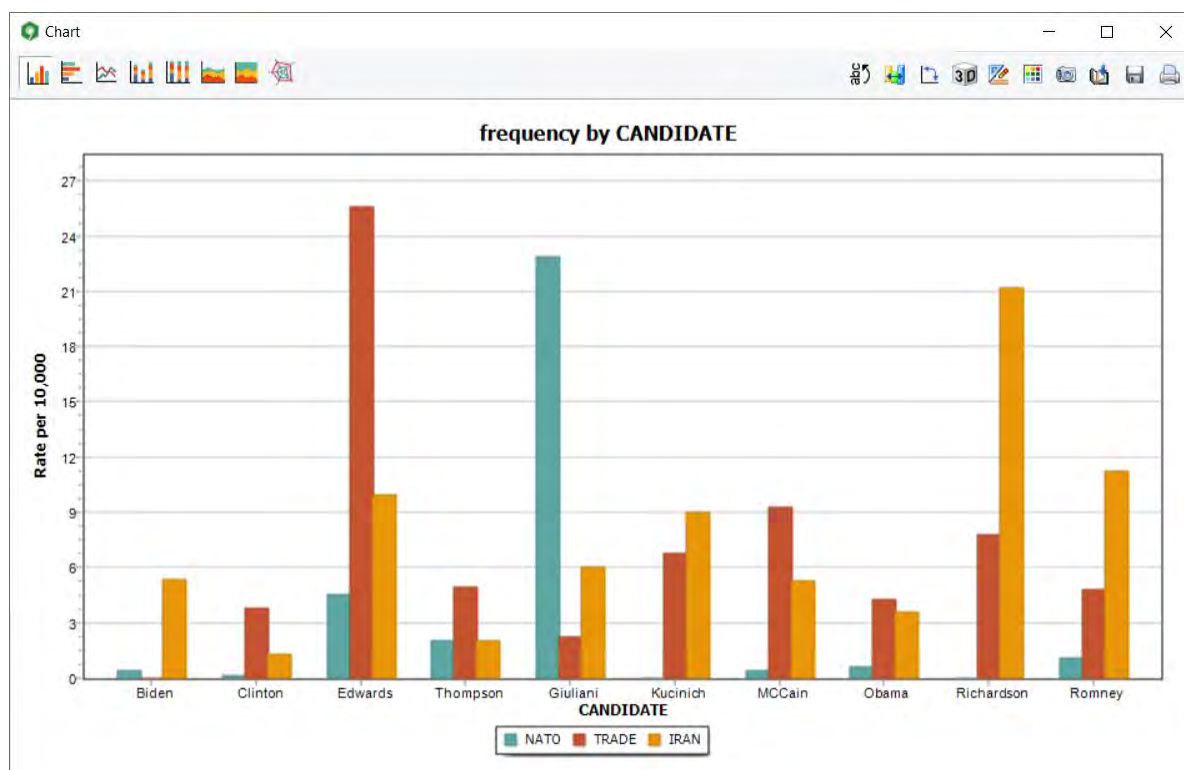
The **stacked area** chart is similar to a line chart with the areas below the lines filled with colors. The stacked version of the area chart allows one to display the contribution of various elements to a total over time.



The **100% stacked area** chart is similar to the stacked area chart with the difference that each value of the horizontal axis is adjusted so that the cumulative area always represents 100%. It is thus appropriate to display the evolution of the contribution of various elements to a total over time.



The **Radar Chart** represents all categories of a variable on a different axes. The origin of each axis of the radar chart is located at the center of the chart so that the higher frequency extends the plotted area farther toward the edge of the chart. The resulting geometric shapes illustrate the overall distribution of all categories.



A quick way to retrieve documents, paragraphs or sentences associated with a specific bar or pie slice is by right-clicking and selecting the keyword retrieval command.

To search for selected text segments:

- Right-click on the bar corresponding to the item and the specific value for which you would like to retrieve associated text segments.
- Select the **search...** command. All retrieve segments will be displayed in a separate dialog box. For more information on the various features available from this dialog box, see [Keyword Retrieval](#).

To remove specific series:

- Right-click on any bar corresponding either to the code or to the specific value you would like to remove.
- The popup menu will display two **REMOVE** commands, one of them for removing a specific item, while the other one for removing all bars associated with a specific value. Select the one associated with what you would like to delete.

The Toolbar

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

Control	Description
---------	-------------



Press this button to vertically display the labels on the bottom axis.



Clicking this button allows you to manually reorder series and bars within a series. It may be used to place bars in ascending or descending orders of frequency, move high frequency series in a back of a 3D chart, so as to make the other lower frequency series more visible. When used with radar charts, it may allow one to produce easier to read geometric shapes.



Press this button to change the default color palette used for various series of the chart.



This button allows the editing of various features of the chart such as the left and bottom axes, the chart and axes titles, the location of the legend, etc. (see [Chart Options Dialog Box](#))



Click this button to turn on/off the 3D perspective for the current chart.



Clicking this button causes the values represented on the bottom axis to be switched with those represented by different lines or bars (legend).



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



Press this button to append a copy of the graphic in the Report Manager. A descriptive title will be provided automatically. To edit this title or to enter a new one, hold down the SHIFT keyboard key while clicking this button.



Click this button to save a chart on disk. Charts may be saved in BMP, JPG or PNG graphic file formats.



Clicking this button prints a copy of the displayed chart.

Barchart and Line Chart Options

The various options on this dialog box allow you to customize the appearance of bar charts and line charts. The options available in this dialog box represent only a small portion of all of the settings available.

To further customize the chart, modify data points, value labels, or series order, click the  button located on the right side of the dialog box.

Left Axis

Minimum / Maximum: WordStat automatically adjusts the vertical axis scale to fit the range of values plotted against it. To manually set these values, type the desired minimum and maximum.

Increment: Increasing or decreasing this value affects the distance between numbers as well as tick marks. Horizontal grid lines are also affected by modification of this value.

Horizontal Grid: This option turns horizontal grid lines on and off. Grid lines extend from each tick mark on an axis to the opposite side of the graph. To increase or decrease the number of grid lines or the distance between those lines, change the Increment value of the axis. A list box also allows a choice among five different line styles to draw the grid lines.

Legend

Location: This option positions the legend. Legends may be placed at Top, Left, Right and Bottom side of the chart.

From top: When the legend is displayed on the left or the right side of the chart, this option specifies the legend's top position in percent of total chart height.

From left: When the legend is displayed on the top or the bottom of the chart, this option specifies the legend's top position in percent of total chart width.

Titles

Proper titles and axis labels are of utmost importance when describing the information displayed in a chart. By default, WordStat uses variable names and labels as well as other predefined settings to provide such descriptions.

The title page allows you to modify the **top** title, as well as the labels on the **left**, **bottom** and **right** axis. To edit the title, select the proper radio button. Enter several lines of text for each title by pressing the **<Enter>** key at the end of a line before entering the next line.

The **Font** button on the right side of the edit box allows changing the font size or style of the related title.

3-D View

Orthogonal: Turning this option off disables the free elevation and rotation of the 3-D chart.

Zoom: This option zooms the whole chart. Expressed as a percentage, increasing the value positively will bring the chart towards the viewer, increasing the overall chart size as the Zoom value increases.

3-D Percent: The 3-D Percent property indicates the size ratio between chart dimensions and chart depth by specifying a percent number from 1-to-100.

Perspective: Use this property with Orthogonal unchecked to modify the 3-D perspective of the Chart. Larger values add more depth perspective.

Bar shadow: Enabling this option add a shadow to the sides of 3-D bars. Turning it off will colors the sides of the bar the same as the front.

Bar width: This option determines the percent of total bar width used. Setting this value to 100 makes joined bars.

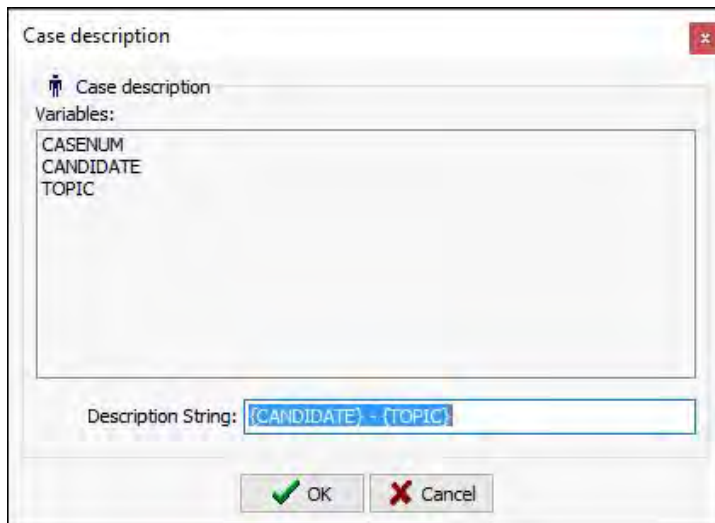
Bar depth: Use this property to limit the depth that each bar series use. By default, bars will take up the part proportional to the number of bar series in the chart so that the back of a bar will join the front of the bar immediately behind it. To insert a gap between series of bars, decrease this value.

Edit Case Descriptors

The **Edit Case Descriptor** command allows you to define a case descriptor that will be used to identify each case based on the values it contains in one or several variables. Such a string is used to identify cases when retrieving segments as well as when analyzing case similarities using clustering, multidimensional scaling or proximity plots.

To edit case descriptors:

- Click the  button in the upper left corner of the main window, scroll down to **EDIT CASE DESCRIPTORS**. A dialog box similar to this one will appear:



This dialog box allows you to specify a label that will be used to describe each case. The label may be changed by editing the text in the **Description String** edit box. To insert the value stored in a specific variable into the description, simply enter the variable name in uppercase letters and enclose this name between braces. Alternatively, you can insert a variable name at the current cursor location by clicking the corresponding item in the **Variables** list located just above the edit box.

If you enter the following string:

{GENDER} subject - {AGE} years old

The {GENDER} and {AGE} strings will be replaced with their corresponding value for this specific case. If the current case contains information about a seventeen-year-old male, the above string will be displayed as:

Male subject - 17 years old

It is also possible to insert the following string:

{CASENUM}

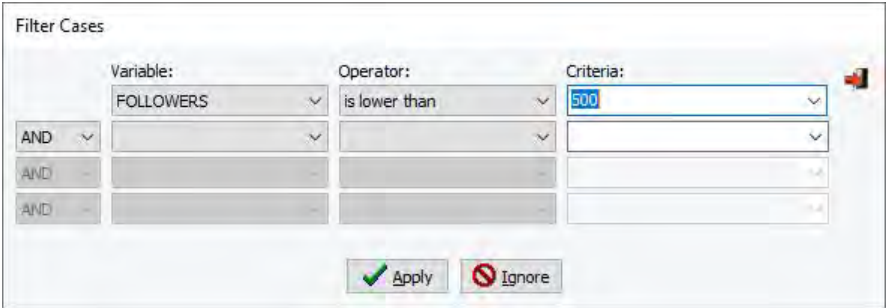
This string will display a unique case number, representing the physical order of this case in the project file.

Filtering Cases

The **Filter Cases** command temporarily selects cases according to some logical conditions. You can use this command to restrict analyses to a subsample of cases or to temporarily exclude some group of subjects. Two types of filtering are available in WordStat: a basic filtering dialog box that provides some guidance for easily building filters, and a powerful xBase filtering tool that allows you to create more complex filtering expressions that contain more advanced mathematical functions, as well as Boolean and relational operators.

