New features of WordStat 2025

Using Generative AI features

Generative AI has emerged as a powerful tool for text analysis, offering advanced capabilities for data transformation, data analysis, interpretation, and post-processing. These models can assist researchers in summarizing content, identifying patterns, categorizing entities, and even refining results. However, despite their impressive capabilities, generative AI models are not infallible—they can introduce biases, produce inconsistent outputs, and lack full transparency in their reasoning. While generative AI can be a valuable tool for text analysis, it struggles with processing large text corpora due to memory limitations, slow processing speeds, and potentially high costs. These constraints make it impractical for tasks requiring large-scale text mining, where more efficient and scalable approaches—such as those found in software like WordStat remain essential for researchers handling extensive datasets. Given these strengths and limitations, a careful and informed approach is essential when integrating AI into research workflows.

Our implementation of generative AI in WordStat is designed to be open and transparent, giving researchers full control over how these models are used. We provide access to multiple AI engines (ChatGPT, Gemini, Claude, and Perplexity, others to come), allowing users to choose the model that best fits their needs. Unlike black-box approaches, we make all prompts publicly visible and fully editable, ensuring that users understand and control how AI interacts with their data. Researchers can modify our predefined prompts or create their own for tasks such as data transformation, content analysis, post-processing of topic modeling or phrase extraction results, and more. This flexibility empowers users to tailor AI assistance to their specific research needs while maintaining transparency and reproducibility in their analyses.

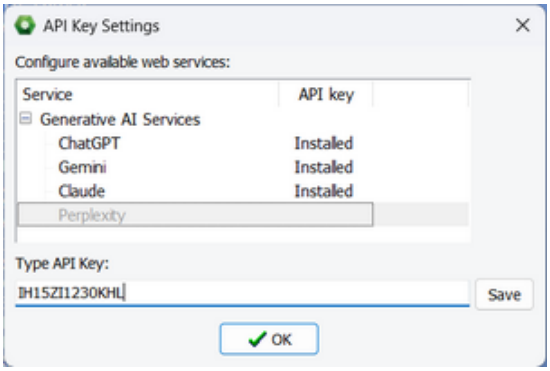
WordStat supports two broad types of AI scripts that can be used and created:

1) Post-Processing Scripts - These scripts operate on specific table outputs generated by WordStat, allowing for further analysis of results such as topic modeling outcomes, extracted phrases, named entities, or lists of potential misspellings. While built-in WordStat post-processing scripts can append additional data to existing tables, user-defined scripts generate responses in text format only. Compared to the second type of script, post-processing scripts are generally easier to create since they do not require instruction specific to the returned data type.

2) **Data Transformation and Analysis Scripts** - Unlike post-processing scripts, these scripts process the original documents rather than the tables generated by WordStat. They can be applied to a single document or multiple documents within a case, with the output displayed in a text editor window where users can review, edit, copy, or save the results. Additionally, these scripts can be executed on all cases in a project, automatically storing responses in new project variables. Extracting AI-generated data into structured variables (e.g., text, numeric, or categorical) requires precise configuration and specific instructions to ensure successful data capture. For this reason, they require more attention and more information when created.

Configuring the API Keys

To begin using AI in WordStat, the first step is to obtain an API key from the engine or engines you wish to use. Once you have the key, click on any AI button displayed at various locations to display the AI ASSISTANT menu. navigate to **SETTINGS | API KEYS SETTINGS**. This will open a configuration panel similar to the one below.



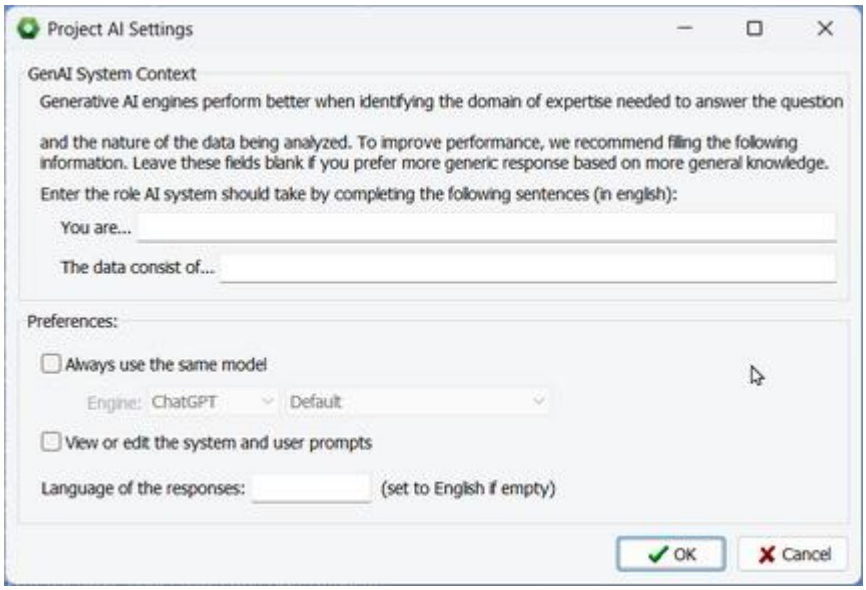
To enter an API key, select the desired AI engine from the list of supported web services, type the key into the **Type API Key** edit box, and then click the SAVE button. The key is encrypted and securely stored in the computer registry. Once a key is successfully added, the word **Installed** will appear next to the AI service name.

If you need to remove a key, access this dialog box again, select the engine, leave the **Type API Key** edit box empty, and click the CLEAR button.

Setting project-specific AI settings

Generative AI engines perform better when they understand the domain of expertise required and the nature of the data being analyzed. Defining these elements helps the AI generate more relevant and precise responses. To improve performance, we recommend specifying those two types of information by accessing the AI **Project Settings** dialog box. The project-specific configuration feature also allows one to set a specific engine and model to be used for all AI operation performed on this project and whether predefined prompts should be shown before being executed, allowing one to edit those if needed.

To access the AI project setting click on any AI button displayed at various locations to display the AI ASSISTANT menu. Then select PROJECT SETTINGS from the SETTINGS menu item. A dialog box similar to the one below will appear:



The following options can be set

You are... - Setting a default role and project description ensures consistency across different analyses, reducing variability in AI-generated outputs. This is particularly useful for research projects that require standardized interpretations. To enter a role, simply complete the sentence **You are...** by typing what kind of expertise the Gen AI engine should have to answer the question. We strongly recommend typing such information in English even if the response provided should be in another language, since this information will be inserted in predefined English scripts. An additional language instruction can be defined if you want the response to be in another language (see below).

The data consist of... - This option lets you give basic information about what the text data in this project is about or where it comes from. When left empty, no information regarding the data is inserted in prompts generated by WordStat. Like the previous option, complete the sentence by typing basic information about the origin and the type of text information the current project contains.

Always use the same model - While multiple AI engines can be configured in WordStat, there may be situations where you want to restrict analysis to a specific engine and model. Enabling this option allows you to set a default AI engine and model for all analyses, ensuring consistency and avoiding the need for selection each time. If left unchecked, a dialog box will appear before executing a script, allowing you to choose the engine (if multiple are available) and model for each operation.

View/Edit the System and Users Prompts - By default, scripts are executed automatically without allowing you to review or modify them beforehand. This option enables you to preview and edit predefined scripts before execution, giving you greater control over AI-generated outputs.

If the **Always use the same model** is enabled and this last option is disabled, predefined scripts will be performed immediately without any dialog box.

Language of the responses - By default, AI-generated responses are in English. To receive responses in a different language, enter the desired language name in the edit box.