To filter cases:

- Click the  button in the upper left corner of the main window, scroll down to **FILTER CASES**. A dialog box similar to this one will appear:



This dialog box consists of two major sections. The upper section of the filtering dialog box allows you to specify up to four filtering conditions joined by logical operators (such as AND, OR). Each condition consists of a variable, an operator and, if needed, some numerical, categorical or string values. The following table presents the various operators available for each data type.

DATA TYPE	AVAILABLE OPERATORS
NOMINAL / ORDINAL	Equals Does not equal Is empty Is not empty
NUMERIC and DATE	Equals Does not equal Is greater than Is lesser than Is greater than or equal to Is lesser than or equal to Is empty Is not empty
BOOLEAN	Is true Is false
STRING	Contains Does not contain Is empty Is not empty
DOCUMENT	Is empty Is not empty Is coded Is uncoded
IMAGE	Is empty Is not empty

The lower section allows you to filter cases based on the presence of words or phrases associated with specific content categories in the document being processed by WordStat. This section is disabled by default and becomes active as soon as a content analysis has been performed. 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 cases that contain words in both categories.

APPEARANCE **OR** ART Selects cases that contain words in either one of these two categories.

APPEARANCE **NOT** ART Selects all cases that contains a word in category APPEARANCE but that do not contain any word found in the ART category.

- Once a filtering expression has been entered, you can apply the filter and leave this dialog box by clicking the **Apply** button. If the filter expression is invalid, an error message will appear and exiting from the dialog box will not occur.
- To temporarily deactivate the current filter expression, click the Ignore button. The filter expression will be kept in memory and may be reactivated by selecting the FILTER CASES command again and clicking Apply.
- To exit from the dialog box and restore the previously active filtering expression, click the **Close** button.

Geocoding

Geocoding is the process by which geographical information such as place names, postal codes or IP addresses are transformed into corresponding geographic coordinates. WordStat can create geographic maps by relating coded segments with existing geographic coordinates such as latitude and longitude, or projected coordinates. While it cannot create maps directly from a city name or a postal code, it offers a geocoding service allowing you to transform such kinds of information into latitude and longitude.

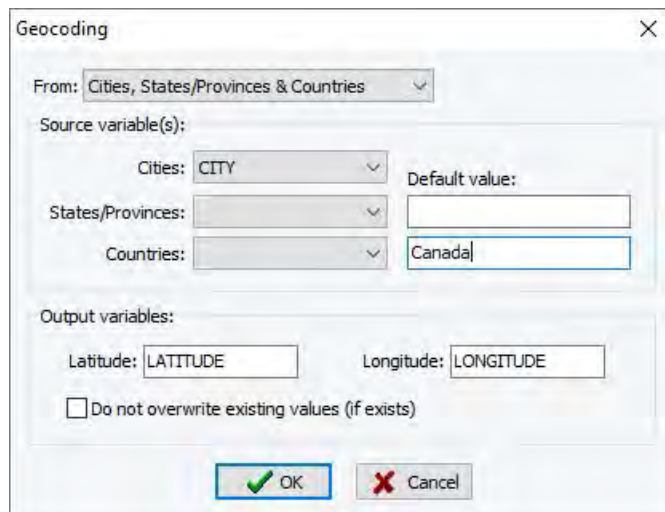
WordStat can geocode four types of data:

- Cities, States/Provinces & Countries
- Country names
- Postal code or zip codes
- IP addresses

This transformation process is performed using a geocoding server. It requires an Internet connection. The result is then stored in the project as numerical variables for the corresponding latitude and longitude. When an IP address is geocoded, you can also obtain a city name, a region and a country name and store the information into string variables. Once the variables are created, mapping can be done without the need of an Internet connection.

To access the geocoding feature:

- Select the **GEOCODING** command from the  menu. You may also access this feature by going through the mapping feature and then through the geographic coordinates setup dialog box. The geocoding dialog box looks like the one below:



- The **From** list box allows you to select the type of geographic information that can be geocoded. Once set, the dialog box automatically adjusts the **Source variable(s)** option box to display the information that needs to be specified. The above example displays the options for geocoding city names. Set the list boxes to the string variables containing the required information. You may also set default values for some fields. These values will be used if no variables are selected, or if a specific case has a missing value for the selected variable. For example, if you want to geocode US zip codes, you may simply set the postal code variable to the string variable containing the zip code and type 'USA' or 'US' in the country default value text box.
- The **Output variables** group box allows you to specify the name of the numerical variables in which the **Latitude and Longitude** information will be stored. If the variables do not exist, the software will ask to confirm their creation. By default, if the variables already exist, their values will be replaced by new values. If you set the **Do not overwrite existing values** check box, only cases containing empty values will be geocoded. Such a feature may be useful when a secondary type of geographic information is used to complement a first geocoding. For example, if you have, for each case in a dataset, a postal code and an IP address, you may start by geocoding all cases with postal codes

and then complement the cases with no postal codes or those for which no geographic coordinates could be found by geocoding their IP addresses.

- Once all the required options have been set up, click the **OK** button to perform geocoding. WordStat will use the current Internet connection to send all the information to be geocoded to a server that will return to WordStat the corresponding latitudes and longitudes. A report dialog will display a summary of the geocoding process, indicating how many items and cases have been successfully geocoded and how many items could not be geocoded. Clicking the **View** button will bring a list of all items that could not be geocoded. This table can then be saved on disk or printed, allowing you to easily identify cases that may need revision.

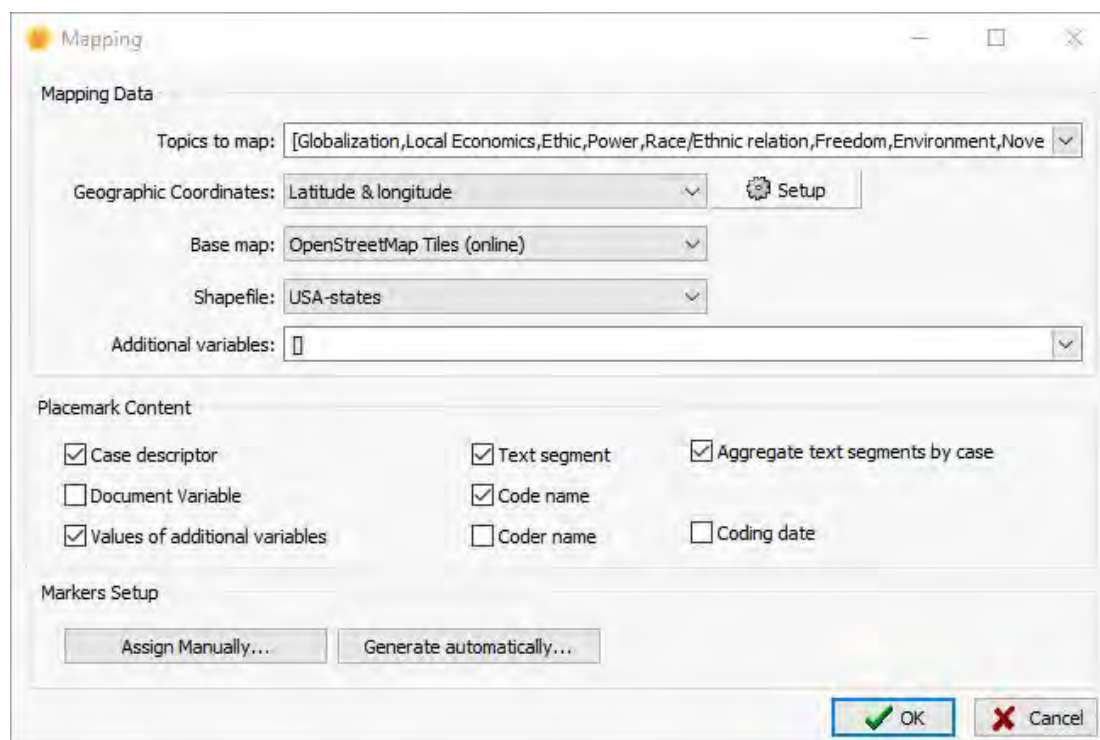
Mapping

The mapping feature of WordSat allows you to analyze the spatial distribution of terms, phrases, content categories or topics. Mapping codes requires relating the codes to geographic location information stored in the project variables in the form either of longitude and latitude variables or of geographic projected coordinates (X and Y). In order to obtain such geographic information, WordStat allows you to compute latitude and longitude data by transforming existing variables such as postal codes, city names, country names as well as IP addresses into their corresponding geographic coordinates. Please see [Geocoding](#) for more information.

The mapping of codes function is available anywhere you see the  icon. You can find this icon on the Frequencies tab, Extraction tab (topics and phrases) and the Crosstab tab.

To map specific text features:

- Select rows of all items you want to map. For the Topic extraction tool, only one topic may be selected, yet several topics may be selected later.
- Click the  button. A dialog box similar to this one will appear:



The image shows a 'Mapping' dialog box with the following sections:

- Mapping Data:**
 - Topics to map:** A text box containing '[Globalization,Local Economics,Ethic,Power,Race/Ethnic relation,Freedom,Environment,Nove]' with a dropdown arrow on the right.
 - Geographic Coordinates:** A dropdown menu set to 'Latitude & longitude' with a 'Setup' button to its right.
 - Base map:** A dropdown menu set to 'OpenStreetMap Tiles (online)'.
 - Shapefile:** A dropdown menu set to 'USA-states'.
 - Additional variables:** An empty text box with a dropdown arrow on the right.
- Placemark Content:**
 - ☒ Case descriptor
 - ☐ Document Variable
 - ☒ Values of additional variables
 - ☒ Text segment
 - ☒ Code name
 - ☐ Coder name
 - ☒ Aggregate text segments by case
 - ☐ Coding date
- Markers Setup:**
 -
 -

At the bottom right are **OK** and **Cancel** buttons.

- The **Topics to Map** checklist box displays all selected topics. You may add additional topics by clicking the down arrow key at the right end of the list box. You will be presented with a list of all available topics. Select all topics you want to map.

- The **Geographic Coordinates** allows you to select which set of variables contains the geographic location information to be used for mapping items. Two types of coordinates can be used: 1) latitude & longitude or 2) projected coordinates. Clicking the **Setup** button will bring a dialog box similar to this one:

You can associate existing numerical variables to geographical properties such as latitudes, longitudes or projected coordinates. In the latter case, you also need to choose the appropriate geographic projection. The dialog box also offers the possibility to geocode existing variables, computing latitude and longitude variables from other types of geographic information (see [Geocoding](#)).

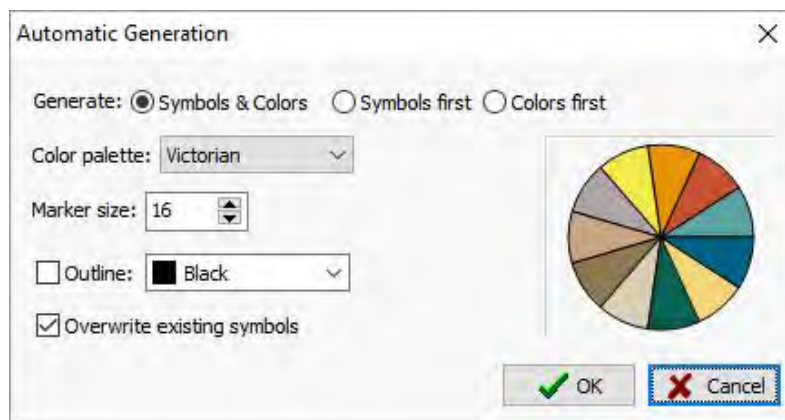
- The **Base Map** list box allows you to choose the background map on which data points will be plotted. It may consist of a web map service (WMS) or a geolocalized raster file (or world files). Web map services (identified with the 'online' suffix) require an Internet connection in order to retrieve relevant map files, while raster files may be used without a connection. Setting this option to **Other...** will allow you to select an image file located on your current PC. This image file may consist of a satellite photo, a street map, a topographical map, or any other representation of a geographic region as long as it is accompanied by a World file that contains the appropriate information to convert the image coordinates to real-world geographic coordinates. When a valid image file is selected, the user will be offered a choice to move this file to the default mapping resource folder or to leave the file where it is.
- The **Shapefile** list box displays the currently available shapefiles. Shapefiles are geospatial vector data format for GIS systems. They typically contain series of polygons, lines and points representing geographical objects, such as countries, states, counties, lakes, etc. Shapefiles are used in WordStat to create thematic maps in which areas are shaded or color-coded in proportion of specific measurements. WordStat provides a few basic shapefiles for representing either all countries or specific continents, as well as some detailed shapefiles for some cities or countries. Thousands of additional shapefile maps can be downloaded for free from various websites. Setting this option to **Other...** will allow you to select a shapefile located on your PC. When a valid shapefile is selected, the user will be offered a choice to move this file (along with all associated files) to the default mapping resource folder or to leave the file at its current location.
- The **Additional variables** checklist box allows you to transfer to the mapping module additional information stored in numerical, categorical, string or date variables. Values of these variables may then be used for analysis, displayed or exported.
- The **Placemark Content** group of options allows you to select how data points will be created as well as what additional information will be used to describe each data point. The following information can be transferred: The **Case Descriptor** which is a user-defined string used to describe the case from which a specific text segment comes from, the **Document Variable** from which this text segment has been extracted, as well as the **Values of Additional Variables** (if any additional variable has been selected). You can also transfer the associated **Text Segment**. Selecting the **Aggregate text segments by case** option will merge all the text into a single text segment and produce a single data point per topic. Additionally, you can transfer the **Code Name**, the **Coder Name** and the **Coding Date**.
- The **Markers Setup** section gives access to two dialog boxes that may be used to configure the appearance of markers (or placemarks). Selecting **Assign Manually** calls a dialog box that allows you to associate specific topics to specific symbols. The associations are stored in the project so that they can be reused later. Selecting **Generate Automatically** allows you to configure the sequence in which symbols are being generated. For more information on manual or automatic marker assignments see [Customizing Markers](#).
- Once all the settings have been adjusted, click the **OK** button to call the mapping module.

Customizing Markers

By default, WordStat automatically assigns various symbols to different topics, cycling through a list of symbols and colors to reduce the likelihood that two topics will have the same marker. You can customize the process by which WordStat assigns symbols and can adjust their visual properties such as their size and outlines. You can also customize symbols for specific topics, so that the same symbol will be used whenever this topic is mapped.

To customize the way symbols are automatically assigned:

- Click the **Automatic Generation** button. A dialog similar to this one will appear.



Generate: This option specifies how WordStat will cycle through symbols and colors. By default, the software changes both the symbol and the color for every consecutive topic. Selecting **Symbols first** will cycle through the seven symbols but keep the color constant, and will only change the colors when all symbols have been used. Selecting **Color First** will keep the symbols constant until all colors of the select color palette have been used.

Color palette: This option allows you to choose a set of colors to be used in the automatic marker generation process.

Marker size: This option lets you adjust the size of the marker.

Outline: When this option is checked, markers will be surrounded by a single line. The color can be set using the color list box.

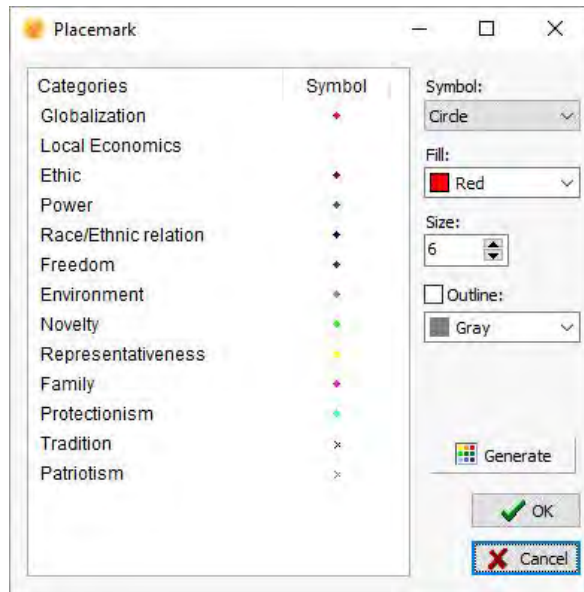
Overwrite existing symbols: When custom symbols have been assigned to specific topics, their symbols and colors will remain unchanged even if the automatic marker generation process is used for assigning symbols and colors to other topics. Enabling this option will cause the custom markers to be overwritten with new symbols and colors.

- Set your automatic generation options.
- Click **OK**.

By default, markers are automatically generated. However, custom markers can be assigned to specific topics, content categories, or words and saved within the project file so that the same marker will be used every time the topics are plotted. The customization feature also allows you to assign a bitmap or a letter to a text element.

To assign specific markers to text features:

- Click the **Assign Manually** button. A dialog similar to the one below one will appear.



- To assign a custom symbol to a word, a content category or a topic, you first need to select it from the list on the left of the dialog box and then adjust any of the following options:

Symbol: This list box displays the various symbols that could be used as markers for this item. **Bitmap...** will bring an open file dialog box that will allow you to choose a bitmap image file (*.bmp; *.bmp or *.bmp). When an image is selected, the following three options are disabled since they only apply for symbol markers.

Fill: This option allows you to choose a color to be used for the marker.

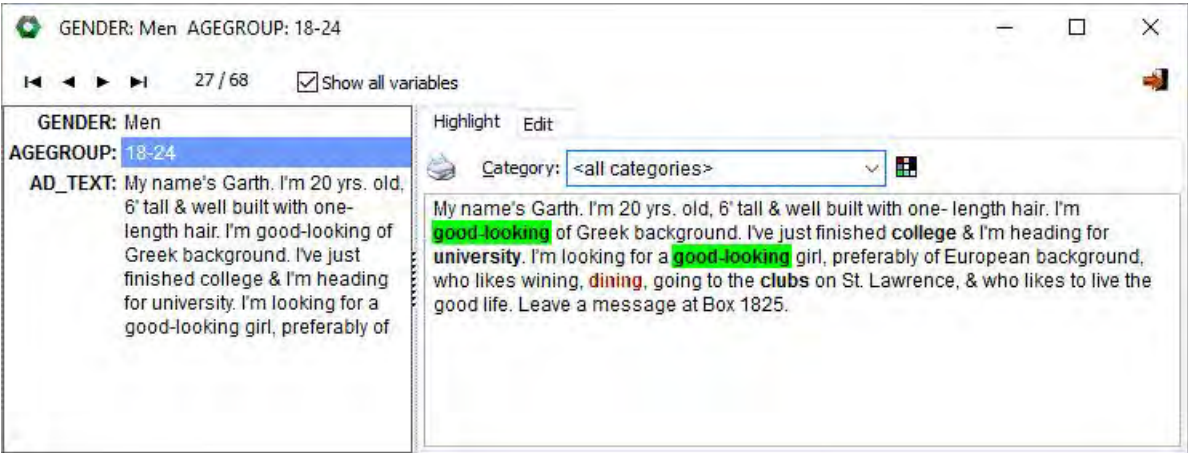
Size: This option lets you adjust the size of the marker.

Outline: When this option is checked, the current marker will be surrounded by a single line. The color of this line can be set using the color list box below.

- To automatically assign initial markers to all items listed, click the **Generate** button.

Text Editor

WordStat's integrated text editor allows browsing and editing alphanumeric variables and documents submitted to content analysis, as well as spell checking of text found in a specific case or in the entire data file. When viewing plain text documents, the editing window consists of a single editing view where text can be edited and keywords are highlighted. When viewing rich text documents, the editing and keyword highlighting features are accessed through two separate views. The keyword highlighting feature allows you to identify all words or phrases 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. You can also jump directly from a selected word to a keyword-in-context table of all instances of this word.



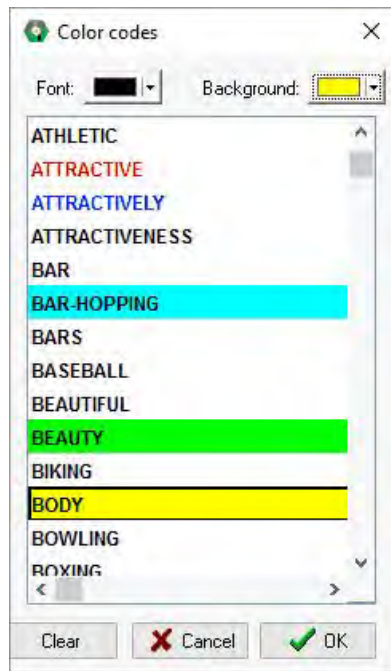
It is also possible from this dialog box to examine and edit all numeric and alpha numeric values stored in other variables of the data file. To view and edit the values for the current case, the **Show all variables** check box should be selected. When enabled, the screen will be split vertically. On the left side, a panel with a list of all variables with their values for this case will be shown. To edit any of the values, press the **F2** key or double-click the value to edit.

The following table provides a short description of buttons and controls of the text editor dialog box:

Control	Description
	This button allows the importation of text from various file formats including plain text file, RTF, MS Word, WordPerfect, MS Write or HTML. If the variable containing the document supports only plain text documents then all formatting options and unsupported features, such as bullets, graphics or headers are removed.
	Export the current document to disk. Plain text document may only be saved as plain ANSI document while RTF documents may be saved in plain text or RTF format.
	Print the current document.
	Cut the selected text to the clipboard.
	Copy the selected text to the clipboard.
	Paste text from the clipboard at the current cursor position.
	Reverse the last action made to the text.
	Search for a specified word or phrase.
	Search for and replace a specified word or phrase.
	Spell-check documents for all cases.
	Spell check only the current document.
	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.
Variable:	This drop-down list box allows you to select the alphanumeric or memo variable 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 shown 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 select all categories using different colors, set this option to <all categories>.
	Clicking this button accesses the color coding dialog box that lets you assign to each category in the dictionary specific font and background colors (see below).
	Move to the first case of the data file.
	Move to the last case of the data file.
	Move to the previous case
	Move to the next case.

Assigning Color Codes to Categories

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



To access the color code dialog box:

- Set the **Highlight** drop-down list to <all categories>.
- Click 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 the **Other** button to define a custom color.

Publishing Categorization Models

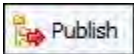
A typical text categorization process may involve any one of the following steps:

- User-defined text preprocessing.
- Automatic lemmatization.
- Exclusion of words and phrases.
- Substitution of words and phrases.
- Categorization of words, word patterns, phrases and coding rules into content categories using a categorization dictionary.
- Specific settings, such as the inclusion of special characters or numerical digits, may also need to be set in order to collect relevant information.