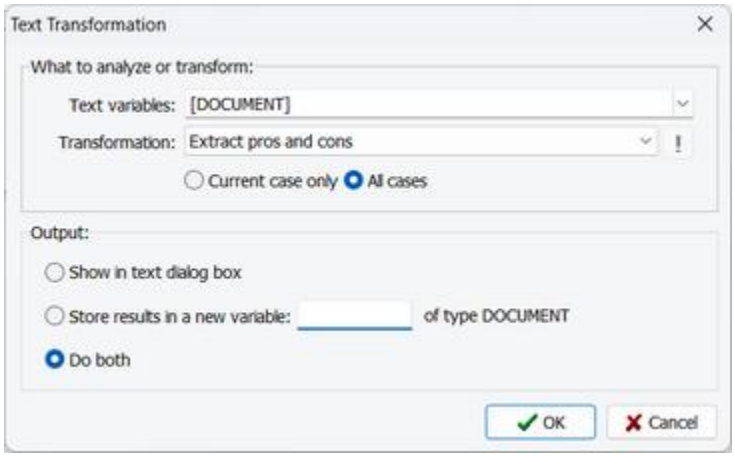
Click **OK** to save the various options.

Running an Existing Data Transformation Script

WordStat offers several predefined text transformation and analysis scripts. These scripts are accessible from the **Data** page, which is visible only when running WordStat as a standalone application. They are not available when using WordStat as an add-on to QDA Miner, SimStat, or Stata.

There are two methods to access those scripts: You may right-click anywhere in the data grid of the **Data Editor** page and select the **AI Assist..**. menu item or click the **AI Assist** button on the **Structure** page.

Then select the **Analyze/Transform Data** button. A dialog box similar to the one below will appear:



Text Variables - Select the text variables to pass to the generative engine for analysis or transformation. You may select multiple variables; the software will combine text from all selected variables into a single input.

Script - Choose the text analysis or data transformation script to apply. Not all scripts are listed by default; additional specialized or user-defined scripts can be added using the Select/edit/create data scripts feature.

Output Settings

The next options allow you to specify whether the script should be applied to the text of the current case only or on all cases. When set to **Current Cases only**, the result will be returned in a text editor. When the **All cases** option is chosen, one has to choose among three output options:

- **Show text in dialog box** will display the engine response in a text editor, allowing you to view, edit, save or copy and paste the response elsewhere.

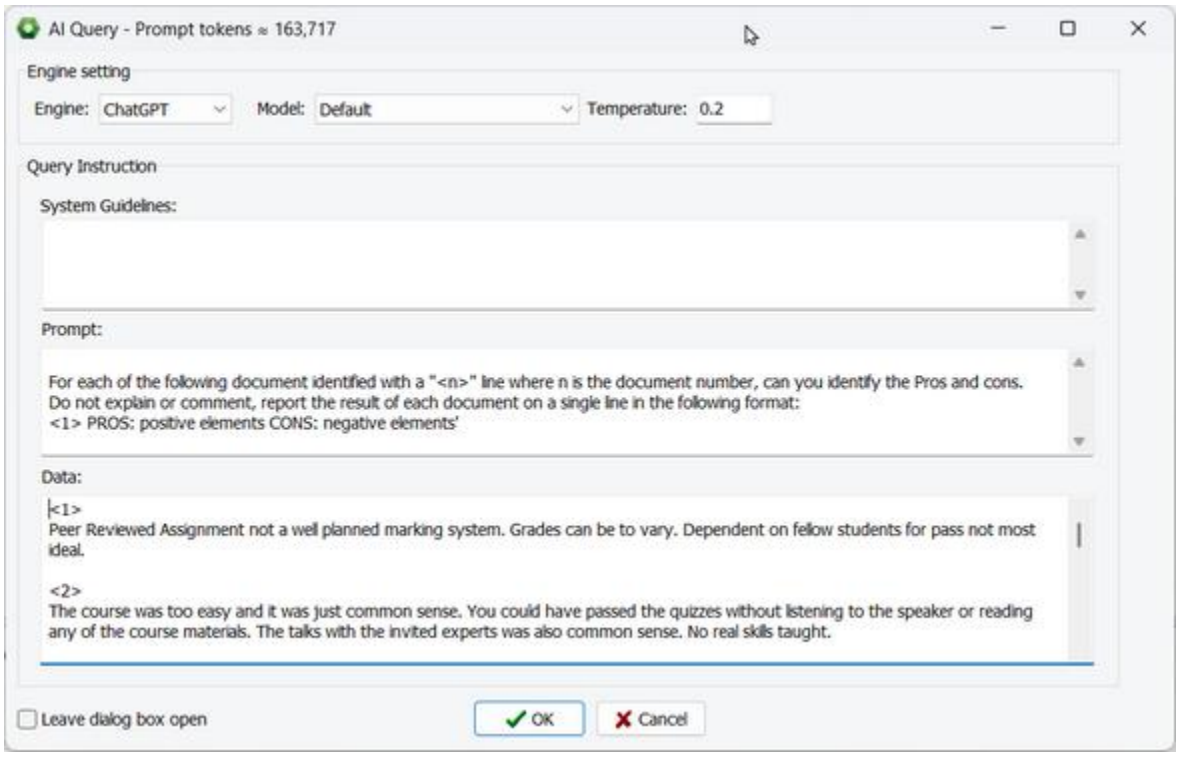
- **Store results in a new variable** will require you to type a variable name. The name must not conflict with any of the currently selected text variables but can overwrite another existing one. If the provided name does not exist, it will

be created. The expected response type and the corresponding output variable type are displayed to the right of the variable name field. In the example above, the response will be stored in a **document** variable.

- **Do both** will store results in a variable as well as display the results for all cases in a text editor.

After configuring these settings, click the **OK** button.

If a specific engine and model have been set for the current project and the option to view and edit the script before execution is disabled, the software will immediately apply the selected script. Otherwise, a dialog box similar to the one below will appear:



The top **Engine setting** panel will be visible if no specific engine and model have been enforced. If multiple engines are configured, the **Engine** list allows you to choose which engine to use for the current script. You can also select the specific model and set the temperature parameter.

The **Query Instruction** section is displayed only if the option to view and edit predefined scripts is enabled. It consists of three text segments:

System Guidelines - This edit box contains general instructions for the system, such as its role or expertise. This helps the AI engine determine the necessary expertise to answer the question or perform the desired task.

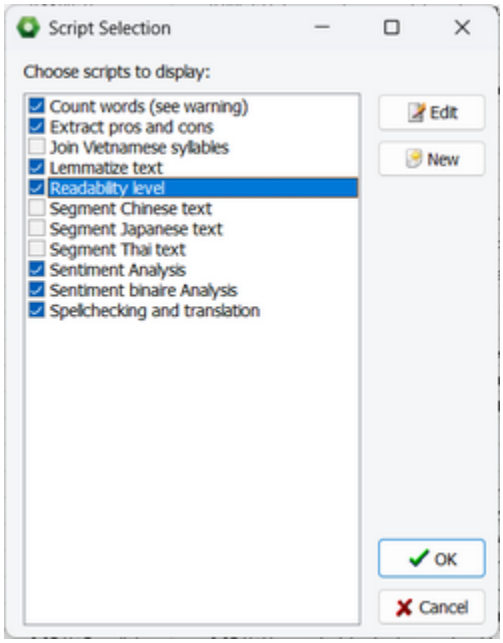
Prompt - This option provides specific text instructions on what the AI engine should accomplish for the current task. It may include a description of the data and its input format, a detailed description of the task, the desired output format as well as instructions regarding the language of the response.

Data - When running a predefined script, this section is read-only to prevent changes to the default format of each case. For user-defined scripts, this section becomes editable.

Selecting, Editing, and Creating AI Transformation

The **Analyze/Transform Data** dialog box may not display all available WordStat scripts, as some specialized scripts might not be relevant to every user. You can customize this list by adding existing scripts or creating new ones to suit your data transformation and analysis needs.

To access this script management dialog box, click on any AI button located throughout the application to open the **AI Assistant** menu, then choose **Select/edit/create data scripts**. A dialog box similar to the one below will appear.