Publishing a model file allows you to access the full text processing features of a WordStat elsewhere including:

- QDA Miner, through the **KEYWORD RETRIEVAL** feature.
- [WordStat Document Explorer](#) utility program,
- Or any other application making use of the WordStat [Software Developer's Kit](#).

To publish a categorization model:

- Set the various analysis options on the **Text Processing** tab necessary to reproduce the required categorization process.
- Click the  button located at the top of the tab. A dialog box will appear asking you for a file name.
- Enter the file name of the model you want to create and click **Save**.

Published categorization model files are saved with a **.WCAT** file extension in the **My Provalis Research Projects\Models** folder .

NOTE: While the information in the exclusion list and the categorization dictionary is all stored in the categorization file, running a categorization model from outside WordStat may still require the availability of some resource files such as language dictionaries or preprocessing libraries (EXE or DLL). This should not cause an inconvenience when applying the models on the same computer as the one used to create the model, since information about the original locations of the resource files is always stored within the model file. However, when attempting to apply these categorization models on another computer, the calling application may have some difficulty locating the needed resource files. The files should be stored either under paths identical to those on the original computer, in the application folder or under specific subfolders. When an attempt is made to apply a saved categorization or classification model for which some resource files are missing, an error message will be displayed providing the list of all missing files, their original location and alternate locations where they might be.

For information on how to apply a saved categorization model, please refer to the sections on [WordStat Document Explorer](#), or to the [WordStat Software Developer's kit](#).

Creating and Using Norm Files

A useful element in interpreting the results of a content analysis is the possibility of comparing the obtained results to some normative data and identifying how similar or dissimilar the observed frequencies are compared to those norms. For example, you may wish to compare the vocabulary of an adult victim of a brain injury to vocabularies of normal adults or the mission statement of a business to a collection of mission statements of Fortune 500 companies. You may also establish the reading level of a school manual by comparing its vocabulary to collections of books read by children of various ages. Normative data are typically computed on a large sample of documents and represent either general norms with data from a wide variety of sources or are computed on a more specific text corpus related to the channel, the domain area or the specific situation being studied. For example, you could compare the speeches of candidates of a presidential election to a large collection of English text from different sources (newspapers, novels, technical documents, etc.) or to a more specific corpus of spoken English or to a collection of political speeches. Comparison of word frequencies to norms established on a general corpus may be especially useful to identify the specific terminology of a set of documents. On the other hand, using a more specific collection may allow one to identify subtler differences or nuances.

WordStat allows you to create normative data on a collection of documents based on either the content of a categorization dictionary or on the frequency of individual words. The norms may be stored on disk and later be compared to the results of a content analysis performed on other documents. When comparing results to norms established using a categorization dictionary, it is highly recommended to use the same dictionaries and the same analysis options as the ones used to create the norms. Using different settings may result in invalid comparisons. To prevent such a situation, WordStat will detect any difference in settings and issue a warning message, pinpointing all differences in the settings. On the other hand, comparing the results to a norm file based on a comprehensive list of word frequencies may provide a more flexible solution than using a norm established using a categorization dictionary since a single word frequency norm file may be used to compare results obtained using various categorization systems. In such a situation, WordStat automatically computes from the words in the norm file the expected frequencies for each content category. However, it is important to remember that if the categorization dictionary contains phrases or rules, the expected frequency will likely be underestimated and may be invalid.

When a comparison to a norm file is performed, WordStat appends four columns to the right of the frequency table and computes each item's expected frequency, the deviation from the observed frequency, the Z value (standardized deviation) and its two-tailed probability.

To create a norm file based on content categories:

- Open WordStat.
- Select the categorization dictionary on which the norms should be computed and set the various analysis options (exclusion list, lemmatization, etc.).
- Move to the **Frequencies** tab to force the computation of frequencies on the content categories.
- Click the  button and select the **SAVE AS A NORM FILE** command. A file-saving dialog box will be displayed.
- Enter the name of the file under which you would like to store the norms (by default, the .wnorm file extension is added to the file name), then click **Save** to create the file.

To create a norm file based on word frequencies:

- Open WordStat.
- If a categorization dictionary is active, disable it and set the various analysis options (exclusion list, lemmatization, etc.). It is recommended to set the minimum frequency or record occurrence to "1" and to disable the option to remove words under a specific frequency or occurrence in order to obtain a detailed frequency list of all words in the normative sample.
- Move to the **Frequencies** tab to force the computation of word frequencies.
- Click the  button and select the **SAVE AS A WORD FREQUENCIES** command. A file-saving dialog box will be displayed.

- Enter the name of the file under which you would like to store the norms (by default, the .wfreq file extension is added to the file name), then click **Save** to create the file.

To compare the obtained frequencies with existing norms:

- Set the required dictionary and options and move to the **Frequencies** tab to instruct WordStat to compute the frequencies of words or of content categories.
- Click the  button and select **COMPARE TO NORM FILE** or **COMPARE TO WORD FREQUENCIES**, depending on whether you would like to compare the obtained frequencies to norms established on content categories or on words.
- Select the norm file to which you would like the comparison to be made and click **OK**.

To remove comparison statistics:

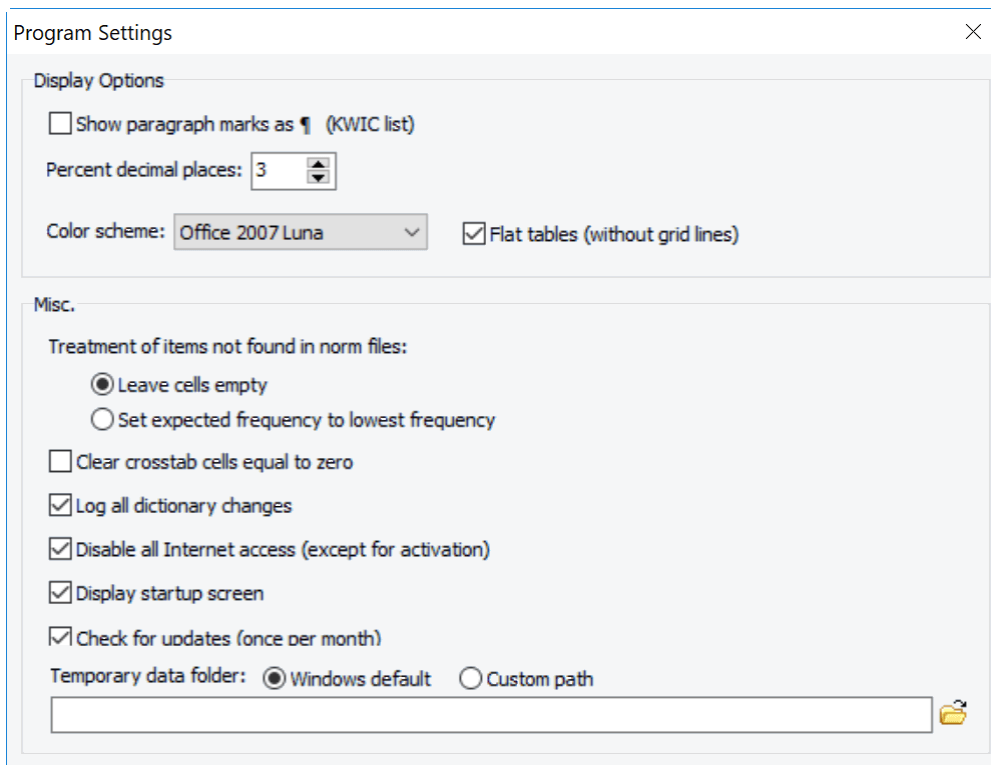
- Click the  button and select the **REMOVE NORM STATISTICS** command.

Program Settings

The **Program Setting** functions cover a variety of options that are program-wide rather than project-specific. These options include display options, dictionary logs and internet access etc.

To access program settings:

- Select the **PROGRAM SETTING** command from the  menu in the upper left side of the Win. A dialog box will appear similar to the one below.



Display Options: These options include the possibility of displaying **paragraph marks** by selecting the checkbox beside this option. You can choose the number of **decimal places** you would like when displaying percentages but typing the number in the field or by selecting the up or down arrow until the desired number is reached. The **color scheme** can be

chosen from the drop-down list and you can choose whether or not to have flat tables by selecting the checkbox beside this option.

Treatment of items not found in norm files: When comparing frequencies of words or content categories in a project to some normative frequency data, some words may be unique to the corpus being analyzed. In such situation, a comparison of the observed frequency to the expected one, based on a reference corpus is not possible. WordStat offers two ways to deal with such a situation: 1) It can leave the cells associated with the expected frequency, deviation, z value and probability empty, or 2) it may set the observed frequency to the lower possible frequency. For example, if the normative frequencies were computed on a reference corpus containing 950,000 words, then the ratio used to assess the expected frequency will be set to 1/950,000.

Clear crosstab cells equal to zero: In a crosstabulation table, empty cells are represented as either 0 or as a percentage equal to 0.0%. Enabling this option leave those cells empty.

Log all dictionary changes: When this option is enabled, all changes to the exclusion list, the substitution list or the categorization dictionary are stored in a log file, allowing one to review changes made previously. This feature is especially useful when more than one person work on a single dictionary file and one needs to review changes made previously.

Disable internet access: WordStat is a desktop text analysis tool. Your text data never have to leave your computer in order to be analyzed. The only situation where data may be sent through the internet is if you use the geocoding features of WordStat to transform text fields such as city names, country names or postal code into latitude and longitude. WordStat may however access the internet to retrieve information about potential updates to the software, to get additional resources, or to access a web mapping service used to generate maps in the GISViewer mode. It also use internet for the initial activation of the software and further validation of the software license. Selecting this feature disable all access to internet, except for the activation and validation of the software licenses.

Display startup screen:

Check for updates:

Temporary data folder:

Mode

Data can be analyzed in two different modes: **Explore** mode and **Expert** mode. The **Explore** mode lets you see at a glance what is in your data set. You will have access to extracted topics and phrases as well as a frequency list. If you would like a more comprehensive analysis, run the **Expert** mode. In **Expert** mode, in addition to all the features available in the **Explore** mode, you can also create dictionaries and crosstabs, and perform cooccurrence analysis etc. All of WordStat's capabilities are available in **Expert** mode.