To modify scripts available for execution:

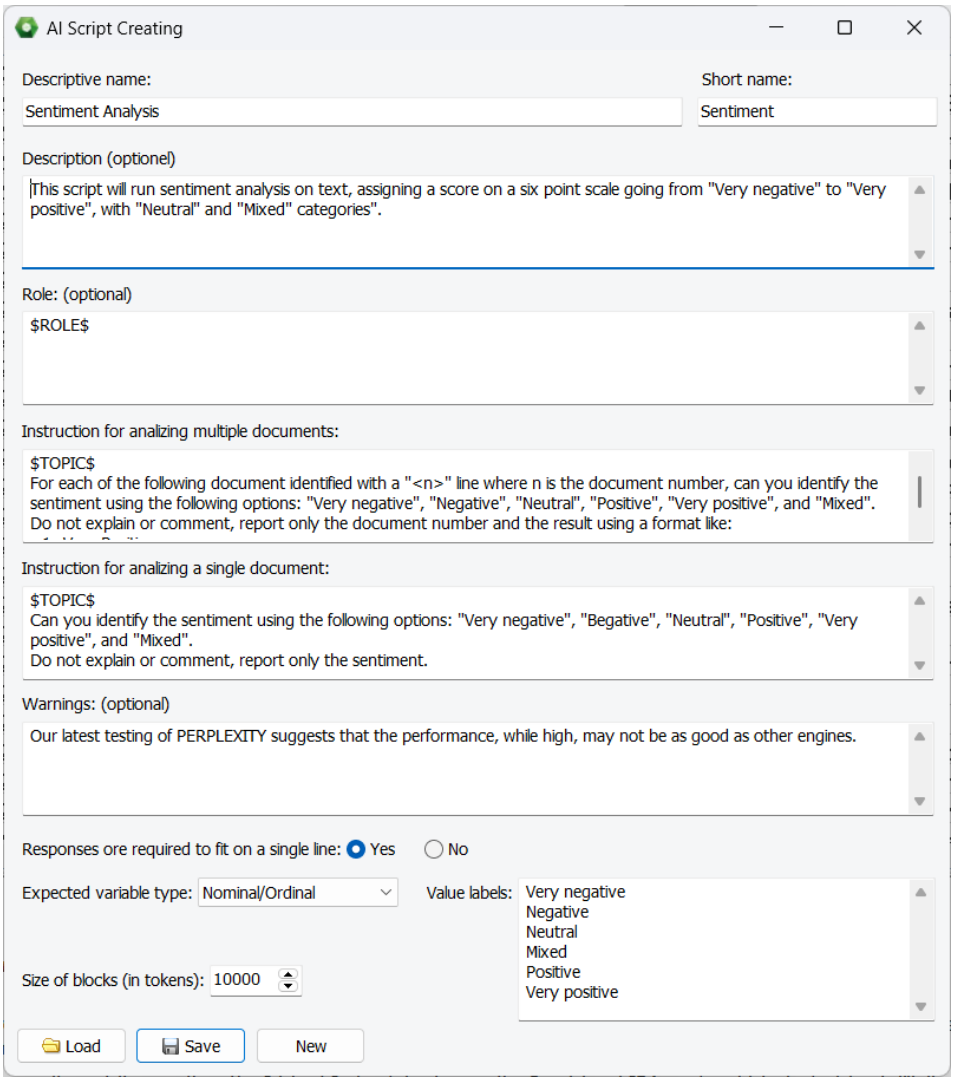
- Place a checkmark beside the script name you want to add to the Analyze/Transform Data dialog box.
- Remove the checkmark beside the script name you wish to exclude from the list.

To edit an existing script:

- Select the script you want to modify from the list and click the **Edit** button to open a script editing dialog box (see below)
- Make the desired changes to the script elements then click the **Save** button.
- If the **Descriptive Name** field remains unchanged, you will be prompted to confirm overwriting the existing script. To save the modified script as a new one without overwriting the original, change the name before clicking **Save**.

To create a new script:

- If you want to create a script that shares some features with an existing one (instructions, description, output format), follow the steps above to edit an existing script and make sure you change the **Descriptive name** option before saving the new script.
- To create a new script from scratch, click the **New** button. A blank script editing dialog box will appear. Refer to the detailed descriptions of each option below for guidance.



Detailed Descriptions of Script Options:

Descriptive Name - This edit box allows you to type a name for the script. This is the text that will be displayed in the drop-down list box of the **Analyze/transform data** dialog box. Make sure you choose a good descriptive name. Each script must have a unique name and the name cannot contain the following characters: **`* : / \ " | ?`**. If you enter the name of an existing script, clicking the **Save** button will overwrite it.

Short name - This option allows you to set a short name that will be used to monitor token usages. You can use a specific name that will be unique to your script or choose instead a more generic name that may be used by other related scripts, allowing you to monitor usage associated with several scripts of similar nature (ex. "data cleaning" for scripts for spelling correction, translation, named entity resolution, etc.)

Description - This text box can be used to offer a detailed explanation of the script's function. This is optional but recommended for clarity.

The next three text fields are used by WordStat to compose the prompt sent to the generative AI engine. They contain instructions that will later be sent to the engine. Predefine text may be inserted in those instructions with the use of the following placeholders:

$ROLE$ - This placeholder will be replaced by the project-specific role that was set by the user in the AI project settings dialog box. The inserted sentence will start with "You are..."

$TOPIC$ - Tying this placeholder in any of the edit boxes below will insert the data description set by the user for the specific project. The inserted sentence will start with "The data consist of..."

$LANGUAGE$ - This placeholder may be used to insert the desired language name in the script. For example, if the user set the language for the project to Spanish, the following prompt:

If the text is in $LANGUAGE$ please fix any spelling mistakes. Otherwise translate it to $LANGUAGE$

will become:

If the text is in Spanish please fix any spelling mistakes. Otherwise translate it to Spanish

Role - This field is optional. If left empty, WordStat will use the default role set by the user (if any). Entering a value here will override the default role for this specific project. Overriding the default role may be necessary if the task requires expertise different from the project's general role. For example, while a project may typically involve a marketing analyst or political scientist, certain scripts may require specialized skills, such as those of a linguist, psychologist, or HR specialist. Use the $ROLE$ placeholder to include in a custom role description the project-specific role set by the user.

Instruction for analyzing multiple documents - When possible, WordStat processes multiple cases at once instead of analyzing them individually. This approach saves time, reduces costs, and minimizes errors caused by excessive requests to the AI service. For this to work effectively, the script must specify how texts from different cases should be distinguished and define the desired output format. A well-crafted prompt should be clear, concise, and provide sufficient context for a relevant response. Note that the text to be analyzed is automatically appended at the end of the prompt, so there is no need to include it manually.

Instruction for analyzing single documents - This edit box allows you to specify the prompt to be sent to the engine when used for analyzing a single document.

Warning - This optional field allows you to display a warning message to alert users or remind yourself of potential issues that may cause the script to return invalid data. Some scripts may not perform adequately with certain AI engines or models. If you are aware of such limitations, it is advisable to include a warning to inform users about possible performance issues with specific data types or AI engines.

If a script definition includes a warning, the button next to the selected script name will become active, allowing the user to view the message.

Response requested to fit on a single line - If the expected response for a case is numerical, categorical, or a short string, WordStat will attempt to read it directly from the line starting with the case number. To ensure correct processing, it is important to instruct the AI engine to provide the response on a single line. Check this box to confirm that the prompt for multiple documents includes instructions to format the response accordingly.

Expected Variable type - When analyzing or transforming data, WordStat needs to know the appropriate variable type to store the AI script's output. This option lets you choose from various data types, including documents, short strings, integers, floating-point numbers, or categorical variables.

If you select **Categorical**, a **Data Labels** edit box will appear, allowing you to enter all expected response values. Each label should be listed on a separate line and properly ordered if the variable represents ordinal values. To ensure consistency, it is strongly recommended to specify these same values in the script instructions. This helps prevent AI-generated responses from including unrecognized variants.

Size of blocks (in tokens) - When processing documents from multiple cases, WordStat may split the data into smaller blocks, passing each block with the relevant instructions. WordStat monitors the total size of the data and the prompt, ensuring that the resulting size does not exceed the limit set by this option. If the limit is exceeded, WordStat will initiate a new prompt with the next case and potentially subsequent ones.