To switch between Expert and Explorer mode:

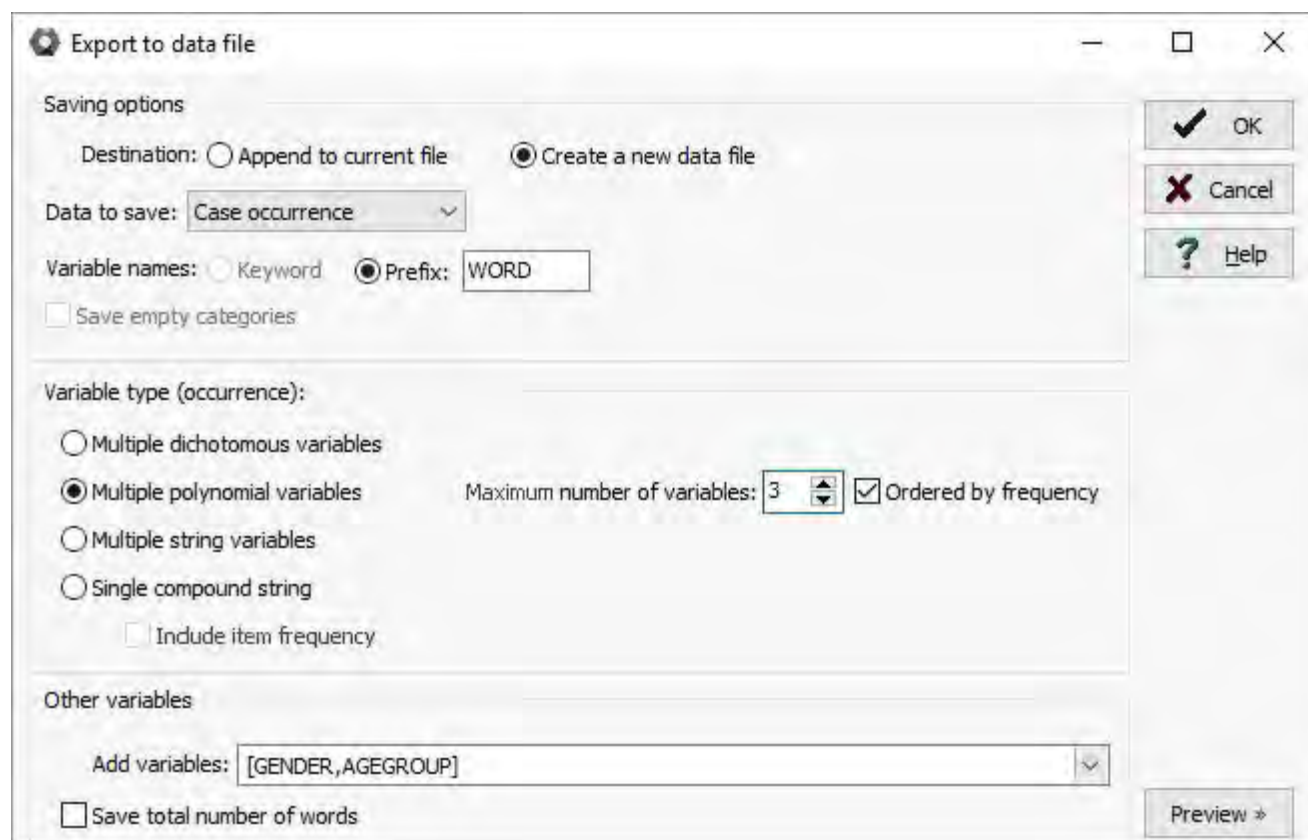
- Select the **MODE** command from the  menu.
- Choose either **Expert** or **Explorer** form the adjacent menu.

Exporting Frequency Data

Various case statistics may be appended to the existing data file or exported to disk in different file formats including SPSS, Stata, Excel, HTML, XML, and tab or comma delimited text files, allowing this data to be further analyzed. The resulting data consist of a matrix where each row represents a case and where the statistics on content categories will be stored in columns along with a few additional variables.

To append or export to disk content category statistics (data matrix):

- Click the  button located at the top of the **Frequencies** tab. A dialog box similar to this one will appear:



Destination: This option allows you to choose whether the new variables should be appended to the current data file or in to 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 variables are created to store the case number and the numerical values of each independent variable. When data is saved in to a new file, clicking **OK** will call a dialog box that lets you specify the name and location of the new data file along with its type. WordStat can export to several types of files, including SPSS and Stata data files, Excel spreadsheets, tab and comma delimited files, as well as HTML and XML files.

Data to save: This option allows you to choose four different kinds of data that may be saved:

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

Variable names: This option sets what method should be used by WordStat to create new variable names. When set to KEYWORD, the program will attempt to use each keyword as the name of a new variable. Illegal characters are

automatically removed and long names are truncated to the first 10 characters. Duplicated variable names are distinguished by the substitution of numerical digits at the end of the name. When this option is set to PREFIX, variable names are created by adding successive numeric values to a user-defined prefix. For example, if the edit box at the right of the prefix option is set to "WORD_", the variable names will be WORD_1, WORD_2, WORD_3, etc. The order of creation of the variables corresponds to the sort order used in the FREQUENCY tab.

Variable type: By default, WordStat saves keyword statistics in as many variables as there are keywords or content categories listed on the frequency tab. For example, if the frequency table contains 100 items, then 100 variables will be necessary to store the statistics associated with each item. When you choose to store the **occurrence** of codes, WordStat offers you the possibility of storing the observed occurrences in a limited number of polynomial (or multinomial) variables. For example, if the maximum number of different content categories per case is no more than 10, then you may instruct WordStat to create 10 numeric variables and store, in each of those, a numeric value representing one of the content categories. If less than 10 categories are found in a specific case, then the remaining variables are left empty. To store values in a limited set of nominal variables, choose the **Multiple Polynomial Variables** option and enter the **Maximum Number of Variables** that should be used for storing the values representing the content categories. To store the name of those categories rather than their numerical values, select **Multiple String Variables**. If the maximum number of content categories found in a single case is higher than the specified number of variables, then a warning message will appear to let you know that some information has been lost and to indicate the maximum number of content categories encountered in the project. To export occurrences as zeros and ones in as many variables as there are codes, select the **Multiple Dichotomous Variables** option.

Add variables: This drop-down checklist box may be used to add the values stored in one or more variable to the exported data file along with the statistics.

Save total number of words: This option appends a numeric variable named TOTWORDS that contains the total number of words processed in each case.

Clicking the **Preview** button displays a grid allowing one to see what the data file will look like.

- Once your options have been chosen select **OK**.
- If you are creating a data file, choose and name, type and place to save your file and then select **Save**.

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 as well as various multivariate analysis techniques (cluster analysis, factor analysis, etc.). The table below illustrates some types of analysis that may be performed with SimStat and, if needed, the required module to perform these analyses.

TYPE OF ANALYSIS	REQUIRED SIMSTAT MODULE
Multiple regression analysis	None
n-way ANOVA/ANCOVA	None
Reliability analysis	None
Factor Analysis (with or without varimax rotation)	None
Principal Component Analysis	MVSP
Advanced cluster analysis	MVSP

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 case in the data file, the occurrence or frequency of specific words or categories.

To create variables:

- Set the various options of the Text Processing tab.
- Perform the frequency analysis by clicking the **Frequencies** tab.
- Click the  button to access the **Export Data** dialog box.
- Set the **Destination** option to **Append To Current File**.
- Set the **Date to save** list box to **Keyword frequency**.
- Set the **Variable names** option to **Keyword** to instruct WordStat to use the names of the categories (or included words) as the names for the new variables.
- Click the **OK** button to proceed to the saving of the data and return to WordStat.
- Click 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 [Exporting Frequency Data](#).

Performing a cluster analysis of keywords:

- 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 box.
- Set the various analysis options to your preferences. (Please take note that in order to perform a cluster analysis on words rather than on cases, the Transpose Data option should be left deactivated).
- Click 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 your preferences.
- Click the **OK** button to perform the statistical analysis.

Performing Analysis on Manually Entered Codes

The QDA Miner software is a computer-assisted qualitative coding tool specifically designed to manually code documents. While WordStat was designed mainly to perform automatic content analysis of textual data, you may also use it to manually assign codes to text. When used for such a purpose, codes need to be manually typed into the text itself. WordStat may then be used to extract the codes and perform various types of analyses (frequency, cross-tabulation, cooccurrences). At least two approaches of coding may be used to achieve this:

Unique Keywords

Unique keywords or code names may be inserted anywhere in the text. These keywords should preferably not be an existing dictionary word. For example, it may consist of an abbreviation of one or several words, includes special symbols (such as #, & ^, _ etc.) or numeric digits. The retrieval of these codes can then be achieved by adding all these keywords to the categorization dictionary. If special symbols have been used, they should also be specified as valid characters on the [Preprocessing](#) tab.

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 into the text and placed within square brackets. By disabling all dictionaries and setting this option, the program simply ignores all text outside the brackets and performs a frequency count of all words found inside square brackets. This method has several advantages over the previous method. First, there is no need to enter all existing codes in the categorization dictionary, since they will be processed automatically. Misspelled keywords will always be extracted and may be easily identified and changed. Also, the processing time and memory required can be much lower than the other approach since WordStat won't process text found outside the brackets.

Besides 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:

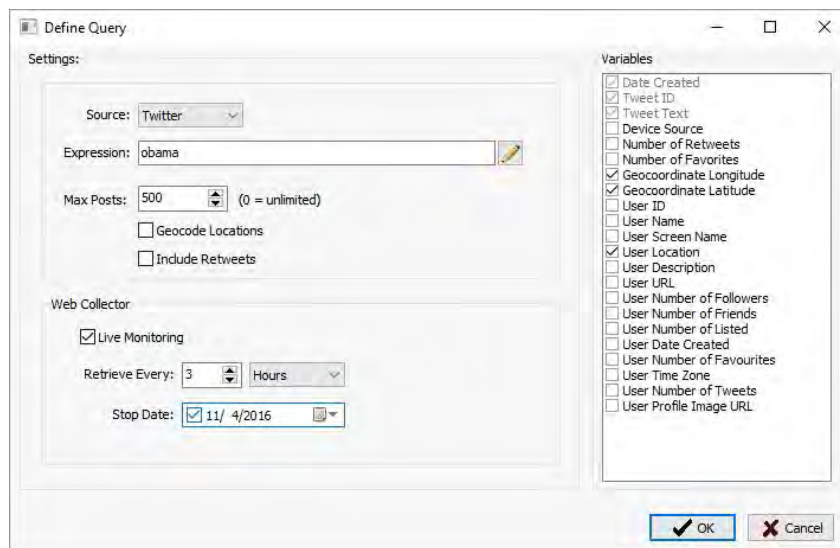
- During the exploratory phase of the analysis, keyword frequency and crosstabulation may be used to identify differences between subgroups in word usages, differences that may have remained unidentified.
- The **KWIC** list may be used to locate and visualize text associated with specific codes either 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 cases containing those words may then help identify instances where the code should have been used.
- A dendrogram of keyword cooccurrences 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 ensure a more systematic application of manually entered codes or to develop and validate a dictionary that will later be used for automatic content analysis.

Web Collector

The **Web Collector** is an external tool that can be accessed through QDA Miner WordStat, through the system tray or the Windows start menu. It monitors RSS feeds, Reddit and Twitter posts and Facebook comments. If you have chosen to live monitor your social media search, the Web Collector will collect newly published posts or comments, which can then be appended to your project.

Setting the Web Collector

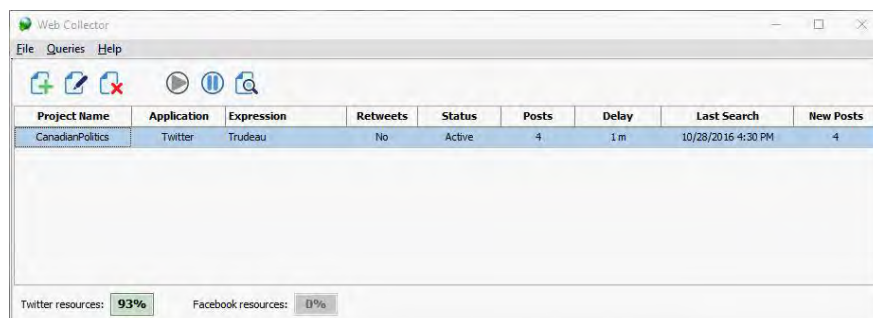
When you create an RSS, Twitter or Facebook project you have the option of monitoring the project over time. This means data will be collected by the Web Collector according to a user defined time interval until a stop date is reached or until the monitoring is disabled or deleted. Data will continue to be collected as long as the Web Collector is running. As the Web Collector is a completely separate application, QDA Miner or WordStat do not have to be running for it to work. Please see the corresponding section on how to import from [RSS](#), [Twitter](#) or [Facebook](#) for more information on creating live monitoring queries.



If you have checked **Live Monitoring** query the Web Collector will run in the background. If you restart Windows, the Web Collector will start automatically and continue collecting data as long as the query remains active.

Accessing the Web Collector

- You can access the Web Collector by clicking on the  icon in the system tray. Alternatively, click on the Web Collector menu item in your Windows start menu and the Web Collector will appear.



The Web Collector displays a list of all of the RSS, Twitter and Facebook projects that you are currently monitoring. It displays the following information:

Project name: The project associated with your query.

Application: The social media platform that is being monitored.

Expression: Either the URL of the RSS feed or Facebook page or your Twitter search query.

Retweets: For Twitter searches, this column will read **Yes** or **No** depending on whether you have chosen to include Retweets when setting up your project.

Status: Lets you know if the data is being collected. Will either read **Active**, **Inactive** or **Searching**.

Posts: The number of posts collected. This will be reset to zero if the Tweets, Facebook or RSS posts are appended to a project.

Delay: The time interval you set for your capture.

Last search: The date and time of the last data capture.

New posts: The number of new posts captured since you last viewed the data.

Twitter resources: The percentage of Twitter resources you have available is displayed in the bottom left of the screen.

Facebook Resources: The percentage of Facebook resources you have available is displayed in the bottom left of the screen.

Web Collector Functions

You can **Edit**, **Add**, **Delete**, **Pause** and **Resume** queries in the Web Collector by selecting the corresponding icon from the tool bar or by selecting the corresponding command from the **QUERIES** menu. You can also **view the data** associated with each query in the same manner.

To add a query:

- Select the query you want to modify.
- Click the  button.
- A **Define/ Edit Query** dialog box will appear.
- Add your query.
- Click **OK**.

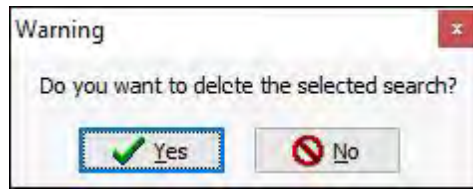
Please see the corresponding section on how to import from [RSS](#), [Twitter](#) or [Facebook](#) for more information on creating live monitoring queries.

To edit a query:

- Select the query you want to modify.
- Click the  button
- A **Define/ Edit Query** dialog box will appear.
- Add your query.
- Click **OK**.

To remove a query:

- Select the query you want to modify.
- Click the  button. A warning dialog will appear.



- Click **Yes** if you want to continue with the deletion. If you do not want to delete the query click **No**.

To pause a query:

- Select the query you want to modify.
- Click the  button
- The query status will change from **Active** to **Inactive** and the Web Collector will cease collecting data.
- To pause all queries at the same time select **PAUSE ALL QUERIES** from the **QUERIES** menu.

To resume a query:

- Select the query you want to modify.
- Click the  button.
- The query status will change from **Inactive** to **Active** and the Web Collector will resume collecting data.
- To resume all queries at the same time select **RESUME ALL QUERIES** from the **QUERIES** menu.

To view data:

- Select the query for which you would like to view the data.
- Click the  button and a **Data View** dialog will open.

Date Created	Tweet ID	Tweet Text	Number of Retweets	Number of Favors
10/7/2016 3:29 PM	784415384250224640	@RMarcelc: impeach obama...i don't care if he has less than 30 days left or whatever...CONGRESS, WHAT'S YOUR PROBLEM? SCARED SHITLESS?	0	0
10/7/2016 3:29 PM	784415400113168389	The 6 key questions for the Obama administration about the Yahoo-HSA cooperation. By @neemaguliani @ACLU https://t.co/eDX3s0BWlh	0	0
10/7/2016 3:29 PM	784415402281533441	Seven Years Ago This Week: Obama's Nobel Peace Prize https://t.co/6k5uaQA7QJ	0	0
10/7/2016 3:29 PM	784415402692489220	@realityinACTION @RepMcCaul @SenRonJohnson @PRyan @CNH_Michael Welcome to Clinton/Obama/Soros Urban Area Terrorist Relocation Program	0	0
10/7/2016 3:29 PM	784415404265381888	@grumpynaomi @purposeholy Well actually this isn't widely known but Obama was actually being Satan at the time.	0	1
10/7/2016 3:29 PM	784415405792264192	https://t.co/Ufz5lRaCp بالتحقيق يطالب : كيري جون الحزب لاجراءم بالنسبة روسيا مع	0	0
10/7/2016 3:29 PM	784415407679627264	#WOW...THIS IS WONDERFUL...#AWESOME...29 PRESIDENT OBAMA WITH HIS GRAND MOTHER IN KENYA...WATCH THIS https://t.co/eiuapY0dJZ	0	0
10/7/2016 3:29 PM	784415408732241920	Obama alcanza un nuevo récord de aprobación en su presidencia https://t.co/tyuHJlH5W https://t.co/PX7cW5Sgraq	0	0
10/7/2016 3:29 PM	784415415942287360	Newly disclosed emails prove Obama officials were in close contact with Hillary Clinton re potential fallout from her emails @KellyannePolls	0	0
10/7/2016 3:29 PM	784415420308656129	#US needs a LEADER that Knows 2 Win Enemies Not Elections ONLY!(e.g.Obama) @DTAP4INSecurity @DomusUSA... https://t.co/nriAFINETG	0	0
10/7/2016 3:29 PM	784415422850314240	Just remember, under Obama and the democrats, the US lost its AAA for the first time ever.	0	0

The table contains the variables you selected to include when you initially created your project. Posts captured since you last viewed the project will be in bold. Posts that have already been viewed will **not** be in bold.

- Select the  button to enable the search function at the bottom of the screen.
- Type your search term into the field and select **Find next** or **Find previous** button.
- Check the **Match case** checkbox if you want the text case to be exactly as you have written it in the search box.

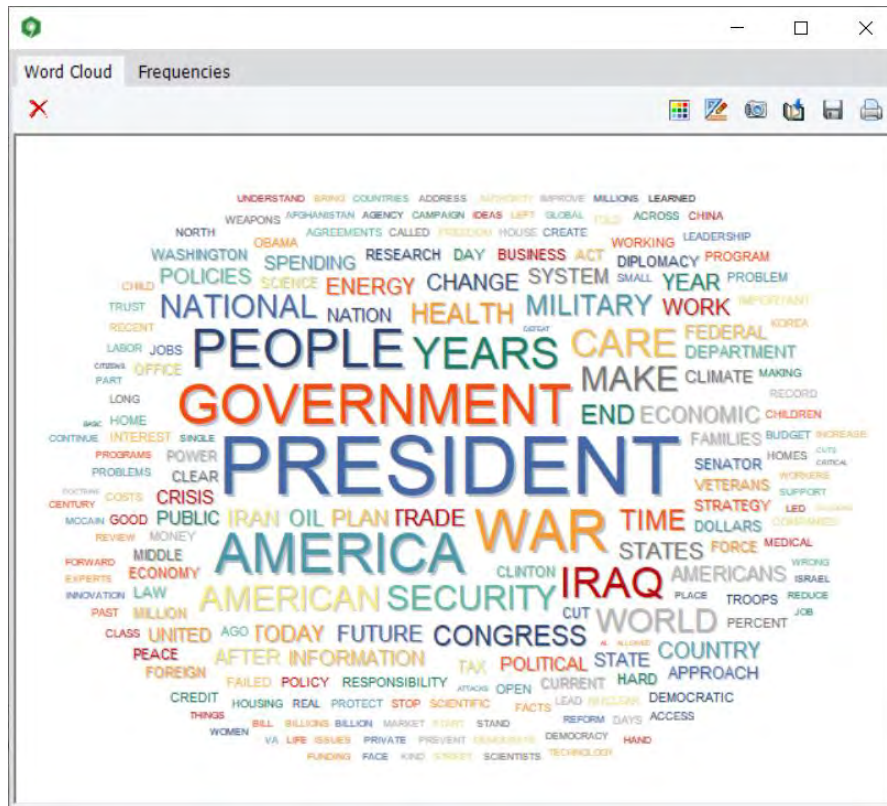
To close the Web Collector:

- Click the **X** at the top right of the dialog box.

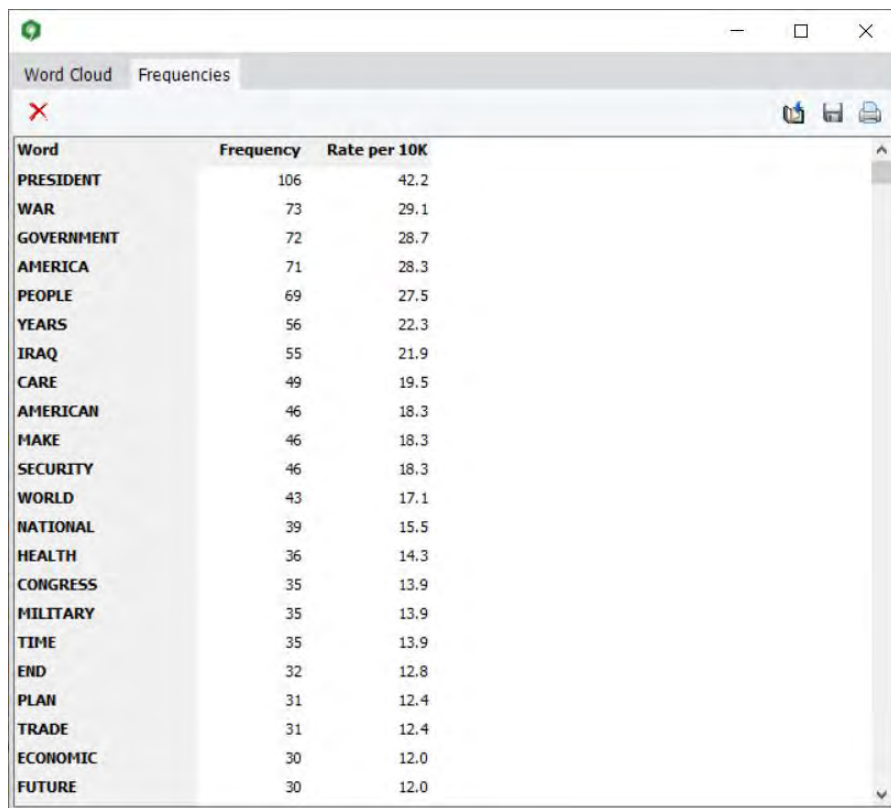
Word Frequency Analysis

WordStat can perform basic word frequency analysis allowing one to quickly get a glance of what are the most common terms in large amounts of text. A word frequency analysis may be useful to identify themes as well as associated terms that may later be used to retrieve segments related to those themes. Such a feature may also be useful to identify co-occurring words, allowing one to expand the search terms used for retrieving relevant text segments. The results are presented both visually, using a word cloud graphical display, and in a table format, for more precise frequency information.

In WordStat, the word cloud represents the 200 most frequent words, where the size of each word is proportional to its relative frequency of use in the text.



The **Frequencies** tab allows one to display the results of all the words in a table format. By default, words are displayed in descending order of frequency. A second column also displays this frequency using a rate per 10,000 words. Such a transformation facilitates comparisons across different text collections by eliminating the effect of different text sizes. To sort the table in alphabetical order, click on the first column header. To sort the table by frequency, click the second column header once for sorting the table in ascending order of frequency and click a second time to sort the list in descending order of frequency.