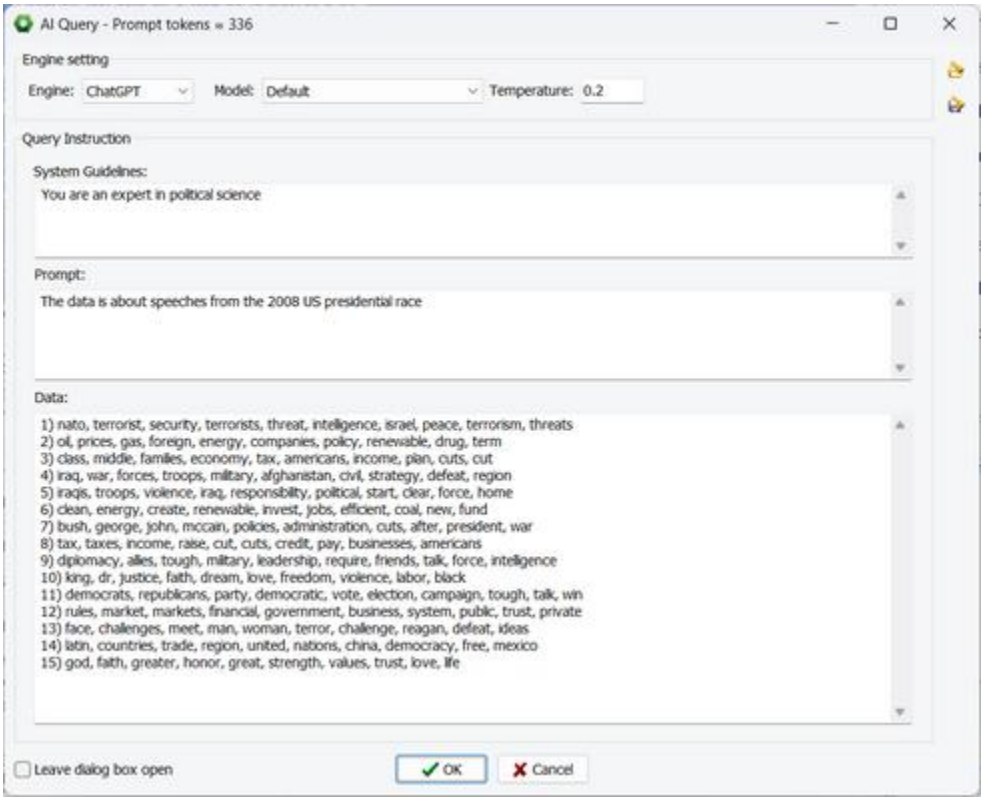
The block size can be set between 10,000 and 100,000 tokens. Smaller block sizes require more API calls but offer faster responsiveness, updating progress more quickly. This comes at the cost of slightly higher token usage (about 2% to 5%) and a small increase in processing time.

It is important to note that the total token limit imposed by the AI engine includes both the prompt and the response. For example, if you request a spelling correction or translation (where the response is likely similar in size to the input), the maximum block size must be less than half the total token limit. For instance, if the model you select has a 128K

token limit, and if you're performing a spelling correction, you should set the block size to less than 64K. For shorter responses, such as sentiment classification, you can set a larger block size (e.g., greater than 64K). Keep in mind that very large blocks may result in longer waiting times for responses from the engine.

Asking Questions and Creating a Post-processing Script

In several areas of the WordStat interface, you can directly ask a question to an AI engine by clicking the AI button and selecting **Ask a Question**. This will open a dialog box similar to the one shown below:



The **Engine settings** section may not appear if a specific engine and model have already been assigned to the project.

Depending on your location within the software, the **Data** section may be prefilled with relevant text. For example, if you access this feature from the topic modeling page, it will contain the 15 extracted topics. This section is fully editable, allowing you to modify or replace the text as needed.

The **System Guidelines** and **Prompt** sections may already include project-specific details. In most cases, you can retain this information and simply append your question or instructions. However, you may remove or adjust these sections as necessary.

Once you have entered your question and instructions, click **OK** to execute the query. The AI service's response will be displayed in a text editor for further review.

To save your question for future use:

- Click the button to open the Save File dialog box.
- Enter a descriptive file name that clearly reflects the task you want to accomplish.
- Click **Save** to store the file.

  NOTE: Script files only contain the **System Guidelines** and **User Prompt**—they do not include the **Data** section. For best results, formulate your question and instructions in a generic way so they can be applied to various text

data. If the project-specific System Role and Data Description are left unchanged, they will be replaced with placeholders before being stored: **$ROLE$** for the system role and **$TOPIC$** for the data description. This ensures the script remains adaptable for different datasets.

To load and execute a saved AI script:

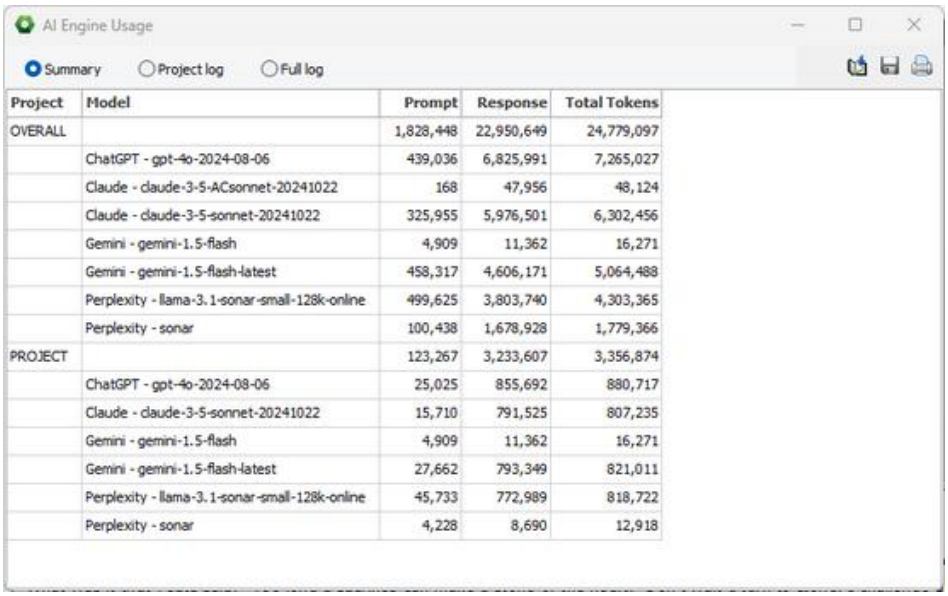
- Click the button to display a file selection dialog box.
- Choose the script file you want to use and click **Open**. This will overwrite the **System Guidelines** and user **Prompt** with the stored instructions.
- If necessary, modify any instruction fields before running the script.
- Click **OK** to execute the script and retrieve the AI response.

  Sometimes, a prompt sent to an AI engine returns unexpected results. Setting the **Leave Dialog Box Open** option will bring back the script editor window after execution, allowing you to refine the prompt and try again.

Monitoring AI Usage Statistics

The cost of using generative AI engines in WordStat can be minimal when post-processing results produced by WordStat, but it will likely be more significant when used for data transformation and analysis on project text data. WordStat tracks token consumption in a log file, allowing you to monitor usage and control costs.

To access information stored in this log, click on any AI button displayed at various locations to display the AI ASSISTANT menu. Then select the AI USAGE menu item. A dialog box similar to the one below will appear:



Project	Model	Prompt	Response	Total Tokens
OVERALL		1,828,448	22,950,649	24,779,097
	ChatGPT - gpt-4o-2024-08-06	439,036	6,825,991	7,265,027
	Claude - claude-3-5-ACsonnet-20241022	168	47,956	48,124
	Claude - claude-3-5-sonnet-20241022	325,955	5,976,501	6,302,456
	Gemini - gemini-1.5-flash	4,909	11,362	16,271
	Gemini - gemini-1.5-flash-latest	458,317	4,606,171	5,064,488
	Perplexity - llama-3.1-sonar-small-128k-online	499,625	3,803,740	4,303,365
	Perplexity - sonar	100,438	1,678,928	1,779,366
PROJECT		123,267	3,233,607	3,356,874
	ChatGPT - gpt-4o-2024-08-06	25,025	855,692	880,717
	Claude - claude-3-5-sonnet-20241022	15,710	791,525	807,235
	Gemini - gemini-1.5-flash	4,909	11,362	16,271
	Gemini - gemini-1.5-flash-latest	27,662	793,349	821,011
	Perplexity - llama-3.1-sonar-small-128k-online	45,733	772,989	818,722
	Perplexity - sonar	4,228	8,690	12,918

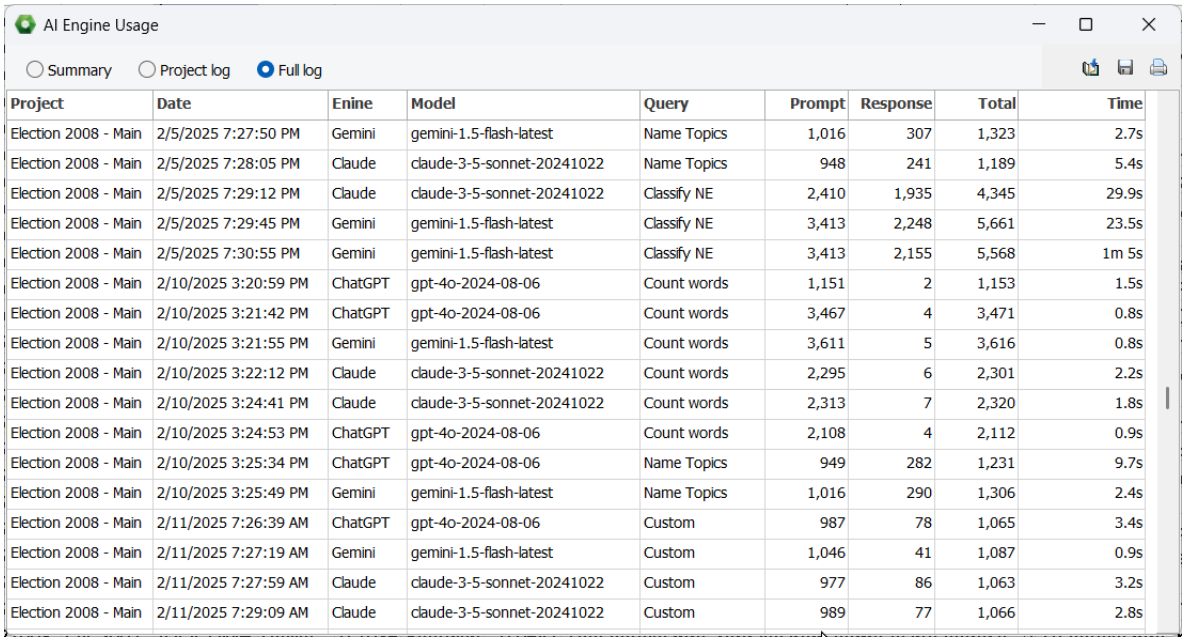
Selecting the **Summary** option will display the number of tokens consumed by the prompt sent to the engine as well as those associated with the AI-generated response. The last column shows the total number of tokens. This information is broken down by engine and model. The table also distinguishes between overall consumption in WordStat and the consumption specific to the current project.

Such detailed information is essential for calculating the overall cost of AI usage, as the cost per million tokens depends on the specific engine and model used for the transformation. AI companies also differentiate costs based on whether tokens are associated with the input (prompts) or the output (responses), with the latter typically being more expensive than the former.

Selecting the **Project Log** option displays detailed information about all operations performed on the current project. This includes the date and time of each operation, the engine and model used, the name of the query, token consumption, and the calculated execution time (see the example below).

Select the **Full Log** option to view a comprehensive log of all operations performed on this computer. The first column identifies the project file name associated with each operation.

By default, the detailed log table is sorted in chronological order. However, you can sort the table by any column by clicking its header. Clicking once will sort the table in ascending order, and clicking again will sort it in descending order.



Project	Date	Enine	Model	Query	Prompt	Response	Total	Time
Election 2008 - Main	2/5/2025 7:27:50 PM	Gemini	gemini-1.5-flash-latest	Name Topics	1,016	307	1,323	2.7s
Election 2008 - Main	2/5/2025 7:28:05 PM	Claude	claude-3-5-sonnet-20241022	Name Topics	948	241	1,189	5.4s
Election 2008 - Main	2/5/2025 7:29:12 PM	Claude	claude-3-5-sonnet-20241022	Classify NE	2,410	1,935	4,345	29.9s
Election 2008 - Main	2/5/2025 7:29:45 PM	Gemini	gemini-1.5-flash-latest	Classify NE	3,413	2,248	5,661	23.5s
Election 2008 - Main	2/5/2025 7:30:55 PM	Gemini	gemini-1.5-flash-latest	Classify NE	3,413	2,155	5,568	1m 5s
Election 2008 - Main	2/10/2025 3:20:59 PM	ChatGPT	gpt-4o-2024-08-06	Count words	1,151	2	1,153	1.5s
Election 2008 - Main	2/10/2025 3:21:42 PM	ChatGPT	gpt-4o-2024-08-06	Count words	3,467	4	3,471	0.8s
Election 2008 - Main	2/10/2025 3:21:55 PM	Gemini	gemini-1.5-flash-latest	Count words	3,611	5	3,616	0.8s
Election 2008 - Main	2/10/2025 3:22:12 PM	Claude	claude-3-5-sonnet-20241022	Count words	2,295	6	2,301	2.2s
Election 2008 - Main	2/10/2025 3:24:41 PM	Claude	claude-3-5-sonnet-20241022	Count words	2,313	7	2,320	1.8s
Election 2008 - Main	2/10/2025 3:24:53 PM	ChatGPT	gpt-4o-2024-08-06	Count words	2,108	4	2,112	0.9s
Election 2008 - Main	2/10/2025 3:25:34 PM	ChatGPT	gpt-4o-2024-08-06	Name Topics	949	282	1,231	9.7s
Election 2008 - Main	2/10/2025 3:25:49 PM	Gemini	gemini-1.5-flash-latest	Name Topics	1,016	290	1,306	2.4s
Election 2008 - Main	2/11/2025 7:26:39 AM	ChatGPT	gpt-4o-2024-08-06	Custom	987	78	1,065	3.4s
Election 2008 - Main	2/11/2025 7:27:19 AM	Gemini	gemini-1.5-flash-latest	Custom	1,046	41	1,087	0.9s
Election 2008 - Main	2/11/2025 7:27:59 AM	Claude	claude-3-5-sonnet-20241022	Custom	977	86	1,063	3.2s
Election 2008 - Main	2/11/2025 7:29:09 AM	Claude	claude-3-5-sonnet-20241022	Custom	989	77	1,066	2.8s

IMPORTANT NOTE. It is important to note that the displayed token consumption is specific to the operations performed on the current computer. The overall and project-specific consumption may be higher if the project data file is accessed from another computer, or if the same API keys are used on a different machine. Additionally, the report may not account for scripts sent to the engine that resulted in error messages or were interrupted, which could affect the actual token consumption.

To verify the overall consumption, you can check the AI engine dashboard. Most AI service providers offer a dashboard where you can review detailed usage statistics, including total token consumption across all devices and API keys. This will give you a comprehensive view of your overall usage, including any consumption that might not be reflected in WordStat's log.

New Topic Extraction Features

The AI button can be used to ask a generative AI engine to name topics or to group topics into themes.

It may also be used to ask other custom questions relevant to the current topic solutions, such topic classification, or the evaluation of their coherence. Additional AI scripts and features will be added in the near future offering more flexibility. See Using Generative AI Features for more information on how to configure various engines and how to create your own question.

The Topic MDS page

This page provides a more intuitive understanding of topic relationships using a multidimensional scaling (MDS) plot. This visualization transforms the correlation matrix into a two-dimensional map where topics are represented as points. The distance between any two topics in this plot approximately represents their correlation – topics that are highly correlated appear closer together, while topics with weak or negative correlations are positioned farther apart. This spatial representation helps identify clusters of related topics and understand the overall structure of relationships in your topic model. The MDS plot complements the correlation table by providing a holistic view of how all topics relate to each other simultaneously.