Word	Frequency	Rate per 10K
PRESIDENT	106	42.2
WAR	73	29.1
GOVERNMENT	72	28.7
AMERICA	71	28.3
PEOPLE	69	27.5
YEARS	56	22.3
IRAQ	55	21.9
CARE	49	19.5
AMERICAN	46	18.3
MAKE	46	18.3
SECURITY	46	18.3
WORLD	43	17.1
NATIONAL	39	15.5
HEALTH	36	14.3
CONGRESS	35	13.9
MILITARY	35	13.9
TIME	35	13.9
END	32	12.8
PLAN	31	12.4
TRADE	31	12.4
ECONOMIC	30	12.0
FUTURE	30	12.0

In WordStat, quick word frequency analysis may be performed on the text segments obtained through several text search tools, such as the keyword retrieval, the keyword-in-context page, as well as from a few additional locations.

Working With the Word Cloud

Items may be removed temporarily from the word cloud. Its appearance may also be customized.

To remove a word temporarily and redraw a new word cloud without it

- Click on the word you want to remove, and then click the  button. You may also right-click on the word and select **Remove & Redraw** from the popup menu.

Working With the Word Frequency table

The same operation to remove words may be achieved from the word frequency table (see section above for specific instructions). However, it adds the ability to perform this operation on more than one word at a time.

Customizing the Appearance of the Word Cloud

The word cloud shape, its color palette, as well as the method used to position words may be adjusted. There are two methods to adjust the colors used to display words.

Clicking the  button will allow you to select a color palette to use from a predefined list. With this method, colors are assigned sequentially and results in colors uniformly distributed across the words irrespective of their frequency. One can also select a single color to be used for all words. Clicking the  button brings a dialog box with an additional possibility to assign a gradient of colors that will be applied based on each word frequency. To enable this option, on the **Text Color** page of this dialog box, select the **Gradient Scheme** radio button and select the colors to be used for high frequency (**From**) and for those with low frequency (**To**). An optional third color may be used to represent **Middle** frequency. The following table presents additional options available.

Font	This option allows you to specify the font to be used to display words.
Shadow	Enabling this option displays a small drop shadow behind each word, creating a subtle 3D effect giving the impression that the words are raised above the background.
Form	This option allows you to choose to display words in inside a rectangular or an elliptical shape.
Starting position	When drawing a word cloud, WordStat starts positioning high frequency words and then attempts to place all the other words by decreasing order of frequency. When the starting position is set to from center it will position the most frequent one at the center of the cloud, and progressively position the following ones around it so that lowest frequency words will most likely be plotted on the edge of the cloud, far from the center. Setting this option to Random will position words in a more random fashion such that high frequency words may be more evenly distributed within the cloud.

WordStat Software Development Kit (SDK)

WordStat Text Classification Process.

Text mining, as performed by WordStat, involves some form of quantification of text data. Such quantification is achieved by applying natural language processing techniques (stemming, lemmatization, removal of stop words, etc.), statistical selection criteria, as well as grouping of words and phrases into concepts using either lexicons or custom content dictionaries. All these procedures can be combined to extract numbers representing the presence or frequency of important keywords, or key concepts. We call this the **categorization process**. WordStat also supports another form of quantification: **automatic document classification**, by which documents are categorized in one of several mutually exclusive classes using some form of machine learning.

Why a Software Development Kit?

The categorization and classification processes are performed by WordStat, which offers a graphical user interface that allows a user to create, validate and refine those processes, apply these to various text collections, perform comparisons, explore, relate and create graphical and tabular reports. While categorization and classification models can be saved to disk and reapplied on a different set of documents, a human operator is still required to perform those analysis, limiting the ability to fully automate the text analysis and reporting operations.

The WordStat software development kit (SDK) provides a solution , allowing models developed with the WordStat desktop tool to be used in other applications written in other computer languages such as C++, Delphi, C#, VB.Net and so on.

An example of such integration would be the application of a categorization model on a company data collection system of customer feedback in order to automatically measure references to specific topics and to classify those feedback as either positive, negative or neutral.

Applying the SDK

All the analysis and text transformation settings set in WordStat are stored on disk in the model files (stemming, lemmatization, categorization rules, selection criteria, etc.). This greatly simplifies the integration of such text processing in other applications by reducing the application of those text analysis process to four easy steps:

1. Load the categorization or classification model file
2. Retrieve the text to categorize or classify
3. Apply the model to the text
4. Retrieve relevant information (frequencies, probabilities, predicted classes, etc.)

A model only needs to be loaded once (step #1), while steps #2 to #4 may be repeated as often as needed.

There are currently no reporting or graphing functions available in the DLL, so it is the task of the programmer to further process the obtained information. Typically, numerical values would be either stored in a database or cumulated to create reports, dashboards, etc..

Technical Details

The SDK consists of a Windows DLL available in both 32 bits and 64 bits versions. The DLL is multi-thread safe, allowing text quantification of multiple documents concurrently. It also supports the simultaneous application of multiple categorization and classification models, allowing one to perform several quantifications of the same documents.

The SDK comes with a sample project with source files illustrating how integration can be achieved. This sample project is currently available in Delphi, C# and VB.NET. Please contact us if you need assistance on how to use the SDK with other computer languages.

Interested?

If you are interested in obtaining more information about the SDK, in getting a trial version (with documentation and sample applications) or to purchase it, please contact developpers@provalisresearch.com.

Report Manager

The Report Manager is a separate application that has been designed to store, edit and organize documents, notes, quotes, tables of results, graphics and images created by QDA Miner or imported from other applications. Items can be added to the Report Manager directly from QDA Miner without needing to run the Report Manager.

The  button, found in many locations in QDA Miner, may be used to copy entire documents, tables and charts to the Report Manager. Selected text segments or image areas may also be appended by clicking the  button.

To access the Report Manager from QDA Miner, run the **REPORT MANAGER** command from the **PROJECT** menu.

The program presents its information as an outline, allowing a hierarchical organization of miscellaneous pieces of information that is ideal for project management, organizing ideas, structuring information, or designing and writing a research report.

The workspace emulates the appearance of Windows Explorer or of a standard Help file with the Table of Contents (TOC) on the left and the Editor on the right.

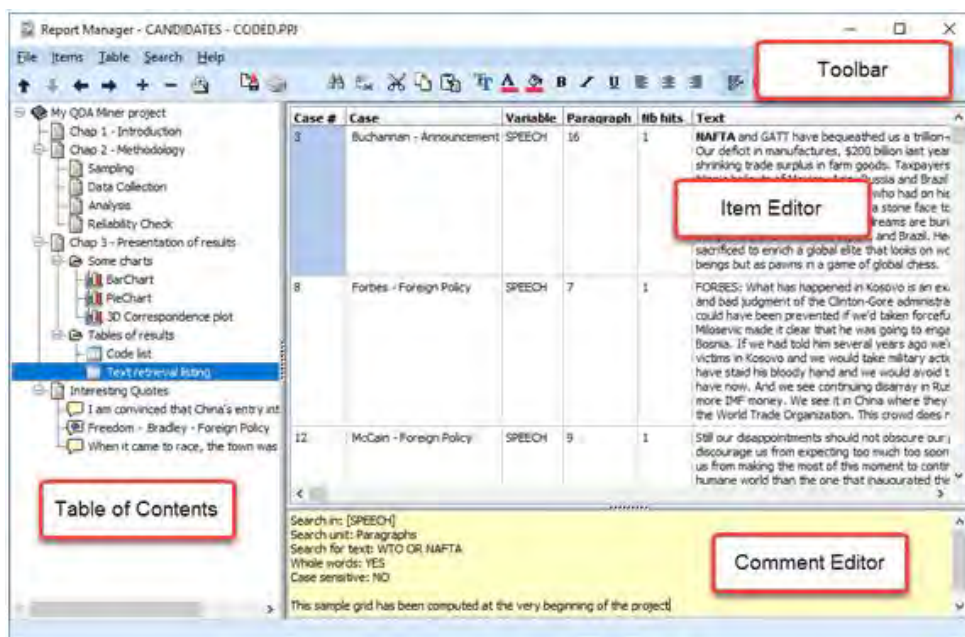


Table of Contents panel

Report Manager files are made up of items or topics that are like chapters in a book. Each item can be thought of as a separate word processor, a table or a graphic file editor or viewer, all of which are stored together in the QDA Miner project file. This panel provides powerful functions for organizing items and structuring the information in a hierarchical manner.

Item Editor

The largest panel on the right of the program window is the Item Editor, which is like a built-in word processor. This is where the item selected in the Table of Contents can be edited. Clicking a Table of Contents item displays its contents for editing.

Toolbar

The Toolbar provides quick access to the most frequently-used functions. Just position the mouse over a tool button and wait for the display of a brief text describing its function.

Comment panel

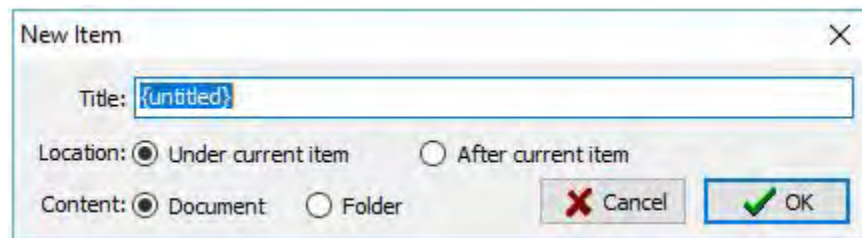
The Comment panel below the Item Editor allows the insertion and editing of comments related to the selected topic. When new items are added to the Report Manager from QDA Miner, a default comment is often already present, providing useful information about the origin of this item.

Working with the Table of Contents

Creating a New Item

To create a new item:

- Select the Table of Contents entry that will be the "parent" or "sibling" of the new item.
- Select the **NEW** command from the **ITEMS** menu or click the **+** button. A dialog box like this one will appear:



- Enter the title for the new item.
- If the new item should be a "child" of the selected item, click **Under Current Item**; if the item should be positioned after the current item, then set it to **After Current Item**.
- Select whether the new item will be a **Document** or a **Folder**. Folders are empty items that are used as containers for other items.
- Click **OK**. The new item will become the current one.

Importing Items from Files

To import items from files:

- Select the item under which the imported items will be stored.
- Select the **IMPORT FILES** command from the **ITEMS** menu or click the  toolbar button. An Open File dialog box will appear.
- Select the type of data you would like to import by selecting the appropriate Files of Type list box option. The Report Manager can import the following data types:
 - DOCUMENTS - Plain text (.TXT), MS Word (.DOC), WordPerfect (.WPD), Rich Text (.RTF) or HTML files (.HTM or .HTML)

- GRAPHICS - Windows Bitmap (.BMP), Windows Metafile (.WMF), JPEG files (.JPG or .JPEG) and Portable Network Graphic files (.PNG)
- CHARTS - QDA Miner or WordStat Charts (.WSX)
- DELIMITED DATA - Tab delimited (.TAB) or Comma Separated Value (.CSV) data files.
- Select one or several files to be imported and click the **Open** button.

Renaming an Item

To rename an item:

- Select the item to be renamed.
- Select the **RENAME** command from the **ITEMS** menu or click the  toolbar button.
- In the Item Title Dialog, change the title.
- Click **OK**.