The MDS plot is interactive – clicking on any topic point reveals a detailed correlation table on the right side of the plot. This table lists all related topics in descending order of correlation strength, along with their correlation values. Using the radio buttons above the table, you can choose to view either all correlations or focus only on positively correlated topics. This feature allows you to quickly identify and explore the strongest relationships for any topic of interest, while the spatial layout of the plot provides context for these relationships within the broader topic structure.

Various other options are available.

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.

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)

Press this button to change the default color palette used for the various data points.

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 value of a third variable. When enabled, one can choose to set the size of each data point to represent the **coherence metric**, to the **estimated frequency** or the **estimated case occurrence**.

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.

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.

.

New Phrases Extraction Features

To categorize phrases according to their syntactic class:

- Click the AI button and select **Find Syntactic Categories**. A dialog box will appear, showing the question that will be sent to the generative AI engine, allowing you to modify it.
- Click **OK** to run the script. An additional column will be added to the phrase table with the proper syntactic category.

  Syntactic categorization of phrases helps analyze the structure of text by grouping words based on their grammatical roles. **Noun phrases** (NPs) represent entities, subjects, or objects in a sentence, making them essential for identifying key concepts and topics. **Verb phrases** (VPs) capture actions, events, processes, or states, helping to understand relationships and dynamics within a text. Other phrase types also contribute valuable information: **adjective phrases** (AdjPs) describe qualities, influencing sentiment and tone; **adverbial phrases** (AdvPs) modify actions or descriptions, refining meaning; and **prepositional phrases** (PPs) provide context, such as time, place, or reason. By categorizing phrases syntactically, we gain a clearer understanding of how information is structured and conveyed in language.

  Once extracted, you may sort the table on those syntactic categories by clicking on their column header.

  See Using Generative AI features for more information on how to set up, run, edit and create AI scripts.

Performing custom post-processing on the phrase list

Beyond syntactic categorization, it is possible to ask your own question on the content of the phrase table. Generative AI engines can analyze phrases by linking them to topics, detecting emotional tones, and classifying them into domain-specific categories like legal or medical terms. It can also identify incomplete or ambiguous phrases, aiding in data cleaning and refinement.

To submit questions regarding the extracted list of phrases:

- Click the AI button and select **Ask a question**. A dialog box will appear, allowing you to type a question or provide instructions to the AI engine.
- By default, the entire list of phrases will be added at the end of the prompt. This section can be edited or be replaced with data of your own, allowing you to ask any type of question.
- Click **OK** to execute the prompt. The response will be returned in a text editor.

Prompts may also be saved to disk and later retrieved.

- To save a prompt to disk, click the button. A save file dialog box will appear, allowing you to name your prompt.
- To retrieve a previously saved prompt, click the button and select the file containing the prompt you want to execute.

New Named Entities Extraction Features

Post-processing named entities using AI

Categorizing named entities

Named Entity Recognition (plays a crucial role in text analysis by identifying and categorizing proper names—such as people, organizations, locations, dates, and product names—within unstructured text. Categorizing named entities allows one to focus on selected type of entities, improving information retrieval and text analysis. LLMs excel at named entity categorization by considering the full context in which an entity appears, enabling them to resolve ambiguities and classify entities with greater accuracy than traditional rule-based or dictionary-driven approaches. This contextual understanding ensures more precise classification, even in cases of polysemy or nuanced distinctions.

WordStat's implementation of named entity classification is performed not on the original text but on the extracted list of named entities, independent of their original context. We found that providing a general description of the data source and instructing the AI engine to consider the full list of extracted entities during classification is often sufficient to achieve high accuracy. This approach also offers significant efficiency gains, performing classification in a fraction of the time required for full-text analysis. To perform NER categorization.

- Click the AI button, and select **Cagegorize Extracted Entities**. A dialog box will appear, displaying the prompt currently used to extract entities.
- No predefined category set has been provided in the question. To restrict AI-generated categories to a specific set, edit the question and include a list of categories for classification.

Click the OK button to execute the prompt.

To submit questions regarding the extracted list of named entities:

- Click the AI button and select **Ask a question**. A dialog box will appear, allowing you to type a question or provide instructions to the AI engine.
- By default, the entire list of named entities will be added at the end of the prompt. This section can be edited or be replaced with data of your own, allowing you to ask any type of question.
- Click **OK** to execute the prompt. The response will be returned in a text editor.

Prompts may also be saved to disk and later retrieved.

- To save a prompt to disk, click the button. A save file dialog box will appear, allowing you to name your prompt.
- To retrieve a previously saved prompt, click the button and select the file containing the prompt you want to execute.

Contrasting Cooccurrences Data Between Groups

Traditional text analysis methods often focus on comparing the frequency of words, topics, or content categories between groups. While this approach can highlight general differences in emphasis, it overlooks how words and concepts are associated with one another within each group. Co-occurrence analysis goes beyond frequency by examining how words or topics appear together, capturing underlying relationships that frequency-based methods may miss. A word or concept may appear equally often in two groups but be associated with very different contexts, meanings, or themes, revealing deeper, more nuanced distinctions.

WordStat already offered ways to explore differences in associations, but these required a more deliberate and manual approach. For example, users could tag text segments about a specific topic in QDA Miner and then analyze only those segments in WordStat, comparing the frequency of associated words using tools like the crosstabluation, the correspondence plot or deviation table. Alternatively, a recently introduced preprocessing feature of WordStat allowed users to filter the documents to only those paragraphs containing a predefined set of words or phrases, bypassing the

need for manual tagging. While effective, these methods required prior knowledge of what to investigate and are limited to analyzing one topic at a time, potentially missing unexpected differences in associations across the entire dataset.

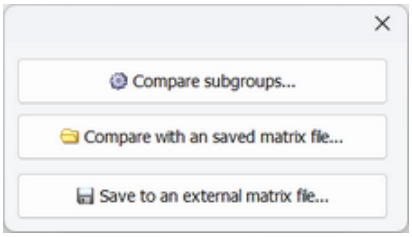
By comparing co-occurrence patterns between groups automatically, WordStat now enables users to uncover unexpected differences that would otherwise go unnoticed. For example, in political discourse, the term "healthcare" may be equally frequent among two parties, but one may associate it with "spending", "budget", and "market" while the other links it to "access", "afford" and "universal." Similarly, in customer feedback, "delivery" may co-occur with "reliable" and "quick" for satisfied users but with "damaged" or "delayed" for unsatisfied ones, indicating underlying differences in customer experience that a simple comparison of frequency may not show. Comparing co-occurrence patterns in incident reports with those in operating manuals can help identify potential human errors by revealing discrepancies between prescribed procedures and real-world practices. Identifying such shifts in associations could provide richer insights, whether in social sciences, business, healthcare, forensic investigations, etc.

This innovative feature in WordStat allows users to explore and detect these unexpected co-occurrence differences automatically. Instead of manually sifting through association patterns, users can quickly identify where and how relationships between words, topics, or categories differ significantly between groups. This capability provides a deeper understanding of contextual variations, making it a powerful tool for discovering hidden insights that frequency-based comparisons alone cannot reveal.

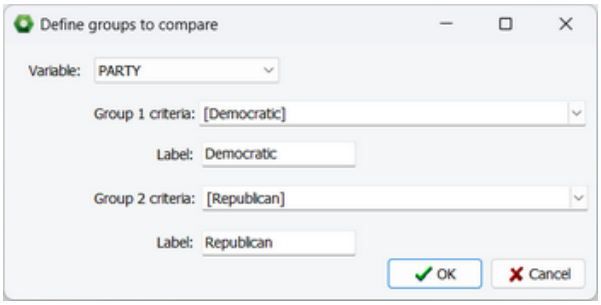
This exploration is achieved by comparing cooccurrence matrices and can be performed using two distinct approaches. One may select two subsegments segments of the current dataset using values of a categorical variable. For example, one may choose to contrast satisfied and dissatisfied customers, compare news from two different media outlets, or speeches from two political parties, The second approach consists of comparing the co-occurrence measured on the current dataset to a similar matrix obtained on a different dataset saved previously. We will first present the situation where one chooses to compare two subsamples of the currently active dataset.