Deleting an Item

To delete an item:

- Select the item to delete in the Table on Contents.
- Select the **DELETE** command from the **FILE** menu or click the  toolbar button.
- You will be asked to confirm that you really want to delete the item. If you're sure, then click **Yes**.

NOTE: Be aware that you cannot undo this if you make a mistake.

Moving an Item

As more items are created and the Report Manager hierarchy grows, it is inevitable that you will want to move items around, either to place one item under another, or to promote one to a higher level.

To move items using drag-and-drop:

The easiest way to move items in the Table of Contents is by using drag-and-drop operations. Using the mouse, you can move an item to a different location or move a group of items stored under a "parent" item by dragging this "parent" item to its new location.

- Select the item to move by clicking and holding down the left mouse button. (Keep the mouse pressed until the drag-and-drop operation is completed.)
- Drag the item to its new location and, only then, release the mouse button.
- The dragged item will now become a "child" of the destination item.
- To move the item to the same level as the item under the cursor, simply hold the ALT key while dropping the dragged item.

To move items using the toolbar:

You can also use menu commands and toolbar buttons to move items. To promote an item is to move it to a higher level in the hierarchy, making the item a "sibling" to its former "parent". To demote an item is to move it to a lower level and make it a "child" of its previous "sibling". After selecting the item you want to move, use one of the following four commands:

- To promote the selected item, click the  button, or select the **PROMOTE** command from the **ITEMS** menu.
- To demote the selected item, click the  button, or select the **DEMOTE** command from the **ITEMS** menu.
- To move the selected item up relative to its siblings, click the  button or select the **MOVE UP** command from the **ITEMS** menu.
- To move an item down relative to its siblings, click the  button or select the **MOVE DOWN** command from the **ITEMS** menu.

Adding or Editing Item Comments

The Comment panel below the Item Editor allows one to insert a new comment or edit an existing one related to a selected topic.

To type a new comment or to edit one, simply click in the yellow region of this panel and start to type. The comment is automatically saved as soon as you move to another item or leave the Comment panel. While there is no menu item or toolbar icon associated with this feature, standard clipboard operations are supported for copy and pasting text. A popup menu with standard editing features can also be obtained by clicking the right mouse button.

To search for text in comments, see the information on the [Global Search](#) command.

Editing Documents

The Report Manager offers many editing features to create and edit both simple text documents and documents with complex formatting, as well as tables and graphics. When a document item is selected in the Table of Contents, a **DOCUMENT** menu appears, displaying all available formatting and editing options. A similar menu can also be obtained by right-clicking anywhere in this document. The toolbar portion directly over the editing area also displays buttons to access the most often-used editing and formatting functions.

Individual documents may also be printed or exported to disk in various file formats such as plain text, Rich Text or HTML format. An **IMPORT** command is also available to read a document file stored in plain text, Rich Text, MS Word, WordPerfect, HTML and a few additional formats. Executing such a command will replace the existing content with the content of the imported file.

Editing Tables

The Report Manager offers many editing features to customize the appearance of a table, to change the text alignment or font setting, to set the cell background color, or to delete entire rows or columns. When a table item is selected in the Table of Contents, a **TABLE** menu become visible displaying all available formatting and editing options. A similar menu can also be obtained by right-clicking anywhere in this table. The toolbar portion above the editing area also displays buttons to access the most often-used table editing and formatting functions.

Individual tables may also be printed or exported to disk in various file formats such as ASCII (*.TXT), tab delimited file (*.TAB), comma delimited (*.CSV), MS Word (*.DOC), HTML (*.HTM; *.HTML), XML (*.XML) and Excel spreadsheet file (*.XLS).

Editing Charts

Many charts saved in the Report Manager may be edited using many of the same options as those available in QDA Miner, such as the multidimensional scaling plot obtained through the **CODE CO-OCCURRENCE** analysis command, the correspondence analysis plots, and the bar charts and line charts created by the **CODING BY VARIABLES** command, as well as the bar charts and pie charts produced by the **CODING FREQUENCY** command. To obtain information on the display options available for those charts, see their corresponding page in this manual. Other charts - such as dendrograms and heatmaps - are stored as image files, so cannot be modified. However, just like other charts, they may be exported to disk in various file formats such as BMP, JPG or PNG graphic files.

Searching and Replacing Text

Two broad types of text search are available in the Report Manager. A local, item-based search-and-replace feature allows one to perform text searches and replacements on individual documents or tables, and a global search engine for searching text patterns in several or all documents, tables and comments in the Report Manager.

To perform a search in a single document or a table:

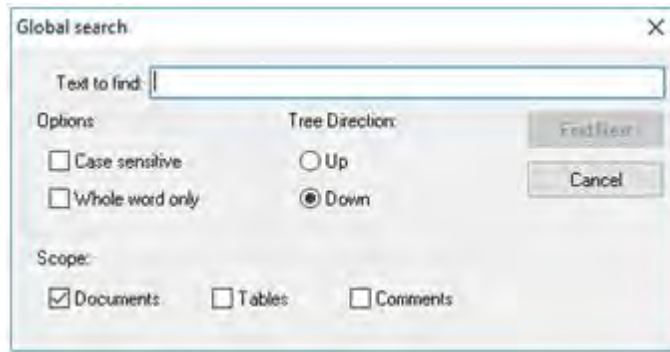
- Select the document or table you would like to search, by selecting its entry in the Table of Contents.
- Position the editing cursor in the document or select the cell in the table where you want the search to begin.
- Select the **FIND** command from the **SEARCH** menu, or click the  button.
- Enter a search expression, set the desired search options and then click **Find Next**.
- To find additional instances of the same text, continue to click **Find Next**.

To replace text in a single document or table:

- Select the document or table in which you would like to perform the text replacement by selecting its entry in the Table of Contents.
- Position the editing cursor in the document or select the cell in the table where you want the search to begin.
- Select the **REPLACE** command from the **SEARCH** menu, or click the  button.
- In **Find What**, type the characters or words you want to find. In **Replace With**, type the text you want to replace it with. Set the desired search options and then click **Find Next**. Click **Replace** to change the selected text. To replace all instances of the text, click **Replace All**.

To perform a global search:

Select the **GLOBAL FIND** method from the **SEARCH** menu. A dialog box similar to this one will appear:



The **Text to Find** edit box allows you to specify the text you want to find. The **Case Sensitive** and **Whole Word Only** options function in the same way as in a standard word processor.

The search starts at the current topic item. Select **Forward** to continue searching items below the current one, or **Backward** to move up and search items above the current document or table in reverse order. To search all items, select the top item in the Table of Contents before using the Global Search dialog box.

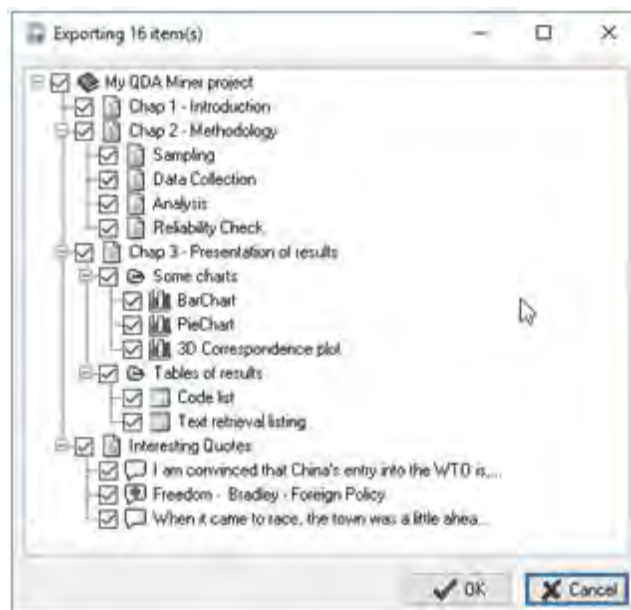
The **Scope** option box is used to specify what is to be searched. You can restrict the search to **Documents**, **Tables**, or **Comments** attached to items, or any combination of these three.

Once the search options have been set, click the **Find** button to start the search as well as to continue searching for additional instances of this text.

Exporting items to HTML or Word

Individual documents, tables, graphics and images may be exported to disk in numerous formats. Such an exportation can be achieved by clicking the  toolbar button or by selecting the **EXPORT** command from the associated menu.

The Report Manager also offers the possibility of exporting the entire content or selected items into a single HTML or MS Word document. The exportation is achieved by selecting the proper command from the **FILE | EXPORT** menu. For example, to export items to HTML, select the HTML command. A dialog box similar to this one will appear.



By default, all items are marked for export. To prevent some items from being included in the exported file, simply remove the check marks beside them. Clicking a "parent" item affects all "children" items in the same way. To unselect all items, uncheck the project item located at the very top of the tree.

Once the selection process is completed, click the **OK** button. A Save File dialog box will be displayed, allowing you to enter a file name and select the location where the file should be saved. After the file is created, you will be asked if you want to view this file. Clicking **Yes** will open a web browser if the exported file is an HTML document or either MS Word or Wordpad if the exported file is a Word document.

References on Content Analysis

Selected references on content analysis

- ALEXA, M. (1997). *Computer-assisted text analysis methodology in the social sciences*. Mannheim, Germany: ZUMA.
- EVANS, W. (1996). Computer-supported content analysis: Trends, tools, and techniques. *Social Science Computer Review*, 14 (3), 269-279.
- KRIPPENDORFF, K. (2019). *Content analysis: An Introduction to its methodology* (4th ed.). Los Angeles, California: Sage Publications.
- LEBART, L., SALEM, A. & BERRY, L. (2011). *Exploring Textual Data*. Dordrecht, Netherlands: Springer.
- NEUENDORF, K.A. (2016). *The Content Analysis Guidebook*. Beverly Hills, California: Sage Publications.
- WEBER, R. P. (2008). *Basic Content Analysis*. Second edition. Quantitative Applications in the Social Sciences, vol 49. Newbury Park, California: Sage Publications.
- WEBER, R. P. (1983). Measurement models for content analysis. *Quality and Quantity*, 17 (2), 127-149.
- WEBER, R. P. (1984). Computer-aided content analysis: A short primer. *Qualitative Sociology*, 7 (1-2), 126-147.

Selected references on text mining

- INGERSOLL, G.S., MORTON, T.S. & FARRIS, A.L. (2013). *Taming Text: How to Find, Organise, and Nanipulate it*. Manning Publications.
- IGNATOW, G. & MIHALCEA, R. (2016). *Text Mining: A Guidebook for the Social Sciences*. Sage Publications.
- IGNATOW, G. & MIHALCEA, R. (2017). *An Introduction to Text Mining: Research Design, Data Collection, and Analysis*. Sage Publications.
- WIEDEMANN, G. (2016). *Text Mining for Qualitative Data Analysis in the Social Sciences*. Springer Fachmedien Wiesbaden.

Multivariate analysis

- GREENACRE, M. (1989). *Theory and Applications of Correspondence Analysis*. London, England: Academic Press.

Other

- GREFENSTETTE, G. (1994). Corpus-Derived First, Second and Third-Order Word Affinities. In W. Martin, W. Meijs, M. Moerland, E. ten Pas, P. van Sterkenburg, and P. Vossen, editors, *Proceedings of EURALEX'94*, Amsterdam, The Netherlands.
- SEBASTIANI, F. (2002). Machine Learning in Automated Text Categorization. *ACM Computing Surveys*, 34(1):1-47.