Comparing Two Subsegments of the Current Dataset

The contrast features are available on the last tab of the Co-occurrence page. When accessing this page for the first time, you will be presented with a dialog box that looks like this:



Click the **Compare subgroups** button. You will be presented with a second dialog box (see below) that will allow you to select and label both groups.



The **variable** list box allows you to select which variable should be used to define the two subgroups that will be contrasted. The available list of variables corresponds to those selected when clicking the Analyze button from the Data page, or from the list of independent variables chosen from a calling application.

If the selected variable contains only two values, all remaining options will be set to correspond to those two values, and you will be ready to run the comparison. If it contains more than two values, you will be asked to provide both the list of values corresponding to each group as well as a label to describe those. While one may select a continuous variable containing large number of numerical values, selecting all the relevant values may be time consuming and prone to error.

It is thus recommended transforming such numerical values into a nominal or ordinal variable using either binning or mathematical transformation of numerical data in order to create a limited number of grouping values.

Once all those options have been set, click the **OK** button to display the result of the comparison.

Comparing the cooccurrence of two datasets

WordStat allows one to store on disk the cooccurrence matrix of a dataset. This allows you to later compare the co-occurrence obtained on another dataset to this previously stored information. This saved you from the necessity of merging two datasets that may otherwise have quite different structures.

Step #1. Creating a cooccurrence matrix file on disk

On the Options page, set the **Occurrence** list box to the option corresponding to the window used to define co-occurrences. Also make sure that the **Type** option is set to **Word Cooccurrence - First order**. The similarity index set for measuring the association has no impact on the saved data since the joint appearance of items are stored in raw frequencies.

Move to the **Contrast** page and click the **Save to an External Matrix file** button. This will bring a Save File dialog box allowing you to choose a name and a location where to store the matrix file. WordStat matrix data files have a .wmtx file extension).

Step #2. Comparing co-occurrence with a saved matrix data file

- Open the dataset you want to compare the joint distribution of items with a previously saved matrix data file.
- Move to the **Co-occurrence** page and set the **Occurrence** option to the same window previously set for the creation of the matrix data file. Selecting another segmentation window will result in a warning message being displayed informing you of the discrepancy in the partitioning method.
- Move to the **Contrast** page and click **Compare with a saved matrix file** button.
- Select the .wmtx matrix datafile you want to compare your current data with.
- A comparison will be performed between the two co-occurrence matrix and will be displayed on the resulting **Contrast** page, allowing you to explore the difference between the two sources.

Using the Contrast Interactive Page

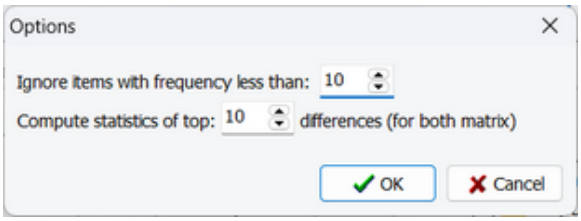
The matrix comparison feature detects differences in co-occurrence patterns using two distinct strategies, each offering unique advantage.

Items (entire row): The first approach focuses on identifying the most significant overall differences across all items. It analyzes the entire row of co-occurrences using the top n positive and top n negative differences in proportion, where *n* is a user-defined value. This method is particularly useful for quickly spotting the most pronounced shifts in associations, providing a high-level overview of how relationships between words, topics, or categories differ between groups.

Single cells: The second approach pinpoints localized differences within individual matrix cells, allowing users to explore where the most substantial deviations occur. This method is beneficial when users need granular insights into specific associations, helping to uncover unexpected anomalies that might be overlooked in an aggregate analysis.

By offering both strategies, this feature provides flexibility for different analytical needs—whether for broad trend detection or detailed, cell-level examination of co-occurrence differences.

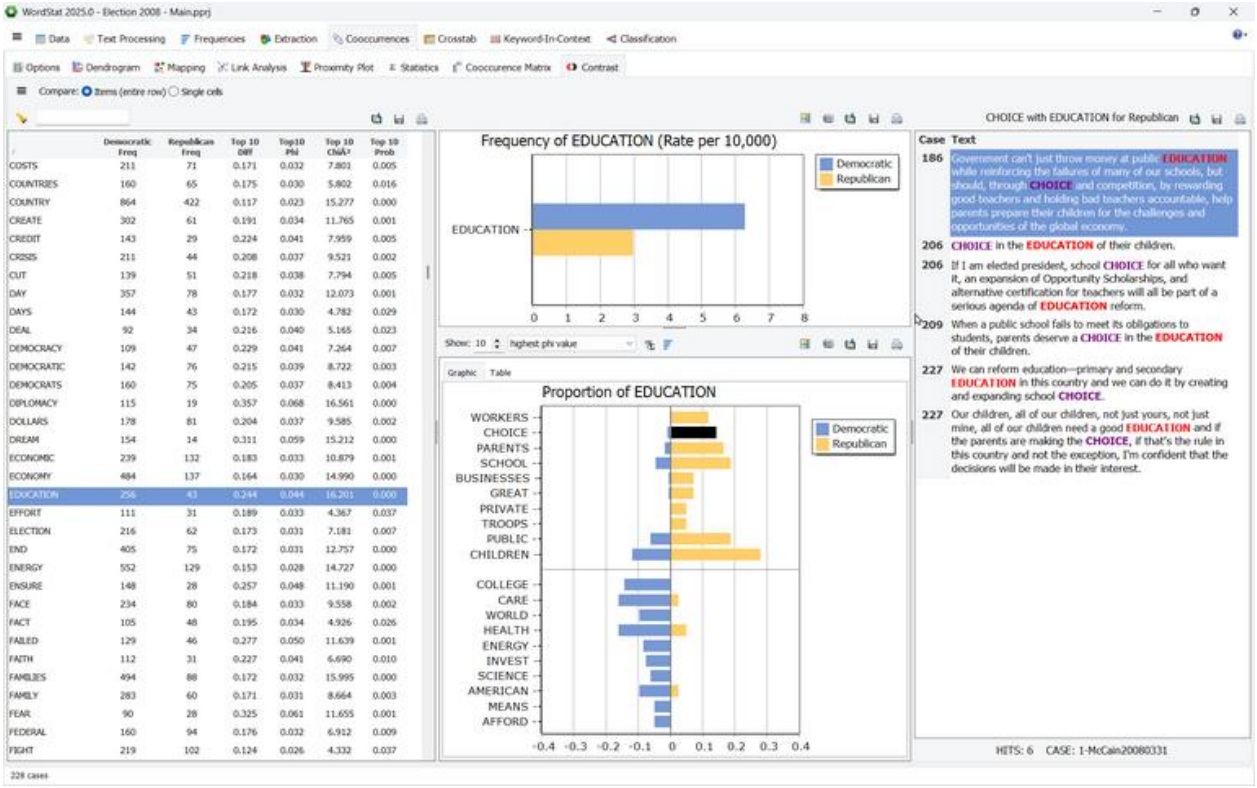
By default, the number of top differences used to establish in which items we can observe the most important differences is set to 10 for each group. Also, WordStat will only display items occurring at least 10 times in each group, preventing interpretation of unreliable co-occurrences information. To change any of those values, click the ☰ button and select Settings... A dialog like this will appear, allowing you to adjust those two values.



Clicking **OK** will result in an update of the displayed information.

Exploring at the item level

The screen shot below shows the interface when the identification is computed on entire rows.



On the left side, a table lists all items that meet a minimum frequency threshold. By default, these items are sorted in descending order of the computed Phi value, an effect size measure calculated based on the top n differences favoring each subgroup. This ranking helps users quickly identify the most significant shifts in co-occurrence patterns. The table includes the frequency of each item in both groups, along with several key statistical measures computed on the top n most relevant differences:

- **Top n Diff** – The average difference in raw proportion of times the item other elements.co-occurs with the selected item.
- **Top n Phi** – The same difference in proportion, but measured using the Phi coefficient, which accounts for association strength.
- **Top n Chi²** – The average chi-square value calculated for the top n differences, indicating the statistical significance of association shifts.
- **Prob.** – The p-value associated with the obtained chi-square statistic, showing the likelihood that the observed differences are due to chance.

By default, the table is sorted in descending order of the computed Phi value. However, users can sort rows by any other column by clicking on the column header—click once to sort in ascending order and click again to sort in descending order.

An incremental search edit box can be used to quickly locate a specific item in the table. As you type, the software moves to the row of the first matching item. Click the button to jump to the next match. If the typed text has no match, the search box text will turn red.

Selecting a row updates the two bar charts in the middle panel, providing more detailed insights into the relative frequency of the selected item and its co-occurrence patterns.

- The top bar chart displays the frequency of the selected item in each group, normalized to a rate per 10,000 words. This normalization accounts for differences in corpus size, ensuring a fair comparison.
- The bottom bar chart highlights the most significant differences in association between groups. If the **Show** option is set to 10, the chart displays up to 20 sets of bars—the first ten represent items more strongly associated with one group, while the next ten show items more frequently associated with the other group.

You can select multiple rows to combine co-occurrence measures in the bottom chart, while the top chart continues to display the relative frequency of each item individually. To select multiple non-adjacent rows, hold down the **Ctrl** key while clicking on additional rows.

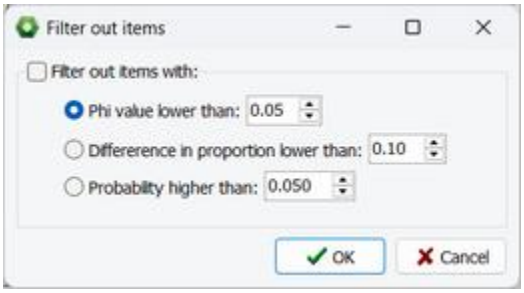
The right panel allows you to view text segments associated with any bar in the two charts:

- Clicking on a bar in the upper chart retrieves all text segments containing the target word for the selected group.
- Clicking on a bar in the lower chart retrieves all text segments where both the selected co-occurring word and the target item(s) appear together, helping you better understand the context and relationship between them.

Several options are available to customize the co-occurrence bar chart.

The **Show** button allows you to choose how many items to display for each group. Additionally, a drop-down menu provides options for selecting which items to include. By default, items are ranked based on their **highest Phi value**. However, you can use the drop-down to select instead items based on their raw **difference in proportion** between groups or to display the most frequently associated items for each group, regardless of differences in association strength.

The number of items displayed may be lower than the set value on **Show** option above if one chose to apply an additional statistical filter. To apply such a filter, click the button. A dialog box similar to this one will appear:

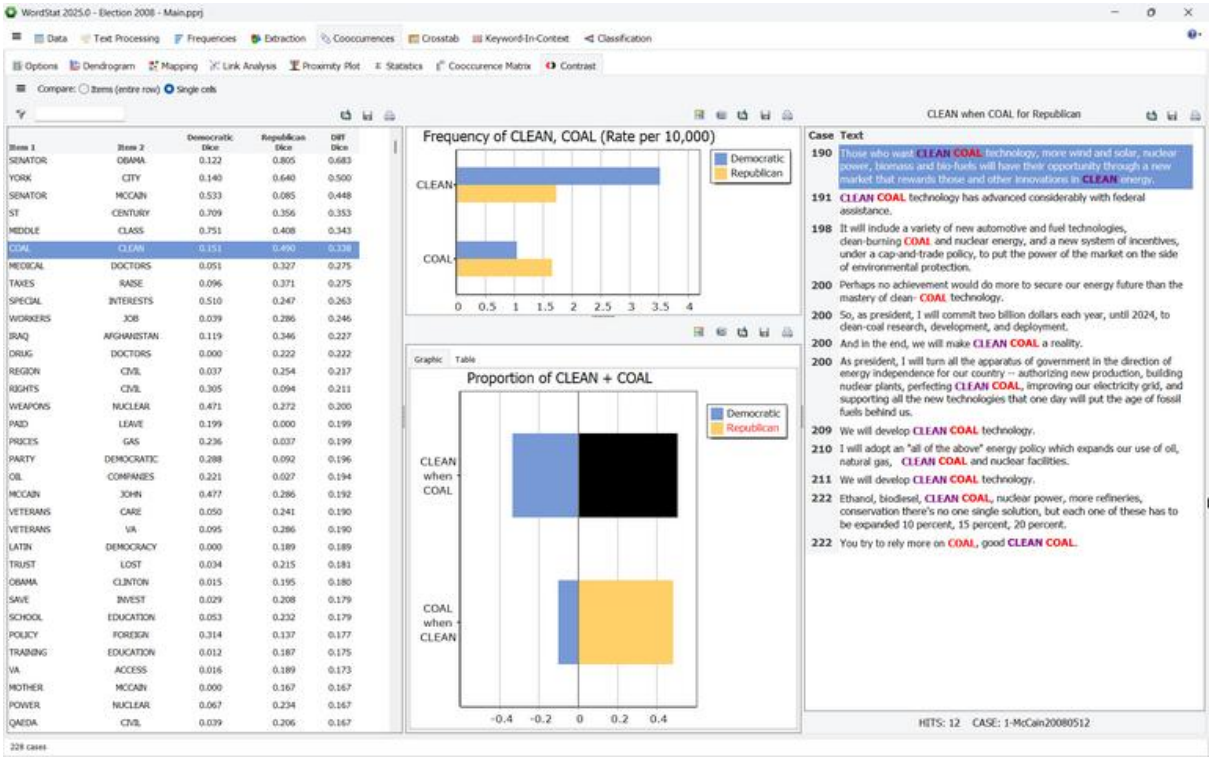


To filter items, set the **Filter out items** check box and choose which criteria should be used to restrict the display, adjust the desired threshold and click the **OK** button.

While bars are usually sorted on the selected statistics, bars may also be sorted in descending order of frequency for the two groups by pressing down the button.

Exploring at the cell level

Exploring difference in co-occurrence at the single cell level pinpoints localized differences within individual matrix cells, allowing users to explore where the most substantial deviations occur offering more granular insights into specific associations. To enable this feature, set the **Compare** option to **Single cells**. The interface of the **Contrast** will appear to be quite similar. We will explain below how different it actually is.



The table on the right displays all pairs of items that meet specific statistical criteria. Difference on the Dice coefficient computed on each matrix is used to identify difference in association between the two groups.

The Dice coefficient is a symmetric measure that quantifies the association between two items based on their co-occurrence. It combines the proportion of times item A co-occurs with item B and the proportion of times item B co-occurs with item A, ensuring that the measure treats both directions equally. By analyzing the difference in the obtained Dice value (or Δ Dice) we can identify patterns of shifting associations, helping to uncover meaningful changes in relationships between the two subgroups.

By default, the table is sorted in descending order of Δ Dice. However, users can sort rows by any other column by clicking on the column header—click once to sort in ascending order and click again to sort in descending order.

A filter edit box can also be used to quickly locate rows containing a specific item. Click the button down to apply the filter. Click it up to remove the filter and see all items. If the typed text has no match, the search box text will turn red.

The first chart in the middle panel displays the relative frequency of both selected items, expressed as a rate per 10,000 words, just as when exploring individual items. However, the lower chart differs in that it breaks down co-occurrence, measured by the Dice coefficient, into its two key components:

- The proportion of times item A co-occurs with item B.
- The proportion of times item B co-occurs with item A.

This breakdown provides a clearer view of asymmetry in their relationship, helping to better interpret differences in co-occurrence patterns.

Clicking on a bar in the top chart retrieves all text segments where the selected word, topic, or content category appears in the chosen group. Clicking on a bar in the lower chart retrieves text segments where both words appear together.

In the example above, the upper chart shows that "clean" is mentioned more frequently by Democrats, while "coal" is mentioned more often by Republicans. Looking at the first bar in the lower chart, we see that Democrats use "clean" when talking about "coal" 33% of the time, but this proportion is higher for Republicans (50%). The second bar reveals that Democrats use "clean" in a much broader context, explaining its lower association with "coal" (10%), while for Republicans, "coal" is more exclusively linked to "clean" (48%). The table view provides exact co-occurrence proportions. Returning to the item-level analysis confirms that Democrats more often associate "clean" with "energy" and "renewable" than Republicans, whereas Republicans link it more strongly with "coal" and "nuclear".